Investigating the Application of Causal Inference Methods for Modelling the Impact of Treatment Sequences in Health Economic Evaluations

Jen-Yu Amy Chang

BSc, MSc, RPh

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

School of Medicine and Population Health

Sheffield Centre for Health and Related Research (SCHARR)

April 2024

Acknowledgement

Reflecting on my PhD journey, I feel an immense gratitude towards those who were instrumental in helping me through it. Being where I am would not have been possible without such an amazing group of people supporting me along the way.

Firstly, I must express my deepest appreciation for my supervisors, Professor Nick Latimer and Professor Jim Chilcott, whose unwavering support has been foundational in my development as an independent researcher—all the while keeping me sane! Their patient and insightful guidance made navigating the unprecedented challenges of the COVID-19 pandemic and unique complexities of this thesis possible.

A heartfelt thank you to my external examiner, Professor Keith Abrams, and internal examiner, Professor Alan Brennan, for their valuable suggestions and opportunity to share and discuss my research. It was an immensely enjoyable and rewarding experience. Special acknowledgment goes to my mentor, Dr. Penny Breeze, for her sage advice in navigating academia, as well as personal support for the challenges which life threw my way. I am indebted to Dr. Carmel Pezaro, Professor Derek Rosario, and Professor Janet Brown, whose generosity in sharing clinical expertise enriched my case studies significantly. My appreciation also extends to those at Flatiron Health, particularly Dr. Blythe Adamson and Dr. Philani Mpofu, for their enthusiastic support in data-related matters and uplifting discussions. I also thank NHS Digital for their provision of data, and extend special thanks to Dr. Amanda Loban for her indispensable guidance in information governance. In the early phase of my PhD, feedback from my confirmation review examiners, Professor Allan Wailoo and Professor Paul Tappenden, along with the mentorship from Professor Petra Meier and Dr. Tessa Peasgood, were paramount. My thanks to Dr. Duncan Gillespie for his kind support in the first year of my thesis. I also thank Professor Mark Strong for his assistance with data contract issues.

The causal inference course on treatment-switching taught by my supervisor Professor Nick Latimer, alongside Professor Uwe Siebert and Professor Ian White, played a demonstrable role in inspiring ideas that shaped my approach to the thesis. This crucial shaping of core methods used within my thesis was further refined with the CAUSALab courses at Harvard University, led by Professor Miguel Hernán and other remarkable scholars in the field. Meeting Professor Jamie Robins there, a foundation of contemporary causal inference methods in epidemiology, was a valued privilege. All of this would not have been possible without Wellcome Trust's generous funding which enabled me to share my research during conferences and attend these pioneering courses.

The highly supportive environment surrounding my project had an invaluable effect on me. I am deeply thankful to Saleema Rex, my causal inference buddy, for her friendship and enriching discussions, as well as to the Causal Inference Journal Club group for their stimulating talks and

discussions. To my peers—Tazeen Tahsina, Paul Schneider, Jen Boyd, and Artur Kowalski—and many fellow PhD students at the Wellcome Trust programme and ScHARR, your camaraderie and support have been pillars of strength. I also thank the ScHARR Library, especially Andy Tattersall, for advice on publishing and media sharing. Sometimes overlooked, but no less important, are the University administrators who facilitated numerous aspects of my journey.

In addition to those who helped throughout my PhD, there are also those guided me towards it. My previous advisors at National Taiwan University, Dr. Chi-Chuang Emma Wang, Dr. Fang-Ju Irene Lin, and Professor Li-Jiuan Rita Shen, helped foment a strong academic foundation which proved invaluable during my career development.

Most importantly, I thank my family, especially my parents, grandma, and aunt—for their unwavering love and support, which has been my sanctuary. Lastly, to my partner Lewis, your selfless care and companionship during the late nights have been both my comfort and strength.

Acknowledgement of Collaborative Work

The thesis includes published works: a manuscript in a peer-reviewed journal and a study protocol as a University of Sheffield SCHARR HEDS Discussion Paper. The candidate (JYAC) confirms that the work submitted in this thesis is her own, except from work that has formed part of jointly authored journal publications. Each authors' contribution is indicated explicitly below.

- Chapter 2: Challenges and Opportunities in Interdisciplinary Research and Real-World Data for Treatment Sequences in Health Technology Assessments. *Pharmacoeconomics*. 2024; 42(5): 487-506.

  Author: Jen-Yu Amy Chang

  Co-authors: James B Chilcott, Nicholas R Latimer

  Author Contributions: All authors contributed to the conceptualisation and design. JYAC performed the literature search and selection, data extraction, evidence synthesis and drafted the original manuscript. JC and NL contributed insights for quality assessment and interpretation of results, revised the manuscript, and provided supervision. All authors reviewed and approved the final manuscript. This work has undergone revisions in response to feedback from external reviewers.

- Chapter 7: Leveraging Real-world Data to Assess Treatment Sequences in Health Economic Evaluations: A Study Protocol for Emulating Target Trials Using the English Cancer Registry and US Electronic Health Records-Derived Database. SCHARR HEDS Discussion Papers (24.01). 2024; 24(01):1-61.

  Author: Jen-Yu Amy Chang

  Co-authors: James B Chilcott, Nicholas R Latimer

  Author Contributions: All authors contributed to the conceptualisation and design of the study. JYAC conducted literature review and drafted the initial protocol. JC and NL reviewed and edited the protocol, and offered supervisory support. All authors approved the final manuscript.

Funding

This is work was supported by funding from the Wellcome Trust [108903/B/15/Z] and the University of Sheffield.

Research Outputs

The following research outputs are based on the work presented in this thesis. The published manuscripts are presented in the form of Author Accepted Manuscripts, which have been adapted in their format to ensure a cohesive integration within the thesis.

Peer-reviewed publication:

Chang JYA, Chilcott JB and Latimer NR. Challenges and Opportunities in Interdisciplinary Research and Real-World Data for Treatment Sequences in Health Technology Assessments. *Pharmacoeconomics*. 2024; 42(5): 487-506.

Study protocol publication:

Chang JYA, Chilcott JB and Latimer NR. Leveraging Real-world Data to Assess Treatment Sequences in Health Economic Evaluations: A Study Protocol for Emulating Target Trials Using the English Cancer Registry and US Electronic Health Records-Derived Database. SCHARR HEDS Discussion Papers (24.01). 2024; 24(01):1-61.

Conference presentations:

Chang JYA, Chilcott JB, Latimer, NR. Exploring Data-Driven Challenges in Modelling the Effectiveness of Treatment Sequences in Health Economic Evaluations: A Systematic Review of UK National Institute for Health and Care Excellence (NICE) Technology Appraisals. The 15th IHEA World Congress on Health Economics, July 8-12, 2023. [Oral Presentation]

Chang JYA, Latimer LR, Gillespie D, Chilcott J. Prevalence, Characteristics and Key Issues of Modelling Treatment Sequences in Health Economic Evaluations: A Systematic Review of Nice Technology Appraisals. Virtual ISPOR Europe 2020, Nov 16-19, 2020. [Poster Presentation]

Dedicated to my grandfather, a man of humble beginnings in a mining village. He devoted his life not only to raising his children to pursue higher education, but also to transforming our community. He selflessly guided many from rural beginnings to urban success, driven by a keen interest in nurturing others towards growth. His passing during my pharmacy internship—my first encounter with the finality of life—deeply impacted me, reinforcing my resolve in health research. His commitment to elevating those around him serves as an inspiration for me to make a positive impact in my life.

"All models are wrong, but some are useful."

— George E.P. Box

Abstract

This thesis explores the complexities of evaluating treatment sequence effectiveness in health technology assessments (HTA), focusing on challenges posed by limited evidence. Clinical trials typically assess individual lines of treatment (LOTs) rather than entire treatment sequences. Existing research recommends merging LOT-specific evidence from different sources to estimate overall treatment sequence effectiveness. However, this relies on a strong assumption about exchangeability between LOT-specific populations, which may not always hold.

Real-world data (RWD) provide a valuable alternative for assessing the comparative effectiveness of treatment sequences. However, generating reliable real-world evidence (RWE) presents significant challenges, notably confounding, which is exacerbated by the longitudinal nature of treatment sequences, leading to time-varying confounding. Conventional statistical methods (e.g. simple outcome regressions) may yield biased estimates, whereas advanced methods grounded in causal inference principles could offer more reliable estimates, provided their assumptions hold and data sources are of sufficient quality and breadth.

This thesis reviews advanced statistical methods and proposes a series of innovative, interconnected proof-of-concept studies to assess the feasibility of deriving unbiased RWE for comparing treatment sequences. It evaluates suitable RWD sources pertinent to English HTA that support these methods. Further, it leverages the Target Trial Emulation framework, a methodology endorsed by the National Institute for Health and Care Excellence (NICE) RWE framework, to mitigate biases in real-world study designs.

This thesis contributes to the field by delineating the challenges associated with treatment sequences and the landscape of English HTA practices. It complements existing treatment-sequencing modelling frameworks by proposing ways to leverage RWD to inform treatment sequence effectiveness and summarising challenges. Additionally, it extends the NICE RWE framework, particularly regarding practical applications and necessary adaptations for treatment sequence comparisons. A case study employing advanced inverse probability weighting methods demonstrates the feasibility of deriving unbiased treatment effectiveness estimates from Flatiron data, benchmarked against an existing trial.

List of Contents

List of Abbreviations

Abbreviation	Full name
A&E	Accident & emergency (dataset)
5-FU	Folinic acid, fluorouracil
A&E	Accident & emergency (dataset)
ACR	American College of Rheumatology
ADHD	Attention-deficit/hyperactivity disorder
ADT	Androgen deprivation therapy
AFT	Accelerated failure time
AG	Assessment Group
AIC	Akaike Information Criterion
AIDS	Acquired immunodeficiency syndrome
ALL	Acute lymphoblastic leukaemia
alloSCT	allogeneic stem-cell transplant
APC	Admitted care (dataset)
ARTA	Androgen receptor-targeted agents
ATE	Average treatment effect
ATS	Adaptive treatment strategies
BADBIR	British Association of Dermatologists Biologic and Immunomodulators Register
bDMARDs	Biologic-disease-modifying anti-rheumatic drugs
BeSt	The Behandel Strategieën (Treatment Strategies for Rheumatoid Arthritis) study
BIC	Bayesian Information Criterion
BMS	Bristol Myers Squibb
BRCA	Breast cancer gene
BRIM-3	A Study of Vemurafenib (RO5185426) in Comparison With Dacarbazine in Previously Untreated Patients With Metastatic Melanoma
BSC	Best supportive care
BSRBR	The British Society for Rheumatology Biologics Register
BSRBR-RA	British Society for Rheumatology Biologics Register for Rheumatoid Arthritis
CA184-024	Dacarbazine and Ipilimumab vs. Dacarbazine With Placebo in Untreated Unresectable Stage III or IV Melanoma
CADTH	Canadian Agency for Drugs and Technologies in Health
CDF	Cancer Drugs Fund
CEA	Cost-effectiveness analysis
CEM	Cost-effectiveness model
CER	Comparative effectiveness research
CHOP	Cyclophosphamide, doxorubicin hydrochloride (hydroxydaunorubicin), vincristine sulfate, and prednisone
CIBMTR	Centre for International Blood and Marrow Transplant Research
CI	Confidence interval
CIF	Cumulative incidence function
CMA	Cost-minimisation analysis
CML	Chronic myeloid leukaemia
CONSORT	Consolidated Standards of Reporting Trials
COSD	Cancer Outcomes and Services Data
COVID	Coronavirus disease
COVD-19	Coronavirus disease
CPRD	Clinical Practice Research Datalink
CRC	Colorectal cancer
CRPC	Castration-resistant prostate cancer
CS	Company submission
CWT	Cancer Waiting Time (dataset)
DAG	Directed acyclic graph
DANBIO	Danish rheumatologic database
DAS	Disease Activity Score
DATA-CAN	Health Data Research Hub for Cancer
DERMBIO	Danish Biological Treatment in Danish Dermatology registry
DES	Discrete event simulation(s)
DMARDs	Disease-modifying anti-rheumatic drugs
DSA	Data sharing agreement
DSU	Decision Support Unit

Abbreviation	Full name
DTR	Dynamic treatment regimens
DWSurv	Dynamic weighted survival modelling
EBCTCG	Early Breast Cancer Trialists' Collaborative Group
ECOG	Eastern Cooperative Oncology Group
EGFR	epidermal growth factor receptor
EHR	Electronic health records
EMA	European Medicines Agency
EMBASE	Excerpta Medica Database
Emtree	Embase Subject Headings
ERAS	Early Rheumatoid Arthritis Study
EULAR	European League Against Rheumatism Responses
ERG	Evidence review group
FAD	Final appraisal document
FDA	(United States) Food and Drug Administration
FOLFIRI	folinic acid, fluorouracil, and irinotecan
FOLFIRINOX	Fluoropyrimidine, oxaliplatin and irinotican
FTA	Fast track technology appraisals
GDPR	General Data Protection Regulation
GnRH	Gonadotrophin-releasing hormone
GUTG-001	Sequencing Abiraterone and Enzalutamide in metastatic castration-resistant prostate cancer (trial)
GVHD	Graft-versus-Host Disease
HAQ	Health Assessment Questionnaire
HbA1c	Haemoglobin A1c
HCC	Hepatocellular carcinoma
HEDS	Health Economics and Decision Science
HES	Hospital Episode Statistics
HIV	Human immunodeficiency virus
HR	Hazard ratio
HTA	Health Technology Assessment
K-RAS	Kirsten-rat sarcoma virus
KM	Kaplan-Meier
ICER	Increased incremental cost-effectiveness ratio
ICH	The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IG	Information governanc
IGARD	Independent Group Advising on the Release of Data
IPCW	Inverse probability of censoring weighting
$IPCW_{txdev}$	Inverse probability of censoring weighting due to treatment deviation
IPTW	Inverse probability treatment weighting
IPW	Inverse probability weighting
ISPOR	The Professional Society for Health Economics and Outcomes Research
ITC	Indirect treatment comparisons
ITT	Intention-to-treat
LHRH	Luteinising hormone-releasing hormone
LOT	Lines of treatment/Line of therapy
LUCADA	The Lung Cancer Data Audit
LY	Life-year
MA	Meta-analysis
MAIC	Matched-adjusted indirect comparison(s)
mCRC	Metastatic colorectal cancer
mCRPC	Metastatic castration-resistant prostate cancer
MeSH	Medical Subject Headings
mHRPC	Metastatic hormone resistant prostate cancers
mHSPC	Mtastatic hormone-sensitive prostate cancer
MLE	Maximum-likelihood estimator
mPC	Metastatic prostate cancer
mRCC	Metastatic renal cell carcinoma
MSD	Merck Sharp & Dohme
MSM	Marginal structural model(s)
MTA	Multiple technology appraisals
mTORi	Mammalian target of rapamycin inhibitor

Abbreviation	Full name
NCCN	National Comprehensive Cancer Network
NCRAS	National Cancer Registration and Analysis Service
NDB	(United States) National Data Bank for Rheumatic Diseases
NDRS	National Disease Registration Service
NHL	Non-Hodgkin's lymphoma
NHS	National Health Service
NHSD	National Health Service Digital
NICE	National Institute for Health and Care Excellence
NLCA	National Lung Cancer Audit
NMA	Network meta-analysis
NMB	Net monetary benefit
nmCRPC	Non-metastatic castration-resistant prostate cancer
nmHRPC	Non-metastatic hormone-relapse prostate cancer
NPBC	Non-platinum-based chemotherapy
NSCLC	Non-small-cell lung cancer
ODR	Office for Data Release
OP	Outpatient (dataset)
OPERAND	The Observational Patient Evidence for Regulatory Approval Science and Understanding Disease project
OS	Overall survival
PartSM	Partitioned survival models
PAS	Patient Access Scheme
PBC	Platinum-based chemotherapy
PBS	(Australian) Pharmaceutical Benefits Scheme
PC	Prostate cancer
PC1	Prostate cancer case study 1
PC2	Prostate cancer case study 2
PC3	Prostate cancer case study 3
PCP	Pneumocystis pneumonia
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed Cell Death Ligand 1
PFS	Progression-free survival
PGR	Postgraduate research
PHE	Public Health England
PLD	Pegylated liposomal doxorubicin
PREVAIL	A Safety and Efficacy Study of Oral MDV3100 (Enzalutamide) in Chemotherapy-Naive Patients With Progressive Metastatic Prostate Cancer (PREVAIL)
PP	Per-protocol
PPcen	Per-protocol censoring—censoring switchers
PPexc	Per-protocol censoring—excluding switchers
PPI	Patient and Public Involvement
PROSPERO	The International Prospective Register of Systematic Reviews
PSA	Prostate-sensitive antigen
PSMA	Prostate-specific membrane antigen
PTE	Privacy, Transparency and Ethics
RA	Rheumatoid arthritis
RCC	Renal cell carcinoma
R-CHOP	Rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone
RCT	Randomised controlled trial
RECORD	Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (trial)
RECORD-3	Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (trial)Phase 3 trial of everolimus for metastatic renal cell carcinoma : final results and analysis of prognostic factors
RHUMADATA	Canadian clinical database of rheumatoid arthritis including patients from Quebec
RMST	Restricted mean survival time
RPSFTM	Rank Preserving Structural Failure Time Model
RR	Risk ratio
RRMS	Relapsing-remitting multiple sclerosis
RTDS	National Radiotherapy Dataset
RWD	Real-world data
RWE	Real-world evidence

Abbreviation	Full name
SACT	Systemic Anticancer Therapy (dataset)
SCHARR	Sheffield Centre for Health and Related Research (previously known as School of Health and Related Research), University of Sheffield
SLA	Service-Level Agreement
SMART	Sequential multiple assignment randomised trial
SNFTM	Structural nested failure time models
SNM	Structural nested models
SOSA	Simulation optimisation via simulated annealing
SPIFD	Structured Process to Identify Fit-For-Purpose Data
STA	Single technology appraisals
STC	Simulated treatment comparisons
STROBE	The Strengthening the Reporting of Observational Studies in Epidemiology (guideline)
TA	Technology appraisal(s)
THIN	The Health Improvement Network
TKI	Tyrosine kinase inhibitor
TMLE	Targeted maximum likelihood estimation
TNFi	Tumour necrosis factor inhibitor
TRE	Trusted Research Environment
TSD	Technical support document
TSEsimp	Simple two-stage estimation
TSEgest	Two-stage estimation with g-estimation
TTE	Target Trial Emulation
UCB	Union Chimique Belge
UK	United Kingdom
UKALL2003	A Randomised Trial Investigating Treatment Intensification for Children and Young Adults with Minimal Residual Disease Defined High Risk Acute Lymphoblastic Leukaema
UKALL2014	A Randomized Trial for Adults with Newly Diagnosed Acute Lymphoblastic Leukaemia
UMiT	University for Health Sciences, Medical Informatics and Technology Tirol, Austria
UoS	University of Sheffield
US	United States
V5FU2-CDDP	5-FU + folinic acid + cisplatin
VARA	Veterans' Affairs Rheumatoid Arthritis (Registry)
VEGFis	Vascular endothelial growth factor inhibitor(s)
YMRS	Young Mania Rating Scale

Chapter 1 Thesis introduction and background

1.1 Chapter overview

This opening chapter lays out the structure of the thesis, spotlighting the main theme—addressing the challenges of evaluating treatment sequences in Health Technology Assessment (HTA)—and research motivation, inspired by a gap in the field and personal interest. This thesis investigates alternative methods for addressing the scarcity of clinical data evidence in HTA treatment sequence evaluations. It begins by defining what constitutes treatment sequences (Section 1.2), proceeds to discuss the significance and implications of evaluating treatment sequences in HTA and the associated challenges (Section 1.3.1 & 1.3.2), and explains the critical issue of clinical effectiveness data scarcity (Section 1.3.3). I then highlight the potential of real-world data (RWD) as an alternative solution for providing unbiased estimates of the comparative effectiveness of treatment sequences, while emphasising the need to investigate the application of causal inference methods to navigate the inherent biases in RWD analysis, especially confounding bias (Section 1.3.4).

Section 1.4 outlines foundational concepts, such as causal inference, confounding, and the distinction between estimand, estimators and estimates. Importantly, I introduce the Target Trial Emulation (TTE) framework[3], a causal inference methodology gaining increasingly widespread recognition, developed by notable figures in the field, Jamie Robins and Miguel Hernán. This framework provides a systematic approach to structuring RWD studies in a way that aids in bias mitigation, which I adopted as the backbone of my proof-of-concept case studies in Chapters 7 & 8. Section 1.5 outlines the research questions and specific aims designed to address data scarcity challenges by leveraging RWD and causal inference methods. Finally, Section 1.6 presents the thesis structure, with each part arranged to systematically address the posed research questions.

1.2 Definition of treatment sequences

With the increasing variety of treatment options, it is common for patients to undergo multiple treatments in managing a single condition. Treatment sequences refer to the ordered series of medical interventions patients receive. The criteria for initiating a subsequent treatment often include, but are not limited to, signs of disease progression and the occurrence of toxicity, which vary widely depending on the disease, such as exceeding certain lab test thresholds, changes in disease severity scores, imaging results, clinical assessments, or after completing a set duration of first-line treatment. Treatment sequences typically embody a form of dynamic treatment regimen (DTR)[4] (e.g. transitioning to a next-line drug depending on patient response) as opposed to time-related static strategies (e.g. two COVID vaccines six months apart) and static strategies (e.g. a single flu shot).[5] I further defined the subtle differences between different types of treatment sequences and their implications for HTA in Chapters 2 and 4.

Treatment sequences may also include non-medical patient-level interventions, such as lifestyle changes to modify behaviours[6,7] (e.g., using nicotine replacement or switching to vaping for smoking cessation), or sequential population-level policies (e.g. progressively implemented alcohol levies), especially within a broader public health context. However, my thesis narrows its scope to medical treatment sequences in HTA, focusing on patient-level interventions to manage the complexity in the methodological exploration of treatment sequence evaluation in HTA, as outlined in Sections 1.3-1.5. Nonetheless, I will discuss the potential broader implications of my research at the end of the thesis (Chapter 9).

1.3 Research motivations: treatment sequences in HTA and their challenges

1.3.1 Why are treatment sequences important?

Incorporating treatment sequences into decision-making becomes increasingly relevant in HTA to ensure effective health resource allocation as altering the order of treatments may impact their overall effectiveness and costs of managing a disease.[1,2,8-10] Health economic evaluations play a key role in supporting reimbursement decisions[11-17] by assessing the cost-effectiveness of introducing a new treatment to the market.[18,19] It is generally accepted that when an intervention can impact long-term outcomes or survival, an economic evaluation should consider a lifetime horizon to evaluate the appraised treatment's cost-effectiveness against specific comparators and for target populations in a given decision problem.[20] Consequently, such an evaluation should account for, whether implicitly or explicitly, the care pathway along with the associated costs and outcomes that follow the intervention being appraised. This often leads to the development of treatment-sequencing decision-analytic models and identification of pertinent effectiveness, costs, and utility evidence for these models to project long-term impacts of integrating the assessed treatment into existing treatment pathways.[1]

Increased attention to treatment sequences in HTA[8,21,22] prompted the National Institute for Health and Care Excellence (NICE) to update their guidelines with generic recommendations for integrating subsequent treatments into technology appraisals (TA) in England and Wales since 2013.[23,24] This shift encourages assessments beyond isolated treatments at discrete times. Similarly, guidelines from the Canadian Agency for Drugs and Technologies in Health (CADTH) adopted comparable recommendations.[25] NICE and CADTH stand at the forefront of global HTA, with NICE's recommendations notably impacting clinical practice in England and Wales due to their legal mandate for implementation, distinguishing it from other nations' HTA bodies.[26] The adoption of the aforementioned forward-thinking guidelines reflects their pragmatic approach to meeting HTA's changing demands, which, in turn, may also drive the changing demands. For example, NICE is piloting the Pathway Approach to streamline the appraisal of new treatments by evaluating treatments in a single disease area with one economic model, aiming to improve efficiency beyond the

conventional single (STA) and multiple technology appraisals (MTA).[27] STA and MTA assess single and multiple technologies at specific treatment positions, respectively. The Pathway Approach expands this by assessing various treatments across different positions within the same pathway in a single review, seeking to address the growing number of new treatments and the limited capacity to evaluate them, thereby facilitating quicker access to new medicines while ensuring they are cost-effective. NICE plans to integrate lessons learned from the pilots, including renal cell carcinoma (RCC) and non-small-cell lung cancer (NSCLC) studies, into the STA process, with further updates expected later in 2024[27-29], possibly shaping future HTA treatment sequence evaluations.

1.3.2 What are the challenges of evaluating treatment sequences?

Evaluating treatment sequences adds complexity to the health economic evaluation process compared to assessing isolated treatments.[1] The core of such complexity lies in the requirement for economic models to accommodate an expanded number of states to represent multiple lines of treatments (LOTs) and real treatment pathways/sequences. Conceptually, this not only demands additional parameters to better reflect the model but also requires extra data to inform these parameters and accurately capture key interactions between effects within the modelled evaluation. While models that better reflect real treatment pathways may reduce some uncertainties, the increase in model parameters can, paradoxically, introduce new uncertainties. Moreover, the increased complexity can lead to a higher computational load.[30,31] Decision-makers, therefore, may encounter challenges in choosing the most pertinent sequences for consideration from all theoretically possible sequences.

The scarcity of pertinent clinical evidence, especially randomised controlled trials (RCTs) that compare treatment sequences, stood out as a prominent challenge among others in treatment sequence evaluation at the outset of my doctoral studies. This challenge was highlighted through my experience of utilising multiple sources of evidence to populate the effectiveness of treatment sequences in a lung cancer treatment economic model, which led to concerns from the HTA body about potential double-counting of death rates due to evidence merging. Despite established methods for integrating treatment sequences into the structure of various types of health economic models[1,2], such as Markov models, discrete event simulations (DES), partitioned survival models (PartSM), and the development of others (e.g., microsimulations, multi-state models)[32,33], obtaining unbiased clinical effectiveness estimates of treatment sequences for these models proves challenging.[34,35] The full scope of this issue, its resolutions, and its connection to other treatment sequencing challenges were underexplored. Thus, my doctoral research began by mapping out the array of challenges in evaluating treatment sequences within HTA, identifying how the challenge of scarce clinical evidence fits into this broader picture (Chapter 2), and examining strategies and extent to which they have been employed to derive

unbiased clinical effectiveness of treatment sequences in English HTA (Chapter 3).

1.3.3 Scarcity of clinical effectiveness evidence for treatment sequences

Several studies have sought to establish a conceptual framework for modelling treatment sequences (see Chapter 2's review for details).[1,34-36] Zheng et al. proposed a generic treatment-sequencing economic modelling framework[1], while Lewis et al. and Huang et al. focused on the challenges of scant clinical effectiveness evidence for treatment sequences in HTA and suggested room for future research.[34-36] The latter two studies were published in parallel to my thesis work, underscoring the progressive advancements in this area and aligning with the central theme my thesis. They recommended leveraging statistical methods for mitigating biases when collating evidence from multiple sources to obtain LOT-specific effectiveness. This includes employing statistical adjustment techniques that account for line-specific effects and indirect comparisons, such as meta-regression and matched-adjusted indirect comparisons (MAIC).[34,35] These strategies provide a pragmatic approach to harness existing evidence, especially in situations where clinical trials, which typically focus on comparing treatment effects for a specific line, are the primary source of evidence, as RWD for the technology under review might not be readily available. Nonetheless, such an approach often relies on critical assumptions that evidence across different LOTs are for comparable populations.

For example, in evaluating treatment sequences A→B versus C→D, consider a situation where first-line treatment effectiveness comes from an RCT comparing A versus C in a first-line setting, while second-line effectiveness is sourced from another RCT comparing B versus D in a later-line setting. Biases can emerge if the second trial's patient population does not accurately represent those progressed in the first trial to receive subsequent treatments, specifically if participants in the second trial have undergone inconsistent lines of previous treatments or received a mixture of different primary treatments before the trial. Attempting to isolate and analyse a subgroup from the second trial—those who only received one prior line of specific treatment (i.e., first-line A (before receiving B) versus first-line C (before receiving D))—risks compromising the trial's original randomisation, leading to biased results without proper statistical adjustments. This simplified example highlights the inherent uncertainties of using LOT-specific evidence and underscores the necessity of cautious assumption-making, such as the impact of any other previous treatments on the effectiveness of B and D being comparable to A and C, or that the effects of B and D remain stable regardless of the number of previous treatment lines. This approach is pragmatic when the assumptions are valid, but these assumptions may not always hold, particularly for survival outcomes. Using confidence intervals from the second trial's estimates as the basis for probabilistic sensitivity analysis will not fully account for the aforementioned uncertainties, given their unknown directions and magnitudes.

1.3.4 Opportunities for real-world data and causal inference methods

Studies have highlighted the capability of RWD to capture treatment sequencing information for HTA, particularly its potential for deriving estimates of the effects of entire treatment sequences rather than LOT-specific effects.[1,34,35] Additionally, RWD offers further benefits such as larger sample sizes and greater generalisability than limited trial populations. However, these studies did not explore specific (statistical) methodologies capable of generating unbiased comparative effectiveness estimates from RWD for treatment sequence comparisons. A notable hurdle in utilising RWD is its susceptibility to confounding[37,38], a challenge that is exacerbated by treatment sequences due to time-varying confounding. Put simply, the lack of random treatment assignment in real-world practice and the tendency for sicker patients to receive particular first-line and subsequent treatments (i.e., treatment sequences) over other treatments suggest directly comparing patients receiving different treatment sequences in RWD can introduce bias, as the groups compared may not be comparable. The presence of time-varying confounding introduces bias in treatment comparisons and cannot be adequately addressed by simple statistical methods, such as multivariate regression models for outcome analysis that simply incorporate time-varying confounders into the regression model, which may in fact introduce additional biases. The need for advanced statistical approaches to tackle this issue is explained in detail in Section 1.4. The use of RWD in health economic evaluations has been explored in various contexts but not specifically for treatment sequences.[39,40] Chapter 3's review of NICE TAs, hence, reveals RWD's use in treatment sequencing HTA and underscores a gap in methods for addressing biases in effectiveness estimates. My research, therefore, centres on developing methodologies to leverage RWD for generating unbiased estimates in treatment sequence comparisons for HTA, covering strategies for assessing fit-for-purpose data sources (Chapter 5), minimising biases in RWD study design (Chapter 6 & 7), employing relevant statistical methods (Chapter 4), and the testing and validation of these approaches (Chapter 8). This focus intends to offer an alternative approach, addressing the lack of clinical evidence on treatment sequences in treatment-sequencing models, where relying solely on LOT-specific evidence may fall short.

In this thesis, "effectiveness"—the real-world effect of treatments—is consistently used to discuss the effects of treatments for simplicity, contrasting with "efficacy", which denotes effects measured under highly controlled-trial conditions.[41] This choice reflects HTA's focus on assessing how new treatments affects real-world clinical practice. Further, although clinical trial data are frequently used to inform treatment effects in economic evaluations, but a single trial is typically not the only evidence source for evaluating treatment sequences. My exploration of RWD led to adopting "effectiveness" and "comparative effectiveness". These terms are associated with comparative effectiveness research (CER), aimed at understanding treatments' relative effects in real-world studies, while sometimes also broadly cover assessments from both RWD and efficacy trials.[42]

Early in my PhD, a course on treatment switching in clinical trials, co-taught by my supervisor Professor Nick Latimer with colleagues at the University for Health Sciences, Medical Informatics and Technology Tirol (UMiT) introduced me to the critical role of causal inference methods in addressing time-varying confounding, shaping my research interests towards the integration of these methodologies within the aforementioned context. Firstly, the shared foundation of causal inference and HTA in assessing counterfactual scenarios to understand the impact of different interventions renders leveraging this method intuitively sensible. Specifically, HTA looks at the effects of introducing a new treatment into the health system and comparing it with not doing so. Further, the parallels between clinical trials with non-randomised unintended treatment switching and treatment sequences sparked my interest in the potential for transferring similarly rooted statistical methods from the former to the latter. In particular, a clinical trial comparing an active drug against a placebo can be seen as evaluating the sequences of active drug → active drug versus placebo → placebo. Thus, undesired switching to the active drug within the placebo group can "contaminate" the intended comparison by unintentionally partially including the effect of a treatment sequence of placebo → active drug. In HTA, statistical methods grounded in causal inference have been adapted and well-established to address unwanted treatment switching in clinical trials, enabling estimation of the placebo arm's effect as if no switching had occurred.[43] A recent NICE Decision Support Unit (DSU) review identified in Chapter 2 also notes the similarities between unwanted treatment switching and the economic modelling of treatment sequences, and therefore suggests the potential for co-developing methods to tackle both issues.[44]

Despite HTA guidelines for analysing patient-level RWD with causal inference methods to derive unbiased treatment effects[40], they have not been accommodated in assessing the effects of treatment sequences (see Chapter 2). Informally, identifying the effect of a specific treatment sequence in real-world settings can be viewed as estimating its effect as though no other sequences or switching had occurred. This concept underpins my research in identifying and adapting relevant causal inference methods to assess treatment sequence effects using RWD. This task brings added complexity compared to handling unwanted treatment switching in clinical trials: Firstly, the diversity of RWD sources poses significant challenges in identifying data sources that can accurately capture the targeted treatment sequences, markedly diverging from the data collection in population-specific clinical trials. Further, the lack of randomisation in RWD may theoretically necessitate additional steps in deploying advanced statistical methods to obtain unbiased estimates. This contrasts with approaches for addressing unwanted treatment sequences in trials, where operations may sometimes hinge on existing randomisation, such as the Rank Preserving Structural Failure Time Model (RPSFTM).[43] Finally, the comparison of two treatment sequences in RWD necessitates adjustments in both arms, contrasting with the singular arm adjustment for tackling unwanted treatment switching

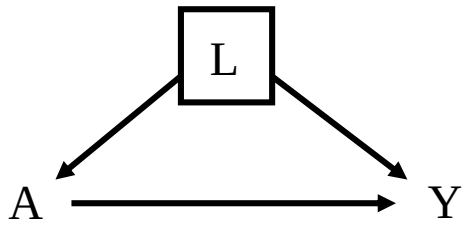
in the placebo arm of clinical trials.[43]

Hence, my doctoral journey has been dedicated to navigating these complexities, assessing the feasibility of using RWD to derive reliable treatment sequence effect estimates (i.e. reliable real-world evidence (RWE)), and suggesting pathways forward. In the upcoming section (Section 1.4), I describe the concept of causal inference in the context of treatment sequencing, highlighting how statistical methods guided by causal inference are crucial in tackling confounding, particularly the need for advanced methods adept at tackling time-varying confounding. Besides exploring statistical methods suitable for treatment sequence effect estimation, I describe how I came to recognise the relevance of the Target Trial Emulation (TTE) framework in RWD study design, subsequently adopting it as the backbone of my case studies. The causal inference-based Target Trial Emulation (TTE) framework[3] is gaining recognition in epidemiology in recent years for its role in minimising biases in RWD study design and execution through structured planning and explicit reporting, recently endorsed by NICE in its real-world evidence (RWE) guideline for HTA.[37]

1.4 Harnessing causal inference for analysing RWD to derive unbiased treatment sequence effect estimates: key concepts

1.4.1 Causal inference and confounding in treatment sequencing

Causal inference seeks to ascertain cause-and-effect relationships between variables beyond associations. The development of causal inference spans various disciplines[45-49], yet my research, with its emphasis on HTA, primarily adopts terminology from epidemiology and statistics for its relevance[5,50-54]. While unadjusted RWD analyses can shed light on local real-world treatment patterns and outcomes[55-57], they cannot answer "what-if" (i.e., causal) questions involving counterfactuals, such as whether a patient group would have experienced better outcomes had they been offered an alternative treatment (sequence) due to confounding. However, answering these counterfactual queries is crucial for TAs that compare different treatment (sequences). Figure 1.1's directed acyclic graph (DAG) illustrates the confounding issue, showing how treatment A's independent (i.e., causal) effect on outcome Y(A → Y) is confounded by a set of variables (L) due to non-randomisation. For example, sicker patients ($L_{\text{intensive care unit stay}} = 1$) are more likely to receive treatment A (A = 1) and also more prone to developing an outcome (Y = 1). Hence, comparing patients who underwent treatment A (A = 1) to those who did not (A = 0) without adjustments would be "unfair" as these two patient groups are not comparable. Adjusting for confounders L allows for the identification of treatment A's causal effect on outcome Y, assuming there are no unmeasured confounders.



A: Single-time treatment
(e.g. antibiotic treatments)

L: Confounders
(e.g. age, sex, intensive care unit stay)

Y: Outcome
(e.g. 7-day mortality)

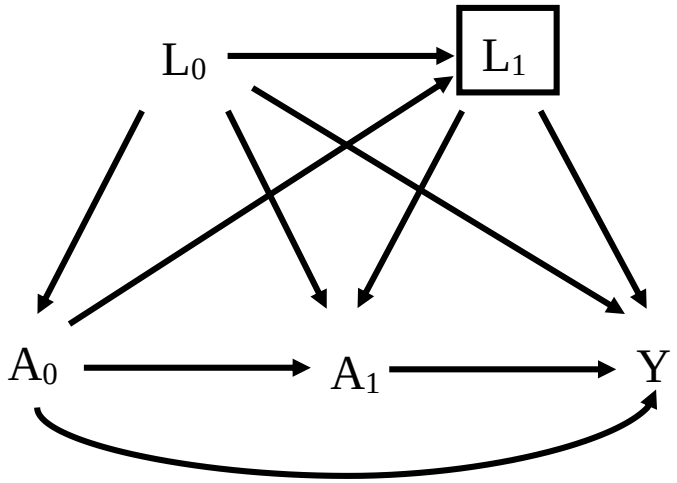
Figure 1.1 Directed acyclic graph (DAG) of an observational study involving a single-time exposure and confounders

The black box around L signifies the conventional method of adjusting for confounding by controlling for confounder L through blocking the backdoor path from Y to A (i.e., Y←L→A).

The equation $Y = \beta_0 + \beta_1 A + \beta_2 L + \varepsilon$ represents a conventional method to adjust for confounding in Figure 1.1's scenario, known as outcome regression, which can be either univariate or more often multivariate depending on the number of confounders. Here, β_0 is the intercept, β_1 and β_2 are coefficients for the effect of treatment and confounders L, respectively, and ε is the error term. This approach assesses the independent effect of treatment A on outcome Y, conditioned on a specific level of L, separating A's impact on Y from L's influence. However, this approach falls short in assessing treatments strategies that may change over time, due to its inability to handle time-varying confounders.[38,58]

Figure 1.2's DAG highlights the complexity of addressing time-varying confounders, specifically in evaluating the causal effects of treatment sequences, where the sequence involves in an initial treatment $A_{t=0} \in (0, 1)$ at time zero ($t = 0$) and a subsequent treatment $A_{t=1} \in (0, 1)$ upon disease progression (i.e., $t = 1$). The aim of evaluating the impact of a specific sequence is to estimate the joint effect of A_0 and A_1 on outcome Y. An example of this is to estimate the effect of the treatment strategy where patients move from an active first-line treatment $A_0 = 1$ to palliative care $A_1 = 0$. Time-varying confounders, such as L_1 (e.g. performance status upon disease progression), can influence both subsequent treatment choices A_1 (i.e., second-line treatment) and outcomes Y, while also being influenced by previous treatments A_0 (i.e., first-line treatment). Including time-varying confounders and treatments in an outcome regression model (i.e., $Y = \beta_0 + \beta_1 A_0 + \beta_2 L_0 + \beta_3 A_1 + \beta_4 L_1 + \varepsilon$) complicates isolating the causal effect of a specific treatment sequence. Specifically, the inclusion of L_1, an intermediate outcome influenced by A_0, introduces bias by adjusting for A_0's indirect effect (mediated through L_1) on Y, conditioning on a specific level of L_1. However, including only L_0 would neglect the time-varying confounding of L_1, which is crucial for understanding the relationship between A_1 and Y. Further, the equation overlooks A_0 and A_1's joint effect. This complicates interpreting the model's results and discerning the true causal impact of a treatment sequence,

underscoring the limitations of conventional methods in handling time-varying confounding and exposures. Hence, advanced statistical methods are required.



A_0: treatment at baseline (time 0)
(e.g. first-line cancer treatment)

A_1: treatment at time 1
(e.g. second-line cancer treatment)

L_0: Confounders at baseline (time 0)
(e.g. performance status, weight)

L_1: Confounders at time 1
(e.g. performance status upon progression, weight upon progression)

Y: Outcome
(e.g. Death)

Figure 1.2 A simple DAG of time-varying confounders and exposures in observational studies

The black box around time-varying confounder L_1 illustrates the conventional method of adjusting for confounding, specifically the backdoor path between A_1 and Y, by controlling for confounder L_1. This conceptually blocks the pathway through which A_0's effect on Y is mediated by L1, since as L_1 is an intermediate outcome that follows A_0.

Historically in epidemiology and medicine, there was skepticism towards deriving causal inferences from observational data.[59] Robins's g-methods for dealing with time-varying confounding revolutionised this perspective in the past 40 years.[54] These methods were rooted in the Neyman-Rubin potential (i.e. counterfactual) outcomes framework.[60] This framework, initially conceptualised by Neyman in the 1920s for randomised experiments[61,62] and later expanded by Rubin in 1970s for observational studies to address confounding[63-65], distinguishes between the outcome that we observe and what could have potentially occurred to determine the causal effect of interventions. The terminology of the "(Neyman-)Rubin causal model" was solidified by Holland in 1986.[50] Building upon the aforementioned principles, Robins introduced the initial version of g-methods for tackling time-varying confounding in his 1986 paper.[54] When introduced, the depth and implications of these methods were not fully grasped by the community. However, many contemporary causal inference methods in epidemiology may trace their origins and relevance to the principles in this paper. Robins and his collaborators have since expanded upon these principles, developing variants of g-methods to address time-dependent confounding with measured variables (e.g., g-formula[54], g-estimation of structural nested models[66,67], marginal structural models (MSMs) with inverse probability weighting (IPW)[58,68,69]). Chapter 4's systematic review aimed to identify relevant causal inference guided statistical methods or their extensions used to address time-varying confounding in the context of

estimating treatment sequence effects. In fact, it revealed that most advanced statistical methods are rooted in or relevant to g-methods. Robins noted that g-methods are not independent of one another, but interconnected across different subsets.[60] He highlighted the pivotal use of g-methods within the increasingly popular TTE framework, which he co-developed with Miguel Hernán, for discerning causal effects in observational studies. The significance of TTE is further discussed in the following sections.

1.4.2 Target Trial Emulation (TTE)

The TTE framework is a methodological framework for using "big data" (i.e., RWD, also known as observational data) to evaluate comparative effectiveness.[3] Specifically, the structure of the TTE seeks to mimic the design of a hypothetical RCT using observational data. This approach is particularly relevant to derive reliable RWE when RCTs are not feasible, ethical, or practical. The emulation of a hypothetical Target Trial (as if an identical actual RCT were possible) helps reduce biases from RWD study design and improve causal inference, making it an important tool in CER.

A standard Target Trial protocol consists of seven key components to resemble a RCT setting: eligibility criteria, intervention strategies being compared, intervention assignment, follow-up period, outcomes of interest, causal contrasts of interest, and analysis plan. The analysis plan outlines the statistical methods, such as those for addressing time-varying confounding, to derive the *estimands* of interest (what constitutes an estimand is further explained in Section 1.4.3). Briefly, an estimand is the precise description of the parameter aimed to be estimated (e.g. treatment effect) in a statistical analysis.

I became aware of the TTE framework's relevance to my research from an Issue Panel at the Professional Society for Health Economics and Outcomes Research (ISPOR) 2020 virtual conference.[70] Hernán, the author of the TTE framework, discussed its capability to facilitate the derivation of "regulatory-grade RWE", enabling reliable direct comparisons of treatments through TTE with RWD, and contrasted with network meta-analysis (NMA), which relies on indirect comparisons by aggregating evidence from multiple trials. Concurrently, my supervisor, Professor Nick Latimer, was developing TTE protocols for pancreatic cancer research, reinforcing its relevance and providing a valuable reference for my research to harness the TTE framework. This led me to anchor the design and validation strategy of my RWD case studies in the TTE framework (Chapters 6 & 7), and to assess the compatibility of candidate advanced statistical methods with Target Trial emulation in my statistical methods review (Chapter 4). Subsequently, I had the opportunity to attend summer courses on Target Trial Emulation (2022) and Advanced Confounding Adjustment (2023) at the Harvard University. These provided me with hands-on knowledge from methods' developers, aiding in the practical implementation of my case studies (Chapter 8).

Technically, well-designed and executed epidemiological cohort studies predating the TTE framework's publication may also be seen as TTE, even without explicit labelling.[71] The TTE framework, however, adds value by systematically defining the research question, thereby standardising the structure of RWD studies and improving their transparency and rigor.[72] Moreover, while Target Trials are typically well-specified cohort studies, a study by Dickerman et al. has shown that the TTE framework can also be useful in reducing bias in case-control study designs.[73]

The explicit specification and precise implementation of TTE help prevent biases inherent in RWD designs, beyond what statistical methods alone can address, such as immortal time and selection biases.[74,75] These biases are well-known in observational studies involving a control group not receiving active treatment, such as surgery versus non-surgery scenarios.[75] Selecting surgical patients from RWD and following them up since surgery inherently carries selection bias, as these patients survived long enough to undergo surgery, thereby introducing an "immortal time" period. This bias challenges the identification of comparable control groups and the accurate timing of their start of follow-ups. Similarly, selecting patients based on treatment sequences received post-initial treatment contradicts the TTE's principles of mimicking RCT procedures, where enrolment should hinge on pre-treatment characteristics. This underscores the need to carefully define "Time Zero"[3,76,77] for assessing eligibility and initiating follow-up in a consistent manner across treatment groups when leveraging TTE to comparing treatment sequences.

A crucial element in employing TTE for comparing treatment sequences is the precise definition of the treatment strategy and causal contrast of interest. A precise treatment strategy defines the treatment or regimen a patient receives and the conditions under which they are still considered to be adhering to the assigned strategy, even when switching or discontinuing treatment (e.g., switching treatment due to intolerance or adverse events which would not be considered as a protocol violation). Imagine emulating a hypothetical RCT that randomises patients to receive different treatment sequences at the outset. One potential causal contrast is the hypothetical per-protocol (PP) effect had all patients fully adhered to their assigned treatment sequences. This differs from the PP effect in a scenario where some patients violate their assigned treatment sequences. In a conventional simple PP analysis, patients who deviate from the protocol are simply censored, and the effects of full adherence are not estimated. However, the hypothetical PP effect—where all patients fully adhered to their assigned treatment sequences—extrapolates to estimate the effects under full adherence. If an actual RCT exist comping the same treatment sequences without protocol violations existed, the hypothetical PP effect in the aforementioned emulated Target Trial would coincide with the intention-to-treat (ITT) effect observed in the real RCT.

Clearly defining what counts as a protocol deviation is crucial in comparing treatment sequences. For example, patients not completing the assigned treatment sequence due to a lack of progression

during the first-line treatment should not be seen as a deviation. Rather, it suggests that the initial treatment within the sequence effectively delayed the time to first-line progression, which is part of the effect of the assigned sequence. What is considered a protocol deviation largely depends on subject knowledge and the specific topic of interest. The detailed considerations of each TTE elements for comparing treatment sequences are outlined in my case study protocol and implementation (Chapters 7 & 8).

During the development of my thesis, NICE endorsed the TTE framework for deriving RWE for HTA in their recently published RWE framework.[37] Although the NICE RWE framework briefly acknowledged TTE's potential to inform the assessment of the effect of treatment sequences, it did not provide specific practical solutions, further highlighting the significance of my investigation. A recent UK health economists' publication highlights TTE's emerging role in HTA, suggesting prioritising RWD quality improvement and developing better analytical methods and guidance, starting with RWD calibration against RCTs (i.e., benchmarking).[78] Their recommendations reflect the investigative approach undertaken in my thesis (Chapter 4-8) and reinforce the rationale behind my proof-of-concept case study design in Chapter 7. The significance of exploring the TTE approach for comparing treatment sequences in relation to other existing TTE studies is further detailed in my scoping review (Chapter 2) and case study protocol (Chapter 7).

1.4.3 Estimands, estimators, and estimates

An *estimand* precisely outlines the parameter one seeks to estimate in an analysis.[79] It represents the target of inference that one wants to make about a population using a sample from that population. This contrasts with an *estimator*, which refers to the mathematical method used to approximate the value of the estimand, such as maximum-likelihood estimator (MLE) and Bayes estimators. An *estimate* is the numerical output derived from applying an estimator to a given dataset. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) provides definitions of these three terms in the context of clinical trials (Table 1.1).

Table 1.1 Glossary of the ICH-E9(R1) addendum[79] on estimands, estimator and estimates

Term	Definition
Estimand	"A precise description of the treatment effect reflecting the clinical question posed by the trial objective. It summarises at a population-level what the outcomes would be in the same patients under different treatment conditions being compared."
Estimate	"A numerical value computed by an estimator"
Estimator	"A method of analysis to compute an estimate of the estimand using clinical trial data"

In November 2019, the ICH updated their E9 guidance on statistical principles for clinical trials, introducing the addendum, ICH E9(R1).[79,80] This update brought forth specific guidelines on defining

estimands and sensitivity analyses, aimed at establishing a clear strategy for determining the treatment effect to be estimated in clinical trials. An estimand is detailed through five attributes: treatment, population, variable, intercurrent events, and a summary measure for the population. Emulating a Target Trial similarly requires the implicit definition of these attributes, reflecting the process of establishing (hypothetical) estimands. The process of specifying an estimand aligns the causal contrast with necessary statistical methods in the analysis to correctly obtain the estimand. Drawing a parallel, the NICE RWE framework signposts to ICH E9(R1)'s approach for handling intercurrent events post-treatment commencement in analysing treatment effect. Strategies for managing intercurrent events informally echo the management of the aforementioned "protocol violations" in TTE (Section 1.4.2), contrasting with ITT analyses that overlook these events.[37] A well-defined estimand is essential for mapping pertinent analytical methods to capture the causal contrast of interest, with its relevance in RWD studies also highlighted in recent research.[81] Hence, I leveraged the concept of estimands in Chapter 4 to assess statistical methods for managing time-varying confounding, examining the estimands generated by each studied method.

1.5 Research questions and specific aims

The scarcity of clinical evidence presents a significant challenge in modelling treatment sequences within HTA. Typically, clinical trials do not involve direct comparisons of treatment sequences. Existing literature that explores the synthesis of clinical evidence for treatment sequences recommends populating treatment-sequencing models with LOT-specific effectiveness evidence, achieved by pooling data from various trials across different therapy lines.[34,35] While pragmatic, this method frequently assumes a comparable patient population across evidence from different therapy lines, an assumption that may not always hold true, particularly when treatments influence survival outcomes. In that context, using statistical methods to adjust indirect comparisons (for evidence collation) or meta-analytic methods, such as meta-regression and subgroup meta-analysis (to account for patient treatment histories in approximating a LOT's effect), can aid in understanding each LOT's impacts, but may not guarantee unbiased estimates for the entire treatment sequence.

In contrast, RWD is adept at capturing entire treatment sequences but is susceptible to confounding biases. Although RWD has been used within HTA[40], its capacity to provide unbiased direct comparisons of treatment sequences has yet to be fully explored. Causal inference methodologies hold significant promise for providing valuable insights in this area, given their relevance to the counterfactual questions that are fundamental to HTA. Even though causal inference methods have been applied in HTA[40,43,44], its potential to facilitate the assessment of treatment sequence effectiveness remains largely untapped. The TTE framework[3], recently endorsed by the NICE RWE framework[37], emerges as a crucial tool for mitigating biases in RWD study designs. Yet,

its utilisation for comparing treatment sequences within HTA contexts has not been extensively investigated. Further, if unbiased direct comparisons of entire treatment sequences were feasible, the need for a separate sequencing model might be bypassed, opting instead for direct summary estimates and thus reducing structural uncertainties. Nonetheless, the extent to which similar approaches have been considered or applied within HTA remains unclear.

Therefore, my research aims to explore the potential and feasibility of using RWD in conjunction with causal inference methods to derive reliable (comparative) effectiveness estimates for treatment sequence comparisons, aiming to provide an alternative solution to the challenges posed by the scarcity of clinical evidence in decision-making for treatment sequences. The investigation is structured around the following guiding questions, with specific study aims designed to address each of these research questions:

(1) What does the landscape look like for evaluating treatment sequences in HTA, and how does the challenge of clinical evidence scarcity fit within the wider array of related challenges?
   - Aim 1: Review methodological literature on evaluating treatment sequences in HTA, summarising challenges and trends in the topic's evolution.

(2) How have treatment sequences been evaluated in English HTA practices, particularly the prevalence of evaluations involving treatment sequences and approaches used for economic evaluation? How have RWD been utilised to support the evaluation of treatment sequences compared to other data sources, and to what extent?
   - Aim 2: Review the prevalence and nature of discussions about treatment sequences in NICE TAs, including modelling approaches, selection and derivation of clinical effectiveness evidence, and the challenges associated with different evidence sources, particularly comparing RWD with other data sources.

(3) How can we leverage RWD and causal inference methods to obtain reliable comparative effectiveness estimates for treatment sequence comparisons in HTA?
   - Aim 3: Review of causal inference methods from broader literature (e.g. statistics, epidemiology) that have been used or could be applied to analyse RWD for treatment sequence comparisons
   - Aim 4: Evaluate RWD sources suitable for treatment sequence comparisons for HTA, assessing their suitability for implementing the identified causal inference methods in Aim 3
   - Aim 5: Design and conduct proof-of-concept case studies using Target Trial Emulation as a backbone, applying the statistical methods and RWD sources identified in Aim 3 & 4 for comparing treatment sequences, and then compare these results against existing RCTs that compared treatment sequences.
   - Aim 6: Summarise findings and limitations in the case studies in Aim 5, and develop a

framework or tutorial on how to use RWD for evaluating the effectiveness of treatment sequences in HTA, incorporating relevant examples as applicable.

1.6 Thesis structure

The thesis is organised into five parts, encompassing nine chapters, with their interconnections depicted in Figure 1.3. This figure employs arrows to demonstrate how insights from each part informs the ones that follow. I briefly explain the purpose of each chapter and highlight their relevance to my research questions and objectives in Section 1.5.

In Part I, I set the stage for the thesis by introducing the concept of treatment sequencing and its evaluation challenges within HTA. Earlier in this chapter, I outlined the thesis's motivation and the recurring theme: the challenges posed by the scarcity of clinical evidence in treatment sequence evaluation. I have briefly discussed how RWD could serve as a promising alternative, despite its inherent challenges, underscoring the potential of leveraging causal inference methods, including TTE and relevant statistical methods, in informing unbiased estimates for treatment sequence comparisons. To guide this investigation, a set of research questions and specific aims were previously detailed in Section 1.5 in the current Chapter. Following this, Chapter 2 examines methodological advances in HTA treatment sequencing literature (Aim 1), exploring how the aforementioned issue intersects with a broader array of challenges, thereby addressing my research question (1).

Part II aims to gauge the significance of the clinical evidence scarcity challenge in treatment sequence evaluation and how it has been approached in English practice. To address my research question (2), Chapter 3 features a systematic review of NICE TAs (Aim 2), focusing on exploring the frequency of treatment sequencing discussions and the manner and extent to which RWD has been used to inform relevant decisions in comparison with other data sources, and its interaction with other challenges.

Parts III to IV collectively tackle the final research question: How can we utilise RWD to obtain reliable effectiveness estimates for treatment sequence comparisons in HTA? In Part III, the thesis investigates solutions, focusing on how biases inherent in RWD may be mitigated through causal inference methods. Hence, Chapter 4 undertakes a systematic review of causal inference-guided statistical methods from other disciplines, primarily statistics and epidemiology, to explore methods potentially capable of deriving unbiased estimates in comparing treatment sequences using RWD, and which can work under the TTE framework (Aim 3). Chapter 5 then explores suitable RWD sources for treatment sequence comparisons, identifying candidate databases for case studies to showcase RWD application (Aim 4). These databases are evaluated for their potential to inform treatment sequencing decisions in English HTA and to accommodate advanced statistical methods identified in Chapter 4. This chapter also addresses the challenges of accessing these databases,

especially from the perspective of an early career researcher. To design case studies using RWD for unbiased treatment sequence comparisons in a manner that allows them to be validated, Chapter 6 reviews RCTs that compare treatment sequences, so they can be used as benchmarks to inform the design of my case studies (Aim 5).

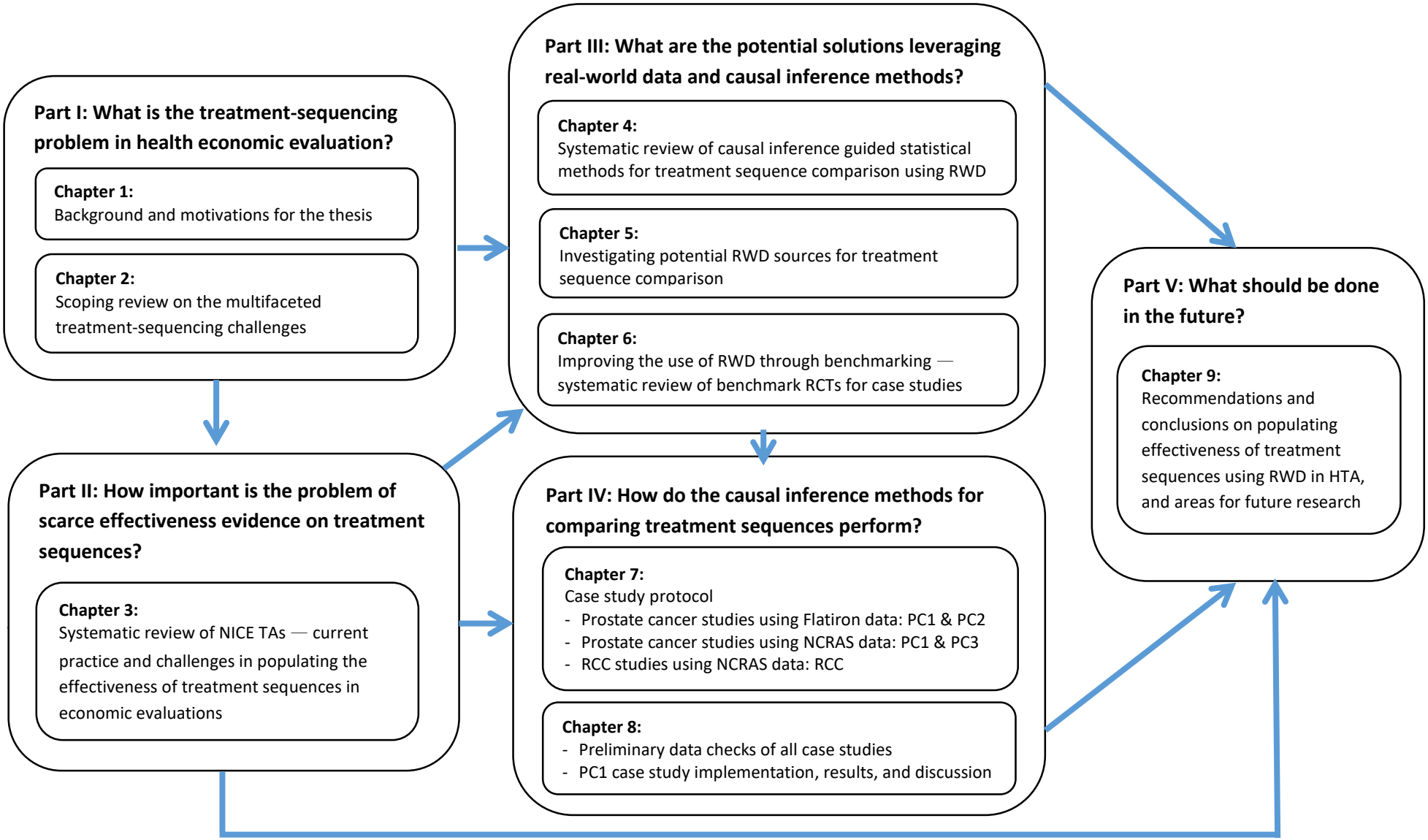


Figure 1.3 Schematic representation of the thesis structure

HTA: health technology assessment; NCRAS: National Cancer Registration and Analysis Service; NICE: The National Institute for Health and Care Excellence; PC: prostate cancer; RCC; renal cell carcinoma; RCT: randomised controlled trial; RWD: real-world data; TA: technology appraisal

The arrows between chapters indicate that one chapter's content leads to the next, with mediation through intermediate sections.

Chapters 5 and 6 are closely linked, starting with Chapter 5's search for pertinent English databases, which discovered the National Cancer Registration and Analysis Service (NCRAS) in England as a pivotal resource.[82] This discovery informed Chapter 6's initial focus of identifying benchmarks suitable for conducting proof-of-concept case studies with NCRAS. However, the absence of suitable benchmarks leads back to Chapter 5's exploration of alternative data sources, extending to international databases akin to the English context. This led to the identification of Flatiron Health database[83] in the United States (US) for direct proof-of-concept studies, which also enables an indirect proof-of-concept study with NCRAS to be conducted. Further, NICE's recent partnership with Flatiron Health to explore the potential of using early access data from the US to inform NICE HTA in oncology highlights its relevance.[83] The interweaving process of selecting candidate databases and benchmark RCTs are detailed in Chapters 5 and 6.

Part IV assesses the performance of causal inference methods for comparing treatment sequences using RWD, importantly addressing not only the methods' applicability but also the adequacy of common RWD datasets (i.e. whether data are of sufficient quality for methods to be successfully applied). It features Chapter 7, which introduces the protocol for a series of TTE case studies (Aim 5), building on insights from Chapters 4, 5, and 6. The first prostate cancer TTE case study (PC1) serves as a direct proof-of-concept study aimed at replicating the GUTG-001 trial[84] (i.e. the benchmark) to compare two treatment sequences in prostate cancer using Flatiron Health data[85], with validation against the benchmark. The execution of PC2 hinges on PC1's success, which expands to compare two alternative treatment sequences that are common in both US and English practice, using Flatiron data. PC2 functions as an exploration of the method's performance in a broader population (i.e., generalisability) and acts as a "bridge" to PC3. PC3 aims to use English NCRAS data to replicate the comparison made in PC2, serving as an indirect proof-of-concept study for the English population. Finally, s separate renal cell carcinoma (RCC) case study was designed to use NCRAS data to replicate a specific treatment sequence from one arm of the RECORD-3 trial[86], serving as a direct, single-arm proof-of-concept study for the English population. Chapter 8 details the implementation and findings of the first case study (PC1) in Chapter 7's protocol. Due to time constraints and the scope of the project, the impact of the COVID-19 pandemic, and delays in data access (explained in Chapter 5)—a significant challenge in using RWD—the remaining case studies are intended to continue under a Wellcome Trust transition fund at the University of Sheffield post-PhD. Despite this, to fully convey the scope and intertwined significance of the case studies, illustrating their collective importance, Chapter 7 retained a comprehensive overview of all studies.

Part V concludes the thesis and outlines future directions. Chapter 9 summarises the contributions from each chapter, addresses the study's limitations, and provides recommendations for using RWD in HTA treatment sequence effectiveness evaluations, drawing from case study insights

(Aim 6). It proposes a preliminary framework, essentially a tutorial, for applying RWD in evaluating treatment sequence effectiveness in HTA. This framework covers initial evaluations of a specific source of RWD's capabilities, approaches to designing RWD studies, analytic approaches to estimate sequence effectiveness, and identifies potential areas for enhancements in data collection. This fulfils the thesis's main objective: exploring whether and how RWD in conjunction with causal inference methods can serve as an alternative means for producing unbiased treatment sequence comparisons, thus supporting HTA decisions in the face of data scarcity. Finally, the chapter offers directions for future research to build upon and address the limitations of my work.

Chapter 2 Treatment sequences in health technology assessments: a scoping review

2.1 Chapter overview (Chang et al. (2024) in *PharmacoEconomics*)

This chapter presents a scoping review aimed at identifying pivotal publications that investigate treatment sequences in health technology assessments (HTA). It focuses on consolidating literature that addresses methodological and conceptual advancements in comparing treatment sequences, setting the stage for the broader scope of the thesis.

Despite recognising the scarcity of clinical evidence as a major challenge in evaluating treatment sequences and the potential for leveraging real-world data (RWD) in this context (Chapter 1), formulating specific strategies to develop relevant methodologies to aid decision-making necessitates a comprehensive insight into the current state of research. This entails a systematic examination of how this issue has been addressed, identifying limitations of current approaches, the overall importance of this issue and its interplay with other related challenges. This chapter's review was therefore designed for such exploration to refine my research objectives at the early stage of my PhD, thereby sharpening the research focus and shaping the strategic direction for my subsequent PhD work, including determining whether it was necessary to undertake further reviews, and informing the design of my case studies, to ensure I could achieve my research goals (i.e., developing strategies for unbiased treatment sequence effect estimation) (Chapter 1).

This chapter was published as an article in *PharmacoEconomics* on 1st April 2024 [Epub ahead of print] by Springer Nature.[87] The publication terms grant the author the right to include the article in a thesis they have authored. This article is reproduced with permission of Springer Nature. The following sections present the version accepted for publication (Section 2.2 to 2.6), which includes background on evaluating treatment sequences in HTA, partially overlapping with Chapter 1's content. The review's structure is detailed in the introduction (Section 2.2). Article numbering has been adapted to align with the thesis format and the article's bibliography is integrated into the thesis bibliography. A summary at the end of the chapter highlights the review's role within the thesis (Section 2.7). Details on publication declarations can be found in the online article.[87]

Title:

Challenges and Opportunities in Interdisciplinary Research and Real-World Data for Treatment Sequences in Health Technology Assessments

Jen-Yu Amy Chang[1] (ORCID: 0000-0001-6660-5246); James B. Chilcott[1] (ORCID: 0000-0003-1231-7817); Nicholas R. Latimer[1,2] (ORCID: 0000-0001-5304-5585)

[1]Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom
[2]Delta Hat Limited, United Kingdom

Contact information of corresponding author:
Jen-Yu Amy Chang, MSc, RPh
Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom
Address: Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK
Tel: +44 (0) 114 222 5454
Fax: +44 (0) 114 222 0749
E-mail: jy.a.chang@sheffield.ac.uk

Running title: Treatment Sequences in HTA: Challenges and Opportunities

Keywords: treatment sequences, health economics, economic evaluation, interdisciplinary research, real-world data, causal inference

Abstract

With an ever-increasing number of treatment options, the assessment of treatment sequences has become crucial in health technology assessment (HTA). This review systematically explores the multifaceted challenges inherent in evaluating sequences, delving into their interplay and nuances that go beyond economic model structures. We synthesised a "roadmap" of literature from key methodological studies, highlighting the evolution of recent advances and emerging research themes. These insights were compared against HTA guidelines to identify potential avenues for future research.

Our findings reveal a spectrum of challenges in sequence evaluation, encompassing selecting appropriate decision-analytic modelling approaches and comparators, deriving appropriate clinical effectiveness evidence in the face of data scarcity, scrutinising effectiveness assumptions and statistical adjustments, considering treatment displacement, and optimising model computations. Integrating methodologies from diverse disciplines—statistics, epidemiology, causal inference, operational research and computer science—has demonstrated promise in addressing these challenges. An updated review of application studies is warranted to provide detailed insights into the extent and manner in which these methodologies have been implemented.

Data scarcity on the effectiveness of treatment sequences emerged as a dominant concern, especially because treatment sequences are rarely compared in clinical trials. Real-world data (RWD) provide an alternative means for capturing evidence on effectiveness and future research should prioritise harnessing causal inference methods, particularly Target Trial Emulation, to evaluate treatment sequence effectiveness using RWD. This approach is also adaptable for analysing trials harbouring sequencing information and adjusting indirect comparisons when collating evidence from heterogeneous sources. Such investigative efforts could lend support to reviews of HTA recommendations and contribute to synthesising external control arms involving treatment sequences.

Key points for decision makers

- There has been a surge in health technology assessment (HTA) research into treatment sequence evaluation. Despite advancements in modelling frameworks, the field faces multifaceted challenges that go beyond economic model structures, including selecting appropriate sequencing comparators, scarcity of clinical effectiveness evidence necessitating simplifying assumptions and statistical adjustments, considering treatment displacement effects, and computational optimisation.

- Cross-disciplinary methodologies, such as statistics, epidemiology, causal inference, operational research and computer science, show promise in addressing these challenges. For instance, real-world data (RWD) have substantial potential for informing estimates of treatment sequence effectiveness but demands analysis with appropriate statistical methods—often adapted from epidemiological and statistical research—to effectively mitigate biases.

- Future research should emphasise harnessing causal inference methods in evaluating the effectiveness of sequences, especially leveraging the Target Trial Emulation approach for sequencing analysis using RWD or clinical trials, and adjusting indirect comparisons. These investigative efforts can better inform reviews of HTA recommendations and the synthesis of external control arms involving sequences.

Acknowledgements

The authors would like to express their gratitude to Professor Allan Wailoo and Professor Paul Tappenden from the University of Sheffield for their valuable feedback on the initial version of the work. The authors would also like to extend their appreciation to the Sheffield Centre for Health and Related Research (SCHARR) library service at the University of Sheffield for their guidance on review methodologies.

2.2 Introduction

Over the last decade, there has been a growing interest in health technology assessments (HTA) in considering treatment sequences alongside the existing focus on evaluating treatments at discrete points in a treatment pathway.[8-10,88,89] Given the increasing treatment options[90-94], most diseases involve a pathway of care, wherein the effectiveness and cost-effectiveness of a treatment can be affected by treatments administered before or after it. That is, alternating the order of treatments can affect the overall effectiveness, such as overall survival (OS). These variations can, in turn, lead to changes in the overall costs of managing a disease. Therefore, to make optimal resource allocation decisions, it is imperative to consider the entire sequences, rather than individual treatments in isolation. Appraising treatment sequences requires the development of economic models that compare sequences. While defining the structure of treatment-sequencing models is recognised as challenging, other aspects, such as selecting an appropriate baseline sequencing-comparator and evidence of treatment effectiveness, are equally critical.[1,2]

To date, only the English[23,24] and Canadian[25] HTA guidelines explicitly acknowledge the potential need to consider the sequencing of technologies[95], albeit without detailed decision-analytic modelling instructions.[23-25] For instance, the 2022 guideline from the National Institute for Health and Care Excellence (NICE)[24] suggests considering the "care pathway" for the appraised technology and its comparator when appropriate, such as the sequences of treatments and diagnostic tests. While the provided guidance is rather generic, two NICE Decision Support Unit (DSU) briefing papers[21,22] that informed the 2013 NICE guideline[23] underscore the ongoing challenges HTA agencies face in numerous appraisals involving treatment sequences, especially regarding the selection of treatment sequences and their associated costs. Furthermore, NICE is currently piloting the Pathway Approach for technology appraisals (TAs) in two disease (i.e., renal cell carcinoma and non-small cell lung cancer)[27], which seeks to streamline the process of economic model development and review for treatments within the same disease, potentially influencing future evaluations of treatment sequences in HTA.

Despite existing reviews highlighting challenges in treatment-sequencing evaluation[1,2], a gap remains in primary research and innovation addressing these challenges. Additionally, the interplay between various challenges and nuances beyond economic model structures is underexplored. Therefore, we aimed to update and consolidate existing literature on methodological and conceptual advancements. We curate a "roadmap" of the relevant literature to provide a holistic view on the current state of treatment-sequencing in HTA, highlighting key challenges and ongoing research areas, laying a foundation for commissioning and developing new initiatives to address the challenges and develop new methods.

Clinical trials are typically the primary source of evidence used in HTA, but these rarely focus

on comparing treatment sequences, or on the effectiveness of a treatment conditional on a specific treatment sequence used prior. Real-world data (RWD) may be more adept at capturing treatment sequences, and there is a trend towards the enhanced use of RWD in HTA—particularly evident in organisations such as NICE for England and Wales.[37] Hence, a key focus of our review investigates whether literature exists that discusses the integration of RWD into HTA in the context of treatment sequencing. We explored the potential and related challenges of this integration and how it fits within the broader literature roadmap. NICE published a framework on the use of real-world evidence (RWE) in 2022[37], which postdates most of the included papers in our study. Consequently, our review aims to contribute fresh perspectives that may enhance the findings of existing studies, by relating their findings to the implications of NICE's RWE framework.

Our paper is structured as follows: we first describe the methods of our review (2. Methods), followed by an overview of the included papers and a narrative synthesis of our findings (3 Results). Finally, we discuss the implications of our findings for HTA and recommend avenues for future research (4 Discussion and 5 Conclusions).

2.3 Methods

2.3.1 The pearl-growing approach

In light of the limitations of conventional systematic review methods for methodological literature, our research employed the pearl-growing approach.[96,97] In particular: (1) a literature search for "treatment sequences" in the absence of established Medical Subject Headings (MeSH) and Embase Subject Headings (Emtree) could lead to biased results. (2) Our objective was to explore methodological advancements that are relevant across diverse diseases. However, during a pilot systematic search (Section 2.3.1.1), we noted a large number of studies that discussed treatment-sequences in a particular disease area (often multiple sclerosis or rheumatoid arthritis (RA)[98,99]), without describing relevant methodology beyond sequencing-model structures. This made the conventional systematic review approach highly inefficient. (3) Furthermore, relying exclusively on searching electronic databases, such as PubMed and EMBASE, could result in overlooking crucial ongoing studies and doctoral theses in this emerging topic area. Given these considerations, the pearl-growing approach appeared to be the best suited strategy.

The pearl-growing approach applied here begins with an initial key paper (i.e., the initial "grit") and expands by reviewing its reference list and studies that cited it (i.e., growing the "pearl").[100] This process is iteratively repeated for each added relevant paper, eventually leading to the identification of a core set of pertinent studies. This approach has proved more efficient for reviewing complex evidence.[96] Variations of this technique exist, including using more than one initial paper[97] or utilising the approach to identify a set of comprehensive search terms rather than studies.[101] For transparency,

we detailed our review procedures in the following sections.

2.3.1.1 Identification of the initial key paper

In October 2019, a pilot review was conducted on PubMed to identify an initial key paper, using search terms: ("economic model(s)" OR "economic evaluation(s)") AND ("treatment sequence(s)" OR "sequence(s)"). The study by Viola et al.[2], focusing on the selection of sequencing comparators, was selected as the initial "grit" for its alignment with the inclusion and exclusion criteria in Table 1.1, and for being the latest relevant publication with treatment sequence in its title.

Table 1.1 Inclusion and exclusion criteria

Definition of treatment sequences	Treatment sequences refer to changes in treatment involving sequential treatments, such as subsequent treatment following disease progression, adverse events (e.g. treatment toxicity), or those part of a planned regimen.
Inclusion criteria	➢ Studies exploring the development or issues of modelling treatment sequences in health economic evaluations, and/or ➢ Studies reviewing or investigating methodological approaches of treatment sequence comparisons in health economic evaluations
Exclusion criteria	➢ Studies solely focusing on treatment sequences in a specific disease area without exploring methodology advances in health economic evaluations*. ➢ Full-text unavailable
Type of studies	Reviews, meta-analyses, and conceptual papers

*Studies that discuss methodological approaches and validate or demonstrate their approach in a certain disease area are not limited by this exclusion criterion.

2.3.1.2 Reference and citation tracking

The first round of reference and citation tracking began by screening the reference and citation list of the initial key paper using their titles and abstracts based on the same criteria from the pilot review (Table 1.1). The reference list was extracted from the full-text of Viola et al.'s study.[2] The citations were identified through Google Scholar Search, as it provides a wider range of results compared with conventional electronic databases. Following screening, we evaluated the eligibility of shortlisted records up to December 2019 by reviewing their full-text.

Upon completing the first round of reference and citation tracking, a new set of key papers for the next round was identified. The same procedure was then performed repeatedly for each new set of literature identified in each round. The process concluded when the literature searches saturated, namely when no more relevant articles could be identified. Any pertinent articles that the iterative process overlooked, but were brought to the authors attention, were manually included.

2.3.2 Review update

To ensure the review remained up to date towards publication, an update of citation tracking was performed in February 2023 for the studies included in the initial review to include additional relevant articles.

2.4 Results

2.4.1 Overview of the included studies

The review included a total of 11 studies, with 6 identified during the initial phase and 5 discovered in the subsequent update. Figure 1.1 illustrates the inclusion process, and Table 1.2 details the objectives of each study. In the initial review, besides Viola et al.'s study[2], three studies were included via reference and citation tracking[1,36,102], and two doctoral theses from Tosh and Kim were added post-hoc.[30,31] Tosh's thesis, focusing on model optimisation methods, led to a derivative publication about economic models in RA.[30] Although this subsequent work was cited by Viola et al.[2], we had excluded it due to its disease-centric focus. Kim's doctoral thesis, which neither referenced nor was cited by other identified papers, was included because of its parallels to Tosh's doctoral work. In the third round (not shown in Figure 1.1 for simplicity), we assessed the references and citations of studies included in the second round as well as those that were manually included. Although we found further studies tangentially related to our review's objective, they had overly broad scopes. Some did not specifically investigate the methodological issues of treatment sequences, but rather broader issues of evaluating chronic diseases involving time-dependent transitions and sequential procedures.[103-115] Others predominantly comprised clinical studies, researching the clinical impacts of sequential treatment strategies and treatment switching, without an emphasis on HTA.[116-123] Further, one study is a conference abstract lacking full-text.[124] Consequently, we excluded these studies, indicating that our literature search had reached saturation. In the updated review, five new studies were included[34,35,44,95,125], all of which cited at least one of the aforementioned six papers. One of the five was a derivative publication[34] of Lewis's doctoral thesis identified in the initial review.[36]

Table 1.3 summarises the characteristics of the included studies. The majority of them (55%) were published between 2020 and 2022[2,34,35,44,95,125], with the remainder from 2015 to 2019.[1,30,31,36,102] Only slightly more than half of the studies (55%) were published in peer-reviewed platforms.[1,2,34,35,95,125] The remaining studies included four PhD theses (36%)[30,31,36,102] and a report from the NICE DSU.[44] Among all non-thesis articles (64%, n = 7), three undertook systematic reviews on methods that can be leveraged to shape the methodological framework for economic evaluation of treatment sequences[1,34,35], one paired a systematic review with a case study on the selection of baseline sequencing-comparators[2], while the remaining three are non-systematic reviews addressing various issues around treatment sequence evaluation.[44,95,125]

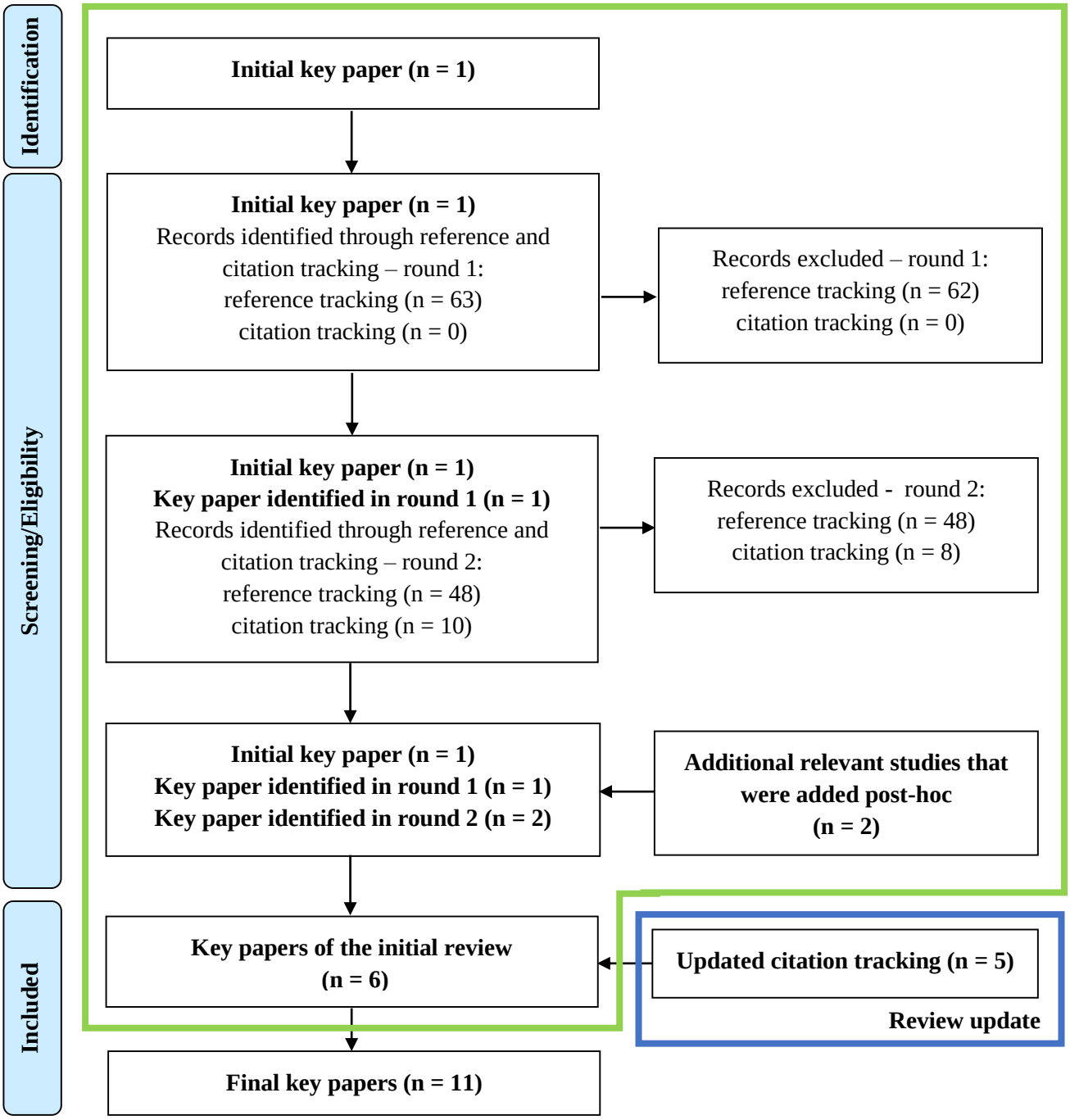


Figure 1.1 The Preferred Reporting Items for Systematic Reviews (PRISMA) diagram

Procedures within the green frame indicate the initial review phase conducted in December 2019, while those within the blue frame denote the updated citation tracking performed in February 2023

Table 1.2 Summary of included studies

Literature	Sources	Main objectives
Initial key paper		
Viola 2020[2]	Identified through the pilot review	To demonstrate that the choice of baseline comparator sequence can significantly impact the cost-effectiveness of treatment sequences and should be carefully considered in health economic evaluations.
Reference and citation tracking – Round 1		
Zheng 2017[1]	Cited by Viola 2020[2]	To provide a conceptual framework for modelling treatment sequences, including recommendations on problem scoping, model selection, and data source considerations.
Reference and citation tracking – Round 2		
Haywood 2018[102]	PhD thesis citing Zheng 2017[1]	To explore the impact of displaced treatments on the cost-effectiveness, equity of cancer care, and to propose strategies to improve the allocation and sequencing of cancer treatments in a way that optimises patient outcomes and resource utilisation.
Lewis 2019[36]	PhD thesis citing Zheng 2017[1]	To develop and evaluate quantitative evidence synthesis methods for assessing the effectiveness of treatment sequences in clinical and economic decision-making.
Other ad-hoc relevant studies		
Tosh 2015[30]	Additional PhD thesis manually added to the literature pool	To develop and apply simulation computational optimisation methods for modelling complex sequential treatments in chronic conditions, using rheumatoid arthritis as a case study.
Kim 2015[31]	Additional PhD thesis manually added to the literature pool	To develop decision models for evaluating the optimal sequencing and timing of different drug treatments in long-term medical conditions, focusing on computational optimisation and using parameterisation sequential drug decisions in primary hypertension as a case study.
Update citation tracking of the above articles (February 2023)		
Welton 2020[44]	Cited Zheng 2017[1]	To critically review and provide recommendations for existing and emerging methods for synthesising evidence on clinical effectiveness in health technology appraisals, including a section specifically about treatment sequences.
Lewis 2021[34]	Subsequent publication of the PhD thesis of Lewis 2019[36]	To review and categorise simplifying assumptions used in quantitative evidence synthesis methods for assessing the effectiveness of treatment sequences, and to provide a taxonomy of these assumptions.
Faria 2021[95]	Cited Zheng 2017[1]	To highlight the importance of problem structuring in economic evaluation, such as how to specify decision options, including selecting relevant comparators for standard of care and sequences of tests and treatments.
Huang 2022[35]	Cited Zheng 2017[1]	To provide a conceptual framework and address methodological challenges for modelling the effectiveness of treatment sequences in oncology.
Simpson 2022[125]	Cited Zheng 2017[1]	To discuss the importance and recent advances in methodologies of using real-world evidence in health technology assessments, including an approach to evaluate treatment sequences.

Table 1.3 Characteristics of the included studies

	Number of publications (%)
Publication date	
2015-2019	5 (45%)
2020-2022	6 (55%)
Peer-reviewed publication	
Yes	6 (55%)
No	5 (45%)
Article type	
Systematic review + framework for modelling treatment-sequencing	3 (27%)
Systematic review + case study	1 (9%)
Review	3 (27%)
PhD theses	4 (36%)
Location of the treatment-sequencing discussion in the article	
Full article	8 (73%)
A designated section in the review	3 (27%)
Areas of interests (each topic area is non-exclusive)	
Sequencing-model structures and problem structuring in economic evaluation	9 (82%)
Shaping the conceptual framework for treatment-sequencing modelling	4 (36%)
Methods to address the effectiveness of treatment sequence, especially in handling evidence from non-randomised data and indirect comparisons	5 (45%)
Potential of real-world evidence	5 (45%)
Potential of model optimisation methods	2 (18%)
Potential of causal inference methods	2 (18%)
Selection of treatment-sequence comparators	1 (9%)
Treatment displacement on resource allocation efficiency and equity	1 (9%)

2.4.2 Key research themes and their development roadmap

We identified several research themes (Table 1.3), including sequencing-model structures and problem structuring (82%, n = 9)[1,2,30,31,34-36,95,102]; developing conceptual frameworks for treatment-sequencing modelling (36%, n = 4)[1,34-36]; methods to tackle the effectiveness of treatment sequences, especially in handling evidence from non-randomised data and indirect comparisons (45%, n = 5)[1,34-36,125]; and the potential for using RWE to inform sequencing analyses (45%, n = 5).[34-36,102,125] Further developing themes included the potential of model optimisation methods (18%, n = 2)[30,31] and causal inference methods (18%, n = 2)[44,125], selection of treatment-sequence comparators (9%, n = 1)[2], and the impact of treatment displacement on resource allocation efficiency and equity (9%, n = 1).[102]

Figure 1.2 presents a "roadmap" detailing the evolution of treatment-sequencing research. It highlights the intersections and shifts among various research topics, revealing an increasing trend towards discussing statistical methodology in the context of treatment sequences. This led to our focus on statistical methods for estimating treatment sequence effectiveness in subsequent sections. We

encourage readers to refer to Figure 1.2 for a visual representation of our narrative synthesis results (Section 2.4.3) and the subsequent Discussion (Section 2.5).

2.4.3 Present landscape and challenges of treatment-sequencing evaluation in HTA

In this section, we provide a narrative overview of the current practice and primary challenges associated with treatment-sequencing in HTA across research themes, drawing from the selected studies.

2.4.3.1 HTA guidance for evaluating treatment sequences

Zheng et al. introduced a framework for treatment-sequencing modelling in 2017, providing step-by-step recommendations.[1] Derived from a systematic review of NICE TAs up to TA321 (2014), Zheng et al.'s framework comprises four steps: (1) model conceptualisation, (2) selecting a suitable modelling approach, (3) considering appropriate data sources for model inputs, and (4) determining computation tools. Zheng et al. indicated that treatment sequences have mainly been incorporated to reflect clinical practice or trial design, determine where to place new treatments in a sequence, or evaluate the impact of placing an additional treatment into existing treatment sequences. They noted an uptrend in TAs involving treatment sequences across various disease areas, including oncology, autoimmune, cardiovascular, neurology/mental health, infectious disease, and diabetes, with oncology featuring the highest number of treatment-sequencing models. A similar pattern was observed in an updated review by Viola et al. extending to TA527 (2018).[2]

Lewis et al.[34,36] and Huang et al.[35] advanced Zheng et al.'s framework by further unpacking the complexities in determining the effectiveness of treatment sequences and suggesting potential solutions. Lewis et al. underscored the importance of leveraging quantitative evidence synthesis methods and scrutinising simplifying assumptions.[34,36] Meanwhile, Huang et al. identified specific challenges inherent to oncology treatment-sequencing and proposed solutions to improve the estimation of treatment sequence effectiveness.[35]

Faria's review offered guidance on problem structuring for economic evaluation[95], with a section highlighting the importance of incorporating relevant sequences in the identification of pertinent decision options. Faria noted challenges in parameterising sequences in economic evaluation due to limited direct evidence on the effectiveness of treatments across all positions of sequences.

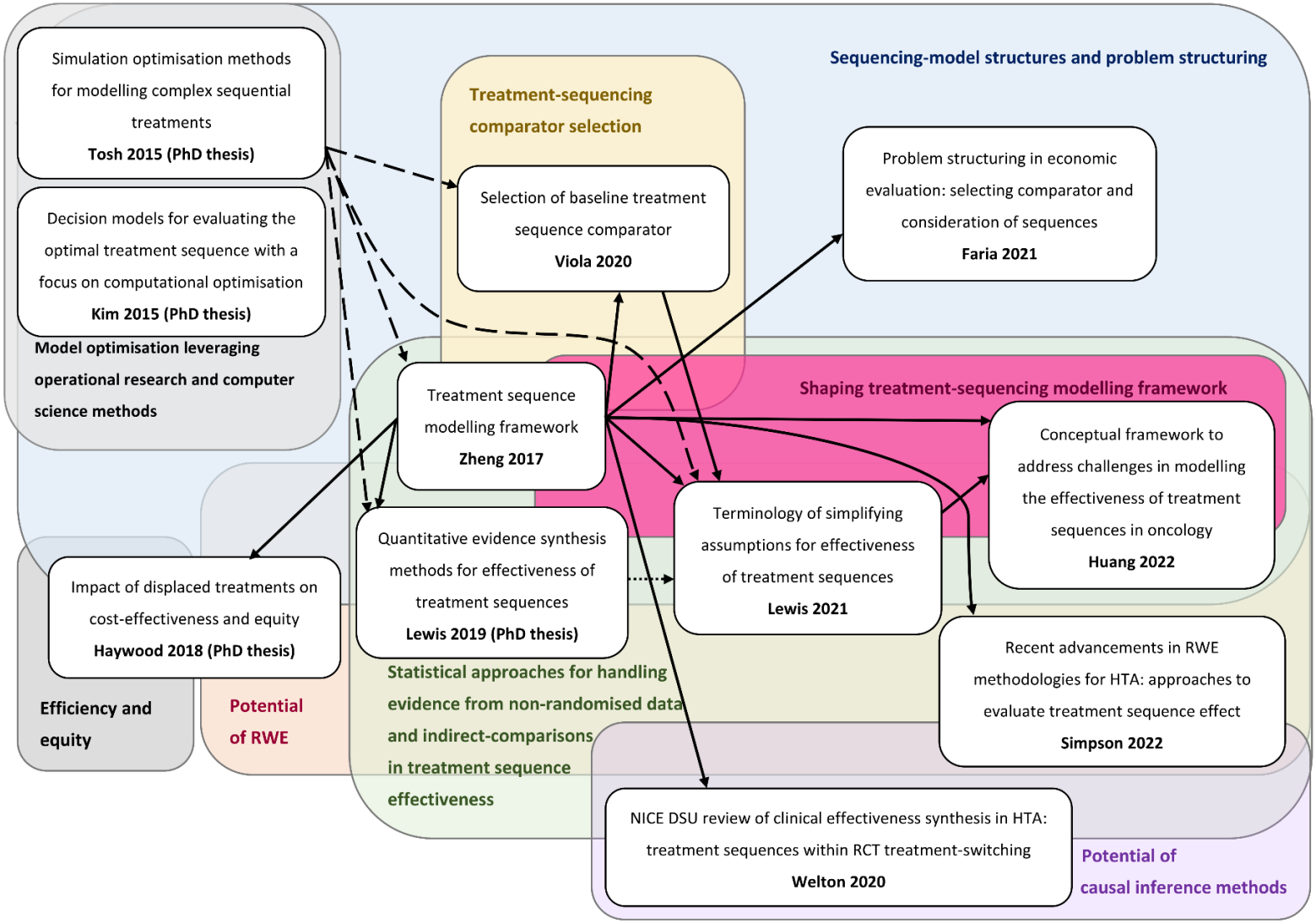


Figure 1.2 "Roadmap" of the treatment-sequencing literature in health technology assessments

Solid lines represent citations; dash lines represent citations of derivative publications; dotted lines indicate derivative publications of PhD thesis.
DSU: The Decision Support Unit, NICE; HTA: health technology assessment; NICE: The National Institute for Health and Care Excellence, United Kingdom for England and Wales; RCT: randomised controlled trial; RWE: real-world evidence

2.4.3.2 Scarce clinical evidence for treatment sequences and necessary statistical adjustments in evidence synthesis

Several studies highlighted the challenges of data scarcity in evaluating the effectiveness of treatment sequences.[34-36,95] Such limitations largely stem from the lack of randomised controlled trials (RCTs) comparing treatment sequences, frequently necessitating the coalescence of evidence of discrete treatment effects (i.e., LOT-specific effect) from different sources.

Lewis et al. identified that while network meta-analysis/meta-analysis (NMA/MA) is frequently utilised in HTA for synthesising treatment effectiveness evidence, there existed a gap in HTA guidance when it comes to adapting these methods for evaluating treatment sequences.[34] They highlighted that the feasibility of conducting a NMA/MA with RCTs comparing treatment sequences was hindered due to the lack of such trials. Furthermore, while it might be feasible to conduct a NMA/MA using observational studies that compare treatment sequences, this approach's viability could be considerably influenced by the inherent biases of observational studies.[34] Driven by these observations, Lewis et al. underscored the potential of adapting meta-analytic methods to reflect the "position effect" in deriving LOT-specific treatment effectiveness, including using meta-regression, stratification, or subgroup analyses in NMA/MA to account for the impact of patient treatment history. Although Lewis et al.'s review of meta-analytic methods was non-disease-specific, only 5 out of 23 studies were cancer-related, while over half (n = 13) focused on RA.

Huang et al., conversely, outlined issues related to data scarcity in modelling oncology treatment sequences.[35] Key challenges include (1) strategies to adapt the effectiveness of a LOT based on its position within a sequence; (2) examining the interplay between the timing of progression, discontinuation of a LOT, and initiation of the subsequent LOT, and their impact on the modelling results; and (3) the availability of head-to-head comparisons evidence for a specific LOT or the entire treatment sequence. Huang et al. reviewed 46 oncology treatment-sequencing models from NICE TAs and PubMed to determine how the aforementioned challenges have been tackled. They stressed the need to include both the time to treatment discontinuation and disease progression for LOT-effect in a treatment-sequencing model, or provide justification for any exceptions. Furthermore, Huang et al. believed treatment-free gaps should be included unless inapplicable, such as in late-stage cancers with minimal treatment gaps.

Huang et al. noted the prevalent practice of combining data from multiple clinical trials across different LOTs to model the effectiveness of a treatment sequence (41 models (89%)), such as merging progression-free survival (PFS) from an earlier LOT trial (e.g. PFS of treatment A in a first-line setting) and OS from a later LOT trial (e.g. OS of treatment B at in second-line setting) to model the OS for an entire treatment sequence (e.g. OS of treatment sequence A → B). However, none of

these “mergers of evidence” adjusted the estimated effectiveness based on the characteristics of patients from different data sources (i.e., the misalignment of patient characteristics between LOTs), primarily due to the absence of individual patient-level data. In contrast, they identified that three models estimated the OS of an entire treatment sequence using information from a single trial[126-128], and two models used RWD to estimate the OS of an entire treatment sequence.[129,130] Given that RWE and single-trial evidence for treatment sequences are often unavailable by the time of the HTA submissions, Huang et al. recommended combining trial evidence from different LOTs with necessary adjustments as a more pragmatic approach.[35]

Huang et al. found that no models applied head-to-head effectiveness evidence for comparing the entire treatment sequences, consistent with findings from Lewis et al.[34] Merely 24% of models applied methods to adjust for indirect treatment comparisons (ITC), which is exclusively limited to comparisons within a single LOT instead of the whole sequence. Huang et al. recommended making ITC adjustments for each LOT in the sequence and further adjusting for patient characteristics at the initiation of subsequent treatments, whenever possible. Overall, Huang et al. concluded that that there is substantial room for improvement in estimating treatment sequence effectiveness.

2.4.3.3 Simplifying assumptions in response to data scarcity

Despite Lewis et al. highlighting the use of advanced meta-analytic methods for generating LOT-specific evidence, many NMA/MA did not employ these methods.[34] Lewis et al. found that such omissions often led to the need for additional simplifying assumptions when integrating such evidence into treatment-sequencing models[34], potentially causing biases and uncertainties. They summarised a taxonomy of these assumptions, such as whether a treatment’s effect is dependent on its position or previous treatments and whether any modifications to the treatment effect should be made depending on its position or disease duration. Lewis et al. underscored the importance of carefully assessing the simplifying assumptions in sequencing models. We encourage readers to refer to Lewis et al.'s study for details.

2.4.3.4 Modelling approaches

Decision-analytic modelling approaches for treatment sequencing are well-established.[1] Among the 63 treatment-sequencing models Zheng et al. analysed, cohort state-transition models (e.g. Markov models, Semi-Markov models) were most prevalent, followed by discrete event simulation (DES), individual state-transition models (e.g. microsimulations), and decision trees. Several crucial factors can influence the choice of modelling approach and the construction of health states, such as patient heterogeneity, number of LOTs, and type of clinical outcomes.[1] Lewis et al. further detailed the advantages and disadvantages of each approach with a comparison table in their paper.[34]

Although partitioned survival models are prevalent in oncology[131], Zheng et al. suggested that

such models were not utilised for treatment sequences due to their inherent methodology and limitations.[1] Contrasting this, an updated review by Viola et al. highlighted several oncology NICE TAs employing partitioned survival models for treatment sequences.[2] Lewis et al. also identified comparable applications.[34]

2.4.3.5 Selection of treatment sequences in a decision problem

Zheng et al. noted that most NICE TAs predominately compared treatment sequences that reflect clinical practice, rather than optimal placement or excluding suboptimal treatment sequences.[1] Viola et al. investigated the impact of baseline sequence-comparator selection on the incremental cost-effectiveness results by conducting a cost-effectiveness case study with a treatment-sequencing Markov model featuring four hypothetical treatments.[2] Their findings revealed that non-cost-effective treatments were never part of an optimal (i.e., most cost-effective) sequence. Viola et al. proposed excluding non-cost-effective treatments from the baseline comparator or placing them later in the sequence to improve health resource allocation efficiency. However, they also acknowledged the limited generalisability of their case study due to the number of treatments and length of the sequences included. Furthermore, their focus was on maximising resource allocation efficiency without posing any constraints on the positions of treatments in a sequence, which could occur in real-world practice due to marketing authorisations. Nevertheless, Viola et al. underscored the need for judiciously selecting baseline sequencing comparators. They cautioned that overlooking the cost-effectiveness of individual treatments when defining the baseline sequencing-comparator could lead to misleading results.[2]

2.4.3.6 Impact of treatment displacement on resource allocation efficiency and equity

Haywood's doctoral thesis delved into the issue of treatment displacement. He defined treatment displacement as where new treatments cause existing treatments to be shifted to later LOTs.[102] If not properly addressed in economic evaluations, Haywood argues that this can result in biased decision-making. Specifically, Haywood argued that continuing to pay the previously agreed-upon price for displaced treatments is unlikely to prove cost-effective. This is primarily due to the uncertainty surrounding the effectiveness of these treatments once they are displaced and used at a later line of therapy, with a high likelihood that effectiveness will be reduced. Consequently, paying the previously agreed-upon price of displaced treatments when they are used at a later LOT can lead to allocative inefficiency. Given the lack of clarity on reduced effectiveness, Haywood proposed adjusting the price of displaced treatments downwards to better reflect their anticipated diminished effectiveness and restore their cost-effectiveness.

To support his argument, Haywood conducted modelling studies to assess the impact of treatment displacement in several cancers[102], including breast, colorectal, and non-small cell lung

cancer. The scarcity of clinical evidence on sequencing led him to undertake a *de novo* RWD analysis using Australian Pharmaceutical Benefits Scheme (PBS) data to understand local treatment patterns, subsequently guiding what treatments and the number of LOTs to be incorporated into his economic models and their associated costs. However, for LOT-specific effectiveness estimates (i.e., PFS), Haywood chose to apply estimates from literature and perform meta-analysis instead of conducting RWD analysis. Haywood's systematic review and meta-analyses of existing evidence revealed that the displacement of a treatment may lead to decreased effectiveness, treatment duration, and increased toxicity per unit time. Haywood's economic modelling results showed that displacement of an existing treatment resulted in an increased incremental cost-effectiveness ratio (ICER), with a reduction in the price of these treatments being necessary to restore the cost-effectiveness. Overall, he recommended carefully considering the impact of displacement on currently subsidised treatments in cancer treatment funding to ensure equity and cost-effectiveness to avoid underestimating the total costs and overestimating the total benefits of a new treatment being introduced to clinical practice.[102]

2.4.3.7 Model computation and tools

Tosh and Kim's theses were among the earliest works on innovative treatment-sequencing methodologies.[30,31] Both emphasised the significance of computational optimisation methods for sequencing-models, particularly in scenarios where the decision problem involves a large number of sequencing-comparators ($n > 1000$). Tosh investigated the use of simulation optimisation methods to identify optimal or near-optimal RA treatment sequences[30], driven by the inconsistencies in existing economic models for comparing RA treatment sequences[8], which may lead to potentially inaccurate cost-effectiveness estimates. He approached the issue as a combinatorial discrete simulation optimisation problem, experimenting with methods such as simulated annealing and genetic algorithms. While Tosh found the simulation optimisation via simulated annealing (SOSA) promising, its time-consuming nature prompted calls for further research on its generalisability. Concurrently, Kim studied similar methods for treatment sequences in primary hypertension.[31] The nuance in Kim's model lay in its emphasis on individualised treatment sequences, where each patient in the model could potentially receive the same "treatment strategy" but with different treatment sequences. For instance, patients exceeding a certain threshold for a particular characteristic would follow one sequence, while those who fell below the threshold would follow a different one. This contrasted with Tosh's approach, which examined the average effect of each treatment sequence on the entire population.

Zheng et al.'s review[1] summarised software that have been used in modelling treatment sequences, including those in Excel VBA, R, Arena and C. While both Tosh and Kim identified VBA and TreeAge as options for modelling treatment sequences, they opted for Simul8 and Matlab in their

respective doctoral these.[30,31]

2.4.3.8 Potential of RWE in informing sequencing analyses

RWD has emerged as a valuable tool in supporting treatment-sequencing decisions.[1,2,34,35,102] Evidence derived from RWD (i.e., RWE) may be used to inform local treatment patterns, sequence selection, and shape model structures.[1,2,34,35,102] RWE can also be leveraged to capture the effectiveness of the entire treatment sequences or LOT-specific effectiveness.[34,35] There are, however, significant challenges that undermine the potential of such application. One primary limitation is the delay in RWD's availability.[1,34,35] Typically, data collection commences only after a new drug is introduced to the market, leading to delays in its inclusion in RWD to be used for HTA submissions. Additionally, studies warned of the inherent issues in RWD that could result in biased findings (e.g. selection bias, confounding) and underscored the importance of applying statistical methods to mitigate biases in RWD analyses.[3,34,102] Despite these challenges, studies advocated for a further understanding of the pros and cons of RWE in treatment sequencing decisions in HTA.[34,35]

A recent review by Simpson et al.[125] highlighted the latest advances in using RWE to evaluate sequencing in HTA. They spotlighted Spelman et al.'s approach of harnessing RWD from multiple countries to tackle the issue of the unmeasured confounding in treatment-sequencing comparative effectiveness studies.[132] Spelman et al.'s study focused on evaluating different treatment schedules for relapsing-remitting multiple sclerosis (RRMS). It is highlighted that, even when earlier observational studies compare "similar" patients based on their measured characteristics[133-135], unmeasured confounders may still exist between patients receiving different treatment patterns in a single health care setting. Spelman et al., conversely, compared patients from Sweden and Denmark, where the treatment options of RRMS patients were comparable but the primary recommended treatment strategy diverged (starting with highly effective disease-modifying therapies versus treatment escalation, respectively).[132] By controlling for various patient characteristics and comparing the clinical outcomes between the two countries, they proxied the impact of different treatment strategies. This method resembles using geography as an instrumental variable, which has faced criticism in oncology.[136] Spelman et al., nonetheless, introduced an innovative approach to comparative effectiveness research[132], which could prove valuable for future HTA submissions involving treatment sequences. Spelman et al.'s study was not directly included in our review because it focused solely on clinical findings, while Simpson et al.'s review[125] shed light on its relevance to HTA.

2.4.3.9 Relevance of causal inference methods in sequencing

A recent NICE DSU report[44] critically reviewed methods for evidence synthesis on clinical effectiveness in HTA, updating methodologies developed since the 2013 NICE guidance.[23] The report by Welton et al. features a section underscoring the potential of causal inference methods, especially in addressing the issue of unwanted treatment-switching in clinical trials, referencing a prior NICE DSU technical support document (TSD) (i.e., TSD 16).[43] Welton et al. highlight the significance of applying causal inference tools in HTA to delineate the disease-treatment pathway, facilitating covariate selection within statistical methods for synthesising evidence of treatment effectiveness.

We included Welton et al.'s report for its insight into the parallels between treatment-switching in trials and treatment sequences, suggesting a new research direction. Welton et al. proposed that statistical methods capable of handling the dynamic treatment changes in RCTs (regardless of whether unintended or permitted) could potentially offer insights into treatment-sequencing. Thus, they recommend joint research to co-develop these methods in tandem with treatment-sequencing modelling. We interpreted this as an opportunity to leverage causal inference principles in developing statistical methods to estimate sequencing effects using RCT data containing treatment-switching, in line with the heavy reliance on these principles in the NICE DSU TSD 16 guidance.[43] While recognising the value of causal inference methods, our review excluded several epidemiological and statistical studies on dynamic or sequential treatment strategies due to their non-HTA focus.[116,117,123] However, these studies may provide further insights into evaluating sequencing effects using causal inference methods, warranting the need for a separate, in-depth review. To highlight this point, we retain Welton et al.'s paper in our roadmap.

2.5 Discussion

Our review underscores the escalating interest in evaluating treatment sequences in HTA and summarises the emerging research themes (Table 1.3, Figure 1.2). These themes have arisen in response to the multifaceted challenges of treatment sequences, encompassing conceptualising the decision problem, choosing suitable approaches for constructing treatment-sequencing decision models, identifying and deriving appropriate clinical effectiveness evidence for treatment sequences in the face of data scarcity, scrutinising effectiveness assumptions and statistical methods for adjustments, and optimising model computations. We found that interdisciplinary research offers promising solutions to these challenges. For example, methods from statistical and epidemiological research were utilised to handle evidence from non-randomised data and perform adjusted indirect comparisons[34,35]; and methods from operational research and computer science were applied for model optimisation.[30,31]

2.5.1 Implications of existing treatment-sequencing research for HTA

Early interest in treatment sequencing effects in specific diseases[21,22,137-139] led to NICE including advice on integrating sequential treatments into comparators in their 2013 methods guide.[21-23] Before Zheng et al. introduced the non-disease specific treatment-sequencing modelling framework[1], modelling guidance provided only general considerations about incorporating subsequent treatments.[20,23,25,140,141] Earlier sequencing-modelling frameworks were either only presented as conference abstracts[124] or had a broader focus on event sequencing beyond treatments, such as the sequencing of diagnostic procedures or whole disease modelling.[108,109] The recently updated 2022 NICE manual highlights the importance of "care pathways"[24], covering the sequence of treatments, tests, and other relevant technologies. In this revision, NICE explicitly requires including all diagnostic technologies in a sequence, but does not stipulate the same for "treatment" sequences. Further, NICE is piloting the Pathway Approach to streamline the review of treatments for the same diseases, potentially impacting future HTA involving treatment sequences.[27]

In estimating the effectiveness of treatment sequences, existing studies primarily recommend refining the derivation of LOT-specific effects and collating evidence from various LOTs, given the difficulties in acquiring data for entire sequences during the appraisal process.[34,35] Such approach hinges on leveraging statistical methods to make adjustments for position effects, and to approximate head-to-head comparisons when collating evidence. Lewis et al.'s findings on adapting meta-analytic methods to account for position effects[34] seem relevant to NICE's guidance for meta-analytic methods (e.g. NICE DSU TSDs 1, 3, and 4)[142-144], while Huang et al.'s recommendations[35] seem to relate closely to NICE DSU TSD 18[145], focusing on population-adjusted indirect comparisons (i.e., matching-adjusted indirect comparison (MAIC) and simulated treatment comparisons (STC)). However, none of these guidelines discuss these similarly rooted methods in the context of treatment sequence evaluation. Furthermore, Huang et al.'s work seems to be anchored to state-transition models, raising questions about the applicability of their insights to partitioned survival models[146], which are prevalent in oncology. Importantly, it remains unclear whether there are other challenges unique to partitioned survival models yet to be identified. For example, "shoehorning" a sequence into a partitioned survival model might require alternative methods to collate evidence from multiple LOTs. For instance, combining survival curves from varied sources for different LOTs into a single OS curve for a sequence can lead to problems, such as unrealistic crossing of cumulative treatment durations and OS curves.

While none of the studies we examined focused on specific methods for generating RWE on treatment sequence effectiveness, several emphasised the importance of applying statistical methods to mitigate biases in RWD analyses.[34,102] These insights mirror those in NICE DSU TSD 17 about using observational data to estimate treatment effectiveness.[40] However, TSD 17 does not touch on

the relevance of these techniques in the context of treatment sequences. Zheng et al. underscored RWE's role beyond deriving effectiveness model inputs — also in corroborating model predictions and assessing uncertainties through sensitivity and scenario analyses. This aligns with the concept of evidence triangulation, prevalent in fields like epidemiology.[147,148] While evidence triangulation has its merits, its utility as a validation tool might be constrained without clear acceptability criteria. Further, it is ambiguous whether Zheng et al. also implied using RWE as targets for model calibration.[149] Should this be the case, the uncertainty inherent in sequencing models could be exacerbated, especially when RWE—known for its susceptibility to confounding bias—is not validated. Specifically, a recent study highlights that the use of calibration targets that are not well-matched between models and data can result in biased outcomes.[150]

In summary, for sequencing effect estimation, we believe it is crucial to incorporate relevant statistical methods for indirect comparisons, evaluating position or sequence effects, and adjusting for potential confounding and selection biases. Further reviews and research are needed to understand the strengths and weaknesses of available statistical methods in the context of treatment-sequencing. Moreover, understanding how these methods should be tailored for varying data sources, model types and outcomes requires further exploration. We therefore provide actionable directions for future research in Sections 2.5.2.1 to 2.5.2.5.

Determining which sequences to incorporate into a decision problem necessitates striking a balance among several factors[2], including conducting an exhaustive review of treatment sequencing clinical evidence[36], considering local clinical guidelines[1], taking into account real-world practice and marketing authorisations[1,95,102], optimising resource allocation efficiency[2,102], and computational feasibility.[30,31] While Viola et al. recommend removing non-cost-effective treatments from a baseline sequencing-comparator to maximise resource allocation efficiency[2], it is noteworthy that NICE typically defines its comparator of interest as the most commonly prescribed treatment sequence rather than the "most cost-effective standard treatment". Lewis raised concerns about the NICE Committee's approach, noting that the their decisions frequently stem from deliberations without comprehensive review of clinical evidence on sequencing.[36] Furthermore, while the majority of appraisals assess fewer than 10 treatment sequences, Tosh and Kim examined exceptions with significantly more sequences.[30,31] The impact of computational feasibility on sequence inclusion remains unknown. To date, there are no explicit guidelines prioritising any specific factor for sequence inclusion.

Regarding the impact of displaced treatments, Haywood's proposal to adjust their costs may be contentious.[102] Displaced treatments typically have shorter durations and reduced effectiveness[102], leading to inherent cost reduction. Hence, implementing additional price cuts could potentially over-penalise. Secondly, reducing the price of displaced treatments in fast-evolving disease areas may raise

fairness concerns and discourage new treatment development, especially for late-stage treatments. In particular, new treatments may emerge quickly, so that the "less new ones" are displaced within a short time. Such dynamic might become apparent through initiatives like NICE's Pathway Project.[27] Thirdly, having treatments with varying prices at different LOTs presents complexities. A potential approach for addressing this issue is to implement mandatory pre-planned re-evaluation when displacement occurs and treatments begin to be used at a later LOT. This could lead to price adjustments where necessary. The potential role of prospective RWE in monitoring the treatment effectiveness and displacement has not been explicitly discussed. These normative issues extend greatly beyond economic evaluation and warrant further discussion.

2.5.2 Future research recommendation

2.5.2.1 Exploring consideration of treatment sequences in HTA beyond sequencing-models

The extent to which treatment-sequencing comparisons that do not necessitate explicit sequencing model structures remains uncertain, as most existing reviews are geared towards treatment-sequencing economic models.[1,2,34,35] For example, the manufacturer in NICE TA387[127] used the OS of abiraterone from the COU-AA-302 trial[151] (abiraterone as a first-line treatment in metastatic castration-resistant prostate cancer) to represent the OS of an entire treatment sequence (abiraterone → docetaxel → best supportive care) in the model, while an underlying three-line model was maintained for the purpose of calculating the costs of subsequent treatments. Therefore, technically, no conventional sequencing structure was required to model the OS. This approach heavily relies on the assumption that the subsequent treatments received by patients in the COU-AA-302 trial are representative of those in England's clinical practice.

Conversely, some HTAs may unintentionally compare the effects of treatment sequences without a sequencing model because of utilising effectiveness data that inherently include sequencing details. For example, when the aim is to compare treatments as if no patients had undergone any unintended subsequent treatment, it becomes problematic to use RCT data with unintended treatment-switching without making appropriate adjustments.[43,152] Exploring the prevalence of these approaches in HTA across disease types is warranted, as they can affect the necessary assumptions required in populating the treatment effectiveness in economic models. While biases in the unadjusted unintended treatment-switching scenario are well-recognised[43], uncertainties in the first scenario—where the effect of a first-line treatment is assumed to represent the effect of the entire treatment sequence (e.g. TA387[127])—may have been overlooked, warranting further investigation.

2.5.2.2 Further research on the derivation of clinical effectiveness for treatment sequences

Data scarcity remains a significant challenge when evaluating treatment sequences in HTA.[1,34,35,125] Despite existing studies outlining potential methods for deriving treatment sequence effectiveness[34,35], there may be a need for an updated overview of the current practice. The meta-analytic methods summarised by Lewis et al. appear to be a viable approach for deriving LOT-specific effectiveness.[34] However, their recommendation primarily stems from RA studies before 2013 and may not be as applicable in oncology due to the complexities of survival meta-analysis.[153,154] Huang et al., conversely, exclusively focused on oncology studies and recommended applying ITC adjustment methods when combining data from various trials.[35]

Both Lewis et al. and Huang et al. focus on methods for deriving LOT-specific effectiveness[34,35], but these approaches rely heavily on simplifying assumptions about effectiveness and population alignment between LOTs. Exploring advanced statistical methods to refine the "degradation effect" parameter (a prevalent strategy in autoimmune disease TAs[155-158]) could potentially mitigate uncertainties stemming from oversimplified assumptions. This approach could be particularly useful for sequences with a significantly higher number of LOTs, where determining LOT-specific effects for each LOT is challenging (e.g. autoimmune HTAs). This approach is suitable where a degradation effect assumption can be made, but less applicable in situations where subsequent treatments may be more effective due to specific prior events, such as treatment-induced mutations that enhance the effectiveness of later therapies. Furthermore, exploring alternative data sources and statistical methodologies to assess the comparative effectiveness of complete treatment sequences remains an underexplored area, and could prove valuable. Reviewing statistical methods originally developed for other purposes but that are potentially applicable to treatment sequence evaluation (e.g. those adapted for tackling treatment-switching in RCTs[43,44]) from both HTA and non-HTA domains, could inspire new strategies to address sequencing challenges in HTA.

2.5.2.3 The role of RWE and causal inference in assessing treatment sequence effectiveness

Despite the limited RWD availability for new treatments during appraisals, considerable research emphasises the significance of exploring RWE's potential in informing treatment-sequencing effectiveness.[1,34,35,102,125] RWD holds importance because of its ability to capture patient treatment trajectories over time. By leveraging causal inference principles, biases in RWD analysis could be mitigated.[5,40,50] While existing HTA guidelines provide methods suitable for non-sequencing scenarios[40,44], tailored modifications may be required for treatment-sequencing. For instance, assessing the comparative effectiveness of two complete treatment sequences is complex due to time-varying confounding, potentially necessitating advanced adjustment techniques, such as g-methods.[54,159] Statistical methods highlighted by Lewis et al. and Huang et al. implicitly tap into the

importance of causal inference[34,35]—for instance, adjusted ITC is used to ensure fair comparisons between two groups. This aligns with HTA's goal to compare the counterfactual outcomes of alternative interventions: one without the new treatment and one with it in the healthcare system. Given that the concept of counterfactual outcomes resonate with causal inference principles, we advocate for their explicit integration into treatment-sequencing HTA, especially when utilising RWE. This perspective largely mirrors the views in the recently released NICE RWE framework.[37]

We identified several niches where RWE can support treatment-sequencing HTA. Each necessitates tailored study designs and statistical approaches to address biases arising from non-randomised treatments. These areas include: (1) deriving the effectiveness of a specific LOT or a segment of the treatment sequence, whether in the control or treatment arm; (2) deriving the overall effectiveness of the treatment sequence in the control arm; and (3) deriving the comparative effectiveness between two (or more) complete treatment sequences.

Application (1) is commonly used to populate the effectiveness of later-line treatments in the model when trial evidence is constrained by short follow-up durations. Conceptually, methodologies described by Lewis et al. and Huang et al. can be considered to address indirect comparisons and ensure population alignment across LOTs.[34,35] However, further nuances and challenges of "plugging-in" RWE into a treatment sequence assessment warrant further exploration. Further, this form of RWE has also been employed as calibration targets for modelling early detection interventions where evidence from long-term follow-up is scarce.[160] Biases can, however, arise if a mismatch between the model and data exists.[150] Additionally, this type of RWE can aid in evidence triangulation in oncology, particularly when scrutinising the plausibility of the extrapolated survival curve of the appraised treatment beyond trial periods. Comparing the late-stage survival curve from RWE against trial extrapolations can reveal any unrealistic extrapolations, given that such extrapolations are greatly influenced by the choice of parametric models.[161,162] However, a caveat is that estimates derived from RWE on late-stage survival might underestimate the effects of the appraised treatment, since this evidence often comes from data gathered before the introduction of newer frontline treatments. Recent studies have explored integrating RWE into trial-based survival extrapolations.[163] Future work in this field may benefit from delving deeper into the nuances of treatment sequencing for such applications.

Application (2) represents a unique case of an external (synthetic) control arm, which received considerable attention in HTA and regulatory bodies[164,165], including the NICE RWE framework[37], even though none of these have specifically addressed the treatment-sequencing scenario. This application's complexity arises from the need to blend data from multiple sources (typically RWD with a trial arm) and employing statistical methods to enable direct comparisons of complete treatment sequences. While the statistical methods may conceptually resemble those used for RWD

analysis, adaptions for extending inferences may be necessary.[166,167] Predictable challenges include unmeasured confounding, the absence of time-varying covariates, and missing information on subsequent treatments in trials. While it seems conceptually viable, a renewed review may be needed to identify existing applications and associated challenges, and further application studies are imperative to assess such application's feasibility.

Application (3) is specifically pertinent in the re-evaluation of appraisals or reimbursement decisions for treatments previously funded through alternative funds (e.g. NICE Cancer Drugs Fund (CDF)).[88,168,169] In such instances, data on previously appraised treatments might become available, presenting an opportunity to harness RWE for direct comparisons of treatment sequences. This is especially relevant in determining the optimal sequencing or positioning of treatments. Notably, clinical trials rarely assess the effectiveness of a new treatment across various treatment lines (e.g., standard of care → new treatment versus new treatment → standard of care), whereas such sequences can co-exist in real-world settings and may serve as relevant comparators in an appraisal.

The Target Trial emulation (TTE) approach in epidemiology[3] seems to be a promising tool for comparing two complete treatment sequences in RWD[170]. The TTE framework[3], rooted in the principle of causal inference, has garnered substantial attention for its potential in explicitly designing observational studies to estimates causal effects and mitigate issues such as confounding, selection bias, and immortal time bias. The U.S. Food and Drug Administration (FDA) has funded projects aimed at examining the Target Trial emulation's feasibility to answer clinical questions for regulatory purposes[171,172], and the NICE RWE framework endorsed the TTE framework and recommends application in HTA wherever relevant.[37]

NICE's RWE framework briefly mentioned the viability of using emulated Target Trial to compare dynamic treatment strategies, such as treatment sequences, without specifying operational details. In our view, treatment sequences can be categorised into different types of treatment strategies[5]: (1) time-related static treatment strategies (where patients transition to the next-line of treatment at fixed intervals)[173]; (2) dynamic treatment strategies (where the timing of patients transitioning to the next-line of treatment is based on specific events (e.g. disease progression)[174,175], with each patient assigned to a specific type of treatment sequence within each treatment arm); and (3) dynamic treatment strategies with individualised treatment sequences (where the timing of patients transitioning to the next-line of treatment is based on patient characteristics, with the possibility of each patient being assigned to different a treatment sequence based on specific events and additional patient characteristics (e.g. biomarker status)).[176] Each of these strategies may require overlapping but distinct analytical approaches. Additionally, instead of emulating Target Trials for the comparison of dynamic treatment strategies, which are more relevant to Applications 2 and 3, a recent clinical epidemiology study by Bujkiewicz et al. performed a bivariate NMA on line-specific

emulated Target Trials from the British RA Register.[177] This approach aimed to bridge the disconnected networks in NMAs when assessing the effectiveness of first and second-line RA therapies, contributing to improving insights for Application 1.

Given the complexities mentioned above, further studies are, therefore, needed to assess the feasibility and validity of utilising Target Trial emulation to compare treatment sequences suitable for HTA purposes under different scenarios.[170] This also echoes a recent review that underscores the need for future research on the application of RWE with Target Trial emulation in HTA re-evaluation.[78]

2.5.2.4 Role of other sources of data harbouring treatment sequencing information

Apart from RWD, other data sources may also harbour comprehensive sequencing information but remain underexplored in prior research. These include (1) trials (RCTs or single-arm trials) with extended-follow up, collecting information on non-randomised subsequent treatments; (2) RCTs that randomise patients to receive different treatment sequences; and (3) sequential multiple assignment randomised trials (SMARTs)[123,176], which involve multiple stages of randomisation at points when treatment changes are indicated, unlike conventional trials where patient randomisation occurs only at baseline. Applying estimates from source (2) is relatively straightforward and requires no adjustment if the treatment sequences being compared align with those in a specific decision problem, although such alignment is uncommon. Utilising estimates from source (3), however, necessitates the use of specialised analytical techniques specific to SMARTs.[116,178] Jamie Robins, known for his contribution in causal inference methods, noted that the key difference between dynamic treatments in real-world settings and SMARTs lies in the randomisation probabilities—unknown in the former while predetermined in the latter.[60] Statistical methods developed for SMARTs share similarities with causal inference methods for real-world dynamic treatments (e.g. g-methods, not limited to sequential treatments), and could be relevant for both assessing treatment sequences in RWD and trials that include sequencing data. Utilising data from source (1) may also require causal inference-guided statistical methods to mitigate confounding bias[44], and facilitate adjustments for indirect comparisons when integrating it with other data sources. Given its prevalence over sources (2) and (3), source (1) holds substantial promise for offering timely insights into treatment sequencing involving the appraised treatment during an appraisal, setting it apart from RWD's role. Thus, methods for utilising data from source (1) for HTA should be investigated.

On a separate note, recent advancements in adaptive trial designs could also potentially be adapted to improve the evaluation of treatment sequence effectiveness.[179-182] While SMARTs focus on developing pre-defined multi-stage adaptive interventions, the design of adaptive trials modify trials over time based on interim data. With the design of dynamic protocols, advancements in adaptive trials may provide transferable statistical methods for assessing sequencing effectiveness.

Specifically, adaptive trials differ from traditional ones by continuously adjusting randomisation probabilities over time. Informally, these (time-varying) adapted randomisation probabilities sit between the known probabilities of receiving treatments in SMARTs and the unknown probabilities in RWD analysis. Thus, statistical methods for analysing adaptive trials theoretically share similarities with those previously mentioned for SMARTs and may offer additional insights in addressing challenges related to sequence analysis.

2.5.2.5 Model computation, utility, costs and equity

While the potential of computational optimisation methods for treatment sequencing decisions attracted early attention[30,31], no further developments have been seen in recent years. This may be associated with a tendency for HTA to focus on individual treatment effects as opposed to optimising treatment sequences. Further, these early studies were hampered by inadequate evidence on LOT-specific effects[30,31], necessitating the use of crude simplifying assumptions that undermined the value of these methods. Nonetheless, advancements in statistical methods to better identify these effects (Section 2.5.2.2) may improve the potential value of these computational methods. A thorough review of recent appraisals might shed light on the recent use of these methods. Although not widely incorporated into routine HTA processes, relevant codes developed by Tosh and Kim are available online for adaption.[30,31] Notably, methods developed in Tosh and Kim's theses were highlighted by the Professional Society for Health Economics and Outcomes Research (ISPOR) Task Force as exemplars when a decision problem is framed as a constrained optimisation problem that accounts for a set of decisions over time.[183]

Finally, several other areas have received limited research attention but may be crucial for further exploration, including equity, utility, and costs within treatment sequence evaluation. Utility, in particular, may encounter similar data scarcity issues as treatment sequences since patient data at specific time points while on a certain treatment is often lacking.

2.5.3 Strengths and limitations of the review

Our study has several strengths. Firstly, our review provides a comprehensive summary of studies focusing on treatment sequencing within HTA, highlighting both challenges and advancements, including early-stage developments from doctoral theses. The compilation reveals emerging research areas and underscores the value of interdisciplinary research. Secondly, our roadmap elucidates how these studies interrelate, enabling researchers with specific interests to swiftly locate pertinent papers and associated references. Thirdly, we bridged the gap between the insights drawn from the selected studies and their relevance to existing HTA guidance[40,142-145] as well as the newly issued NICE RWE framework.[37] This highlights the current gaps in HTA guidelines pertaining to tailored methods for treatment-sequencing and offers actionable research areas for

enhancement, aiming to complement the existing guidelines.

Despite its strengths, our study also comes with limitations. First, by focusing on reviews, methodological, and conceptual papers, the scope of our review may be deemed limited. However, given the frequent citations to Zheng et al.'s work[1] across the majority of the identified literature, we are confident that we have captured the field's pivotal paper and its associated literature. Secondly, with our non-disease-specific and HTA-oriented focus, we might have missed other relevant methodological papers that bring additional insights. However, for readers interested in specific modelling studies or innovative methods from other fields, these can likely be found in the references of the papers we included. Thirdly, we may not fully capture all disease-specific nuances of their interplay with each of the research topics. Readers are encouraged to further explore the implications of subject knowledge. Lastly, we may have missed articles discussing treatment sequences if they did not directly mention such concepts in their titles or abstracts.

2.6 Conclusions

Our research highlights the multifaceted challenges in evaluating treatment sequences in HTA, including problem structuring, modelling approaches, selection of treatment-sequencing comparators, model optimisation and data scarcity in evaluating the effectiveness of treatment sequences. Each of these aspects presents complex, interrelated issues that warrant further investigation. This field is rapidly evolving, with the issue of data scarcity in treatment sequences being particularly salient, causing decision uncertainties, arising from inherent biases in data sources, effectiveness simplifying assumptions, and the appropriateness of the adjustment methods utilised. We found that interdisciplinary research offers promising solutions to these challenges. This includes applying causal inference principles from epidemiology and statistical research to manage evidence from non-randomised data and perform adjusted indirect comparisons. Additionally, insights from operational research and computer science contribute to model optimisation.

Several avenues for future research are identified. Firstly, given the remarkable potential of RWD in capturing treatment sequences, it is imperative to investigate appropriate strategies for leveraging RWE in treatment sequence evaluation. In particular, it is prudent to explore how emerging causal inference methods, such as Target Trial emulation, can be employed to harness RWD for generating comparative effectiveness of treatment sequences. This is especially pertinent in the context of re-evaluating treatment appraisals or synthesising an external treatment-sequencing control arm. Secondly, it is equally crucial to delve into the application of causal inference methods in analysing clinical trials that harbour sequencing information. This approach holds promise for providing timely estimations of the appraised treatment's effectiveness within a sequence, offering valuable insights during the appraisal process. Thirdly, under-researched areas should be explored,

including the generalisability of model optimisation methods, equity, utility, and costs in treatment sequence evaluation. Finally, an updated review of treatment-sequencing economic evaluation applications may provide valuable insights into the evolving field.

2.7 Chapter summary

My review reveals an expanding literature on synthesising clinical evidence for treatment sequences, with existing studies suggesting the use of meta-analytical methods and ITC adjustments to collate trial evidence from diverse sources to derive line-of-therapy (LOT)-specific treatment effects.[1,34-36] Yet, these methods significantly rely on an assumption that patient populations align consistently across different data sources. The review explored the perception of RWD in current treatment sequencing studies: RWD was valued for its ability to capture treatment sequencing details but necessitates further investigation to better understand its utility and to tackle inherent challenges. This leads to an in-depth review in Chapter 3 on the use of different data sources for assessing treatment sequencing effectiveness within English HTA, with a focus on the existing practice of employing RWD.

Chapter 3 Comparative effectiveness of treatment sequences in health economic evaluations—a systematic review of NICE TAs

3.1 Overview

In this chapter, I systematically reviewed NICE technology appraisals (TAs) to delineate the current practice and challenges of comparing treatment sequences in health economic evaluations. The aim was to investigate how frequently sequencing questions were raised in NICE TAs, how they are addressed, and specifically, how effectiveness estimates for treatment sequences were derived in economic models. This investigation builds on the findings from Chapter 2, where deriving effectiveness estimates was highlighted as a key challenge in HTA involving treatment sequences.

The initial review in Chapter 2, which informed the required subsequent reviews, shows that the methodological literature on treatment sequences initially focused on the structure and computational optimisation of treatment-sequencing models.[1,2,30,31] Most did not discuss in detail the challenge of deriving unbiased comparative effectiveness estimates for treatment sequences, despite this being highlighted as a key aspect requiring further exploration.[1,36,102] One exception was a PhD thesis by Lewis highlighting the importance of making necessary simplifying assumptions in combination with meta-analysis approaches to derive LOT-specific effectiveness estimates.[34,36] Randomised controlled trials (RCTs) are often considered as the most reliable source to reflect comparative treatment effectiveness in cost-effectiveness models (CEMs). However, it is self-explanatory that a RCT comparing all possible treatment sequences in a decision problem is often non-existent due to the requirement of a larger sample size, longer follow-up, and ethical considerations.

In view of the clinical evidence scarcity issue, I became aware that non-RCT data sources, such as real-world data (RWD), can be useful in informing treatment-sequencing models, especially in reflecting local treatment patterns (Chapter 1 & 2).[184,185] However, the use of alternative data sources to inform economic evaluations in treatment-sequencing models—especially with data from non-randomised patients—remains unclear. It is important to understand what alternative data sources of effectiveness evidence have been used in treatment-sequencing models, and what assumptions and adjustments methods have been used in conjunction with them.

In this chapter, I conducted a systematic review to address my research questions, with two main objectives: (1) To investigate how prevalent treatment sequences have been discussed within NICE TAs regardless of whether a treatment-sequencing model was used, and their primary intentions; (2) To identify challenges and common sources of data for selecting appropriate effectiveness estimates to compare treatment sequences. I considered a systematic review of NICE TAs to be sufficient and appropriate to address my research questions as they are representative HTA examples with national resource allocation implications in England. Moreover, the rigorous and well-established NICE

appraisal process often leads to reports involving innovative methods to learn from.

The specifics of review objectives are detailed in Section 3.2. Section 3.3 outlines the review design. Section 3.4 presents the review results, Section 3.5 discusses and compares these findings with the literature, and Section 3.6 summarises with the review's key insights and their implication for the rest of my thesis.

3.2 Objectives and review questions

In this section, I outline two main objectives for the review, aimed at expanding both the breadth and depth of previous NICE TA reviews on treatment sequences.[1,2] Hence, the review is split into two parts, Part A and Part B, respectively. First, previous reviews focused on TAs that explicitly incorporated treatment sequences as part of the decision-analytic model structure, overlooking instances where treatment sequences are relevant but not explicitly modelled. An example is the use of partitioned survival models (PartSM) that reply on effectiveness evidence of front-line treatment with subsequent treatments reflective of those used in the NHS without necessitating an explicit economic model structure for multi-line treatments. These nuances were not captured in previous NICE TA reviews by Zheng et al. and Viola et al.[1,2] Therefore, Part A of this review aims to provide an overarching understanding about the prevalence of TAs involving treatment-sequencing considerations, not just those with a sequencing model. Specific review questions about this are listed in Table 3.1, Part A.

Table 3.1 Review questions of the NICE TA treatment-sequencing review

Part A: What is the prevalence of treatment-sequencing considerations in NICE TAs?
➢ What is the prevalence of treatment-sequencing considerations in all NICE TAs? ➢ In what context have treatment sequences been considered in NICE TAs? ➢ What is the prevalence of economic models with a treatment-sequencing structure? ➢ What are the key disease areas involving treatment-sequencing issues?
Part B: What is the current practice of deriving treatment-sequencing comparative effectiveness? What are the characteristics of models (i.e. structure), effectiveness estimates (i.e. model assumptions, adjustments, data sources), challenges and justifications? *This part of the review only includes TAs that has a purpose of comparing alternative treatment sequences
General characteristics and treatment-sequencing models ➢ In what position (i.e. line) was the treatment being appraised? ➢ Has a de-novo treatment sequencing model been used? If yes, what are the details? ***Characteristics of treatment-sequencing effectiveness estimates*** ➢ If treatment sequences were considered, has the effectiveness of a whole treatment sequence been estimated? [Yes/No] ✧ If not, has it stated why not? ✧ If yes, how was this done? Document data sources, assumptions, adjustments methods, and challenges if applicable.

TA: technology appraisal

Next, to extend the depth of previous reviews, Part B of the review focuses on describing current practice in generating effectiveness inputs when comparing treatment sequences. This topic has only been touched on as data scarcity issue in previous reviews.[1,2] Several techniques have been used to combat the scarcity of line-specific treatment effect, such as applying an effect modifier to reflect the potential effect degradation of subsequent treatments. However, applying these methods may rely on strong assumptions about the interaction of previous and subsequent treatments, which are often without any data to validate. As such, I formulated detailed review questions in Table 3.1, Part B, aiming to document the current practice and challenges in populating effectiveness estimates for comparing treatment sequences in HTA, aiming to identify common assumptions, adjustment methods, and data sources in supporting decisions making. Due to the complex nature of NICE TA documents for practicality, Part B was restricted to TAs that explicitly compare two or more treatment sequences. The rationale behind this choice is detailed in Sections 3.3.2 and 3.3.3. Figure 3.1 illustrates the scope of Parts A and B in relation to earlier reviews.[1,2]

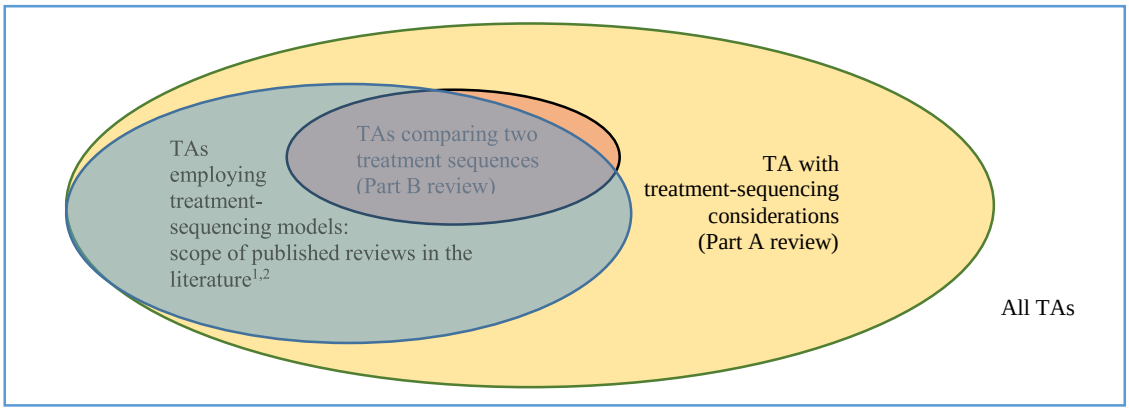


Figure 3.1 Schematic expression of the scope of NICE TA treatment-sequencing review

TA: technology appraisal

3.3 Methods

3.3.1 Methods overview

In the methods section, I begin by detailing NICE TAs I selected for the review (Section 3.3.2), and how I defined the subsets of TAs for each part of the review (Section 3.3.3). Given the complex nature of TAs compared to journal articles, Section 3.3.4 details the specific documents reviewed within each TA. Section 3.3.5 presents the screening strategies developed to identify relevant TAs, while Section 3.3.6 addresses the review's challenges and details the tailored process for screening, eligibility assessment, and data extraction. Section 3.3.7 introduces synthesis strategies for summarising the extracted information. Finally, Section 3.3.8 describes how I automated the process

to systematically retrieve full-text TA documents (i.e. a type of grey literature) to minimise bias.

3.3.2 Searches

I included all NICE TAs updated before November 30, 2019, in my screening, compiling the final list of TAs from NICE's website on December 1, 2019, to ensure transparency in light of possible redactions to some TAs later.[186] This review was registered at The International Prospective Register of Systematic Reviews (PROSPERO) prior to kick-off.[187]

3.3.3 Review subsets

Part A covers all TAs that mentioned considerations around treatment sequences, while Part B narrows its focus to only TAs that listed at least two treatment sequences as comparators. This approach, chosen for its practicality due to the extensive number of appraisal, fits well with Part B's in-depth review of how effectiveness for treatment sequences is derived. Specifically, it is likely to be more relevant for appraisals that clearly defined comparisons between treatment sequences.

3.3.4 Types of TA documents to be reviewed

Reviewing NICE TAs presents unique challenges compared to reviewing journal articles due to the NICE HTA process's complexity, necessitating clear definitions of which documents within each TA to review. Basic NICE TAs include single TAs (STA), multiple TAs (MTA) and fast track TAs (FTA). Here, I outline the publicly available documents for each TA subtype and specify which ones were selected for inclusion in this review.

For single STAs, each TA includes a company submission (CS) and a comprehensive evidence review from an independent academic centre (i.e. the evidence review group (ERG)) that scrutinises the CS report. The CS and ERG reports together with other expert opinions (e.g. clinicians and patient groups) will then be considered by the NICE Appraisal Committee to assist drug reimbursement decision making. For MTAs, there are CS from each manufacturer followed by an Assessment Group (AG) report that reviews each of the CS reports. Unlike ERGs, the AG not only critiques what the companies have done, but also undertakes a systematic review of efficacy and safety, and builds their own model to address the decision problem. That is, the role of the independent group is expanded from reviewing CS reports, to actually doing the modelling itself and undertaking a full systematic literature review on the topic. In summary, each TA is comprised of at least an original CS report at the consultation stage, an ERG or AG report that scrutinises the CS report, and a final appraisal document (FAD) that summarises the decision of NICE’s Appraisal Committee. FADs capture most of the important discussions related to the final recommendations of a TA, while CS and ERG/AG reports offer additional details to an appraisal, including a full report of clinical- and cost-effectiveness analysis (CEA).

I focused on three main publicly available documents per TA: (1) CS, (2) ERG or AG report, and (3) FAD. When a full CS was unavailable, I reviewed the executive summary of the CS report instead (if available), often the case for MTAs. I primarily reviewed the initial ERG/AG report for comprehensive details. If the TA process included up to 2 consultations, I also reviewed addenda and company responses wherever relevant. In instances of multiple FADs due to appeals, I reviewed all FADs. Other documents like matrices of consultees/commentators were not reviewed.

3.3.5 Pilot review and screening strategies

From January to February 2020, I undertook a pilot TA review to develop a list of "sequence terms" for TA screening. Prior to my pilot review, I compiled an initial set of sequence terms from studies included in Chapter 2's initial review.[1,2,8,105] I then reviewed 35 randomly chosen TAs to expand these strategies, finalising the list with my supervisors. Both the initial and final screening strategies are outlined in Table 3.2.

Table 3.2 Development of the screening strategy for treatment sequences

<table>
<tr><th></th><th colspan="2">Sequence terms</th></tr>
<tr><td>Initial set</td><td colspan="2">➢ optimal sequence, treatment sequence, discrete treatment options, optimal position, subsequent therapies, subsequent treatments, sequential therapy, trends of multimodality, pathways, changing points, switching
➢ excluding "consequence", "sequencing" (if referring to gene sequencing or randomization sequencing)</td></tr>
<tr><td rowspan="15">Final set</td><td>Terms</td><td>Potential equivalent variations</td></tr>
<tr><td>treatment sequenc*</td><td rowspan="4">sequence of treatments, sequence of lines of treatments</td></tr>
<tr><td>therapy sequenc*</td></tr>
<tr><td>optimal sequenc*</td></tr>
<tr><td>optimum sequenc*</td></tr>
<tr><td>sequential treatment*</td><td rowspan="4">sequential strategy</td></tr>
<tr><td>sequential therap*</td></tr>
<tr><td>subsequent treatment*</td></tr>
<tr><td>subsequent therap*</td></tr>
<tr><td>treatment switching</td><td rowspan="5">switching therapy, switching treatment</td></tr>
<tr><td>therapy switching</td></tr>
<tr><td>clinical pathway</td></tr>
<tr><td>treatment pathway</td></tr>
<tr><td>care pathway</td></tr>
<tr><td colspan="2">Look-alike but should-be-excluded terms:
➢ Sequencing referring to gene sequencing or randomization sequence generation.
➢ Sequence that is a part of words like "consequence".
➢ Clinical pathway that refers to biochemical pathway e.g. inflammatory, or receptor signaling pathway.
➢ Treatment pathway or care pathway that only serve as a subtitle in the report (often seen in company submission template).
■ Describing the position of intervention in the care pathway of a disease without discussing the impact to the whole treatment sequence pathway
■ Only to encourage incorporating TA guidance into local care pathway</td></tr>
</table>

3.3.6 Screening, eligibility assessment and data extraction process

The pilot review highlighted significant challenges in conducting a broad, non-disease-specific review of NICE TAs (Section 3.3.5). Firstly, there was a significant rise in the number of published TAs since the treatment-sequencing reviews by Zheng et al. and Lewis's thesis.[1,36] Before 2015, the annual publication of TAs ranged from 1 to 30; from 2015 to 2019, it increased to an average of 50 to 60, the major period of record retrieval for this review. Secondly, the document volume for each TA greatly exceeded that of standard journal articles. For example, TA375 involved seven CS reports—six above 200 pages and one over 1100 pages, an AG report exceeding 700 pages, and a FAD over 90 pages.[188] Reviewing just TA375 thoroughly was a challenging task, let alone all published TAs, which number over 600. Third, TA documents lack structured abstracts for screening, necessitating full-text screening. Fourth, no systematic review platforms or citation managers currently support NICE TA formats.

Hence, to enhance the efficiency and effectiveness of this review in addressing my research questions, I developed a set of strategies to navigate its complexity. Given the limitations of traditional review methodologies, a conventional two-stage abstract screening and full-text eligibility assessment were unsuitable for this context. As a result, I developed a tailored four-stage screening and assessment process, outlined below and visually represented in Figure 3.2:

- Stage 1 (Screening): I began by excluding TAs that did not feature any sequence terms defined in Section 3.3.5 through full-text screening of CS, ERG/AG reports, and FADs via the Ctrl + F function. This significantly narrowed the pool of TAs needing full-text review.
- Stage 2 (Full-text eligibility assessment—FAD Review): FADs, being the shortest documents, were reviewed first to filter out TAs not meeting the inclusion and exclusion criteria, further reducing the number of TAs requiring comprehensive review of CS and ERG/AG reports.
- Stage 3 (Full-text eligibility assessment for Part A: reviewing CS and ERG/AG documents): I then assessed each TA's eligibility against inclusion and exclusion criteria via full-text review of CS and ERG/AG reports, identifying TAs for Part A by the end of this stage.
- Stage 4 (Full-text eligibility assessment for Part B: reviewing CS and ERG/AG documents): an additional full-text review of CS and ERG/AG reports was performed to determine which TAs were eligible for the Part B review.

The definitions for treatment sequence and criteria for inclusion and exclusion are outlined in Table 3.3. Ambiguities were resolved through discussions with my supervisors for consensus. Data extraction details for each stage, as illustrated in Figure 3.2, are provided in Table 3.4 for Part A and Tables 3.5-3.6 for Part B

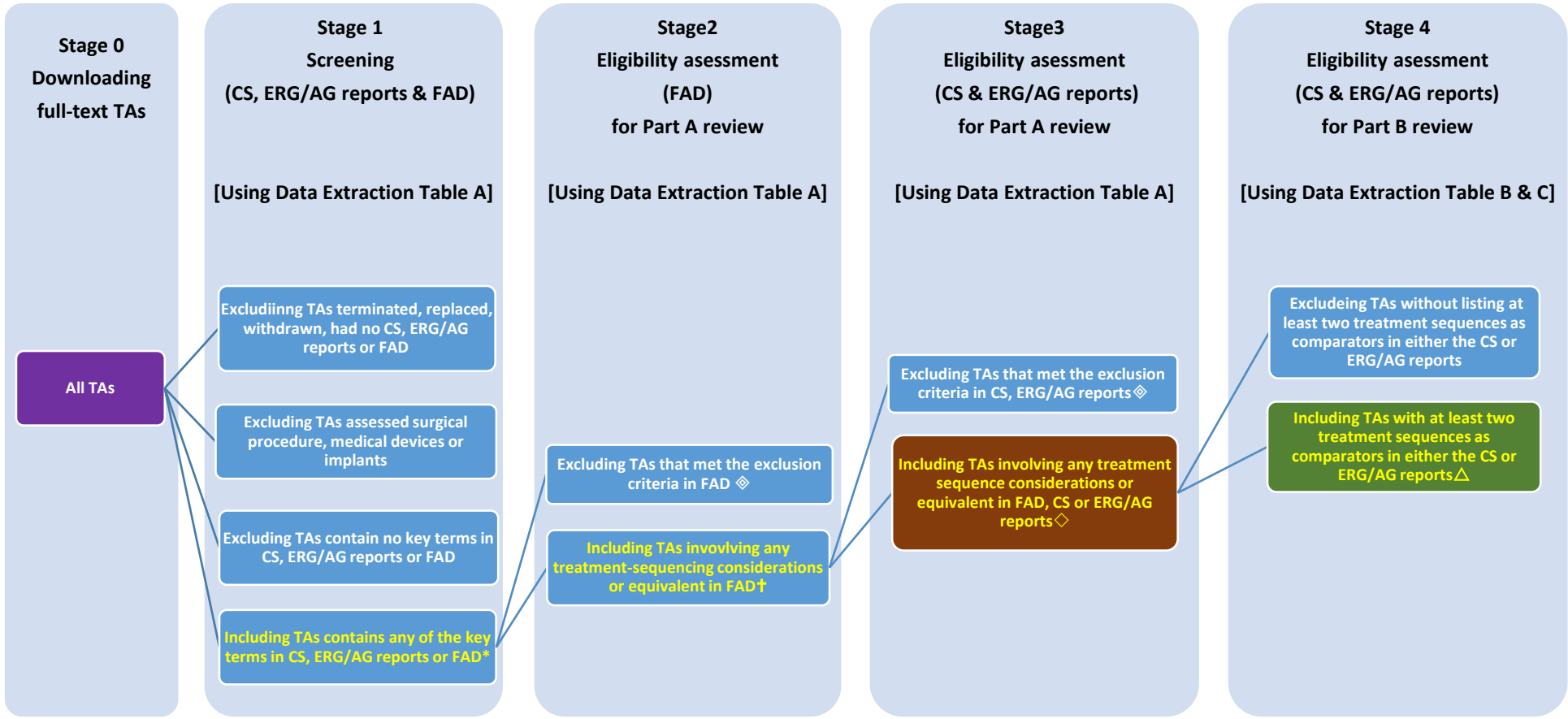


Figure 3.1 Schematic of screening and data extraction process

AG: assessment group; CS: company submission; ERG: evidence review group; FAD: final appraisal determination; TA: technology appraisal

Purple square: TAs entering demographic analysis; Brown square: TAs included for Part A review - to what extent has treatment sequences related issues been discussed; Green square: TAs included for Part B review - what techniques were used to compare treatment sequences; Text highlighted in yellow: TAs included at each stage of the screening and eligibility assessment

* Screened full-text, FAD, CS, and ERG report. Included a TA only if any of its three main documents (i.e. CS reports, ERG report, FAD) contains any of the relevant key terms or terms that deemed to be relevant. Extracted TA basic information in Data Extraction Table A.

† Reviewed full-text FAD, extracted treatment-sequencing relevant information in Data Extraction Table A.

◇ Reviewed full-text CS and ERG report, added additional information of treatment-sequencing relevant information (if not extracted from FAD) in Data Extraction Table A.

◈ Excluded TA that included key terms but did not discussed about treatment-sequencing in any way (based on full-text review of FAD, CS, and ERG report), or met other the exclusion criteria in Section 3.3.6 Eligibility Assessment.

△ Extracted detailed information from CS, or ERG model (or both). For MTA, only info of AG model will be extracted: information extracted are slightly different depending on whether a treatment-sequencing model structure is incorporate

Table 3.3 Inclusion exclusion criteria of the systematic review

Definition of treatment sequence	The definition was applied from a previous review of Zheng et al 2017[1], namely if there is any treatment switching due to clinical reasons, such as disease progression, adverse events, non-compliance or others. A sequence will not be considered if it is only assessing a single drug used in different clinical condition (e.g. the timing of initiating a drug).
Part A review: Inclusion criteria	All TA that contain any of the stakeholders' (company, ERG/AG, the NICE committee) discussion or opinions about treatment sequences (i.e. order of treatments).
Part A review: Exclusion criteria	➢ Exclude TAs if the only mention of a treatment sequence in any of the appraisal documents does not relate directly to the appraisal in question. e.g. treatment switching mentioned when describing one source of evidence included in the clinical-effectiveness assessment, but was not used to justify any sequence of the assessed new technology. ➢ Exclude sequences referring to fixed treatment regimen bundle (e.g. dose escalation) or established chemotherapy regimen to be given as a single line of treatment Exclude a TA if sequence is only mentioned in the reference list. Excluded TAs assessing medical devices, implants, or surgical interventions as they were less comparable to other TAs in the context of this review.
Additional exclusion criteria for the subset of Part B review	Part B of the review is a subset of part A: TAs without any specification of at least two specific treatment sequences (not treatment mix) as comparators in either the CS, ERG/AG report were excluded in Part B of the review.

AG: assessment group; CS: company submission; ERG: evidence review group; FAD: final appraisal determination; TA: technology appraisal

Table 3.4 Data Extraction Table A: basic TA information and treatment-sequencing discussion extraction for TA screening and Part A review

	Items for data extraction	
Documents	**For all TAs**	**Additional questions for TAs included in the Part A review**
General questions	➢ TA number [number] ➢ TA name [free-text] ➢ Disease [free-text] ➢ Disease area [free-text] ➢ Last published date [year]	➢ Primary reasons for mentioning treatment sequences in TAs [free-text, up to three reasons] *
CS	➢ File name [free-text] ➢ Page number [number] ➢ Included key terms? [Yes/No] ➢ Page number/key term	➢ Has a treatment-sequencing model? [Yes/No] ➢ At least two treatment sequences as comparators? [Yes/No]
ERG/AG	➢ File name [free-text] ➢ Page number [number] ➢ Included key terms? [Yes/No] ➢ Page number/key term	➢ Has a treatment-sequencing model? [Yes/No] ➢ At least two treatment sequences as comparators? [Yes/No]
FAD	➢ File name [free-text] ➢ Page number [number] ➢ Included key terms? [Yes/No] ➢ Page number/key term	-

AG: assessment group; CS: company submission; ERG: evidence review group; FAD: final appraisal determination; TA: technology appraisal

Table 3.5 Data Extraction Table B: data sources and statistical methods for comparing treatment sequences (Part B review)

General Information

- TA number [number]
- MTA or STA [number]
- Are there at least two treatment sequences as comparators in either or both CS or ERG/AG reports? [Yes/No] If Yes, then fill in the following questions.

Company Submission and AG report: treatment-sequencing effectiveness estimates (information from Company Submissions in MTAs are extracted from AG report if available)

- Completeness of the CS report? [full CS report, extracted from AG report or others]
- How many lines of treatments are there in a treatment sequences? [number]
- Are treatment sequences with different length (i.e. total lines of treatment) being compared? [Yes/No]
- In what position (i.e. line) is the treatment being appraised being placed in the treatment sequence? [number]
- Type of switching to different lines of treatment [free-text: disease relapse, others]
- Type of effectiveness outcome? [free-text; absolute treatment effect (e.g. scores), number or proportion of responders, relative risk, time to event parameters: survival, time to event parameters: treatment duration, treatment duration), hazard ratio, transition probabilities to the next health state, others]
- If treatment sequences are considered, is effectiveness of a whole treatment sequence for different sequences estimated? [yes/no]
    - If not, is it stated why not? [free-text; only considered impact on costs, unspecified, others]
    - If yes, how was this done?
        - Degradation style assumptions [yes/no]
            - If yes, for which part of the model and details? [free-text]
        - Using pivotal-trial RCT data [yes/no]
            - If yes, for which part of the model and details? [free-text] and;
            - What are the features of the pivotal trial? [free-text; RCT with two or more arms, single arm trials, SMART design]
            - What type of statistical methods were used and details? [free-text: subsequent treatments in trial matches those in the appraisal/NHS, subgroup analysis, adjustment for treatment-switching, sensitivity analysis, external source to guide the choice of treatment sequences as comparators, others]
        - Using other trial data [yes/no]
            - If yes, for which part of the model? [free-text] and;
            - What are the features of the trials? Specify separately if multiple sources were used. [free-text; trials comparing sequential treatments, effectiveness of a specific line, effectives of blended-line, SMART design]
            - What type of statistical methods were used and details? [free-text; model inputs: subsequent treatments in trial matches those in the appraisal/NHS, subgroup analysis or adjustment for treatment-switching or indirect comparisons, external source to validate model inputs or to guide the choice of treatment sequences as comparators, others]
        - Using NMA/MA data [yes/no]
            - If yes, for which part of the model? [free-text] and;
            - What are the features of the NMA/MA? Specify separately if multiple sources were used. [free-text; line-specific NMA/MA, treatment-history specific NMA/MA, unspecified-line NMA/MA]
            - What type of statistical methods were used and details? [free-text; model inputs: subsequent treatments in trial matches those in the appraisal/NHS, subgroup analysis or estimating the effect

of a treatment modifier (e.g. discontinuation rate), external source to validate model inputs or to guide the choice of treatment sequences as comparators, others] • Using RWD [yes/no] ▪ If yes, for which part of the model? [free-text] and; ▪ What are the features of the RWD? Specify separately if multiple sources were used. [free-text; sequential treatment study, effectiveness of specific lines] ▪ What type of statistical methods were used and details? [free-text; model inputs: subsequent treatments in study matches those in the appraisal/NHS, subgroup analysis or estimating the effect of a treatment modifier (e.g. discontinuation rate), external source to validate model inputs/model assumptions or to guide the choice of treatment sequences as comparators, others] ✧ What are the key challenges reported for estimating the comparative effectiveness of treatment sequences? ✧ Is the focus on estimating the effectiveness of sequences likely to be used in NHS clinical practice (i.e. adjusting for subsequent treatments), or on comparing the effectiveness of alternative sequences, or others if any? [free-text; effectiveness of treatment sequences is not estimated in anyway, sequences likely to be used in the NHS, clinical practice, to reflect the trial setting, to reflect evidence seen in registry or RWD, expert opinions, inconsistent with previous TAs, to assess which treatment lines new treatment belongs as the most cost-effective treatment, others] ➢ Besides effectiveness, does treatment-sequencing impact any elements (other than costs and quality-of-life measures) that should be noted [free-text: scope of appraisal, others]
ERG/AG comment on treatment-sequencing effectiveness estimates in the CS reports
➢ Has ERG or AG made any criticism of the company's submission regarding effectiveness estimates of treatment sequences? [yes/no] If yes, what are the comments and criticisms? [free-text] ➢ Any additional analyses undertaken or requested by ERG regarding estimating comparativeness of treatment sequences? If yes, document the details [free-text; sensitivity analysis, scenario analysis]
FAD
➢ What issues did the Committee raise regarding treatment sequences in addition to those mentioned by the company and the ERG [free-text]?

AG: assessment group; CS: company submission, ERG: evidence review group; FAD: final appraisal determination; MA: meta-analysis; MTA: multiple TA; NHS: National Health Service; NMA: network meta-analysis; RCT: randomized control trial; RWD: real-world data; SMART: The Sequential Multiple Assignment Randomized Trial; STA: single TA; TA: technology appraisal

Table 3.6 Data Extraction Table C: treatment-sequencing model information (Part B review)

Model features	Data extraction
TA number	[number]
Disease	[free-text]
Treatment-sequencing model structure	➢ base-case [Yes/No] ➢ additional analysis [Yes/No]
Model source	[CS or ERG/AG]
Position of the new treatment in the treatment pathway (i.e. LOT(s))	[free-text]
Time horizon	[free-text]
Model Type	[free-text; e.g. Markov model]
Treatment sequences evaluated in the economic evaluation	[free-text; list all treatment sequences if possible]
Model health states/model health state transitions (only document the longest in the model)	[free-text]

AG: assessment group; CS: company submission; ERG: evidence review group; LOT: line of treatment

3.3.7 Strategies for data synthesis

The section details data synthesis strategies for addressing review questions from Section 3.2, organised into the following categories:

(1) Inclusion and exclusion process

- Summarising the selection process with A PRISMA flow diagram[189]
- Summarising the reasons for each excluded TA in Part A of the review.

(2) Part A review: Prevalence of treatment sequences issues

- Summarising the number of TAs with treatment-sequencing and stratifying them by the year of publication and type of diseases
- Summarising the number of treatment-sequencing discussions and treatment-sequencing models among all TAs and comparing them against findings from previous literatures[1,2]
- Summarising the primary reasons for mentioning treatment sequences in TAs
- Summarising the number of TAs containing treatment-sequencing models and stratifying them by disease areas and the TA process (i.e. STA versus MTA)

(3) Part B review: Comparative effectiveness estimates of treatment sequences

- Summarising the data sources, major assumptions, relevant statistical adjustment methods and justifications used in each TA using tables or narration, whichever is applicable
- Narratively describing the key challenges for deriving comparative effectiveness of treatment sequences and categorising them by disease areas. Further elaborating distinctive examples and illustrating them in graphs if applicable.
- Summarising the features of treatment-sequencing models of each included TA, if applicable

If possible, propose a preliminary framework that summarise the pros and cons, and potential hurdles in employing different data sources to derive comparative effectiveness of treatment sequences, drawing upon insights from the TA review and established HTA guidelines.

3.3.8 Systematic full-text retrieval

The review of each TA necessitates reading multiple documents (see Section 3.3.4) not retrievable by standard literature software, such as Endnote. To address this, an R script with web-scraping capabilities was developed for systematic document download from the NICE website (i.e., web scrapping), minimising manual bias arising from manual operation. For the R script, see Appendix 3.1.

3.4 Results

3.4.1 Inclusion and exclusion

Figure 3.3, a PRISMA diagram, outlines the inclusion and exclusion process. Initially, 460 TAs

(up to TA613) were identified. Title screening led to the exclusion of 49 TAs without available documents and 30 assessing medical devices, implants, or surgeries. Furthermore, 59 TAs were excluded during full-text screening, with an additional 25 during full-text assessment, resulting in 297 TAs considered for Part A, focusing on detailing treatment sequence considerations. Appendix 3.2 outlines the exclusion reasons for each excluded TA.

For Part B, 262 TAs that did not explicitly a comparison of at least two treatment sequences in their CEA were further excluded, resulting in 35 for inclusion. Sections 3.4.2 and 3.4.3 report the findings for Parts A and B, respectively. Appendix 3.3 contains a summary table detailing information on each TA included in Part A, such as the presence of treatment sequences as comparators and the existence of any de novo treatment-sequencing models.

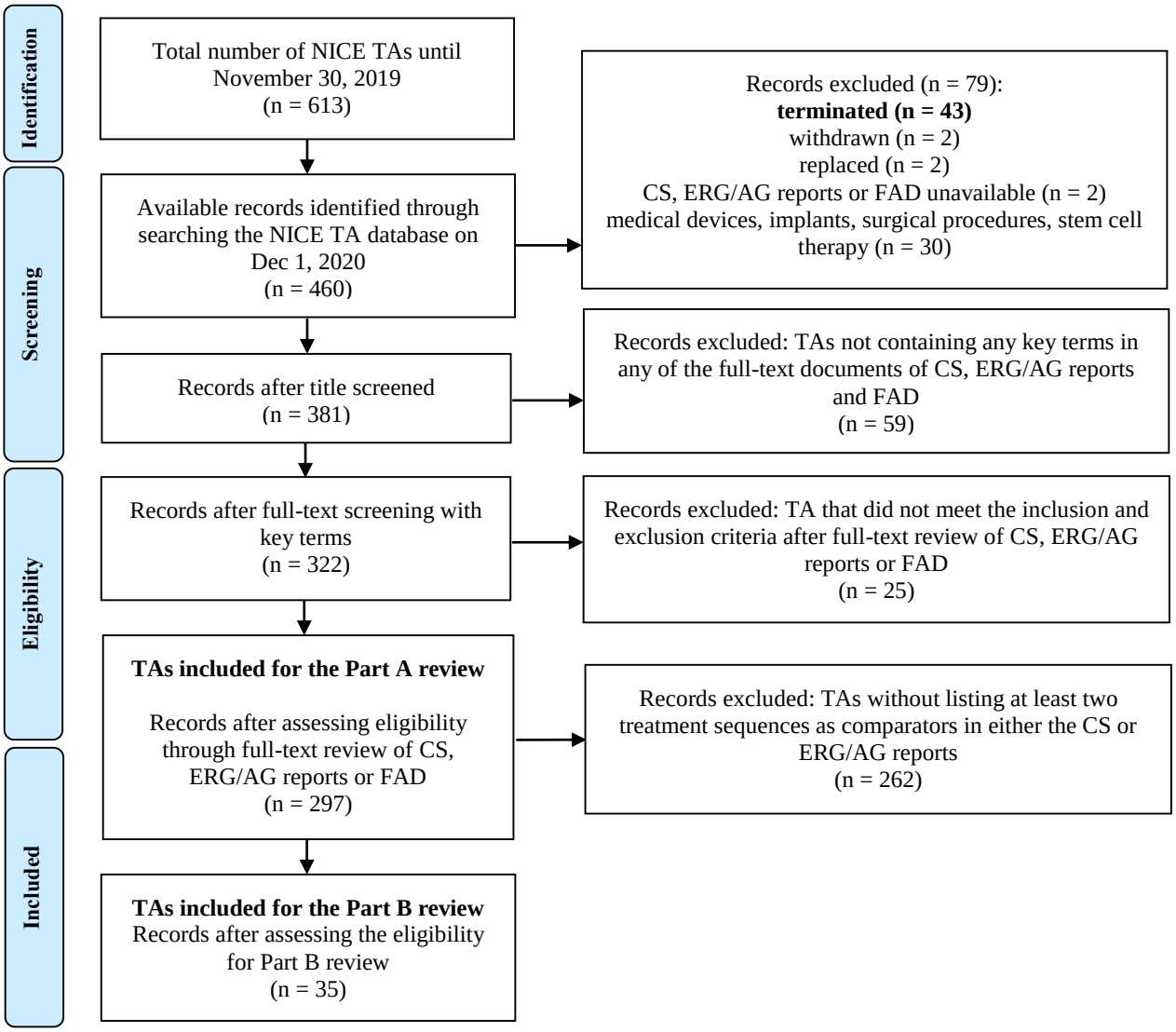


Figure 3.3 The Preferred Reporting Items for Systematic Reviews (PRISMA) diagram for the systematic review of treatment sequences in NICE TAs

AG: assessment group; CS: company submissions, ERG: evidence review group; FAD: final appraisal determination; TA: technology appraisal

3.4.2 Part A review: treatment-sequencing considerations in NICE TAs

3.4.2.1 Prevalence of treatment-sequencing considerations

Figure 3.4 shows the trend in TA publications by year, separating those with treatment-sequencing considerations (included in Part A) from those without. It reveals an uptrend in the topic over the last two decades, especially from 2015 onwards. This rise, however, might be overstated due to missing early documents and could also reflect the broader increase in TA outputs since 2015 (i.e. over 50 TAs per year).

Over three-quarters (297 out of 381, 78.2%) of the TAs discussed treatment sequences or equivalent concepts, with a notable increase in the number of TAs employing treatment-sequencing models, listing treatment sequences as comparators, or both between 2016 and 2018 (16-18 TAs per year). This prevalence of treatment-sequencing discussions varies across disease areas, with oncology TAs leading in mentions of treatment sequences (Figure 3.4). However, few oncology TAs explicitly compared the cost-effectiveness of treatment sequences. For instance, non-small cell lung cancer (NSCLC) TAs had the highest mention of treatment sequences among all diseases, with 25 TAs discussing them. However, only two of these NSCLC TAs employed treatment-sequencing models, and none made direct comparisons of treatment sequences. This trend was consistent across various oncology diseases, including leukaemia, breast cancer, lymphoma, multiple myeloma, melanoma, renal cell carcinoma, colorectal cancer, prostate cancer and ovarian cancer. In contrast, autoimmune disease TAs more often made explicit sequence comparisons and employed treatment-sequencing models.

To compare my findings with existing NICE TA reviews[1,2], I categorised them by three themes and seven broad disease areas as shown in Table 3.7. These themes are: (1) including treatment-sequencing considerations, (2) employing treatment-sequencing models, and (3) listing at least two treatment sequences as comparators. Theme (2) aligns closely with the scope of previous literature[1,2], while themes (1) and (3) correspond to the focus of Parts A and B of my review. My review found a higher percentage of treatment-sequencing models being used across various disease areas, notably in oncology, autoimmune diseases, diabetes, and infectious diseases, compared to previous reviews by Zheng et al. and Viola et al.[1,2] Autoimmune diseases featured the most frequent use of these treatment-sequencing models (50.8%) and treatment sequence comparators (36.1%), with diabetes next (44.4% and 11.1%). Oncology, despite a high discussion prevalence regarding treatment sequences (76.0%), showed lower rates of treatment-sequencing model usage (18.8%) and sequence comparisons (3.1%). However, with more than half of the TAs discussing treatment sequences (174 out of 297) in oncology, this field had the highest count of treatment-sequencing models used (n = 43).

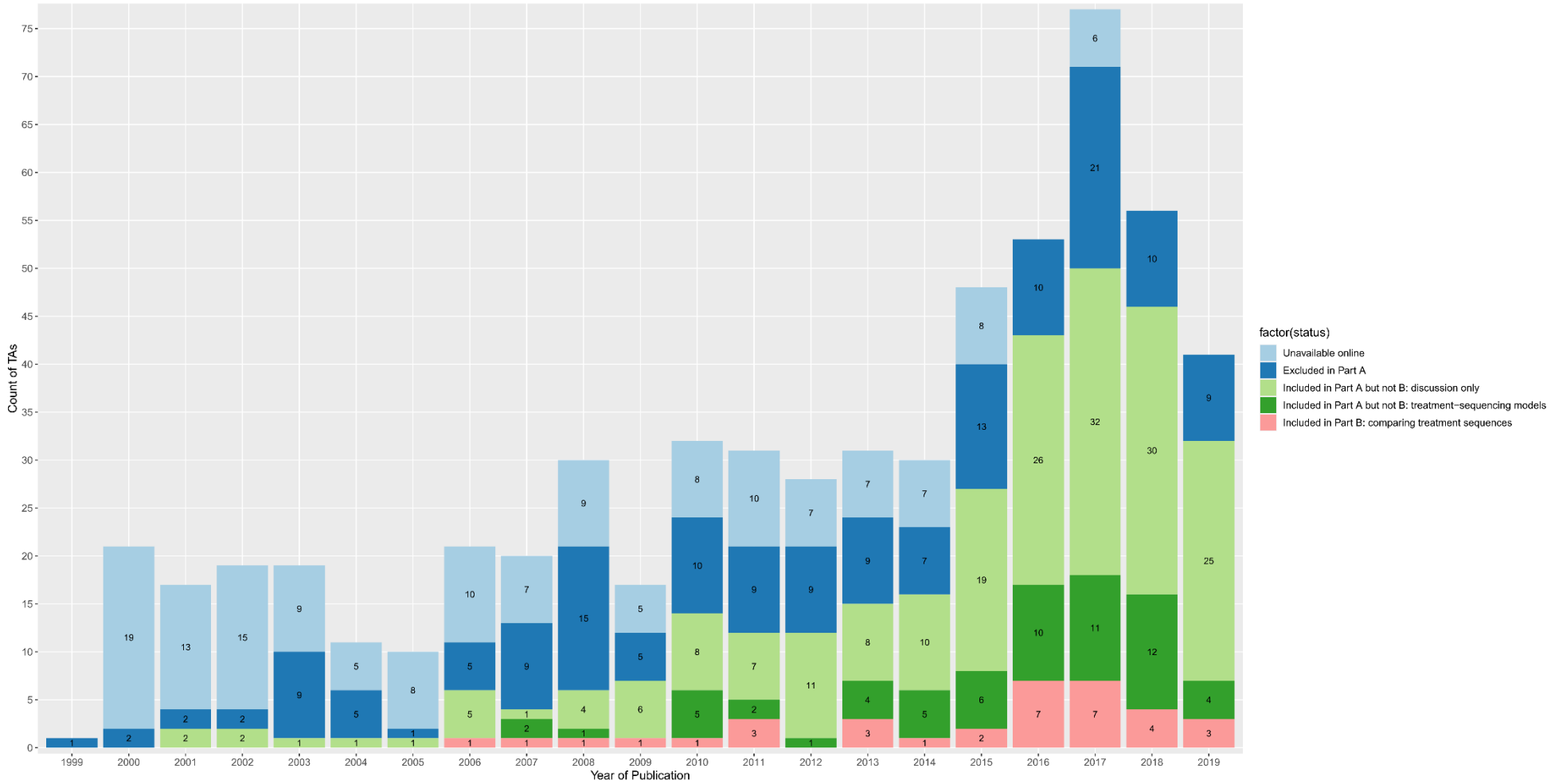


Figure 3.4 Summary of NICE TAs with treatment sequence discussions by year

Unavailable online: TAs used to exist, but was removed from the NICE web content; Excluded in Part A: none of the three key documents (i.e. FAD, CS report, ERG report) contains discussion about treatment sequences; Included in Part A but not B: discussion only: TAs that contain discussions about treatment sequences, but did not have a treatment-sequencing model nor compare treatment sequences. Included in Part A but not B: treatment-sequencing models: TAs that contain at least a treatment-sequencing model in the CS or ERG/AG report. Included in Part B: TAs explicitly specify at least two treatment sequences as comparators in the description of cost-effectiveness evaluation in either or both CS and ERG/AG reports. It should be noted that not all of these TAs have a treatment-sequencing model (but most of them). TAs published in December 2019 were not included in this review, therefore number of TAs may have been underestimated for this calendar year.

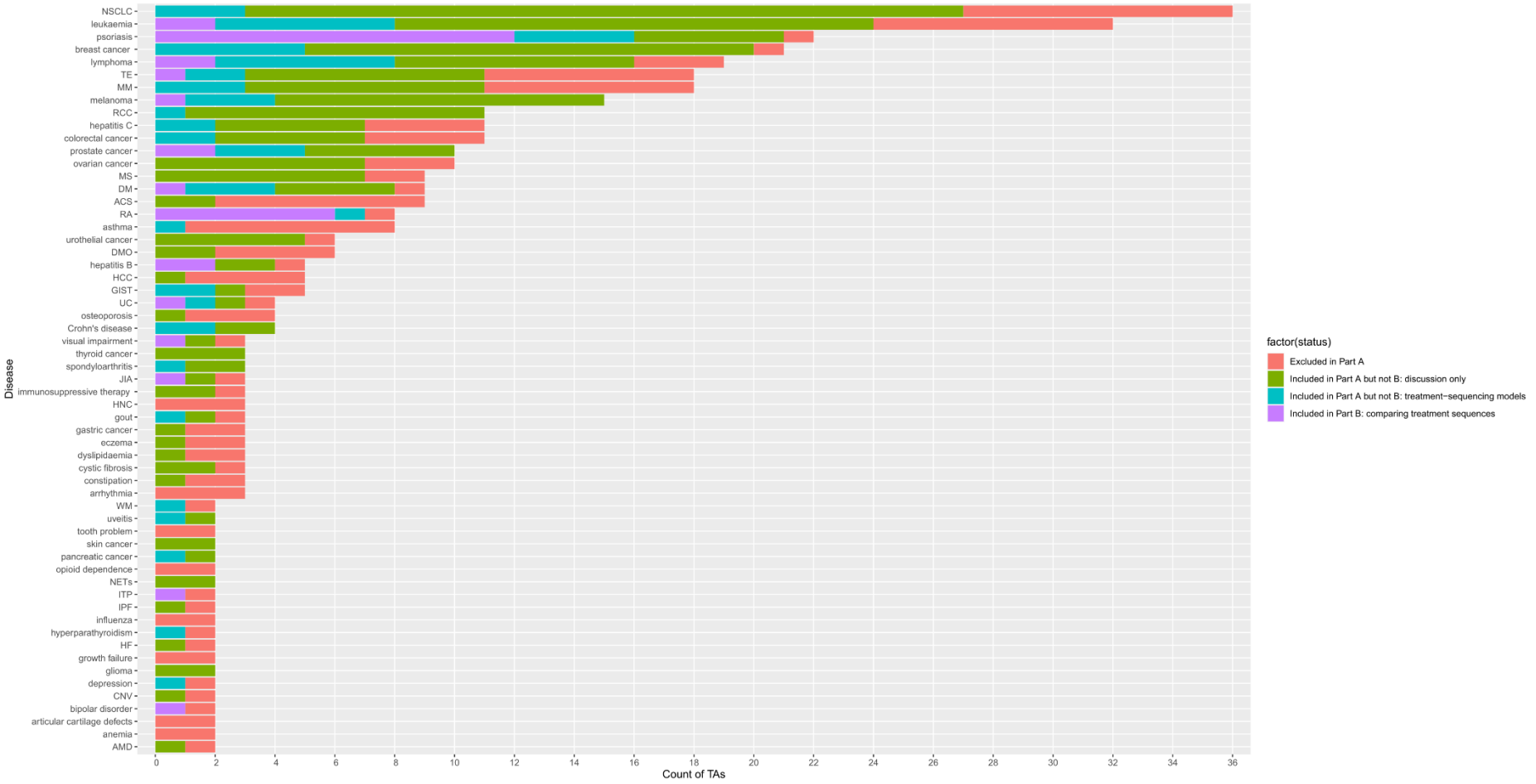


Figure 3.5 Summary of NICE TAs with treatment sequence discussions by diseases

GIST: gastrointestinal stromal tumour, HCC: hepatocellular carcinoma, HF: heart failure: HNC: head and neck cancer, IPF: idiopathic pulmonary fibrosis, ITP: immune thrombocytopenia, JIA: juvenile arthritis, MM: multiple myeloma, MS: multiple sclerosis, NETs: neuroendocrine tumours, NSCLC: non-small cell lung cancer, RA: rheumatoid arthritis, RCC: renal cell carcinoma, TE: thromboembolism, UC: ulcerative colitis, WM: Waldenstrom macroglobulinemia.
The y axis are the diseases of the TAs, while length of the bars on x axis depicts the total number of TAs within each disease. Red bars show the number of TAs excluded from the review. Green bars and blue bars represent TAs that were included in Part A of the review but not Part B of the review. The only difference of the green and blue bars is that TAs depicted by blue bars have treatment-sequencing models. The purple bars depict TAs that were included in Part B of the review, namely TAs that explicitly listed at least two treatment sequences as comparators in their cost-effectiveness analysis. It should be noted that although most of the TAs included in Part B has a treatment-sequencing model but this is not always the case. Only TAs that were available on the NICE website by the time of data extraction are included in this diagram. Due to limited graphic size, only diseases with at least two TAs were plotted. Diseases in this plot are ranked from top to the bottom by their total number of TA

Table 3.7 Summary of the treatment-sequencing NICE TA review in comparison with previous relevant reviews

Source of review	Zheng et al. 2017*	Viola et al. 2020*	The current review			
Range of TAs	TA1 – TA326 (until Oct 2014)	TA1 – TA527 (until June 2018)	TA1 – TA613 (until November 2019)			
Type of documents reviewed	For MTAs, only the ERG model was included	For MTAs or where the ERG built a de novo model different from manufacturers, only the ERG's model was considered	All key documents (i.e. CS, ERG, FAD) are considered, and all models in any of the key documents are considered. Both Zheng and Viola excluded studies without a complete CS report for reviewing models, I extracted the CS model information if sufficiently mentioned in the ERG report.			
Definition of treatment switching	Treatment switches due to clinical reasons, such as loss of efficacy (disease progression), intolerability (e.g. side effects or toxicity), non-compliance					
Criteria of treatment-sequencing models	Explicitly having a treatment-sequencing model structure	Explicitly having a treatment-sequencing model structure, but excluded models with only two lines of treatment or pre-post progression models	-	Same as Zheng 2017	-	-
Topic of the review	Examining TAs that employed treatment-treatment sequencing models	Examining TAs that employed treatment-treatment sequencing models and explored the selection of an appropriate baseline treatment-sequencing strategy	Part A of the review: Exploring TAs involving discussions about treatment sequences	A subset of Part A review: Identifying TAs that employed treatment-treatment sequencing models	Part B of the review (a subset of Part A review): Exploring TAs that listed at least two treatment sequences as comparators†	-

Number of TAs by disease areas (n/total n, %)			**Number of TAs by disease areas (n/total n, %)**		**Number of TAs by disease areas (n, %)**						**Total TAs**
Oncology	13/93	14.0%	23/206	11.2%	174	76.0%	43	18.8%	7	3.1%	229
Autoimmune	7/33	21.2%	14/53	26.4%	52	85.2%	31	50.8%	22	36.1%	61
Cardiovascular	6/26	23.0%	2/37	5.4%	16	43.2%	4	10.8%	1	2.7%	37
Neurology/mental health	4/13	30.8%	5/15	33.3%	6	60.0%	3	30.0%	1	10.0%	10
Infectious disease	2/12	16.7%	1/19	5.3%	12	60.0%	4	20.0%	2	10.0%	20
Diabetes mellitus	2/8	25%	3/10	30.0%	8	88.9%	4	44.4%	1	11.1%	9
Other	6/63	9.5%	2/94	2.1%	29	30.9%	7	7.4%	1	1.1%	94
Total	40/248	16.1%	50/434	11.5%	297	64.6%	96	20.9%	35	7.6%	460

* These statistics were drawn from those published in the Zheng et al. [1] and Viola et al.[2] study

† These observations do not always have a treatment-sequencing model structure

3.4.2.2 Features of treatment-sequencing considerations

I mapped the key reasons for considering treatment sequences in TAs into six categories, as shown in Table 3.8. Each TA was classified based on the first key term detected, adhering to a set priority derived from terminology variations, including "treatment sequences," "subsequent treatments," "treatment pathways," "treatment positions," and "treatment-switching". I extracted relevant descriptions from each TA and combined them to form finer reason categories based on subjective qualitative judgments. Discussions with supervisors helped refine and consolidate these into the final categories presented. For each reason, I provided up to five TAs in Table 3.8 as examples.

It was found that over three-quarters of TAs mentioned subsequent treatments, significantly more than other terms, while more than 10% mentioned the term treatment pathway/treatment positions. This likely stems from alignment with the terminology used NICE guidelines for TAs.[23] Hence, broad terms such as "subsequent treatments" and "treatment pathways/treatment positions" triggered further record for additional reasons when they were the sole focus, allowing up to three reasons to be noted for each TA's treatment sequence considerations.

The most common reasons include "considering subsequent treatments available within the NHS" (n = 229)[190-194], "assessing a new technology in different treatment sequences" (n = 28)[195-199], and "assessing the appraised technology using separate scopes (populations) within the TA depending on its candidate positions in the treatment pathway" (n = 23). Additionally, a few TAs set the comparison of treatment sequences as their main objective (n = 5)[155,200-203], with occasional discussions on the optimal sequence for appraised technology, including "considering the optimal treatment sequence of appraised technology in the treatment pathway" (n = 1)[204] and pointing out "uncertain optimal treatment sequences" (n = 3).[198,199,205] Beyond evaluating entire treatment sequences, several TAs focused on identifying the optimal placement for the technology under review. This involves: choosing a comparable position for the appraised technology in the treatment pathway relative to its comparators in a cost-comparison analysis (n = 1)[206], "exploring the treatment position of a new treatment" (n = 1)[207], and "assessing whether the appraised technology will change the current treatment pathway" (n = 1)[208].

In terms of economic modelling, a few TAs explicitly mentioned establishing de novo treatment-sequencing models, rather than adapting non-treatment-sequencing models from previously approved TAs (n = 3)[207,209,210]. Beyond the structure of treatment-sequencing models, one TA attempted to model survival curves for different sequences.[211] Several barriers to establishing treatment-sequencing models were highlighted, including "treatment sequences should be modelled, but outside of ERG remit" (n = 1)[212], "modelling treatment sequences will result in high uncertainty due to the lack of data" (n = 1)[213], and "no data for modelling treatment sequences" (n = 1).[214]

Table 3.8 Primary reasons for mentioning treatment sequences in TAs (Part A review)

Reasons for mentioning treatment sequences*	Number of TAs	Examples (listed up to 5)
Treatment sequences – assessment		
Assessing a new drug in different treatment sequences	28	TA197, TA409, TA426, TA154, TA171
Compare different treatment sequences	5	TA195, TA238, TA292, TA293, TA319
Considering different treatment pathways/sequences based on trial evidence	4	TA137, TA187, TA460, TA462
Treatment sequences/sequential treatments should be explored in future research	4	TA365, TA413, TA430, TA499
Uncertain optimal treatment sequences	3	TA154, TA171, TA178
Uncertain trial evidence of treatment sequences	2	TA391, TA396
Uncertain treatment sequences in NHS	1	TA359
Assessing two or more appraised treatments in treatment sequences within a MTA	1	TA535
Considering the optimal treatment sequence of appraised technology in the treatment pathway	1	TA212
Conducting treatment-sequence network-meta analysis	1	TA456
Treatment sequences - modelling		
Using a treatment-sequencing model rather than using previously approved models within NICE TAs	3	TA448, TA504, TA533
Modelling the survival of different treatment sequences	1	TA496
Treatment sequences should be modelled but outside of ERG remit	1	TA164
Modelling treatment sequences will result in high uncertainty due to the lack of data	1	TA455
No data for modelling treatment sequences	1	TA180
Subsequent treatments		
Considering subsequent treatments available in the NHS	229	TA605, TA607, TA610, TA611, TA612
Considering subsequent treatments available in the NHS based on trials	3	TA458, TA471, TA498
Considering the comparison of subsequent treatments available in the NHS and in trials	2	TA496, TA572
Considering the sequential use of treatments in clinical practice	2	TA164, TA199
Considering subsequent treatments available in trials	1	TA183
Treatment pathways		
Assessing whether the appraised technology will change the current treatment pathway	1	TA534
Treatment positions		
Assessing the appraised technology using separate scopes (populations) within the TA depending on its candidate positions in the treatment pathway	23	TA137, TA178, TA190, TA288, TA546
Choosing a proposed treatment position	2	TA190, TA388
Choosing a comparable position for the appraised technology in the treatment pathway relative to its comparators (cost-comparison analysis)	1	TA497
Exploring the treatment position of a new treatment	1	TA533
Treatment-switching		
Taking treatment-switching in clinical practice into account	14	TA161, TA266, TA298, TA345, TA482
Taking treatment-switching in trial into account	5	TA322, TA357, TA399, TA432, TA550
Taking treatment-switching in clinical practice and in trial into account	1	TA320

ERG: evidence review group, NICE: The National Institute for Health and Care Excellence, MTA: multiple TA, NMA: network meta-analysis, TA: technology appraisals, NHS: national health services.

* Only the primary reasons (up to three) regarding treatment sequences discussions of every TA were recorded. Treatment-switching in trials was a common issue in oncology TAs. Therefore, treatment-switching in trial was only recorded as a primary reason in the current review only if no other treatment-sequence related discussions were mentioned in a TA

With regards to clinical evidence, a subset of TAs discussed using existing evidence to ensure that appropriate treatment sequences have been considered, such as "considering different treatment pathways/sequences based on trial evidence" (n = 4)[215-218], "considering subsequent treatments available in the NHS based on trials" (n = 3)[219-221], "considering the comparison of subsequent treatments available in the NHS and in trials" (n = 2)[211,222], "considering the sequential use of treatments in clinical practice" (n = 2)[212,223], and "considering subsequent treatments available in trials" (n = 1)[224]. Treatment-switching associated considerations were towards utilising treatment-switching information from trials (n = 14)[225-229], clinical practice (n = 5)[230-234], or both (n = 1)[235] to inform what treatment sequences should be included for decision making. Advanced statistical analysis on existing data was also found to synthesise effectiveness evidence for treatment sequences (n = 1)[236] (i.e., treatment-sequence network-meta analysis). Despite the need of evidence, the scarcity of evidence to inform decisions involving treatment sequences was highlighted in a small number of TAs, including concerns about "uncertain trial evidence of treatment sequences" (n = 2)[237,238], "uncertain treatment sequences in NHS" (n = 1)[239], and "treatment sequences or the use of sequential treatments should be explored by future research" (n = 4)[240-243]. These coincide with the aforementioned clinical evidence scarcity challenges of in modelling treatment sequences.

3.4.3 Part B review: comparative effectiveness of treatment sequences in NICE TAs

Part B examined 35 TAs, predominantly STAs (n = 30), including a Cancer Drugs Fund (CDF) rapid review[197] and two rapid reviews with company Patient Access Scheme (PAS) submissions.[157,244] Additionally, it comprised four MTAs[155,188,245,246] and one FTA.[247] The TAs covered a range of disease areas, including 7 in oncology, 22 in autoimmune diseases, and 6 in other diseases.

I summarise the key findings in several tables: Table 3.9 includes essential information about each TA, including the disease topic, an indicator of any existing treatment-sequencing model, and data sources used to inform the effectiveness of treatment sequences. Details for treatment-sequencing comparisons for each TA are presented in Appendix 3.4, including the candidate position(s) of the appraised technology and treatment sequences compared. For TAs employing treatment-sequencing models, I documented their model specifications, including time horizon, model type, and model states. Finally, I synthesised the current practice and challenges to populate the effectiveness of treatment sequences in Table 3.10. The findings from these tables are further discussed in the sections below, including: (1) an overview of treatment sequence compactors (Section 3.4.3.1), (2) features of treatment-sequencing models (Section 3.4.3.2), (3) data sources used to inform the comparison of treatment sequences (Section 3.4.3.3), and (4) current practice and challenges to populate the effectiveness of treatment sequences (Section 3.4.3.4)

Table 3.9 Data sources used to inform comparative effectiveness of treatment sequences in cost-effectiveness analyses (Part B review)

	TA number	Sequencing-Model†	Data sources to inform the comparative effectiveness of treatment sequences				
			Effectiveness simplifying assumption	Pivotal trials	Other trials	NMA/ MA	RWE
Oncology							
Leukaemia	408	✓	✓	✓	✓	X	✓
	426△	X (CMA)	✓	NA	NA	NA	NA
		✓ (TA251)	X	✓	✓	X	X
Lymphoma	137	✓	X	✓	✓	X	✓
	462	X, but ✓ in amendment	X	✓	X	X	✓
Melanoma	319	✓	✓	✓	✓	X	✓
Prostate cancer	377	✓	X	✓	✓	X	✓
	580	✓	X	✓	✓	X	X
Subtotal (n)	***7***	***6 (86%)***	***3 (43%)***	***7 (100%)***	***6 (86%)***	***0 (0%)***	***5 (71%)***
Autoimmune							
ANCA-associated vasculitis	308	✓	✓	✓	X	X	✓
Immune thrombocytopenic purpura	293	✓	✓	✓	✓	X	✓
Systematic Juvenile Idiopathic Arthritis	238	✓	✓	✓	✓	X	✓
Psoriasis	103* AG	✓	✓	X	X	✓	✓
	419	✓ (PAS)§	✓	✓	X	✓	X
	433	✓ (PAS)§	✓	✓	X	✓	X
	442	✓	✓	✓	✓	✓	✓
	445* AG	✓	✓	X	X	✓	X
	UCB	✓	✓	✓	X	✓	X
	475	✓	✓	✓	✓	✓	✓
	511	✓	✓	X	X	✓	✓
	521▣	X (FTA)#	NA	NA	NA	NA	NA
	537	✓	✓	X	X	✓	X
	543	✓	✓	X	X	✓	X
	574	✓	✓	✓	X	✓	X
	575	✓	✓	X	X	✓	X
Rheumatoid arthritis	195 AG	✓	✓	✓	✓	X	✓
	Abbot¶	✓	✓	NA	NA	NA	NA
	Wyeth¶	✓	NA	NA	NA	NA	NA
	S-P¶	✓	NA	NA	NA	NA	NA
	Roche¶	✓	✓	NA	NA	NA	NA
	BMS¶	✓	NA	NA	NA	NA	NA
	225	✓	✓	✓	✓	✓	✓
	375 AG	✓	✓	X	X	✓	✓
	Abbvie¶	✓	NA	NA	NA	NA	NA
	BMS¶	✓	NA	NA	NA	NA	NA
	MSD¶◈	✓	NA	NA	NA	NA	NA
	Pfizer¶	✓	NA	NA	NA	NA	NA
	Roche¶	✓	NA	NA	NA	NA	NA
	UCB¶	✓	NA	NA	NA	NA	NA
	415	✓	✓	✓	✓	✓	✓
	466	✓	✓	✓	X	✓	X
	485	✓	✓	X	X	✓	✓
Ulcerative colitis	547	✓	✓	X	✓	✓	X
Subtotal (n)	***22***	***21 (96%)***	***21 (96%)***	***12 (55%)***	***8 (36%)***	***17 (77%)***	***12 (55%)***

	TA number	Sequencing-Model†	Data sources to inform the comparative effectiveness of treatment sequences				
			Effectiveness simplifying assumption	Pivotal trials	Other trials	NMA/ MA	RWE
Other diseases							
Cardiovascular							
Thromboembolism	249	✓	✓	✓	X	✓	✓
Neurology/mental health							
Bipolar disorder	292	✓	✓	✓	X	✓	X
Infectious disease							
Hepatitis B	154	✓¶	✓	✓	X	X	X
	173	✓¶	✓	✓	✓	✓	✓
Endocrine disorder							
Diabetes Mellitus	418	✓	✓	✓	X	✓	✓
Ophthalmology							
Macular oedema	409	✓	✓	✓	X	✓	✓
Subtotal (n)	***6***	***6 (100%)***	***6 (100%)***	***6 (100%)***	***1 (17%)***	***5 (83%)***	***4 (67%)***
Total (n)	***35***	***33 (94%)***	***30 (86%)***	***25 (71%)***	***15 (43%)***	***22 (63%)***	***21 (60%)***

ANCA: antineutrophil cytoplasmic antibodies, BMS: Bristol Myers Squibb, CDF: cancer drug funds, CMA: cost-minimisation analysis, DSU: Decision Support Unit, FTA: fast track appraisal, MA: meta-analysis, NA: not available, NMA: network meta-analysis, PAS: Patient Access Schemes, RWD: real world data, S-P: Schering-Plough, TA: technology appraisals
Only AG report are counted in subtotal and total.
* Comparing two or more sequences was only considered in AG report, but not in some of the company submissions (i.e., TA445: secukinumab (Novartis); TA 103: efalizumab (Serono) and etanercept (Wyeth))
▣ Effectiveness of full treatment sequences was not estimated in the company's submission because it was not feasible within FTA cost-comparison. However, costs of different treatment sequences were calculated.
△: The NICE DSU report considers two CDF review at the same time (TA 241 and TA 251 for TA 425 and TA 426, respectively). However, comparing two or more sequences was only considered in the original report of Company submissions and AG in TA 251, and the DSU CDF review report of the TA251 part. Data sources used in TA251 were informed by summary provided in TA426 as TA251 was no longer available on the NICE website. A cost-comparison analysis was conducted by the company in the CDF review (TA426). The company assumed that health outcomes for all treatments are equivalent, including treatment durations. Effectiveness of full treatment sequences was not estimated.
† : ✓= at least one treatment-sequencing model from either company submission report, ERG report or AG report; X = There was economic model, but none of the models contain an explicit treatment-sequencing structure.
§ Patient Access Schemes submission: modelling details unavailable.
# No economic models are required for FTA submission.
⁋ Information unavailable due to the unavailability of full CS submission.
◈ MSD prepared their submission for two drugs: golimumab and infliximab, so two CS reports (both containing comparisons of treatment sequences and sequencing models) were submitted.
¶ The sequence of drug treatments was not explicitly modelled in the structure, but subsequent transplantation was. Nevertheless, clinical impact of subsequent treatment was incorporated as time-varying probabilities

3.4.3.1 Treatment sequence comparators

This section offers an overview of treatment sequences comparators across disease categories, including oncology, autoimmune, and others. For detailed information, see Appendix 3.4 and the first two columns of Table 3.9.

3.4.3.1.1 Oncology treatment sequence comparators

The majority of oncology TAs were in haematological malignancies (n = 4, 57%)[197,215,218,248], while three (43%) were in solid tumours (Table 3.9).[126,203,249] The majority of TAs examined two to three lines of therapy (LOTs), though a few explored up to four (Appendix 3.4). The ERG/AG sometimes conducted scenario analyses with alternative treatment sequences when they disagreed with those proposed by manufacturers.[126,197] Regarding treatment positions in manufacturer's analyses (Appendix 3.4), two TAs (29%) involved comparisons of treatment sequences with the

appraised technology only as a first-line treatment.[197,218] For example, in TA462[218], the manufacturer's scenario analysis assessed nivolumab as a first-line treatment for refractory Hodgkin Lymphoma with the following treatment sequences:

- **Nivolumab** → allogenic stem-cell transplantation → best supportive care (BSC)
- Standard of care → allogenic stem-cell transplantation → BSC

By contrast, in the remaining five oncology TAs[126,203,215,248,249], the appraised technology was positioned at various LOTs within an incremental CEA. Despite this, most aimed to evaluate the appraised technology primarily as a first-line treatment for treatment-naïve or -refractory patients.[126,203,249] For example, TA580 assessed enzalutamide for non-metastatic hormone-relapse prostate cancer (nmHRPC) as a first-line option, with comparisons including enzalutamide at different LOTs[249]:

- **Enzalutamide** → androgen deprivation therapy (ADT) → ADT or docetaxel → BSC
- ADT → **enzalutamide** → ADT or docetaxel → BSC

Similarly, in TA377[126], enzalutamide was appraised as a first-line treatment for metastatic hormone-resistant prostate cancers (mHRPC) and allowed as a third-line standard of care. Consequently, the manufacturer's base-case analysis compared the following treatment sequences:

- **Enzalutamide** → docetaxel → palliative
- BSC → docetaxel → **enzalutamide** → palliative
- Abiraterone → docetaxel → palliative

In contrast to the enzalutamide TAs (TA580[249] and TA377[126]), TA408[248] and TA137[215] not only positioned the appraised technology at multiple points within the treatment sequences but also sought approval for its use at these varied positions within the same appraisal. For example, in TA408[248], pegaspargase was considered for both first and second-line treatment in acute lymphoblastic leukemia (ALL), but consecutive use was prohibited. Consequently, the comparisons made were:

- **Pegaspargase** → Erwinase
- Native *E.coli* asparaginase → Erwinase
- Erwinase → **pegaspargase**
- Erwinase → native *E.coli* asparaginase

Additionally, TA137[215] assessed rituximab's cost-effectiveness for follicular non-Hodgkin's lymphoma (NHL), permitting its use at various sequence positions, including consecutively. The manufacturer's aimed to demonstrate rituximab's cost-effectiveness as induction therapy (i.e., first treatment position), standalone maintenance therapy (i.e., second treatment position), or both, resulting in comparisons of the following sequences:

- **R**-CHOP (**rituximab** plus cyclophosphamide, doxorubicin, vincristine, and prednisone) → **rituximab maintenance**
- **R**-CHOP → observation
- CHOP → **rituximab maintenance**
- CHOP → observation

3.4.3.1.2 Autoimmune treatment sequence comparators

Autoimmune TAs were primarily in psoriatic disease[157,158,244-247,250-255] (n = 12, 55%) and rheumatoid arthritis (RA)[155,188,256-259] (n = 6, 27%). The remaining ones, with one each, covered antineutrophil cytoplasmic antibodies-associated vasculitis[156], immune thrombocytopenic purpura[202], systemic juvenile idiopathic arthritis[200], and ulcerative colitis (Table 3.9).[260] Treatment sequences in autoimmune diseases vary from two to ten LOTs and may result in more than eight treatment sequences within an incremental CEA (n = 6)[155,188,202,244,251,254] due to the availability of numerous treatments (Appendix 3.4). All four MTAs reviewed were autoimmune TAs, for which the treatment-sequencing comparisons from CS and AG reports were summarised separately (Table 3.9, Appendix 3.4). In two MTAs[188,245], the AG allowed treatment sequences containing appraised technologies from different manufactures, while in another two MTAs[246,261], the AG only compared appraised technologies from different manufacturers at comparable positions in separate treatment sequences.

Autoimmune treatments were often assessed for use in populations with diverse prior treatment histories (i.e, number of LOTs). Consequently, many TAs assessed the appraised technology among populations with different treatment histories (i.e., different scopes/populations) in separate CEAs within the same appraisal (Appendix 3.4). An example is TA543[253], which assessed tofacitinib for psoriatic arthritis in three specific populations, each with a different number of prior treatments:

- Patients who did not respond to at least two non-biologic-disease-modifying anti-rheumatic drugs (non-bDMARDs)
  - **Tofacitinib** → ustekinumab → BSC
  - Apremilast → ustekinumab → BSC
  - Certolizumab pegol → ustekinumab → BSC
  - Etanercept → ustekinumab → BSC
  - Golimumab → ustekinumab → BSC
  - Infliximab → ustekinumab → BSC
  - Secukinumab 188 mg → ustekinumab → BSC
  - BSC
- Patients who did not respond to non-bDMARDs and at least one tumour necrosis factor

inhibitor (TNFi)

- **Tofacitinib** → BSC
- Secukinumab 300 mg → BSC
- Ustekinumab → BSC
- BSC

- Patients who were contracted or intolerant to an TNFi
  - **Tofacitinib** → BSC
  - Secukinumab 188 mg → BSC
  - Ustekinumab → BSC
  - BSC

In TA543[253], tofacitinib was appraised for use at different positions within the disease pathway, but was specifically assessed at a certain position (i.e., first-line) within each CEA, corresponding to each CEA's distinct scope/population (e.g. three slightly different target populations in TA543). This approach was also found in six other STAs[158,252,256,258-260] and several CS reports within two MTAs (TA445: Union Chimique Belge (UCB); TA375: Abbvie, UCB)[246,252]. In contrast, five STAs[156,157,200,202,257] and a CS report in MTA[188] (TA375: Pfizer) included treatment sequences with the appraised health technology placed at different LOTs within the same incremental- or paired-CEA in base-case or sensitivity analyses.

Another five STAs[247,250,251,254,255] and several CS reports in two MTAs[155,188] (TA195: Wyeth, Schering-Plough, Bristol Myers Squibb (BMS); TA375: BMS, Merck Sharp & Dohme (MSD) infliximab, MSD golimumab) assessed the appraised technology only at a specific LOT in a single incremental CEA. Finally, TA433[244], an STA, uniquely performed multiple incremental CEAs, consistently applying the appraised technology in the same position within the same population (i.e., first-line following non-biologic systematic treatment), with variations in the primary comparator at first-line and the number of LOTs for each CEA.

3.4.3.1.3 Treatment sequence comparators in other diseases

The six TAs in other diseases covered treatments for thromboembolism (n = 1)[262], bipolar disorder (n = 1)[201], hepatitis B (n = 2)[198,263], diabetes (n = 1)[264], and macular oedema (n = 1)[196]. Treatment sequences in these TAs had shorter LOTs, typically up to three lines, similar to oncology TAs (Appendix 3.4). Four of these TAs included comparisons of treatment sequences with the appraised technology used at different LOTs within the same incremental CEA.[198,201,263,264] Similar to TA543 in autoimmune diseases[253], TA249 also assessed the appraised technology (dabigatran) at different positions in the treatment pathway using separate incremental CEAs.[262] In TA 418[264], dapagliflozin was only considered as a first-line therapy for type 2 diabetes in the CS report, but the

ERG preferred an alternative analysis where patients remained on dapagliflozin throughout the entire treatment sequence as patients were more likely to receive later-line treatments as add-ons rather switch to a different treatment.

3.4.3.2 Treatment-sequencing models

In Table 3.9's third column, I summarise whether any treatment-sequencing model was employed in each CS and AG report. Three TAs lacked treatment-sequencing models, despite identifying several treatment sequences as comparators.[197,218,247] For TA462[218], initially, the manufacturer developed a non-treatment sequencing model for two-line treatment comparisons in the CS report, then added a treatment-sequencing model for an alternative scenario analysis in a revised submission to compare three-line treatment sequences. Two FTAs, TA426[197] and TA521[247], lacked economic models because they focused solely on cost-minimisation analyses (CMA). Although TA426 did not employ any treatment-sequencing model, the TA it reviewed (TA251[265]) was likely to contain Markov model or PartSM involving sequencing information. This assumption is based on the nature of TA251 as a MTA that evaluates the use of dasatinib, nilotinib, and imatinib as first-line treatments for chronic myeloid leukaemia (CML). Considering that these therapies may be sequentially, it is logical to assume that the economic models might incorporate sequencing information, aligning with the typical approaches in oncology, as outlined in Section 3.4.3.1.1. However, model structure details were unavailable in TA251. Apart from the aforementioned TAs without any treatment-sequencing models, treatment-sequencing models were omitted in several CS reports in two autoimmune MTAs.[245,246] Specifically, two manufacturers in TA103 (Serono, Wyeth) and one manufacturer in TA445 (Novartis) did not build a treatment-sequencing model because they did not list treatment sequences as comparators. Thus, information from these three CS reports were not reported in Table 3.9. However, TA103 and TA445 were still included in the Part B review since at least one document—either a CS or AG report—in each TA listed at least two treatment sequences as comparators.

Appendix 3.4 shows that Markov models were predominantly used for treatment-sequencing in oncology (4 out of 6, 67%)[126,215,218,248] and autoimmune diseases (22 out of 33, 67%)[156-158,188,200,202,244-246,250-257,259,260]. Two of these models in oncology had variations: one was a semi-Markov model[218], while another was attached to a decision-tree.[248] State-transition models (e.g. Markov model) were also the most prevalent structure for treatment-sequencing models in other diseases (5 out of 6, 83%)[196,198,201,262,263]. In oncology, PartSMs were the second most utilised model structure (2 out of 6, 33%)[203,249], always in combination with Markov or Semi-Markov models. Conversely, for the remaining treatment-sequencing models in autoimmune diseases (11 out of 33, 33%)[155,188,258] and other diseases (1 out of 6, 17%)[264], patient-level simulation models, such as discrete

event simulations (DES), were utilised. For MTAs, models from each CS and AG were counted individually.

3.4.3.3 Data sources to inform the comparison of treatment sequences

This section provides an overview of data sources used to inform the effectiveness of treatment sequences in CEAs. Table 3.8 shows which data sources were used in each TA, including pivotal and other trials, NMA/MA, RWE, and effectiveness simplifying assumptions. For simplicity, discussions about MTAs were based on the AG report unless otherwise specified, while discussion about other TAs involved a combination of CS and ERG considerations. Two TAs did not contain sufficient information about data sources due to conducting cost-minimisation analyses (TA426, TA521) and assuming non-inferiority between the appraised treatment and its comparator for being used at identical positions in the treatment sequences.[197,247] The majority of TAs (n = 30, 86%) combined evidence from an array of data source types to determine the effectiveness of treatment sequences, with the exception of four psoriatic disease TAs[246,252,253,255] using only NMA/MA evidence (noting, however, NMA/MA inherently pools multiple sources) and one hepatitis B TA[198] using only pivotal trial evidence. Most TAs specified using effectiveness simplifying assumptions to some extent, with only four oncology TAs not explicitly specifying any.[126,215,218,249] Note, the "use of data" here broadly refers to informing parameters or model structures that can impact treatment sequence effectiveness estimates in the economic model. It is not confined to inputs for LOT-specific effectiveness (though which it primarily is) or the effectiveness estimates of entire treatment sequences.

For oncology TAs, pivotal trials (100%) and other trials (86%) were the most common evidence sources to determine effectiveness of treatment sequences. None of the oncology TAs applied evidence from an NMA/MA. In contrast, most of the autoimmune TAs (77%) used NMA/MA results to inform their decision models, while the direct use of pivotal (55%) or other trial evidence (36%), rather than being used as sources for NMA/MA, was relatively low. TA in other diseases tend to combine the direct use of pivotal trial evidence (100%) and NMA/MA evidence (83%), while one TA in hepatitis B[263] incorporated evidence from non-pivotal trials. Non-pivotal trials and RWE were always used in combination with other data sources in CEAs, never alone. RWE complemented trial or NMA/MA evidence in over half of all TAs, most notably in oncology (71%). These include informing parts of the treatment sequence's effectiveness, other treatment-sequencing model structures, or key model assumptions. These are discussed in detail in the next section (Section 3.4.3.4).

3.4.3.4 Current practice and challenges to populate the effectiveness of treatment sequences

Table 3.10 summarises current practices and challenges in determining the effectiveness of treatment sequences. Each row of the table represents an approach for populating the effectiveness of

treatment sequences, with Section (A) detailing the effectiveness simplifying assumptions that have been used, and Section (B)-(E) outlining approaches for adapting evidence from pivotal trials, other trials, NMA/MA, and RWD. I also included an additional Section (F) to address challenges related to treatment-sequencing comparisons that did not fit into the other categories but were described either by the appraisal committee, the company, or the ERG. Each column of the table lists TAs that applied a particular approach in different disease types.

It is worth noting that the table covers not only approaches that directly determine the effectiveness of entire treatment sequences or parts of them (e.g. LOT-specific effectiveness), but also those that populate parameters that indirectly affect the effectiveness of treatment sequences (e.g. death rate) or to inform key model assumptions. Moreover, I also included approaches where data sources were used for external validation or to inform treatment-sequencing model structures. The following paragraphs describe the findings in accordance with the structure of Table 3.10. Statements related to specific approaches are labelled with their corresponding numbers from Table 3.10.

3.4.3.4.1 A. Effectiveness assumptions

An approach to determine the effectiveness of treatment sequences is to assume that all treatment sequences have equal effectiveness (A1), which was seen in an oncology TA where the manufacturer performed a CMA for a CDF rapid review (TA426).[197] The ERG, however, critiqued that such an assumption is only applicable when comparing strategies with the same number of LOTs. Approach A2 assumes equal effectiveness for all subsequent LOTs (second-line onwards), irrespective of prior treatments (i.e., first-line treatment). This approach was applied in a diabetes TA (TA418)[264], treating the full sequence of subsequent treatments after any first-line treatment as equally effective in managing blood sugar levels. Approach A3 assumes a treatment's effectiveness remains unchanged when being use at any LOT, including the first-line. While common in autoimmune and other disease TAs, it was not observed in oncology. Although several TAs that made such an assumption acknowledged the possibility that a treatment may be less effective when given later in a sequence, no evidence was found to adjust for this.[200,201,251,253,258] In TA485, despite ERG warnings in TA485 about potential inaccuracies in comparing different LOTs, the equal effectiveness assumption was maintained.[259]

Table 3.10 Challenges and approaches in deriving comparative effectiveness of treatment sequences

Challenges and justifications used for each type of data source	Oncology	Autoimmune	Others
A. Effectiveness assumptions			
A1. Assuming the effectiveness of different treatments sequences are equal	TA426	-	-
A2. Assumes equal effectiveness for all subsequent LOTs (second-line onwards), irrespective of prior treatments (i.e., first-line treatment)	-	-	TA418
A3. Assuming a treatment has the same effectiveness when being used in different LOTs	-	TA293, TA238, TA419, TA433, TA445(AG, CS: UCB), TA475, TA511, TA543, TA574, TA575, TA195(AG), TA375(AG), TA415, TA466, TA485, TA547	TA249, TA292, TA154, TA173
A4. Assuming that a treatment's effectiveness varies when used in different LOTs, and using a treatment modifier to approximate the degradation effect in later lines	TA319	TA308, TA419, TA442, TA195(AG)	-
A5. Assuming treatments that have similar treatment mechanisms have the same effectiveness or bundled-treatment effectiveness	-	TA308, TA195(AG, CS: Abbot, Roche), TA225, TA415	-
A6. Assuming maintained long-term treatment effect after the last-line of treatment	-	-	TA409
A7. Assuming there is a constant non-treatment-specific disease severity progression rate	-	TA225	-
A8. Assuming there is a non-treatment-specific and non-LOT-specific withdrawal rate	-	TA238, TA103(AG), TA419, TA433, TA442, TA537, TA225, TA415	-
A9. Assuming there is a treatment-specific but non-LOT-specific treatment withdrawal rate	-	-	TA249
A10. Assuming there is a non-treatment-specific risk of death after x years	TA319	-	-
A11. Assuming there is a non-treatment-specific increased risk of death post-treatment failure	TA408		
B. Pivotal trials			
Data being used as effectiveness inputs			
B1. Using data from a two-stage randomised SMART to inform the effectiveness of treatment sequences as those used in the SMART	TA137	-	-
B2. Using the OS of the first-line treatment to represent the OS of a whole treatment sequence	TA251[#], TA319, TA377	-	-
B3. Estimating OS based on cumulative duration of each LOT in trials	TA251[#] (scenario analysis)	-	-
B4. Informing a specific LOT-effectiveness (e.g., PFS/response rate of first-line treatment) in the model with effectiveness estimates from the same LOT in pivotal trials	TA319, TA377, TA580	TA308, TA293	-
B5. Due to lack of head-to-head comparison of the appraised technology and its comparator, the effectiveness of the comparator is assumed to be equivalent to that of the appraised technology in the pivotal trial (LOT-specific, e.g. The appraised technology and the comparator are assumed to have equal effectiveness when used as the same LOT.)	TA408	TA293	-
B6. Using effectiveness estimate from patients with similar treatment history to inform the effectiveness estimate of a specific LOT in the model	-	TA308, TA445(CS: UCB), TA574, TA195(AG), TA466	-
B7. Using the effectiveness estimate from a specific LOT or unspecified LOT to inform a treatment's effectiveness across multiple different LOTs	-	TA238	TA249, TA292, TA154, TA173

Data being used as inputs that may indirectly affect treatment effectiveness			
B8. Informing the proportion of patients moving to the next LOTSS (e.g., first-line to second-line)	TA377	-	TA409
B9. Informing the treatment-specific discontinuation rate (non-LOT specific)	-	TA419, TA433, TA442, TA475., TA195(AG)	TA292, TA418
B10. Informing patient demographics in the model (non-LOT specific)	-	TA442, TA225, TA415	TA418
Data being used as external validation			
B11. Acting as an external source to cross-validate the effectiveness of a specific LOT, which has been modelled using other data sources	TA580		
Data being used to inform model structures			
B12. Informing the need of modelling treatment sequence	-	-	TA409
B13. Informing model assumptions	-	TA442	-
C. Other trials			
Data being used as effectiveness inputs			
C1. Using the OS of the first-line treatment from a comparator's first-line pivotal trial to represent the OS of a whole treatment sequence for the comparator	TA251#, TA319, TA377, TA580	-	-
C2. Using LOT-/treatment-history-specific effectiveness evidence to inform LOT-specific effectiveness	TA319, TA377, TA580		-
C3. Using effectiveness estimates from a specific LOT or unspecified LOT to inform the effectiveness of multiple LOTs	TA408	TA293, TA238, TA415, TA195(AG), TA225, TA547	-
Data being used as inputs that may indirectly affect treatment effectiveness			
C4. Informing treatment-specific discontinuation/resistance rates (non-LOT specific)	-	TA442	TA173
Data being used as external validation			
C5. Acting as external source to guide the choice of treatment sequences	TA408	-	-
C6. Acting as an external source to validate or justify model assumptions	TA408, TA319	-	-
D. NMA/MA			
Data being used as effectiveness inputs			
D1. Using effectiveness estimates from patients with similar treatment history to inform effectiveness estimate of a drug being used as a specific LOT or multiple LOTs	-	TA442, TA445 (CS:UCB), TA537, TA543, TA574, TA575, TA415, TA466, TA547	-
D2. Using the effectiveness estimates of a drug from unspecified LOTs to inform the effectiveness estimates of it being used in multiple LOTs	-	TA103(AG), TA419, TA433, TA442, TA445(AG), TA511, TA225, TA375(AG)	TA249, TA173, TA409
Data being used as inputs that can potentially have an indirect effect on treatment effectiveness			
D3. Informing the non-treatment-specific discontinuation rate	-	TA537	-
E. RWE			
Data being used as effectiveness inputs			
E1. Using existing RWE to inform the effectiveness of subsequent lines as a whole (e.g. post-progression survival)	TA319, TA462	-	-
E2. Using existing RWE to inform treatment effectiveness in unspecified LOTs	-	TA293, TA238, TA195(AG), TA375(AG)	TA249

Data being used as inputs that may indirectly affect treatment effectiveness			
E3. Informing duration until discontinuation/discontinuation or death rates (non-LOT specific)	TA319, TA408	TA308, TA293, TA238, TA103(AG), TA475, TA511, TA195(AG), TA225, TA415, TA466, TA485	TA292
E4. Informing patient demographics in the model	-	TA238, TA195(AG)	TA249, TA418
E5. Using RWD to estimate an effect modifier of subsequent or long-term treatment	-	TA511	TA409
E6. Using RWD to calibrate for subsequent events	-	-	TA154
E7. Informing the proportion of different clinical events within a health state	-	-	TA418, TA409
E8. Mapping different response types of those in trials and models	-	TA308, TA433, TA475, TA375(AG), TA415, TA466	-
Data being used as external validation			
E9. Acting as external source to guide the choice of treatment sequences	TA137	TA547	-
E10. Acting as an external source to validate or justify model assumptions	TA319, TA377	TA308, TA442, TA375(AG)	TA249, TA292, TA173
E11. The AG was granted access to multiple source of registry data to assess model parameters and correlations	-	TA375(AG)	-
F. Additional key challenges related to comparing treatment sequences (that have been explicitly mentioned in TAs)			
Challenges in modelling treatment sequences related to data			
F1. A lack of sequence-specific effectiveness data	TA319	TA574, TA195(AG)	TA292
F2. A lack of access to relevant UK data to inform treatment pathways and to evaluate treatment effectiveness.	TA408	TA293	TA292
F3. Paucity of time-varying data	-	TA433, TA442, TA537, TA575, TA415	-
F4. PFS and time to the next line of treatment are sometimes used to proxy each other when one of them was unavailable	TA319	-	-
F5. Model structure may not be appropriate to address the decision problem with lack of evidence available on treatment sequences	TA319	-	-
F6. Challenges of using evidence from real-world treatment-sequencing studies	-	TA195(AG)	-
F7. Awaiting results from ongoing studies exploring the comparison of treatment sequences	TA319, TA377	-	-
Other data-related challenges			
F8. Lacking head-to-head evidence, using data from indirect comparison of trials may introduce bias	-	TA293, TA238, TA195(AG), TA225, TA415	-
F9. Longer-term data are required	-	TA103(AG), TA225, TA375(AG)	TA292, TA154, TA173
Challenges related to selecting treatment sequences			
F10. Challenges include all possible treatment sequences as comparators	-	TA442, TA511, TA537, TA574, TA575, TA485, TA547	-
F11. The rapidly evolving treatment landscape makes it difficult to model treatment pathway.	TA377	-	-
F12. Potential omissions of other relevant treatment sequences as comparators	TA319, TA377	TA308, TA293, TA238, TA419, TA433, TA442, TA575, TA415	TA292, TA418
F13. Treatment sequences may depend on patient characteristics at clinical practice rather than following the average population pattern	-	TA293, TA103(AG)	TA292
F14. Additional administrative considerations need to be taking into account when considering what treatment sequences to be included in the comparison	TA377, TA293	-	TA292, TA409
F15. Concerns of including treatment sequences with comparators that are not cost-effective	-	TA475, TA511, TA574, TA575	-
F16. Cost-effectiveness results may be sensitive to the choice of what treatment sequences have been included	-	TA442	-

in comparison			
F17. Challenges in deciding the number of LOTs that should be modelled	-	TA238, TA537	TA249, TA292
F18. Concerns of having different length of treatment sequences (LOTs) as comparators	-	TA308, TA433, TA475, TA466	-
F19. The assumption that treatment duration equals to time to relapse do not always hold	-	TA308, TA238	-
Other modelling-related challenges			
F20. Clinical data does not support the choice of model structure	-	TA308	-
F21. Challenges in revising models to accommodate alternative treatment sequences upon ERG request	-	TA433	TA249
F22. Challenges in modelling survival and treatment duration in relation to model structure	TA377	-	-
F23. Inconsistent cost and effectiveness inputs in the treatment-sequencing model structure	TA377	-	-
F24. Inconsistent timing of receiving subsequent treatments in real-world practice and that modelled based on clinical trial outcome measurements	-	-	TA292
F25. Model results may be sensitive to the time horizon	-	TA475	TA249
F26. Unmatched population between the data source and those intended to be modelled	-	TA442	-
F27. Technical modelling error	-	TA419	-
Decision-making concerns			
F28. Pair-wise comparison versus incremental analysis when comparing treatment sequences	-	TA442	-
F29. Requiring disaggregated LY results of treatment sequences for decision making	-	TA442	-
F30. Considering the appraised technology as a replacement to an existing technology or an extension to the current treatment sequences	-	TA308, TA419, TA238	-
F31. Difficulties in making suggestions regarding the position of the appraised treatment in the treatment pathway	-	TA419, TA433	TA292, TA409
F32. Assessing the appraised technology's position in the treatment pathway through separate CEAs for different target populations/scopes, instead of comparing its multiple sequence positions in one CEA	-	TA225, TA375(AG)	-
F33. The danger of TA approval leading to the sequential use of clinical practice (that lacked robust clinical evidence)	TA319	-	-
F34. The Committee making statements that they were only considering the scenario of placing the new technology in the sequence compared to another drug in the sequence rather than depicting the optimal treatment sequence.	-	TA415	-

LOT: line-of-therapy; OS: overall survival; PFS: progression free survival; SMART: sequential multiple assignment randomized trial;
TA251 is a part of the original submission of TA426, which is a Cancer Drug Fund review for dasatinib and other health technologies

Approach A4 uses an effect modifier to adjust for the degradation effect of a treatment being used in later treatment lines, a common method. Sources to inform these modifiers include expert opinions, literature, or subgroup analyses of trials or NMA/MA. For instance, in TA308[156], a relative risk ratio between the treatment-naïve subgroup and recurrent disease subgroup in the RAVE trial was applied to assume the remission rate of the second course of rituximab treatment. Subgroup analyses were useful in approximating the effectiveness of later LOTs as the RAVE trial only assessed the effectiveness of a single LOT, but recruited patients with different prior treatments. However, these analyses may compromise the randomisation of the trial and limit the strength of evidence.

Approach A5 assumes that treatments with similar mechanisms share the same effectiveness or exhibit bundled-treatment effects. This assumption, often applied where treatments with theoretically interchangeable effects as part of the sequence, was prevalent in autoimmune TAs. For example, in TA485[259], all TNF inhibitors were considered to have equal effectiveness across all LOTs. In one of the incremental CEAs of T485[259], this assumption can be considered equivalent to approach A2, with all sequences from second-line (i.e., any TNFi) onwards assumed to have equal effectiveness, regardless of the first-line treatment

- Sarilumab → TNFi bundle (i.e., representing any type of TNFi) → BSC
- Tocilizumab intravenous injection + methotrexate → TNFi bundle → BSC
- Tocilizumab subcutaneous injection + methotrexate → TNFi bundle → BSC
- TNFi bundle → TNFi bundle → BSC

Approach A6 assumes that the treatment effect was maintained in the long-term even after patients stopped the last-line of treatment, as seen in TA409, where the compared treatment sequences involved a combination of laser surgery and medication to maintain visual acuity in patients with visual impairment caused by macular oedema.[196] This assumption is particularly relevant to this TA, as visual impairment may be largely irreversible.

The rest of the assumptions were found to indirectly impact the effectiveness of treatment sequences, including assuming a constant disease severity progression rate (A7), a specific treatment withdrawal rate (A8 & A9), and risks of death (A10 & A11). Assumptions about the rate of disease severity progression and treatment withdrawal assumptions were frequently seen in autoimmune TAs, while risks of death related assumptions were more relevant in oncology TAs. Regardless of being treatment-specific or non-treatment specific, applying an annual risk of disease progression, treatment withdrawal or death in a CEM might alter the total number of LOTs that a patient can receive and thus affect the results of CEA. The choice of time horizon may also affect the CEA results in a similar fashion and thus scenario analyses were conducted in many TAs, as shown in Appendix 3.4.

These aforementioned effectiveness assumptions share similarities with the concepts outlined in Lewis's doctoral thesis and its subsequent publication, which presents a taxonomy for simplifying

assumptions in the synthesis of quantitative evidence on treatment sequence effectiveness.

3.4.3.4.2 B. Pivotal trials

The approaches for using pivotal trial data to inform the effectiveness of treatment sequences varied depending on disease area. I categorised these approaches into four categories, including using information as direct effectiveness inputs (B1-B7), as inputs that may indirectly affect effectiveness (B8-B10), as sources for external validation (B11), and as inputs to inform model structures (B12-13).

Approach B1 involved using evidence from a two-stage sequential multiple assignment randomized trial (SMART) to directly populate the effectiveness of treatment sequences in the CEM. This was only observed in TA137[215], where evidence from the pivotal trial of rituximab in treating relapsed follicular non-Hodgkin's lymphoma (EORTC 20981) was applied. The EORTC 20981 trial randomised patients to receive one of the two different induction therapies (rituximab-CHOP or CHOP) followed by either rituximab maintenance therapy or observation, resulting in four subgroups of patients receiving different treatment sequences as illustrated in Figure 3.6. These four subgroups matched the treatment sequences compared in the CEM in TA137 as described in Section 3.4.3.1.1.

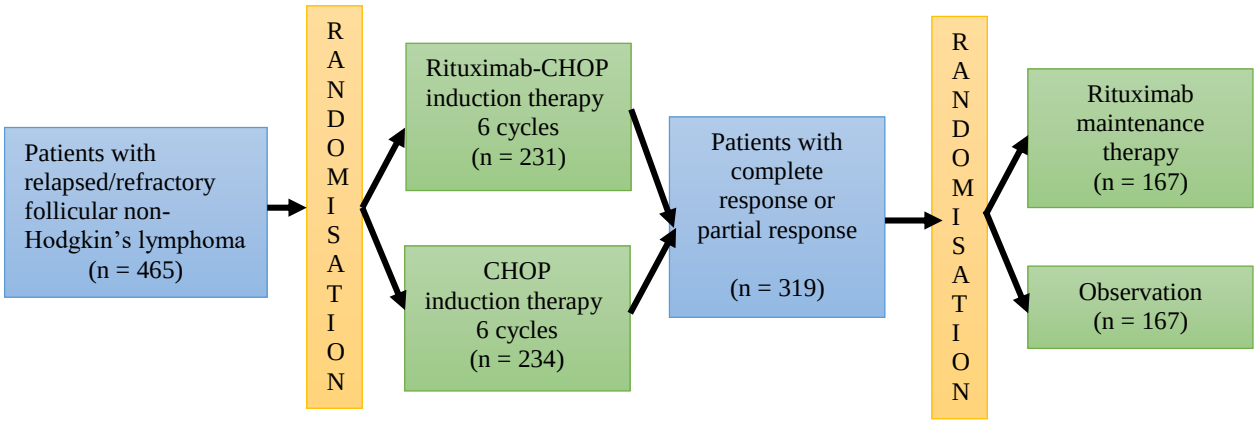


Figure 3.6 An example of two-stage sequential multiple assignment randomized trial (SMART) from the EORTC 20981 trial

CHOP: cyclophosphamide, doxorubicin, vincristine, and prednisone

The trial assessed the use of rituximab to the standard of care at two stages, including as induction therapy (versus CHOP) and as maintenance therapy (versus observation).

The manufacturer conducted intention-to-treat (ITT) analyses to obtain the OS and progression-free survival (PFS) for each treatment sequence subgroup, which were then used to estimate the Markov transition probabilities for each subgroup in the CEM. For example, the response rate for patients becoming eligible for maintenance treatment was informed by the outcome of induction treatment. The manufacturer was criticised by the ERG for only calibrating their parametric survival models on within-trial data, which may not remain valid when later-line treatments are introduced. Although the manufacturer recommended adding more LOTs to the model, the limited data and

extensive assumptions deemed it unfeasible Additionally, the consulted by the appraisal committee raised concerns about the use of rituximab as an induction-maintenance treatment in relapsed patients, as it was licensed and recommended for earlier use as first-line therapy in TA110, while patients in the EORTC 20981 trial were rituximab-naïve.

Approach B2 and B3 relied on pivotal trials that assessed a certain line of treatment (LOT) to determine the OS of entire treatment sequences, which were only seen in oncology TAs. Conversely, approach B4-B7 applied evidence from LOT-specific pivotal trials only to populate the effectiveness of certain segments of the treatment sequences. Approach B2 employed the OS of first-line treatments to represent the OS of complete treatment sequences populated in the CEM. This approach often required making the assumption that subsequent treatments after the first-line treatment in the trial can represent those used in the NHS. Three TAs were found to use approach B2, including TA319, TA377 and the original submission of dasatinib in TA426 (TA251).[126,203,265] For TA319[203], the manufacturer developed a de novo semi-Markov PartSM to compare the following treatment sequences that included the appraised technology ipilimumab:

- BRAV V600 mutation-negative patients
  - **Ipilimumab** → BSC → BSC
  - Dacarbazine → **ipilimumab** → BSC
- BRAV V600 mutation-positive patients:
  - **Ipilimumab** → vemurafenib → BSC
  - Vemurafenib → **ipilimumab** → BSC
  - Dacarbazine → **ipilimumab** → BSC

To populate the OS for treatment sequences, various approaches were taken in conjunction with Appraoch B2 in TA319. In the case of ipilimumab and dacarbazine as first-line treatments, the manufacturer adopted the Kaplan-Meier (KM) curve of pooled data from chemotherapy-naïve patients (including the pivotal trial CA184-024) for the first 0-104 cycles of the model (Approach B2). For cycles 105-206, a parametric curve was fitted. The OS for treatment sequences with vermurafenib as first-line treatment were derived from other trials as no head-to-head comparison between first-line ipilimumab versus vermurafenib were available (see approach C1). These estimates were used for cycles up to 260. After this point, however, the OS estimates derived from registry data were used to approximate the OS under palliative care (see approach E1).

Similarly, in TA580[249] (see Section 3.4.3.1.1 for treatment sequence details), OS estimates for treatment sequences starting with enzalutamide and BSC for mHRPC were derived from the pivotal trial of enzalutamide, PREVAIL, where patients were randomly assigned to receive enzalutamide or placebo as first-line treatments (Approach B2). In contrast to TA319[203], the manufacturer adjusted for

treatment-switching in the trial using methods such as inverse probability of censoring weighting (IPCW) and a two-stage estimation. Here, treatment-switching referred to patients who switched to a second-line treatment that was not part of the standard treatment pathway in England. Both the manufacturer and ERG conducted a series of sensitivity analyses using different data cut-off and treatment-switching adjustment methods to explore the uncertainty of OS extrapolation. Instead of relying on the OS estimates, Approach B3 involved using the accumulative duration of successive treatments in the pivotal trial to populate OS in the model. This approach was used as a scenario analysis in the original company submission in TA251 which was later reviewed by TA426.[197,265]

Approach B4 involves using the effectiveness estimate from a LOT in the pivotal trial to inform the effectiveness of the same LOT in the model. For example, in TA580 (see Section 3.4.3.1.1 for treatment sequence details)[249], parametric curves of metastasis-free survival upon receiving the first-line treatment of enzalutamide versus androgen deprivation therapy was fitted from the pivotal PROSPER trial to determine transition probabilities from non-metastatic HRPC (nmHRPC) (first-line) to mHRPC states (second-line onwards) as the solid straight arrows illustrate in Figure 3.7. Transitions from subsequent health states were informed by evidence from clinical trials that studied subsequent lines of therapy. Similarly, TA319 and TA377 applied the PFS estimates of first-line treatments that had head-to-head comparisons in pivotal trials to inform the PFS for them in the model as described in Approach B2 (ipilimumab versus dacarbazine, enzalutamide versus BSC, respectively), while using other evidence for post-first-line-progression-survival. It is worth nothing that approach B2 and B4 are not mutually exclusive. In fact, TA319 combined both approaches to populate the PFS and OS in a PartSM.

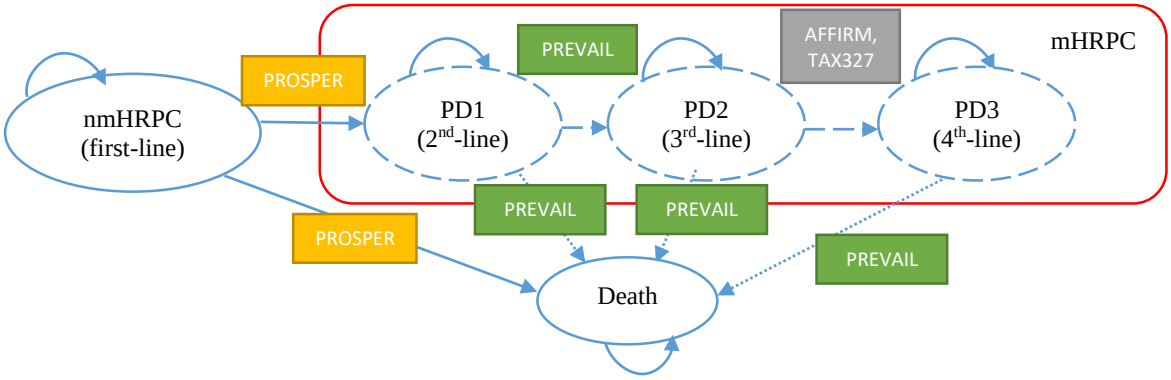


Figure 3.7 Model structure of TA580 (adapted from Company Submission)

nmHRPC: non-metastatic hormone-resistant prostate cancer; mHRPC: metastatic hormone-resistant prostate cancer; PD: post-progression states

Boxes highlighted with colors (yellow, green, grey) represent trials that were used to derive the transition probabilities or survival curves in the model. The three progressive diease states (PD1, PD2, PD3) after first-line treatment were seen as a collective state as mHRPC. Survival from the nmHRPC state and mHRPC followed a partitiond surival approach, while transitions within the mHRPC were Markov models.

In TA580, validation with pivotal trial evidence (Approach B11) was also undertaken.

Specifically, the pre-progression survival for enzalutamide versus ADT were populated using evidence from PROSPER, where first-line treatments in patients with nmHRPC were evaluated. The post-progression survival was populated by the OS of enzalutamide versus ADT from PREVAIL, a different trial that assessed treatments in patients with mHRPC. To illustrate, the dotted lines in Figure 3.7 represent the (post-progression) survival in nmHRPC states (PD1, PD2, P3), which were seen as a collective state and populated by OS estimates in PREVAIL. Since the OS was informed by two data sources, the OS curves from PROSPER were used to cross-validate the post-progression survival informed by PREVAIL through visually comparing their overlap.

Approach B5 is almost identical to B4 in that it involves using effectiveness estimates from a specific LOT to inform the effectives of the same LOT in the model. The only difference is that, for a treatment lacking head-to-head comparisons in the pivotal trial, its effectiveness was assumed to be equivalent to one of those in the pivotal trial. For instance, in TA408 (see Section 3.4.3.1.1 for treatment sequence details)[248], the effectiveness of E. coli asparaginase as first-line treatment was populated with the effectiveness of the appraised technology, pegaspargase. When specific treatment effectiveness of a particular LOT is unavailable, an alternative approach (Approach B6) is to use the effectiveness estimates of patients with a similar treatment history in the trial to represent the effectiveness of that LOT in the model. To illustrate, in TA308, the probability of achieving remission for the second course of cyclophosphamide was unavailable in the pivotal trial RAVE (see approach A4). Therefore, a subgroup analysis was conducted in patients having relapsed disease to proxy this probability. This approach was frequently seen in autoimmune TAs. Approach B7 involves using effectiveness from pivotal trials to inform the effectiveness of multiple LOTs. This approach may be used with or without a degradation modifier, as described in approach A3.

Pivotal trial evidence can also be used to inform parameters that may indirectly affect the effectiveness of treatment sequences, such as the proportion receiving a next-line of treatment (Approach B8), treatment-specific discontinuation rate (Approach B9) and patient's baseline characteristics in the model (Approach B10). It is worth noting that the discontinuation rate and patient characteristics observed were all non-LOT-specific.

Approach B12 involves using pivotal trial evidence to identify the need to model treatment sequences in the CEM. For example, in TA409[196], the pivotal trial VIBRANT showed that treatment switching was allowed in patients who began with first-line treatment of laser versus aflibercept. Therefore, subsequent treatments after first-line were modelled. Approach B13 used pivotal trial information to inform model assumptions, as seen in TA442. In the base-case analysis, the appraised technology, ixekizumab, was evaluated as a first-line treatment in a treatment sequence, while in a scenario analysis, it was assessed as a second-line treatment following adalimumab, a biologic treatment. The effectiveness estimate of ixekizumab was derived from NMA results to inform its use

in different lines of treatment (approach D1). One of the pivotal trials, UNCOVER-2, confirmed the assumption that biologic treatment does not modify the effect of ixekizumab, and thus no adjustment to the NMA was required for ixekizumab in the scenario analysis.

3.4.3.4.3 C. Other trials

Approach C1 is similar to B2 in that they both used the first-line OS from trials to represent the OS of a whole treatment sequence. However, approach C1 leverages evidence from a comparator's first-line trial when a head-to-head comparison of the appraised technology and its comparator is not available. For example, in TA377[126], which appraised enzalutamide, the OS estimates for the comparator abiraterone treatment sequence was obtained from abiraterone's pivotal trial COU-AA-302, while the OS evidence for the enzalutamide and BSC sequences were derived from the pivotal trial of the appraised technology enzalutamide (PREVAIL) (see Section 3.4.3.1.1). In fact, all oncology TAs that applied Approach B2 also applied Approach C1 because they involve comparisons of more than two sequences, while most of the RCTs have only two arms. The application of both approaches results in an indirect comparison of OS. In TA377[126], the manufacturer performed a naïve comparison between COU-AA-302 versus PREVAIL as well as a scenario analysis using an adjusted indirect comparison for abiraterone OS. Although not applied in the CEM, the ERG undertook a separate Bucher indirect comparison using data from PREVAIL and COU-AA-302. Similar to approach B2, treatment-switching adjustments may also be applied in Approach C2 where subsequent treatments in trials deviate from those in the NHS. Although the manufacturer in TA377 acknowledged the treatment-switching issue in using evidence from COU-AA-302, they were unable to adjust for abiraterone OS due to lacking access to patient-level data.[126]

Similar to Approach B4-B6, non-pivotal trial evidence can also be used to inform LOT-specific effectiveness (Approach C2). For instance, in TA580[249], treatment durations of enzalutamide as first-line treatment in the PREVAIL trial were used to inform the probability of progression from second-line treatment to third-line treatment in the CEM as illustrated in Figure 3.7. This approach was taken as patients in PREVAIL received a first-line treatment for mHRPC, which is comparable to patients who progressed from nmHRPC to mHRPC and received second-line treatment. Probabilities of receiving a fourth-line treatment was derived from the AFFRIM trial for patients who progressed on third-line enzalutamide and the TAX327 trial for those who those who progressed on third-line docetaxel and ADT, as these patients were in comparable disease states.

Approach C3 is comparable to Approach B7 with the difference of using estimates from non-pivotal trials and was also frequently seen in appraisals for autoimmune diseases. Evidence from non-pivotal trials have also been used to inform non-LOT specific treatment discontinuation rates and resistance rates (Approach C4). Instead of informing on the need of model treatment sequences, non-

pivotal trial evidence was found to be used as an external source to guide the choice of treatment sequences (Approach C5). For example, in TA408[248], the manufacturer referenced two UK studies, UKALL2003 and UKALL2014, suggesting that Erwinase should be the only second-line treatment after hypersensitivity has occurred to a first-line treatment (see Section 3.4.3.1.1). Although pegaspargase was historically given to patients following hypersensitivity to native *E.coli* asparaginase, this has no longer been the standard of practice due to risk of reactivity and subsequent hypersensitivity.

In line with Approach B13, non-pivotal trial evidence may also be used to support model assumptions around the effectiveness of treatment sequences (Approach C6). For example, in TA319[203], treatment sequences were compared in two different populations, BRAF V600 positive and negative, whereas the pivotal trial recruited patients with mixed BRAF V600 expressions. The OS of dacabazine, a key comparator as first-line treatment, was found to be similar in the BRIM-3 and CA184-004 trial. The manufacturer stated that this finding supports the use of the same OS for treatment sequences starting with the same first-line treatment in patients with different BRAF V600 expressions. In particular, the BRIM-3 trial compared vemurafenib versus dacabazine among BRAF V600 positive patients, while CA184-004 compared ipilimumab versus dacabazine among patients with mixed expressions. Additionally, the manufacturer visually compared the dacabazine arm in the pivotal trial CA184-024 (ipilimumab versus dacabazine) and the BRIM-3 trial to justify the direct application of vemurafenib OS curves from a non-pivotal trial (BRIM-3) without the need for indirect comparison adjustment (see Approach C1).

3.4.3.4.4 D. NMA/MA

The use of NMA/MA evidence was mainly limited to deriving the effectiveness of parts of the treatment sequences (Approach D1-D2) and parameters that indirectly affect treatment sequences (Approach D3). None of the TAs used NMA/MA results for external validation or to inform model structures. The most common application of NMA/MA results was to use effectiveness estimates from patients with similar treatment history to inform the effectiveness of a treatment at a specific LOT or multiple LOTs (Approach D1). NMA subgroup analyses were often performed to derive effectiveness in patients with different treatment histories. For example, in the CS submitted by UCB in TA445[246], the NMA was restricted to studies that included treatment-naïve or treatment-experienced patients to inform the treatment effectiveness in populations with different treatment histories regardless of LOTs. Conversely, in TA442.[158] the manufacturer argued that subgroup NMA could not be used to proxy effectiveness as later-line treatments due to a lack of evidence, but a degradation modifier was applied in their scenario analyses.

Subgroup analyses of NMA/MA can sometimes be problematic if inconsistent with other

assumptions made in the TA. For instance, in TA537[252], the manufacturer performed subgroup NMA/MA to obtain first-line treatment effectiveness in different populations, including biologic-naïve and biologic-experienced, for comparison of the following treatment sequences in patients with mild-to-moderate psoriasis:

- Biologic-naïve: no psoriasis and mild-to-moderate psoriasis
  - **Ixekizumab** → ustekinumab → secukinumab → BSC
  - Adalimumab → ustekinumab → secukinumab → BSC
  - Apremilast → ustekinumab → secukinumab → BSC
  - Certolizumab pegol → ustekinumab → secukinumab → BSC
  - Etanercept → ustekinumab → secukinumab → BSC
  - Golimumab → ustekinumab → secukinumab → BSC
  - Infliximab → ustekinumab → secukinumab → BSC
  - Secukinumab → ustekinumab → adalimumab → BSC
- Biologic-experienced: no psoriasis and mild-to-moderate psoriasis
  - **Ixekizumab** → BSC
  - Ustekinumab → BSC
  - Certolizumab pegol → BSC
  - Secukinumab → BSC

However, the manufacturer also assumed that the effectiveness of secukinumab remains the same as first-line treatment when used as later-line treatment in the biologic-naïve group, which was pointed out by the ERG as contradictory to not using a different set of evidence to estimate the effectiveness of secukinumab in the biologic-experienced group. The manufacturer also conducted a meta-regression in their sensitivity analysis to adjust for the baseline risk of placebo response for within the biologic-naïve NMA subgroup.

Approach D2 utilised NMA/MA effectiveness estimates to inform treatment effectiveness in multiple LOTs without explicitly examining patient history in the studies included in the NMA/MA. In TA375, this approach was used by the AG to conduct NMA to populate the Disease Activity Score (DAS)-based European League Against Rheumatism (EULAR) responses in the CEM. However, not all clinical trials assess EULAR responses, so the AG had to map EULAR responses from some trials measuring the American College of Rheumatology (ACR) responses. The relationship between EULAR and ACR was established using data from the Veterans' Affairs Rheumatoid Arthritis (VARA) Registry, which the AG received from the United States (US) (see approach E9).

Similar to approach B9, NMA/MA results may also be used to inform treatment discontinuation rates (C4). However, in the example of TA537[252], the discontinuation rate was derived as non-

treatment-specific.

3.4.3.4.5 *RWE*

Approach E1 involves leveraging RWE to inform the effectiveness of subsequent lines as a whole, such as in TA319 and TA462.[203,218] In TA319[203], the duration and breakdown of BSC treatments after first-line treatments were determined using RWE from the MELODY study. In TA462[218], OS estimates from a real-world study by Cheah et al.[266] were used to inform the post-progression survival after first-line treatments in a scenario analysis assessing the effectiveness of nivolumab as a bridging agent to receive allogeneic stem-cell transplant (alloSCT) for refractory Hodgkin lymphoma, which may potentially cure the disease (see Section 3.4.3.1). The Cheah 2016 study evaluated the effectiveness of brentuximab vedotin in patients with relapsed Hodgkin lymphoma. In the Cheah 2016 study, a proportion of patients who responded to brentuximab vedotin and received alloSCT showed a non-significantly superior OS. The manufacturer made an assumption that patients who survived 6 months after the first-line treatment were qualified for alloSCT. Consequently, they used PFS from RCT Checkmate 205 for the nivolumab-sequence arm and Cheah 2016 for the standard of care-arm to populate survival after 6 months. However, the ERG expressed that this setup might double count the survival benefit of nivolumab. In contrast, the Committee believed that the proportion of patients receiving alloSCT should be higher than those in Cheah 2016 and Checkmate 205, based on feedback from clinical experts.

Approach E2 utilised existing RWE to inform treatment effectiveness in unspecified LOTs where evidence from trials is not available. This approach was observed in autoimmune and other diseases, but not in oncology. In TA249[262], for instance, the manufacturer used RWE from retrospective cohort studies to determine the relative risk of disability following ischemic stroke between the appraised technology (dabigatran) versus warfarin, regardless of when the event occurred. Additionally, the manufacturer applied RWE to inform the effectiveness of dose-adjusted warfarin due to the narrow therapeutic window and variability in its real-world use.

RWE can also inform discontinuation and death rates (Approach E3), as well as patient demographics (Approach E4). For example, in TA511[251], the British Association of Dermatologists Biologic and Immunomodulators Register (BADBIR) provided information on treatment discontinuation rates, supporting the assumption that the same discontinuation rate applies to all drugs, assuming that treatment effectiveness does not vary across treatment lines. In TA195[155], estimates from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis (BSRBR-RA) were used by the AG to inform patient baseline characteristics in the model (i.e., LOT-effect modifier) (Approach E4). While many autoimmune TAs utilised evidence from UK-based studies using the above-mentioned data, evidence from other countries were also observed, such as studies using the

Danish Biological Treatment in Danish Dermatology registry (DERMBIO), the Danish rheumatologic database (DANBIO), the Canadian clinical database of rheumatoid arthritis including patients from Quebec (RHUMADATA).

Approach E5 involved using RWD for effect modifier estimation (Approach E5) and model calibration (Approach E6). In TA511[251], an effect modifier was derived from a Danish registry to account for reduced treatment effectiveness among patients with previous exposure to biologic treatments (Approach E5). It was noted that this effect modifier may suffer from selection bias, as patients who were harder to treat may have switched to another treatment earlier (i.e. had previous treatments) and thus more of these patients were included in the estimation of the modifier. However, the model may require an averaged modifier involving patients that do not necessarily switch treatments at an early stage. In my view, an alternative design of the RWD and longer follow-up may mitigate the issue. In TA154[198], the risk of developing compensated cirrhosis and hepatocellular carcinoma (HCC) among Hepatitis B patients was calibrated to match those reported in two prospective cohort studies among Taiwanese population (Approach E6), as these long-term outcomes were not measured in the pivotal trials.

Approach E7 involves determining the proportion of different clinical events within a health state. For instance, in TA418[264], various clinical events may occur when receiving treatments for diabetes, such as myocardial infarction, stroke, hypoglycaemia, nephropathy and cardiovascular death. Each of these events has a different impact on a patient's life expectancy and quality of life and thus, dictates the total LOTs that a patient may receive. However, these events are often measured as a composite event in clinical trials. Hence, the manufacturer used estimates from a Health Improvement Network (THIN) database analysis to inform the proportion of each sub-event. Approach E8 involved mapping different outcome types in trials and CEMs, such as the EULAR and ACR response described in Approach D2.

Approach E9 involved RWE acting as an external source to guide the choice of treatment sequences. In TA137[215], the manufacturer used evidence from a market research report of 50 haematologists to justify their selection of CHOP as the only relevant comparator for induction therapy. In another case (TA547[260]), the ERG identified a published CEA study suggesting that certain treatment sequences may be optimal in the UK, but was not modelled by the manufacturer. Approach E10 acted as an external validation to support model assumptions. In TA319[203], OS estimates derived from long-term American registry data were used to model patient survival after 5 years since the first-line treatment. The manufacturer considered this approach conservative because it assumed that the appraised technology (ipilimumab) produced no further relative survival benefit (compared to its comparator arm) beyond its pivotal trial (CA184-024). The manufacturer justified this assumption by comparing the RWE with other evidence from pooled trials.

Approach E11 involved the AG being granted access to registry data for assessing key model parameters and correlations, which differed from previous approaches in Section E that mostly relied on de novo RWD analysis by the manufacturer or using existing RWE. In TA375[188], the AG received data provided by the BSRBR, the Early Rheumatoid Arthritis Study (ERAS) and the United States National Data Bank for Rheumatic Diseases (NDB) to gather additional information about Health Assessment Questionnaire (HAQ) score progression. The trajectory of the underlying HAQ score progression may interact with the outcome at each LOT and thus affect the total number of LOTs that a patient may receive. The AG acquired the data after the NICE Decision Support Unit (DSU) was requested to provide additional information on HAQ progression for patients treated with non-biological therapies following the initial Committee meeting. The AG viewed the use of large observational databases to derive model parameters as a strength of their model and favoured this method over using data from small RCTs with restricted follow-up.

3.4.3.4.6 Additional key challenges related to comparing treatment sequences

In addition to the previously discussed challenges, several key challenges not directly related to any specific type of data source were identified and categorised into five types.: challenges in modelling treatment sequences related to data (F1-F7) and selection of treatment sequences (F10-F19), other data-related challenges (F8-9), other modelling-related challenges (F20-27) and decision-making concerns (F28-34).

Challenges F1-F4 are related to data scarcity from different perspectives. Challenge F1 deals with the absence of sequence-specific effectiveness data, and in fact, all data scarcity-related challenges can be viewed as subtypes of this in comparing treatment sequences, even though they were not always explicitly mentioned. While this issue was explicitly mentioned in a few TAs, it was an implicit challenge in almost all TAs. In TA574[254], the ERG found it challenging to evaluate the treatment sequences in the absence of data on sequence-specific effectiveness. The ERG concluded that the variation in outcome of different treatment sequences might primarily be due to discounting and mortality, rather than a genuine disparity in effectiveness. Thus, the ERG suggested that treatment-sequencing in the absence of real sequence-specific effectiveness data was highly problematic and could be used to suggest that one sequence is superior than another. F2 concerns the lack of UK-specific data for informing local treatment pathways and evaluating treatment effectiveness. In TA292[201], for instance, there was no UK-specific data available to inform the treatment pathway for treating paediatric bipolar disorder. Challenge F3 is comparable to F1 but differs in the sense that lacking time-varying data on other parameters, such as time-varying discontinuation rates, can also make it difficult to model treatment sequences. Challenge F4 notes that PFS and time-to-the-next-line of treatment may need to be used as a substitute for one another

when one of them is unavailable. This substitution may introduce additional uncertainties to a treatment-sequencing model.

Challenge F5 highlighted the inadequacy of using a treatment-sequencing model in addressing a decision-problem when there is a lack of evidence on the effectiveness of sequences. This limitation was noted in TA319 where the ERG found the model structure to be inappropriate.[203] However, it was possible to deactivate the treatment-sequencing part of the model, resulting in a direct comparison of first-line treatments (ipilimumab versus darcabazine versus vemurafenib) in previously untreated malignant melanoma patients. Challenge F6 refers to the challenge of applying RWE from treatment-sequencing studies to the decision problem. The Committee in TA195 discussed the Behandel Strategieën (Treatment Strategies for Rheumatoid Arthritis) (BeSt) study[261], which investigated the effectiveness of various treatment sequences of biological and conventional DMARDs in patients with early rheumatoid arthritis. Nonetheless, the AG argued that the evidence from the BeSt study was not applicable to the current MTA as it did not account for the clinical effectiveness of individual DMARDs and the study population did not represent those with established rheumatoid arthritis.

Challenge F7 involves the uncertainty surrounding ongoing studies exploring treatment-sequencing comparisons, which can be relevant to decision making. For example, in TA319[203], the Committee was aware of ongoing trials exploring treatment sequences involving vemurafenib and ipilimumab, including a randomized phase III trial of ipilimumab followed by vemurafenib versus vemurafenib followed by ipilimumab planned for BRAF V600 positive patients (ECOG E1612 trial). The Committee addressed the need of future research to better understand the position of ipilimumab within the treatment pathway for patients with malignant melanoma. Similarly, in TA377[203], the Committee also took note of several ongoing trials investigating treatment sequences in prostate cancer, though no specific reference was given.

In multiple TAs, the lack of head-to-head evidence was identified as a significant challenge (Challenge F8). The use of indirect comparison data from trials was seen as potentially biased. Challenge F9 refers to the necessity for longer-term data. For example, in TA375[188], the AG analysed large observational databases to obtain parameters for HAQ progression on conventional DMARDs, but the need for longer-term evaluation was still emphasized. In TA292[201], additional long-term data were required to assess the prevention of recurrent acute manic episodes in bipolar disorder in children and adolescents with aripiprazole and its comparators, due to the younger population.

Selecting all possible treatment sequences as comparators can be challenging in autoimmune TAs (Challenge F10). TA511 highlighted the difficulty, as it involves the ordering of treatments and the varying number of LOTs.[251] Some TAs added more treatment sequences in scenario analyses. The rapidly evolving treatment landscape in prostate cancer was identified as a challenge in TA377 (Challenge F11).[126] A few years ago, there were only limited treatment options available for mHRPC

patients, but many new treatments became available by the time of TA377. This rapid evolution makes it difficult to determine how a new treatment will affect the use of existing treatments in the new pathway and what proportion of patients will receive a specific subsequent treatment.

In Challenge F12, potential omissions of relevant treatment sequence comparators were noted in TAs across various diseases. In TA433[244], although the manufacturer used ordering rules summarised from previous TAs to determine the treatment sequences to be modelled, the ERG identified missing relevant treatment sequences. Challenge F13 discusses the challenge of determining the cost-effectiveness of specific treatment sequences due to the variation of patient characteristics in clinical practice. In TA103[245], individual characteristics, such as treatment history, renal and hepatic function, and associated response, may influence the optimal treatment sequence. Therefore, providing cost-effectiveness estimates for specific treatment sequences may not be relevant to an individual patient.

F14 pertains to challenges in comparing suitable treatment sequences due to administrative restrictions. TA377[126], for example, faced this issue with enzalutamide and abiraterone, both of which were available through the Cancer Drugs Fund (CDF) for mHRPC patients. However, using one treatment was not recommended for patients who had already received the other. The Committee determined that using both treatments was not common practice and recommended against using abiraterone after enzalutamide, except in cases experiencing toxicity without disease progression. Despite ongoing appraisal of abiraterone for a similar indication as the appraised technology enzalutamide (i.e. first-line treatment in mHRPC), it has not been incorporated into the NHS. Thus, even though both the manufacturer and the ERG considered abiraterone a relevant comparator, it was not included in the treatment sequences considered in the FAD. This exclusion was because primarily because the sequential use of abiraterone and enzalutamide was not permitted in the NHS, given their similar mechanisms. Additionally, the support for abiraterone by the CDF made its inclusion unnecessary.

F15 concerns that including treatment sequences with comparators that are not cost-effective may be problematic. For example, the ERG in TA575 stated that modelling selective treatment sequences could result in misleading cost-effectiveness results, particularly when including treatments that are not cost-effective in a sequence. The ERG conducted exploratory analysis to address this issue, which involved comparing individual treatments with best supportive care (BSC) using pairwise incremental CEA and net monetary benefit analysis (NMB) as well as changing the order of treatments based on efficacy and efficiency (NMB ranking).

Challenge F16 highlights that cost-effectiveness results may be sensitive to the selection of treatment sequences included in the CEA (Challenge F16). Determining the appropriate number of LOTs of treatment sequences (Challenge F17) and whether all treatment sequences should have the

same number of LOTs (Challenge F18) can be particularly difficult in autoimmune diseases where numerous treatment options were available. Additionally, the assumption that the duration of treatment equates to the duration of time to relapse is not always valid (Challenge F19). For instance, the ERG in TA308 criticized the manufacturer's assumption that all relapses result in immediate treatment with the next-line therapy, whereas patients who experience minor relapses may not require major relapse treatments.[156]

The choice of model structure may not always be supported by available clinical data, which can create challenges (Challenge F20). In the case of TA308[156], the ERG found that the KM curves for relapse outcomes in the RAVE pivotal trial crossed for the two first-line comparators, indicating that an alternative parametric model might be a better fit. However, this type of model may not be compatible with a standard Markov model structure, so the ERG suggested that the manufacturer's use of a Markov model might not be appropriate. The ERG but was unable to assess the impact of this limitation. In addition, F21 highlighted that manufacturers may face difficulties in modifying their CEMs to incorporate alternative treatment sequences as per ERG's request.

It can be difficult to model survival and treatment duration with respect to model structure (F22). In TA377[126], the ERG found that the extrapolated time-to-treatment-discontinuation curve crossed the OS curve for both abiraterone and enzalutamide, indicating that patients died before experiencing disease progression. The company assumed that patients stopped treatment at the point of crossing curves, which may be problematic as subsequent treatments after that point cannot be modelled, potentially leading to biased results. In the same TA, the costs and effectiveness were inconsistent between different treatment-sequencing arms (F23). The manufacturer did not include the costs of post-docetaxel treatment (i.e. BSC costs) in the enzalutamide-sequence and abiraterone-sequence arm, but included the post-docetaxel treatment costs in the BSC arm. According to the ERG, this inconsistency tended to improve the cost-effectiveness estimates for enzalutamide. Challenge F24 refers to the inconsistency between the timing of subsequent treatment in real-world practice and those modelled using clinical trial data. In TA292[201], the Young Mania Rating Scale (YMRS) response was the only suitable tool to model the decision of switching to subsequent treatments based on clinical trial data. However, in real-world practice, receiving a next-line treatment may not depend solely on the YMRS response.

Challenge F25 refers to the issue of modelling results being sensitive to the time horizon, which is a straightforward concept as the duration of time taken into account can influence the total number of treatments patients can receive in a treatment-sequencing model. Challenge F26 refers to the issue of having a population in clinical evidence that does not match the population intended to be modelled. In TA442[158], the manufacturer identified the population in their base-case analysis as biologic-naïve, but this was inconsistent with the patients included in the pivotal trial, UNCOVER, and other studies

used in the NMA. F27 concerns the issue of technical errors in the model, which is a frequently encountered challenge that is not limited to comparisons involving treatment sequencing.

There are various methods to modify the CEA that can provide different viewpoints to decision-making in a treatment-sequencing context. Challenge F28 involves evaluating pair-wise comparisons in addition to incremental CEA when assessing treatment sequences. Although the exclusion of relevant treatment sequences could result in misleading outcomes, in the case of TA442[158], the Committee examined pair-wise comparisons of relevant treatment sequences. This comparison was conducted to demonstrate that the sequences containing the appraised technology ixekizumab were superior to all other treatment sequences, rather than identifying the most cost-effective treatment sequence. In the same TA, the ERG requested that the manufacturer provide disaggregated life-year (LY) gained for each treatment sequence (Challenge F29). The ERG believed that although the total LYs across different treatment sequences may not vary, the disaggregated LYs are more likely to differ and are important in interpreting the QALY gains. In several autoimmune TAs, the ERG discussed whether the appraised technology should be considered as a replacement for an existing technology or as an extension of the current treatment sequences (Challenge F30). This consideration may influence the options and length of treatment sequences that should be compared.

In several TAs, the Committee faced challenges in providing concrete recommendations on the position of the appraised technology in the treatment pathway despite having evaluated a complete incremental CEA (Challenge F31). This was due to uncertainties in the model, for which more evidence is required for further evaluation. When a new technology was appraised for use at various LOTs, it may be appropriate to evaluate the technology's use for different positions by examining them in different populations (e.g. treatment naïve versus experienced), rather than comparing the same technology used at different positions within the same incremental CEA (F32). However, the arrangement should depend on the decision problem.

In TA319[203], the Committee expressed reservations about approving the assessed technology for use in treatment sequences, as there was a risk that it could lead to the sequential use of treatments that lacked robust clinical evidence (Challenge F33). In particular, there was no clinical trial data available to compare the sequential use and relative effectiveness of treatments used in sequence for previously treated advanced melanoma, despite ongoing studies mentioned in Challenge F7. Additionally, the company's submission oversimplified the treatment sequencing process by making arbitrary decisions regarding the choice of OS at different LOTs, both within and between treatment sequences. This raised concerns from the ERG that the results obtained for different treatment-sequencing groups may not be comparable, as there appeared to be no clinical rationale underpinning the choice of curves at LOT. As with Challenge F33, the last Challenge F34 also pertained to the Committee's caution. The Committee in TA415 made statements indicating that they were only

considering the scenario of placing the appraised technology in the treatment sequence compared to another drug, rather than attempting to depict the optimal treatment sequence.[257]

3.5 Discussion

This review aimed to complement prior reviews by offering a detailed analysis of the challenges and data sources utilised in determining the effectiveness estimates of treatment sequences in English HTA. It aligned with previous literature by highlighting a growing trend in the adoption of treatment-sequencing models, particularly in oncology, autoimmune diseases, and diabetes.[1,2] Furthermore, the review emphasised that nearly all TAs consider treatment sequences to some extent, regardless of whether they explicitly employ a treatment-sequencing economic model. This is conceptually fitting for managing chronic diseases where a variety of treatment options exist, often necessitating lifetime economic models to evaluate long-term outcomes and costs.

Although treatment sequences are relevant for nearly all TAs, each appraisal identifies challenges, either directly or tangentially, related to the lack of evidence on the effectiveness of treatment sequences. Only one TA in my in-depth review utilised effectiveness evidence from a SMART that randomised patients to different sequences (TA137).[215] This SMART technically might require advanced statistical analysis because not all participants received a second-line treatment (see Section 4.5.1 of Chapter 4). However, the manufacturer did not conduct such analysis, which complicates applicability in the economic model, but was not addressed in the TA. Furthermore, several TAs referenced ongoing trials that randomised patients to different treatment sequences; however, their evidence was not available at the time of appraisal. This scarcity of head-to-head trial evidence comparing treatment sequences often led to merging different LOT-specific evidence to estimate the overall effectiveness of treatment sequences in most TAs. Such an approach raises several concerns, including biases stemming from indirect comparisons and inconsistent patient populations across LOTs within the same treatment sequences. These insights broadly align with findings from a review published by Huang et al. during my PhD, which was included in the updated review in Chapter 2.[35]

Each appraisal utilised a variety of evidence sources to some extent, including clinical trials, NMA/MA, RWE, and effectiveness simplifying assumptions, to evaluate the effectiveness of treatment sequences. Notably, 77% of autoimmune TAs primarily depended on NMA/MA for their decision models, typically assuming constant treatment effectiveness across different LOT. Some even added modifiers for effect degradation or employed subgroup NMA/MA to account for effects in later treatment lines. In oncology, the most common method (43%) involved using data on PFS or time to treatment failure from LOT-specific trials to determine the cumulative treatment duration across LOTs. Typically, estimates of OS for an entire treatment sequence were based predominately

on results from trials that assess the effectiveness of first-line treatments, with adjustments for treatment switching when subsequent treatments in the trial did not align with local clinical practices.

LOT-specific treatment effect simplifying assumptions were predominantly used in autoimmune disease appraisals, likely because these treatments do not directly affect patients' survival. The reflects the fact that the simplifying taxonomy summarised by Lewis et al., along with their recommendations on employing NMA/MA techniques to better estimate LOT-specific effects, were primarily derived from the autoimmune disease literature. Therefore, these approaches may not be entirely applicable to oncology. Specifically, my review highlighted that merging LOT-specific survival evidence was considered problematic in the majority of oncology TAs, highlighting the need for distinct strategies in populating effectiveness of treatment sequences between the two topics.

Although RWE has been used to support the effectiveness of treatments in later lines, estimate modifiers for subsequent treatments, or as an external guide for selecting treatment sequences, or to inform death or treatment discontinuation rates that may indirectly affect later line treatment effectiveness, it has not been used to directly compare different treatment sequences. Given its retrospective nature, RWE holds potential for this, as discussed in Chapter 2. However, no TAs have delved into the statistical methods needed for such analyses. Therefore, this review reinforces the approach that is planned to be taken in this thesis: to explore how the strengths of RWD can be harnessed to deliver reliable RWE that informs the effectiveness of treatment sequences. This approach would provide an alternative way to overcome the challenges of merging evidence, but it would necessitate exploring pertinent advanced analytical methods, identifying relevant RWD sources, and designing RWD studies specifically tailored for this purpose. In one MTA[188], the AG were granted access to registry data to inform parameters in their treatment-sequencing model for rheumatoid arthritis, although not specifically for comparing treatment sequences. However, it is unclear what other RWD sources, especially in other disease areas, might be relevant for English HTA in the context of informing the effectiveness of treatment sequences. This includes considerations such as accommodating advanced statistical adjustment methods and whether the data are sufficiently longitudinal, suggesting that further exploration is needed.

3.6 Chapter summary

This chapter re-emphasised the growing importance of evaluating treatment sequences in HTA and dissected the clinical evidence scarcity problem across multiple facets by reviewing English HTA practices in this area. Specifically, it highlighted the distinct primary evidence sources utilised in autoimmune diseases and oncology, the major diseases where considerations of treatment sequencing are most prevalent. A major challenge is the lack of head-to-head trial evidence comparing treatment sequences. In oncology, the challenge is further compounded by the fact that survival is often a critical

outcome measure, which complicates the common approach of merging of evidence from LOT-specific sources to populate the effect of treatment sequences. Despite the potential of RWE, no TAs have yet utilised it to compare entire treatment sequences. This gap directs the focus of the thesis towards further investigating the use of RWD to inform the effectiveness of treatment sequences. This investigation begins with exploring advanced statistical adjustment methods suitable for analysing RWD to compare treatment sequences (Chapter 4). This focus arises because the review did not identify any directly applicable statistical methods currently in use.

Chapter 4 A systematic review of statistical methods for evaluating the effectiveness of treatment sequences with real-world data

4.1. Overview

This chapter presents a systematic review focused on identifying statistical methods for assessing time-to-event outcomes of treatment sequences using real-world data (RWD). It delves into methods that may be compatible with the Target Trial Emulation (TTE) framework[3,267], designed for mitigating biases in observational studies by explicitly defining the study population and estimands in a structured manner, as introduced in Chapters 1 and 2.

Chapters 2 and 3 highlighted the challenge of scarce effectiveness evidence for treatment sequences in health technology assessments (HTA) involving sequences, suggesting RWD as a potential alternative for estimates (i.e., real-world evidence (RWE)). However, despite the potential of RWD, its application has largely been limited to supporting evidence for the effectiveness of specific lines of therapy (LOTs) rather than directly comparing the effectiveness of different treatment sequences (Chapter 3). This approach implicitly assumes that the populations across different LOTs from distinct data sources are consistent, whether it involves merging RWE or evidence from LOT-specific trials. This assumption, however, may not always hold true. On the other hand, comparing treatment sequences using RWD faces concerns about potential biases inherent in analysing RWD, a concern not yet tackled in existing English technology appraisals (TAs) (Chapter 3). Hence, this chapter aims to review which advanced statistical methods are compatible with the TTE and are adept at generating unbiased estimates for comparing the effectiveness of treatment sequences.

Section 4.2 further details the review's rationale (Section 4.2.1), highlighting the issues conventional statistical methods encounter in adjusting for confounding, a significant challenge in establishing causal effects from RWD, —in the context of comparing treatment sequences. It then sets forth the detailed review questions (Section 4.2.2). Section 4.3 elaborates on the review's methods and procedures, including key search terms selection, literature identification, eligibility assessment, and data extraction and synthesis strategies.

Section 4.4 reports the literature search results and characterises the included studies, leading to the introduction of a taxonomy in Section 4.5. This taxonomy addresses the types of studies including treatment sequences (Section 4.5.1, briefly introduced in Chapter 2), the variations of treatment sequences (Section 4.5.2), and the categories of statistical methods found in the included studies (Section 4.5.3). Section 4.6 presents a narrative synthesis of the identified statistical methods, while Section 4.7 moves into a discussion, comparing these methods and discussing their applicability in a HTA context. The section concludes by outlining the methods chosen for the case studies in Chapter 7, which are implemented in Chapter 8.

4.2. Rationale and key questions the review

4.2.1. Rationale

This review aims to address the lack of clinical trials comparing the effectiveness of treatment sequences, an issue that poses significant uncertainties in health economic evaluations (Chapters 2-3). Real-world evidence holds promise as sequencing information can often be captured in RWD, yet analysing RWD presents challenges, primarily the need for statistical methods to address confounding and selection bias due to its non-randomised nature (Chapter 2). Therefore, this review aims to identify statistical methods for reliably estimating the effectiveness of treatment sequences using RWD. This review focused exclusively on methods applicable to time-to-event outcomes due to the emphasis on survival outcomes in my case studies (see Chapters 7-8).

Section 1.4.1 of Chapter 1 briefly introduced the concept of causal inference and the challenges of confounding, especially time-varying confounding, in comparing the effectiveness of treatment sequences. Despite studies attempting to compare treatment sequence using observational data, they often fail to consistently account for time-varying confounding, rendering them inadequate for addressing counterfactual questions (i.e., causal inference).[268-270]

This leads to a core question: "*what statistical methods are available to derive appropriate effectiveness estimates of treatment sequences using observational data?*". Here, "appropriate" refers to methods capable of counterfactual comparisons—thereby offering unbiased estimates similar to those from a RCT that randomises patients to different treatment sequences—and suitable for HTA. Expected methods might overlap with those correcting for unwanted treatment-switching and non-adherence in clinical trials, such as marginal structural models (MSMs) with inverse probability censoring weighting (IPCW) and structural nested failure time models (SNFTMs). However, a comprehensive review of their application in the context of treatment-sequencing remains lacking. To address this, I conducted a literature review to identify and assess methods suitable for HTA, with Section 4.2.2 delving into the detailed aspects of this core question.

In addition to time-varying confounding, my review also aimed to explore the functionality of each method within a Target Trial framework[3], which is endorsed by the NICE RWE framework[37] to further mitigate biases in RWD study design, such as immortal time bias and selection bias.[74,75] Further rationale for using RWD within a TTE framework to mitigate these biases has been detailed in Section 1.4.2 of Chapter 1. The compatibility of each method identified in this chapter with the TTE framework will be explored in the methods discussion (Section 4.7) and further in the implementation chapter (Chapter 8).

The field of causal inference continues to expand across various disciplines (e.g., statistics[50,51], epidemiology[5,52-54], sociology[45], psychology[46], computer science[47], political science[48] and

economics[49]), each employing unique approaches and terminologies. This review primarily concentrates on methodologies developed in epidemiology and (bio)statistics, given their relevance to HTA. This includes approaches that, while originating in other fields, have found application within epidemiology and (bio)statistics. Theoretically, these fields provide methodologies that may be particularly suited for analysing medical interventions, promising that the findings of this review are likely to be more readily adaptable for HTA.

4.2.2. Review sub-questions

In this section, I break down the core question from Section 4.2 into a series of sub-questions. To expand the range of methods considered, I incorporated the terms *"dynamic treatment regimens (DTR)"*, *"adaptive treatment strategies (ATS)"*and treatment pathways in my literature search (i.e., sub-question (2)), given their relevance to treatment sequences. Section 4.3.2 elaborates on how I became aware of these terms. These terms facilitate the identification of relevant statistical methods, with their link to treatment sequences further elaborated Section 4.5's taxonomy of treatment sequences, where treatment sequences are described as a specific type of DTR (this concept was also briefly touched on in Section 2.5.2.3 of Chapter 2). My goal is to assess the "appropriateness" (as defined in Section 4.2.1) of each method based on principles of causal inference[5,63,271-273], aiming to understand their applicability within HTA. The investigation is guided by the following sub-questions:

(1) What statistical methods have been proposed to estimate the (comparative) effectiveness of treatment sequences using observational data?
(2) Alternatively, what statistical methods have been used to estimate the effect of DTR, ATS, or any similar types of sequential treatment pathways in observational data, and may potentially be used to derive the (comparative) effectiveness of treatment sequences?
(3) If the statistical method was not originally applied in the presence of treatment sequences, what issue did it originally address (e.g., adherence, treatment cross-over)?
(4) What types of estimates were derived (e.g. time-to-event, hazard ratios (HRs), risk ratios)?
(5) How were the results interpreted? What were the estimands (see Section 1.4.3 of Chapter 1 for the relevance of estimands)?
(6) Can each method potentially be applied within the framework of Target Trial Emulation?
(7) What is the performance of each method, and what are the key issues encountered when applying them?
(8) Are they suitable for use in an HTA context? For example, can the method be used to determine the effect of a specific treatment sequence for the entire population, rather than the effect of a broader dynamic treatment strategy including various sequences (See Section 4.3.2 and 4.5.1 for taxonomy)?

4.3. Review methods

In Section 4.3.1, I outline the overarching approach of this review, encompassing a systematic search combined with citation-reference tracking. The systematic search strategy and eligibility criteria for evaluation are specified in Sections 4.3.2 to 4.3.4, while the approaches for data extraction and synthesis are explained in Sections 4.3.5 to 4.3.7.

4.3.1. Overarching approach

The primary objective of this review was to identify and evaluate statistical methods for comparing the causal effects of different treatment sequences and understand the challenges encountered in their application. Furthermore, I sought to assess the evidence on the performance and relevance of these methods for HTA and determine which methods can be applied to analyse RWD in my case studies. As per Section 4.3.1, I was interested in two types of literature: (1) methods proposed in the methodological research or reviews for estimating the causal clinical effects of treatment sequences, and (2) those related to broader DTR, ATS, or treatment pathways with potential applicability to derive effects of static treatment sequences.

The evolving terminology in methodological research complicates reliance on fixed-term systematic searches, risking either overly narrow or overly broad results. To address this, I adopted a two-step literature search strategy, starting with a systematic search to establish a foundational set of papers, followed by pearl-growing to expand this pool through citation and reference tracking. The latter mirrors the strategy employed in Chapter 2. Reference checking could efficiently lead to upstream methodological papers, while citation checking may identify downstream papers describing method extensions or applications. Given that my ultimate goal was to identify most relevant and promising methods (expected to be primarily found through systematic searches, with pearl-growing likely revealing method extensions and origins) for testing their applicability in my cases studies (Chapter 7-8) that serve as initial examples on this subject in HTA, I considered one round of citation-reference checking adequate for this exploration.

To ensure the review's robustness, a second reviewer, Saleema Rex (a peer PhD student), independently screened 100 random papers against the eligibility criteria. Disagreements were resolved through consensus, and any outstanding disagreements were discussed with my supervisors, Professor Nick Latimer and Professor Jim Chilcott. The final pool of included papers was reviewed with my supervisors to ensure no major relevant papers were overlooked.

4.3.2. Golden bullets

"Golden bullets" refer to papers that should undoubtedly be included in a literature review, serving to validate the search strategy and ensuring that relevant studies are identified.[101] In this section, I describe how I identified the golden bullets for this methods review.

Before starting the review, I was not aware of any literature providing a comprehensive summary of methods for estimating sequencing effect. However, I came across several ongoing research topics that shared similar concepts related to changing treatments, such as DTR (also known as ATS and treatment policies)[274-276], treatment-switching[43,277] and treatment non-adherence[278]. A DTR involves sequential decision rules for individual patients, where their characteristics are considered at each decision point throughout the disease course to optimise individual treatment trajectories. Notable examples of DTR include personalised strategies in treating human immunodeficiency virus (HIV) infection[279,280] and diabetes mellitus [275]. For HIV, decisions to change a treatment for an individual are contingent on their response to previous treatment (e.g. viral load). Optimising HIV treatment strategies does not always involve switching to an alternative medication. Instead, it could involve determining whether to continue or intensify the current treatment, as well as the optimal timing to initiate treatment in patients who receive active monitoring. In diabetes management, treatment typically begins with metformin and may be supplemented with additional drugs (e.g. sulfonylurea, DPP-4 inhibitors, SGLT2 inhibitors, insulin) based on the individual's response (e.g., HbA1c levels). Such DTR takes into account the sequential arrangement of treatments, allowing personalised sequences within a specific strategy to meet individual needs. In theory, methods used to assess the effect of individualised DTR could potentially be adapted to estimate the effect of (static) treatment sequences, commonly seen in HTA, where a single sequence is applied across the population.

With this prior knowledge, I came across Mahar et al.'s review (2021), which provides an overview of statistical methods used to estimate the effect of DTRs using observational data.[274] Their study identified commonly used methods for determining optimal DTRs, including inverse-probability weighting (IPW), parametric G-formula, G-estimation, and Q-learning for conditions such as HIV, diabetes and beyond. To some extent, Mahar et al.'s review has a wider scope than my intended review as it presents methods for optimising DTR beyond those related to switching treatments, such as methods concerning the timing of treatment initiation and dose adjustments. Additionally, a recent study by Simoneau et al. introduced dynamic weighted survival modelling (DWSurv)[275], a method that estimates optimal DTR for survival outcomes, demonstrated through a diabetes case study with UK Clinical Practice Research Datalink (CPRD) data. Consequently, both studies[274,275] are considered as golden bullets for my review.

It is crucial to distinguish between the effect of a (static) treatment sequence and a personalised DTR. An optimised DTR may involve tailoring different treatment sequences to each patient to achieve the optimal population effect. In HTA, comparing treatment sequences at a population level is common, with less focus on individualised treatment sequences (Chapter 2 & 3).[201,202,245] While my thesis does not specifically target the varied treatment effects based on patient characteristics, exploring DTR methods may still be beneficial because of the overlapping concept of sequential

treatments.

On the other hand, statistical methods addressing treatment-switching in clinical trials[43] and adherence-dependent effectiveness[43] may also share common ground with those in treatment sequencing and personalised DTR, given the time-varying nature of exposures. However, methods used in adherence studies often focus on understanding the effectiveness under specific levels of adherence, which may not be directly transferable to the treatment-sequencing context. Methods used to adjust for treatment-switching in trials typically aim to understand the effectiveness of a single LOT rather than sequential treatments.[43] Latimer et al. recently proposed an improved two-stage estimation method for adjusting for treatment-switching, factoring in different waiting times before moving to an active drug post-initial treatment failure (i.e., immediately after progression or delayed).[152] This method bears similarities to Huang et al.'s consideration of treatment-free periods between LOTs oncology treatment-sequencing economic modelling (Chapter 2).[35] However, in specific contexts like advanced cancers, which are the focus of my case studies (Chapters 7 and 8), these nuances might be less relevant due to the likely brief gaps between LOTs. Hence, I did not consider any treatment-switching and non-adherence studies as golden bullets. Nonetheless, two statistical method reviews in these areas by my supervisor, Professor Nicholas Latimer, and his co-supervised student, Dr. Abualbishr Alshreef, provided valuable structural insights for my methods review.[278,281]

In summary, examining whether the literature search includes the two golden bullets—the DTR review by Marhar et al.[274] and the study by Simoneau et al.[275], which estimate optimal DTR for survival outcomes—helps validate the effectiveness of the literature search.

4.3.3. Search strategy

To enhance the search results' breadth while maintaining specificity, I refined my core search strategy of "treatment sequence" and "statistical methods" by incorporating variations of "treatment sequence" (Chapter 3, Table 3.2) and adding terms from Mahar et al.'s search strategy[274], including DTR, ATS and their variations. While Mahar et al.'s review focused only on clinical application studies using observational data (i.e. excluding pure methodological papers, simulation studies, and reviews), my goal was to identify relevant methods across all study types, so I did not fully follow their strategy. Furthermore, I assumed that statistical methods intended for causal inference would rely on constructing a statistical "model" and likely to mention "time-varying" or "causal relationship(s)". Hence, I refined my search by replacing "statistical methods" with ("causal + model" OR "time-varying + model").

Following a pilot search and discussions with supervisors, the final search terms were established. A thorough PubMed search on March 13, 2022, identified the two "golden bullets"

specified in Section 4.3.2. Equivalent strategies were then replicated for EMBASE/MEDLINE. Specific search terms are outlined in Appendix 4.1 and Appendix 4.2 (PubMed and EMBASE/MEDLINE, respectively).

4.3.4. Screening and eligibility assessment

The screening and eligibility assessment process involved comparing records from database searches and reference-citation tracking with the criteria outlined in Table 4.1. Initially, titles and abstracts were screened, followed by a full-text review for final inclusion. To optimise the pool of papers via citation and reference tracking, I included all papers describing relevant methods in the initial systematic search, whether they presented the method in reviews, simple examples, or full application studies. This included theoretical papers lacking practical case studies, as they may be cited by application studies that demonstrate their utility.

This review did not aim to include upstream methodology papers that did not attempt to estimate the clinical effect of treatment sequences, DTR, or similar. However, foundational methods cited in eligible studies were discussed in the context of their theoretical principles and origins in the narrative synthesis (Section 4.6). Further, the review focused on individual patient-level data methods, excluding aggregated data techniques such as meta-analysis and meta-regression. The review extended beyond health economic literature to encompass all pertinent statistical methods across various fields, particularly biostatistics and epidemiology, mirroring insights from Chapter 2.

Table 4.1 Inclusion and exclusion criteria for eligibility assessment

Inclusion criteria	➢ Methodological papers* that describe statistical methods to estimate the (comparative) effectiveness of different treatment sequences ➢ Methodological papers* that describe statistical methods to evaluate decisions involving the effect of dynamic treatment, adaptive treatment strategies, or similar may be included if the paper provided explanations for calculating outcomes of a particular treatment sequence. ➢ Application studies (either practical example using real data, or simulation studies) that applied statistical methods to estimate the (comparative) effectiveness of different treatment sequences. ➢ Application studies that applied statistical methods to evaluate decisions involving the effect of dynamic treatment regimens, adaptive treatment or similar may be included if the paper provided methods for calculating outcomes of a particular treatment sequence. ➢ Methods reviews that summarise statistical methods that can be used in the context of treatment sequences or estimating the effectiveness of dynamic treatment regimens adaptive treatment strategy or similar. ➢ Paper published from databases inception to date
Exclusion criteria	➢ Full-text unavailable ➢ Non-English papers ➢ Conference abstracts or paper with insufficient description of the statistical methods ➢ Book Chapters and Thesis ➢ Methods based on aggregated data such as meta-analysis or meta regression ➢ Papers that apply a statistical method but without description and justification of the method (Papers of methods application with detailed justification will be included) ➢ Original** or upstream papers that did not contain any application of the methods in a treatment-sequencing, dynamic treatment regimen, adaptive treatment strategy or similar context
* Theoretical papers with no application (regardless of using actual data or simulated data) will be included; however, papers without an application need to at least offer practical exposition. ** The first paper of a statistical method or the first paper using an existing method in a new context (e.g. treatment-sequencing) will be considered as “original”.	

4.3.5. Data extraction

The data extraction process, detailed in Table 4.2, was adapted from previous reviews on methods for DTR, treatment-switching, and non-adherence.[274,278,281] It started by gathering key details from each study, focusing on basic information and methodological origins. Papers employing similar methodologies were grouped together and classified as "original," "method extension," or "application only." Papers introducing a new statistical method or applying an existing one in a new context, such as treatment sequencing, were labelled "original". Theoretical suitability for each study was assessed, followed by the extraction of application study details where applicable.

4.3.6. Data synthesis

Quantitative results from multiple choice or binary questions in Table 4.2 were first summarised. Then, a qualitative narrative synthesis was performed for each methodological group—comprising original, method extension, and application only papers. Each group of papers were collectively assessed and discussed using the methods appraisal framework in Table 4.3, adapted from Latimer's doctoral thesis and Alshreef et al.'s research.[281,282]

In the synthesis, I first noted the complexity arising from the variations of data sources and treatment sequences, each adding nuances to the statistical methodologies applicable in different contexts. For clarity, I summarised several taxonomies before moving into the narrative synthesis of results in Section 4.5. These taxonomies include: the types of studies involving treatment sequences (Section 4.5.1, and as briefly outlined in Sections 2.5.2.3-2.5.2.4 of Chapter 2); the variants of treatment sequences (Section4.5.2, and as briefly outlined in Section 2.5.2.3 of Chapter 2); and the categorisation of simple versus non-simple methods (Section 4.5.3).

Here, simple methods broadly refer to (1) statistical analyses that do not adjust for confounding bias, potentially suitable in scenarios where randomisation eliminates such bias, yet biased where confounding exists, notably in (but not limited to) RWD analysis; (2) conventional approaches such as simple outcome regression that include both baseline and time-varying confounders, aiming to adjust for confounding bias but may inadvertently introduce additional bias. Additional bias introduced by such techniques was clarified with directed acyclic graphs (DAG) in Section 1.4.1, Chapter 1. Additionally, (3) RWD analyses may attempt to restrict to certain patient groups or observation periods, whether incorporating the conventional adjustment approach (2) or not, to mitigate bias from treatment-sequencing strategy deviations. However, this risks introducing immortal time or selection bias by selecting patients based on characteristics observed after enrolment, as briefly explored in Section 1.4.2, Chapter 1. Contrary to simple methods, non-simple methods here refer to advanced statistical techniques that are capable of comparing the counterfactual outcomes of different treatment sequences. Section 4.5.3 presents a more detailed taxonomy of non-simple

methods, preparing readers for the narrative synthesis in Section 4.6.

Table 4.2 Data extraction table: basic information and methodological origin

Dimension	Items	Info extraction
Article information (all papers)	➢ Title	[Free-text]
	➢ Author	[Free-text]
	➢ Journal	[Free-text]
	➢ Year of publication	[Number]
	➢ Has an application study?	[Yes/No]
Methodological origin (all papers)	➢ Method (Which statistical methods have been used?)	[Free-text]
	➢ Method acronym	[Free-text]
	➢ Methods group e.g. parametric G-formula, G-estimation and Q-learning, others [free-text]	[Multiple choice/ Free-text]
	➢ Type of methodological contribution. e.g. original, methodology extension, or application only (informing clinical practice)]	[Multiple choice]
	➢ Was the method originally developed to estimate the effectiveness of treatment sequences?	[Yes/No]
	➢ If not (i.e. an application or extension of another method), what is the original method and in what context?	[Free-text]
	➢ Following the last question, how has the method been adapted?	[Free-text]
Theoretical suitability (only "original" or "method extension" papers*)	➢ How does the method work?	[Free-text]
	➢ What key assumptions does the method make?	[Free-text]
	➢ What are the potential biases?	[Free-text]
	➢ Why might the method not be appropriate?	[Free-text]
	➢ What are the advantage or disadvantages of the method?	[Free-text]
Application (all papers with applications)	➢ Has the method been applied to estimate the (comparative) effectiveness of different treatment sequences or to evaluate decisions involving the effect of dynamic treatment regimens, adaptive treatment or similar in a case study? If yes, then answer the question below:	[Yes/No]
	➢ Has the method been applied to estimate the (comparative) effectiveness of different treatment sequences or to evaluate decisions involving the effect of dynamic treatment regimens, adaptive treatment or similar in a simulation study? If yes, then answer the question below:	[Yes/No]
	➢ In what diseases/conditions has the method been applied in the case/simulation study? E.g. Oncology, HIV/AIDS.	[Multiple choice/free-text]
	➢ What is/are the intervention(s) assessed in the case/simulation study?	[Free-text]
	➢ Outcome(s) assessed? E.g. Time-to-event, Binary, Continuous	[Multiple choice]
	➢ Used within a Target Trial emulation framework?	[Yes/No]
	➢ Are there multiple analysis approaches	[Yes/No]
	➢ Following the last question: If multiple model/analysis were compared how do they differ?	[Free-text]
	➢ What were the results compared to traditional approaches (ITT, AT) if available?	[Free-text]
	➢ What are the possible estimands in the application? How were they interpreted?	[Free-text]
	➢ Has model performance been evaluated?	[Yes/No]
	➢ Following the last question: If yes, how? E.g. cross-validation, AIC/BIC	[Free-text]
	➢ Have the authors specified methods used for covariate selection?	[Yes/No]
	➢ How are the covariates selected? DAG, expert opinion, statistical significance?	[Free-text]
	➢ Have the authors mentioned any violations of the key assumptions of the methods?	[Yes/No]
	➢ Following the last question: If yes, what are they?	[Free-text]
	➢ Have any methods been used to tackle missing data issue?	[Yes/No]
	➢ Following the last question: How was missing data tackled? E.g. complete case analysis, multiple imputation.	[Free-text]
	➢ Any sensitivity analysis?	[Yes/No]
	➢ Following the last question: If yes, what were those? E.g. truncated inverse probability weights?	
	➢ Data type. E.g. registry, electronic health records, claims	[Free-text]
	➢ Software. E.g. SAS, STATA, R, Unavailable	[Free-text]
	➢ Key challenges identified by the authors	[Free-text]

AIC: Akaike Information Criterion; AIDS: Acquired Immunodeficiency Syndrome; BIC: Bayesian Information Criterion; DAG: directed acyclic graph; AT: as-treated; HIV: Human Immunodeficiency Virus; ITT: intention-to-treat

* Papers within the same method group were labelled as "original", "method extension" or "application only". The first paper of a statistical method or the first paper using an existing method in a new context (e.g. treatment-sequencing) was be considered as "original".

Table 4.3 Methods appraisal framework

Dimension	Considerations
Origin	➢ Has the method been used to estimate the effect of treatment sequences? ➢ If not, in what context has the method been applied and what is the similarity of its use compared to its potential use in analysing the effect of treatment sequences? ➢ In what disease area has the method been frequently used? ➢ Does the method represent an extension to another method used to estimate the effect of treatment sequences?
Theoretical Suitability	➢ How does the method work? ➢ What are the key assumptions of the method? ➢ What major concerns may invalidate the use of the method, causing bias? ➢ What are the advantages and disadvantages of the method? And are these in relation to data quality, data type or methodological assumptions? ➢ Can the method be used in combination with other methods wherever relevant? ➢ What are the similarities and differences of the method compared to other methods identified?
Application	➢ Has the method been applied to estimate the (comparative) effect of treatment sequences or to evaluate decisions involving the effect of dynamic treatment regimens, adaptive treatment or similar in a case study or simulation study? If yes, summarise the case/simulation study ➢ What were the results compared to simple methods (e.g. ITT/AT), if compared? ➢ Was any evidence presented to allow an evaluation of the "successful use" of the method? If yes, how? E.g. Compared to a benchmark trial or estimates from clinical expert opinions ➢ If there was any "successful" example, was it in a treatment sequence setting? And in what data type has it been applied to? Can it potentially be applied to other types of data? ➢ What are the key challenges associated with the application of the method and how could these be addressed? These may include key assumptions, covariate selection and missing data problem. ➢ What are the possible estimands for each method? How were they interpreted? What were the differences between each possible estimand? ➢ Are there open source statistical programmes or packages operationalising the method? ➢ Is the method suitable to be applied in a HTA context? Why and Why not?
Others	➢ Any other relevant characteristics of the method that need to be considered?

AT: as treated; ITT: intention to treat

4.3.7. Data management

I utilised Covidence, an online systematic review platform, to collaborate with my peer reviewer, Ms. Rex, on screening and evaluating full texts, as well as eliminating duplicate entries. All included studies were managed using EndNote (version X8) for data synthesis and referencing.

4.4. Results

4.4.1. Search results

The review identified 2,208 records. Figure 4.1's PRISMA diagram shows the eligibility assessment, citation-reference tracking, and a post-hoc paper addition process. Reasons for excluding papers are categorised. Sections 4.4.1.1 to 4.4.1.3 provide further insights on the inclusion process, while Section 4.4.2 summarises characteristics of the included papers.

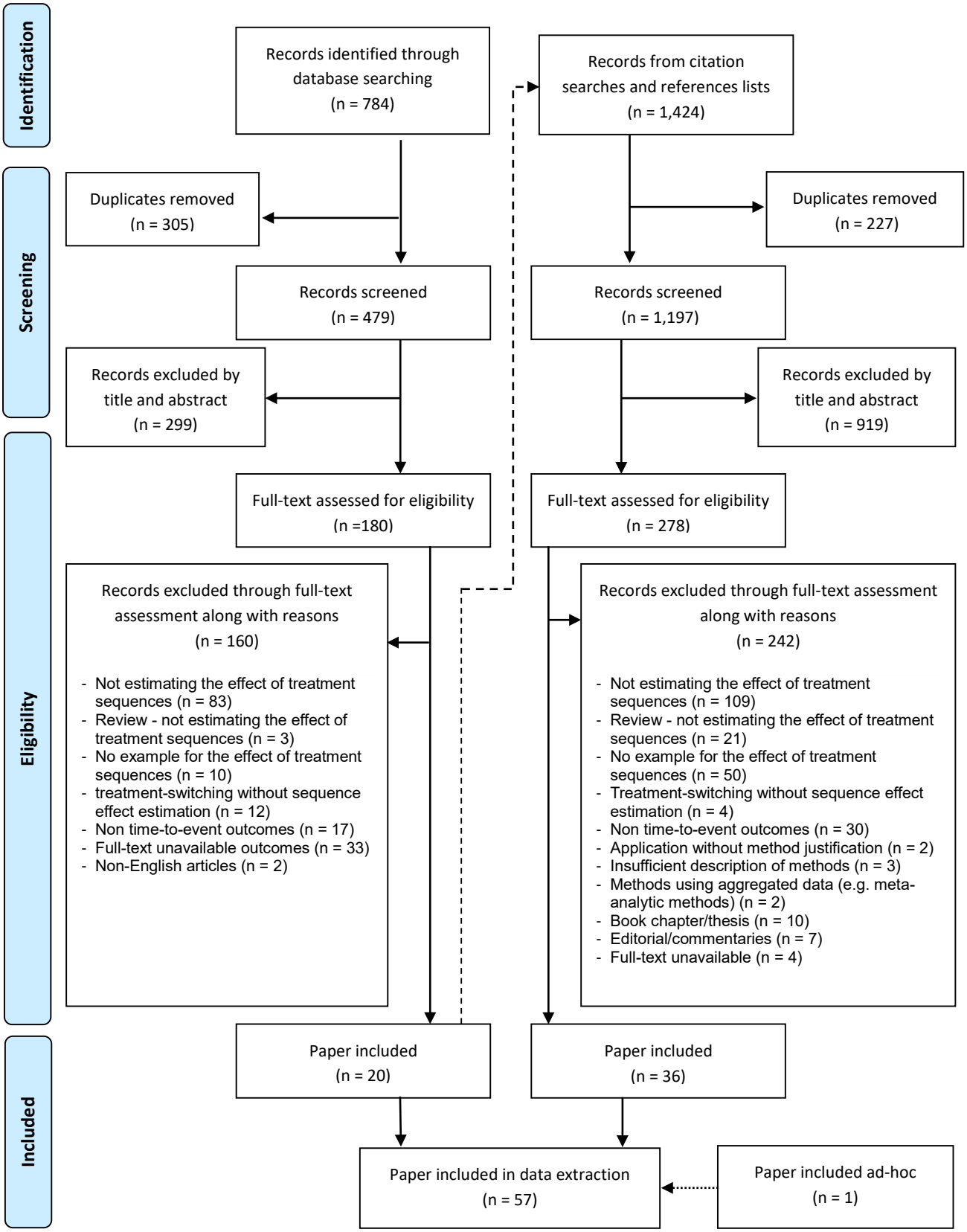


Figure 4.1 PRISMA flow diagram

The solid lines illustrate the inclusion and exclusion process through duplicate removal, title/abstract screening, and full-text assessments. The dashed line represents the one-time citation and reference tracking process, while the dotted line indicates the papers added based on expert suggestions.

4.4.1.1. Papers included through systematic database search

The database searches resulted in 784 records: 363 from PubMed, 249 from EMBASE, and 172 from MEDLINE (Figure 4.1), including the two "golden bullets" referred to in Section 4.3.2.[274,275] After deduplication (n = 305), 479 records were screened by title and abstract, leading to 299 records being excluded. Following a full-text review, another 160 did not meet eligibility criteria, resulting in 20 papers being included for data extraction. Ms. Rex, my peer reviewer, independently reviewed 100 random records out of the 479 non-duplicated ones for quality assurance. We resolved any disagreements through thorough discussion, as mentioned in Section 4.3.1. The same evaluative logic was then consistently applied to assess the eligibility of the remaining papers.

The 20 papers included at this stage served as the "grit" for citation-reference tracking in the next stage of the review (i.e., pearl growing). Among these were 16 methodology papers and 4 methodological reviews. These include the "golden bullet" papers by Simoneau et al. and Mahar et al., as outlined in Section 4.3.2.[274,275] The four methodological reviews cover diverse topics: causal methods enabling predictions under hypothetical interventions[283], causal inference in RCT analysis[284], Bayesian nonparametric statistics in cancer research with an example of DTR[285], and DTR methods in observational studies.[274] I selected these review papers not only for their potential in leading to additional relevant methodological studies through citation/reference tracking but also because they may provide clearer explanations and comparisons of different methods.

The second column from the left in Table 4.4 shows the number of papers included in Stage 1, with parentheses indicating those using non-simple methods (as defined in Section 4.3.6). Five studies explored non-simple methods in analysing RCTs involving sequencing information, including two reviews.[284-288] Six studies discussed non-simple methods for analysing RWD, including two reviews.[274,275,283,289-291] Another two studies investigated non-simple methods for analysing sequential multiple assignments randomised trials (SMARTs)[292,293], while the remaining seven studies were application studies comparing sequences using simple methods across various data types.[173,294-299]

4.4.1.2. Papers included through citation and reference tracking

In stage 2, 1,424 records were identified through citation/reference tracking (June 9, 2022), as shown in Figure 4.1. While Google Scholar yielded a higher number of citations, the excess mostly came from non-peer-reviewed or non-English publications. Consequently, I exclusively utilised Web of Science for tracking citations. After de-duplicating 227 records, 1,197 underwent abstract and title screening, leading to the exclusion of 919 and leaving 278 for full-text review. Many papers (919 out of 1,197) were excluded during the abstract and title screening for their sole therapeutic focus, which lacked relevance to statistical methodology. These papers primarily consisted of references or citations from Stage 1's method application only studies, which employed simple methods in specific

therapeutic areas, or were merely cited to provide disease backgrounds. Hence, it was deemed unlikely that further insights into advanced statistical methods could be gleaned from these papers. After full-text review, 242 records were excluded, leaving 36 for data extraction, including a review by Chakraborty and Murphy on DTRs for its comprehensive overview on the topic.[4]

The third column from the left in Table 4.4 presents the count of studies included in Stage 2. Most non-review studies included at this stage focused on non-simple methods in analysing SMARTs (n = 16)[123,178,300-313], alongside application-only studies that utilised simple methods for sequence comparison across various data types (n = 17) [314-330], and two studies that employed non-simple methods to analyse RWD or RCT with sequencing information.[331,332]

Table 4.4 Characteristics of included papers

	Publication count (Publication count excluding those using simple methods)			
Characteristics	**Stage 1: systematic search**	**Stage2: reference-citation tracking**	**Ad-hoc**	**Total**
Year of publication				
1996-2000	1 (0)	1 (0)	-	**2 (0)**
2001-2005	2 (0)	3 (3)	-	**5 (3)**
2006-2010	3 (3)	5 (3)	-	**8 (6)**
2011-2015	6 (3)	17 (8)	-	**23 (11)**
2016-2020	4 (4)	7 (4)	1 (1)	**12 (9)**
2021	3 (2)	2 (1)	-	**5 (3)**
2022	1 (1)	1 (0)	-	**2 (1)**
Type				
Methods paper with application study, including those with or without simulation studies	8 (8)	15 (15)	1 (1)	**23 (23)**
Methods paper with only simulation studies	-	2 (2)	-	**2 (2)**
Application only	8 (1)	18 (1)	-	**26 (2)**
Review of methods	4 (4)	1 (1)	-	**5 (5)**
Disease area*, excluding review publications				
Cancer (blood)	5 (5)	8 (8)	-	**13 (13)**
Cancer (solid tumour)	7 (3)	25 (8)	-	**32 (11)**
Diabetes	1 (1)	-	-	**1 (1)**
Glaucoma	1 (0)	-	-	**1 (0)**
Hepatitis B	1 (0)	-	-	**1 (0)**
HIV infection	1 (0)	-	-	**1 (0)**
Cardiovascular disease	-	-	1 (1)	**1 (1)**
None (simulation only)	-	2 (2)	-	**2 (2)**
Type of data used, excluding review publications				
RCT	2 (2)	-	-	**2 (2)**
Sequential treatment RCT	2 (0)	5 (0)	-	**7 (0)**
Pooled sequential treatment RCTs	1 (0)	1 (0)	-	**2 (0)**
Simulation studies + RCT	1 (1)	1 (1)	-	**2 (2)**
RWD	6 (2)	12 (1)	-	**18 (3)**
Simulation studies + RWD	2 (2)	-	-	**2 (2)**
RWD toy example	0 (0)	0 (0)	1 (1)	**1 (1)**
SMART	-	5 (5)	-	**5 (5)**
Simulated SMART	-	4 (4)	-	**4 (4)**
Simulation studies + SMART	2 (2)	7 (7)	-	**9 (9)**
Total	**20 (13)**	**36 (19)**	**1 (1)**	**57 (33)**

DTR: dynamic treatment regimen; HIV: human immunodeficiency virus; RCT: randomised controlled trial; RWD: real-world data; SMART: Sequential Multiple Assignment Randomized Trial

4.4.1.3. Paper included ad-hoc

During the ad-hoc phase, I added a briefing paper by Hernán on the use of the clone-censor-weight method on survival outcomes, featuring a RWD toy example.[333] Despite this method increasingly being used in various scenarios over the past five years[75,333-335], Hernán's paper was selected for its practical illustration of comparing aspirin treatment durations (e.g., none, 1 year, 2 years), which subtly mirrors the comparison of several treatment sequences (e.g., no-aspirin → no-aspirin, 1 year aspirin → 1 year no aspirin, 1 year aspirin → 1 year aspirin). This paper was not identified in the systematic search and reference/citation tracking as it technically did not compare different treatment sequences (Table 4.1 eligibility criteria). The relevance and conceptual parallels of such type of time-related static treatments, which, to an extreme extent, could arguably be considered a type of DTR, only became clear to me after attending CAUSALab courses at Harvard University on Target Trial Emulation in June 2022 and Advanced Confounding Adjustment in June 2023, which was beyond the timeframe of my initial review. Among the materials and references reviewed in those courses, this paper stood out as one of the most analogous and conceptually approachable ones for comparison. In fact, methods in the majority of studies included in my review share similar roots with those discussed in the CAUSALab courses. This indicated I have likely identified the most relevant advanced statistical methods for comparing treatment sequences, including those utilised in broader contexts of DTR. Hernán paper's briefing paper is distinguished by its use of the recently increasingly recognised "cloning" technique and practical example in data organisation[333], proves to be a valuable and fitting addition to my review (see Section 4.6.1.2.2). With this paper's inclusion, the total number of studies in the final review reached 57 (n = 20 + 36 + 1).

4.4.2. Characteristics of the included papers

Table 4.4 highlights that over half of the studies published in the last 15 years. There has been a growth of non-simple methods in comparing treatment sequences in the last decade compared to the previous one. Among non-review articles, non-simple methods (n = 28) are predominantly linked to oncology studies (n = 24, solid tumour n = 11, haematological malignancy n = 13), with exceptions including a diabetes-focused study (i.e., the golden bullet by Simoneau et al.[275]) and the cardiovascular disease ad-hoc paper by Hernán.[333] Simple methods have been broadly applied across studies in oncology, glaucoma, hepatitis B, HIV, and cardiovascular disease, with the majority them being application-only studies.

All methods used to analyse sequential treatment RCTs—where patients were randomised to different treatment sequences—employed ITT analyses without any adjustments (n = 9), regardless of whether the sequential treatment RCTs were pooled from multiple sources. For RWD studies, the majority used simple methods for analysis, with only 6 out of 21 (29%) employing non-simple

methods. All studies analysing SMARTs (n = 18) utilised non-simple methods.

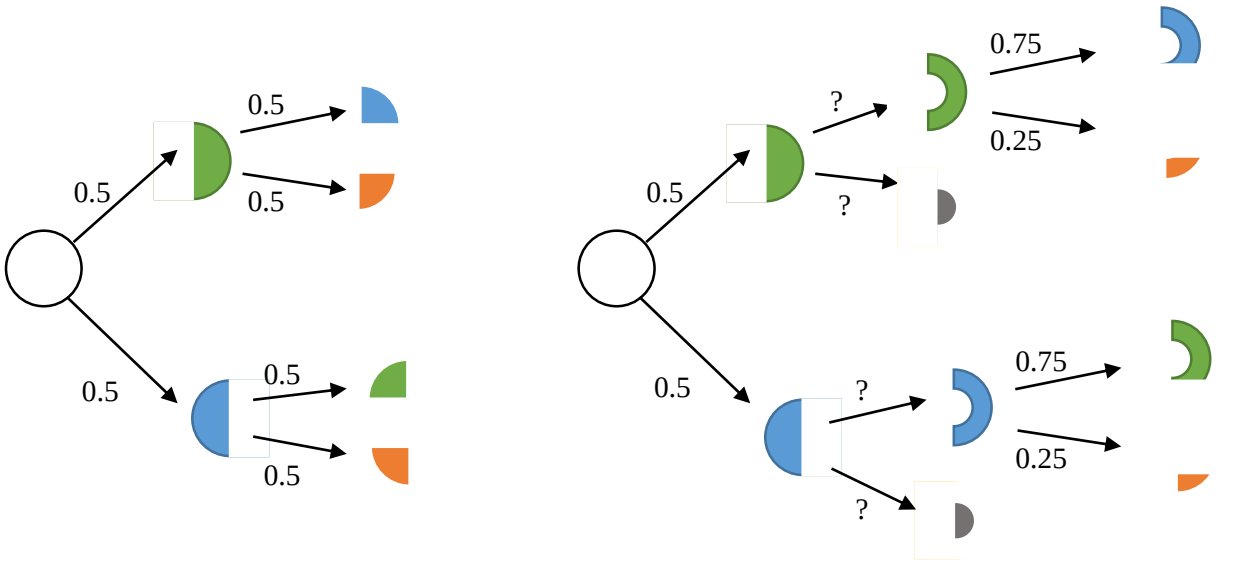
4.5. Taxonomies

4.5.1. Types of studies involving treatment sequences

Figure 4.2 highlights the parallels across different data sources harbouring sequencing information, as briefly introduced in Section 2.5.2.3-2.5.2.4 in Chapter 2.

A. SMART with two-stage randomisation, in which all participants undergo second-stage randomisation

B. SMART with two-stage randomisation, in which second-stage randomisation is only for criteria-meeting participants (e.g. progression survivors)



C. RCT randomising patients to two different treatment sequences where only criteria-meeting participants receive subsequent treatments (e.g. progression survivors)

D. RWD study where patients may undergo various treatment sequences, with those meeting specific criteria receiving further treatments (e.g., progression survivors)

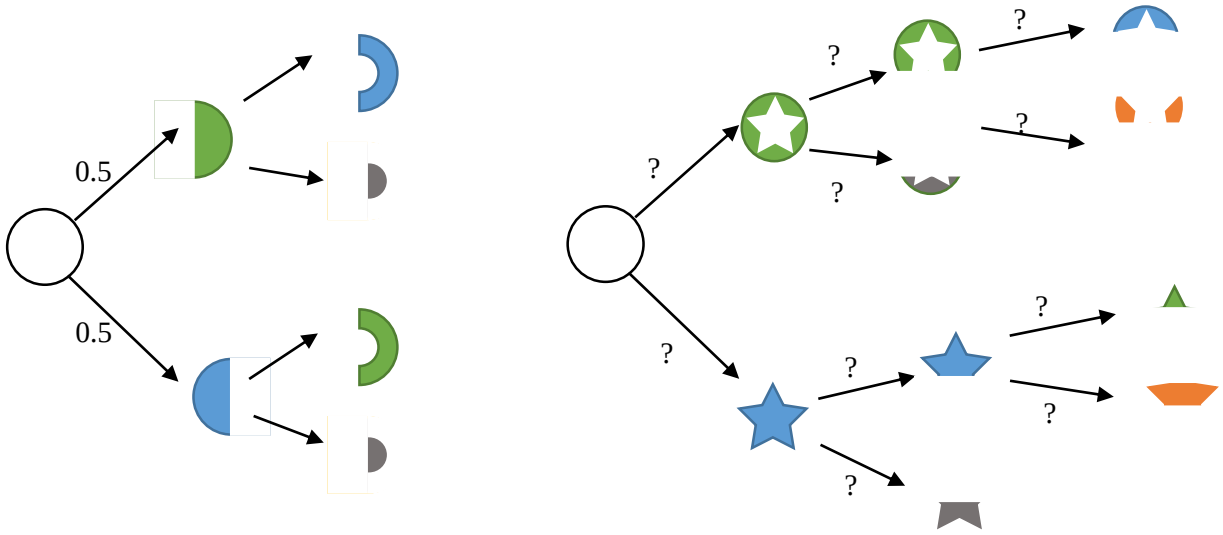


Figure 4.2 Illustration of parallels among various data sources containing treatment sequences
RCT: randomised controlled trial; RWD: real-world data; SMART: Sequential Multiple Assignment Randomized Trial

In the series of figures in Figure 4.2, arrows numbered with values indicate the probability of randomization to a specific group, while a question mark "?" next to an arrow signifies a non-randomized process. Figure 4.2A depicts a SMART with two-stage randomisation, assigning participants with a 0.5 probability at each stage to one of two treatments. This results in each 25% of participants (a quarter of the circle) experiencing one of four sequences: green → blue, green → orange, blue → green, or blue → orange. For causal effect comparisons among these sequences, theoretically, no adjustment is needed since all patients had equal chances of receiving one of the treatment sequences due to all participants undergoing the second-stage randomisation (i.e., each group is comparable).[4,336]

However, more advanced analysis may become essential if not every participant undergo each stage of randomisation, as shown in Figure 4.2B. This complexity arises because these deviations can affect the probability of receiving a specific treatment sequence (i.e., not equal circle sizes for each one of the four treatment sequence group), leading to biased conclusions if direct comparisons are made between patients receiving different sequences. For example, the SMART design may only proceed to a second randomisation for participants meeting specific criteria, such as survival upon disease progression (Figure 4.2B). Despite an initial 50/50 randomisation ensuring baseline comparability, varying survival rates to the second randomisation (first-line treatment: green versus blue) necessitate advanced analytical adjustments to accurately compare the effects of different sequential treatment strategies. When assessing the effect of the green → blue strategy by following all patients who start with green and either excluding, or censoring those who initiated second-line orange treatment at second-line initiation, it may seem to eliminate the "contaminated effect" from the orange second-line treatment. However, this approach will disproportionately represent those unable to proceed to second-line treatment (due to death or unfitness), resulting from the "artificial removal" of a significant proportion of data post-second-line. Hence adjustment with advanced statistical methods is needed.[336]

Conceptually, one solution is to apply weights to patients. For example, to estimate the marginal effect (i.e., average treatment effect (ATE)) of the green → blue strategy, patients who completed this sequence could be weighted by the inverse probability of receiving the second-line blue treatment (e.g., 1 over 0.75) to simulate a scenario where all patients eligible for a second-line treatment received blue. Meanwhile, patients who only received the first-line green treatment and did not progress to the second line would receive a weight of 1, as they were still considered as adhered to the green → blue strategy. Note, this is a conceptual example, and while analyses using different estimators may vary slightly, this captures the essence. This weighted group then represents what the outcomes would have been had all patients initially been randomised to the green → blue treatment path. Given the initial 50/50 randomisation between the first-line green and blue treatments, this

estimation also represents the counterfactual effect where all participants in the SMART would have received the green → blue strategy. This example is applicable for analysing restricted mean survival time, whereas other survival analysis might censor patients at the point they deviate from the treatment strategy of interest and upweight those who did not deviate to represent those who did, only from the point where someone else in the group deviated (See Sections 4.6.2.1.1-4.6.2.2.2 for details). As mentioned earlier in Chapter 3, only one TA were found to leveraged evidence from a SMART (Section 3.4.3.4.2, Chapter 3).[215] Specifically, the SMART utilised in TA 137 followed a structure similar to that shown in Figure 1.2B.[215] However, the manufacturer relied solely on a simple statistical method, intention-to-treat (ITT) (see Section 4.6.1.1), to assess the overall survival (OS) and progression-free survival (PFS) across different treatment strategies. This approach may not effectively derive the counterfactual outcomes of varying treatment sequences. Nonetheless, the evidence review group (ERG) did not explicitly raise any specific concerns about this.

Figure 4.2C represents a two-arm sequential treatment RCT, where patients were randomised to two distinct treatment sequences (green → blue versus blue → green), and only those who survived upon disease progression were eligible for subsequent treatments. This setup reflects two of the four groups in the SMART in Figure 4.2B that have been adjusted with the aforementioned method to estimate treatment sequence strategy's ATE for the entire trial population. Participants in Figure 4.2C underwent a 50/50 randomisation to a fixed treatment sequence without a second randomisation point, so no adjustment is needed to assess the causal effect of the two sequential treatment strategies. This highlights the efficiency of the SMART design in comparing sequential strategies with fewer participants (though more advanced analytical techniques may be required).[337] Contrarily, to compare four sequences using a traditional RCT approach, a four-arm design would be necessary, requiring a larger participant pool, such as the BIG 1-98 breast cancer trial.[321]

Figure 4.2D shows the parallels between a SMART (Figure 4.2B) and a RWD study. The key distinction lies in the known randomisation probabilities in a SMART versus the unknown probabilities of receiving treatments in RWD studies.[60] Despite these differences, RWD may allow for the estimation of treatment probabilities with assumptions (i.e., in a sense, probabilities that can be leveraged to emulate randomisation), such as treatment decisions being determined by a patient's performance status, weight, age, and comorbidities. This means the approaches to estimating treatment sequence effects from RWD may theoretically mirror those in SMART analysis, but depending on the complexity of SMART (simpler ones like in Figure 4.2A, may not require advanced analytical techniques). With its typically larger sample sizes, RWD may be leveraged to perform counterfactual comparisons between any paired treatment sequences, similar to those in the two-arm sequential treatment RCTs in Figure 4.2C.

4.5.2. Variants of treatment sequences and their relevance for HTA

Table 4.5 classifies different types of effect of treatment sequences: uniform sequence for the entire population, optimal uniform sequence for the entire population, and optimal individualised sequences. Each of these can involve either time-related static treatment strategy or dynamic treatment strategy (i.e., DTR). In the context of treatment sequences, a time-related static strategy involves switching treatments based on a predefined duration, whereas a DTR entails switching treatments in response to events that may occur at varying times for each patient.

Based on the TA review in Chapter 3, the most relevant and commonly used estimand of treatment sequence effectiveness for HTA is the average treatment effect of a uniform treatment sequence within a population (bottom of the second column in Table 4.5). This estimand does not necessarily yield an optimal population effect (i.e., maximum total quality-life adjusted life years). This is because, when treatment sequences exist in HTA, the appraised treatment is often placed in a sequence and compared against one or several pre-defined reference treatment sequences. However, none of these sequences may necessarily lead to an optimal overall population effect but instead represent the most commonly used treatment sequences in clinical practice.

On the other hand, research focusing on computation optimisation of treatment-sequencing models, such as Tosh's doctoral thesis, aimed to identify the optimal uniform treatment sequence for the entire population (see Section 2.4.3.7, Chapter 2), aligning with the concept of optimal sequences.[30] Further, Kim's doctoral thesis focused optimising decision-making models for sequential treatments given patient characteristics, reflecting the concept of optimal individualised treatment sequences, where each patient receives a specific treatment sequence based on their unique characteristics.[31] While both cases focus on optimisation of cost-effectiveness rather than effectiveness, their concepts are analogous to optimal treatment sequences (i.e., the right two columns in Table 4.5). Identifying the most cost-effective or most effective treatment sequence is currently not the primary focus of HTA. Some TAs model a mix of subsequent treatments, but these usually designed to reflect market share rather than being optimised for individual patients.

All simple methods identified compared the effects of two or more uniform treatment sequences for the entire population. Conversely, non-simple methods not only addressed these comparisons—most relevant to HTA—but also identified the optimal uniform sequence and explored treatment strategies with optimised individual sequences (see the right two columns in Table 4.5). Although not directly relevant, insights from these studies were valuable and retained in the review (see Section 4.6.2), though they received less focus in my narrative synthesis. Most included studies focused on comparing sequences that are dynamic rather than time-related static strategies (see the bottom two rows of Table 4.5). This aligns with Chapter 3's findings, where most treatment sequences compared in TAs are dynamic treatment strategies. However, insights from both types of studies are transferable.

Table 4.5 Types of treatment sequences

Type of treatment strategy	Type of treatment sequence		
	Uniform treatment sequence for the entire population	**Optimal uniform treatment sequence for the entire population**	**Optimising treatment sequences for each individual in the population**
	These are sometimes referred to a subtype of "static treatment sequence" in the literature (see the cell below), as all patients within the same population receive the same treatment sequence.		This is often referred to as "individualised treatment strategy" or "individualised DTR" in the literature or "dynamic treatment sequence"
Static strategies	Not applicable in the treatment sequence context; typically refers to one-time treatment or sustained single treatment. Example: o One-time flu vaccine o Always treated with treatment A versus treatment B.		
Time-related static strategies	Switching to a subsequent treatment in a sequence depends only on the pre-defined duration of the previous treatment.		
	Example: In the BIG 1-98 breast cancer trial[321], four treatment sequences compared were: o Letrozole for 2 years → letrozole for 3 years o Tamoxifen for 2 years → tamoxifen for 3 years o Letrozole for 2 years → tamoxifen for 3 years o Tamoxifen for 2 years → letrozole for 3 years	Theoretical example: Identifying a uniform optimal sequence for the entire population from the left cell's example, such as letrozole for 2 years → letrozole for 3 years, which leads to longer overall survival.	Theoretical example: Identifying an optimal treatment sequence tailored to individual patient characteristics at initiation and predefined switching points. For example: o Letrozole for 2 years → letrozole for 3 years could be optimal for patients who are at high risk of early recurrence o Letrozole for 2 years → tamoxifen for 3 years could be optimal for patients who cannot tolerate long-term letrozole and might benefit from different mechanisms of treatment The optimal individualised treatment strategy then involves starting on letrozole, with a subsequent switch to a treatment suited to each patients' needs at a predefined point. Note: Individualised treatments require considering patient characteristics, making them technically DTR, although the timing of treatment switches is only static and time-related rather than event-dependent.
Dynamic treatment strategies	Switching to a subsequent treatment in a sequence based on an event that may occur at variable times for each patient		
	Theoretical example: Comparing the population causal effect of treatment sequences A→ A, A→ B, B→ B, B→A. This strategy is dynamic, as subsequent treatments are administered based on time-varying criteria such as disease progression (e.g., patients receiving only A without disease progression are considered to have followed the dynamic strategy of A→B). For simplicity, a dynamic treatment strategy involving a uniform treatment sequence for the entire population is referred as the average treatment effect of a uniform treatment sequence within a population.	Theoretical example: Identifying a uniform optimal sequence for the entire population from the left cell's example, such as A→ B and estimating its population effect.	Theoretical example: Estimating the population effect of a DTR where each individual's treatment sequence is optimised based on specific triggers. For instance, the treatment sequence A → B might be optimal for patients achieving complete remission after first-line treatment A, while A → C could suit those showing disease progression. Thus, the optimal DTR involves individualised treatment, with A followed by either B or C, depending on patient characteristics before initiating the second-line treatment.

DTR: dynamic treatment regimen

4.5.3. Methods identified to compare treatment sequences

Table 4.6 and Table 4.7 provides a taxonomy of methods with the potential to evaluate the effectiveness of treatment sequences from RWD, specifically within the context of time-to-event outcomes. This review focused exclusively on methods applicable to time-to-event outcomes due to the emphasis on survival outcomes in my case studies (see Chapters 7-8). This categorisation aims to improve the understanding of each method's theoretical principles and their relationship with other methods. Additionally, these tables specify the types of data and instances (e.g., specific RCT or RWD studies) where the methods have been applied. Feedback from supervisors was applied to refine this taxonomy. It is important to acknowledge that although this taxonomy primarily adapts terminology from reviewed papers and studies citing them, it also reflects my interpretation of how each study relates. The proposed taxonomy broadly categorises methods into five groups.

(1) Simple methods (Table 4.6): Some of these may provide unbiased estimates for analysing RCTs that randomise patients to different treatment sequences (i.e., ITT analysis when patients minimally deviate from their assigned treatment sequences). However, they often fail to adjust for time-varying confounding in RWD settings (and in RCTs that involve non-randomised subsequent treatments) and may introduce additional biases such as immortal time bias by only including patients who survive to receive subsequent treatment. While not the main focus of this review, understanding the applicability of simple methods in specific RCT settings is crucial.

(2) Generalised methods (g-methods) (Table 4.6): Originating from Robins's 1986 paper[54], these methods were developed to adjust for time-varying confounders and treatments when analysing RWD to answer causal questions. These methods harnessed the Neyman-Rubin's counterfactual outcomes framework[50], which was originally developed to assess causal effects of time-fixed treatments in RCT and RWD scenarios, respectively (see Section 1.4.1, Chapter 1 for introduction of causal inference). Not surprisingly, all methods identified in this review were either adaptations of g-methods, or closely related to them. These include adaption of marginal structural models (which is frequently used with inverse probability weighting)[58,68,69] and g-estimation of structural nested failure time models (SNFTMs).[66,67] While most studies in this category examined the effect of a uniform treatment sequence across the entire population, one study utilised adaptations of SNFTM to identify an optimal uniform sequence for the population.[290]

(3) G-methods for identifying optimal individualised DTR (Table 4.6): These are an extended application of g-methods incorporating individual characteristics into identifying optimal treatment strategies, allowing for tailored decisions during treatment switches rather than a fixed subsequent treatment for the entire population. Methods in this category were used to identify optimal treatment strategies that could involve different treatment sequences for each individual

in the population. In addition to the g-methods mentioned in (2), adaption of g-formula (also known as, parametric g-formula, g-computation, or (g-)standardisation))[54] were also identified in this category.

(4) Q-learning (Table 4.6): This is a distinct category featuring methodology originating from reinforcement learning. However, the model structures in all identified Q-learning studies borrowed concepts from, or paralleled, SNFTM. All identified Q-learning studies were used to determine optimal individualized DTRs.

(5) Methods for analysing SMARTs (Table 4.7): According to Robins, there are parallels between SMART designs and RWD. The review confirmed that all methods for analysing SMARTs are derived to some extent from g-methods, yet they exhibit more variation in terms of estimations compared to other identified methods, including g-formula, IPW, and Q-learning.

Table 4.6 Methods used in analysing RCT and RWD

Method Groups	Method subcategory	Method/Extension	References	Case study data details	Data used for demonstration	
					RCT	RWD
Methods for comparing uniform treatment sequences for the entire population						
Simple methods	Simple methods for analysing sequential RCTs	ITT	Recht 1996[316] AGIS Investigators 2002[295] Derks 2017[324]	Each study conducted its own sequential treatment RCT	✓ (appl. only)	-
			The BIG 1-98 Collaborative Group 2009[321] Regan 2011[173] Chirgwin 2016[328]	The BIG 1-98 breast cancer sequential treatment RCT		
			EBCTCG 2015[314] Benchalal 2005[296]	Pooled data of breast cancer sequential treatment RCTs		
		ITT, with adjustment of variables at randomisation due to pooling data from various sequential treatment RCTs	Rabaglio 2021[317]	The BIG 1-98 breast cancer sequential treatment RCT	✓ (appl. only)	-
		PP	(None)	-	-	-
	Simple methods for analysing RWD	AT, limited to patients who received all treatment lines	Lu 2022[315] Jonasch 2014[318] Stenner 2012[322] Herrmann 2011[323] Dudek 2009[326] Ishihara 2018[327] Li 2014[329] Paglino 2013[330]	Each study conducted its own RWD study, restricted to patients who underwent the entire sequence of treatment lines	-	✓ (appl. only)
		AT with multivariate adjustment, limited to patients who received all treatment lines	Busch 2011[297] Busch 2013[298] Iacovelli 2013[319] Giuliani 2012[320] Buchler 2012[325]		-	✓ (appl. only)
		AT with time-dependent model with multivariate adjustment, limited to patients who received all treatment lines	Van Leeuwen 1997[294] Kumada 2021[299]		-	✓ (appl. only)

Method Groups	Method subcategory	Method/Extension	References	Case study data details	Data used for demonstration	
					RCT	RWD
G-methods	Inverse probability weighting (IPW)	IPTW*, or IPCW	Huang 2006[289]	Simulation + Soft tissue sarcoma RCT: Cormier 2004 cohort analysis[338]	-	✓
			Nishino 2013[332]	RWD study of 335 Japanese lung cancer patients[332]	-	✓ (application only)
			The BIG 1-98 Collaborative Group 2009[321]	The BIG 1-98 sequential treatment RCT[173]	✓ (appl. only)	-
		(Clone)-censor-weighting	Hernán 2018[333]	A toy example comparing the effect of aspirin treatment durations	-	✓
	Adaption of SNFTMs without g-estimation	Joint AFT models with backward induction with adaption to Q-learning*	Huang & Ning 2012[290]	Simulation + Soft tissue sarcoma: Cormier 2004 RWD study[338]	-	✓
Methods for identifying optimal DTR involving multiple treatment sequences within a population						
G-methods	Inverse probability weighting	Inverse probability of a treatment regimen†	Wahed & Thall 2013[287]: reference method Xu 2016[331]: reference method (IPTW & augmented IPTW)	Simulation + Leukaemia: Estey 1999 4-arm RCT[339] (non-randomised salvage treatment)	✓	-
	G-formula adaption	Likelihood-based (parametric) approach with AFT models	Wahed &Thall 2013[287]: main investigated method	Simulation + Leukaemia Estey 1999 4-arm RCT[339] (non-randomised salvage treatment)	✓	-
		Bayesian nonparametric models: dependent Derichilet process prior and a Gaussian process base measure (DDP-GP model)	Xu 2016[331]: main investigated method		✓	-
	G-estimation	G-estimation under optimal structural nested models for optimal sequential decisions	London 2010[286] (Robins 2004[340], Murphy 2003[341], Brumback 2008[342])	Children neuroblastoma: P9462 phase II RCT[286]	✓ (application only)	-
Backward induction methods	Q-learning	Q-learning adaption in a singly robust framework	Huang 2014[288]	Leukemia: Estey 1999 RCT[339] (non-randomised subsequent treatments)	✓	-
		Q-learning	Krakow 2017[291]	GVHD treatment in Leukemia RWD study using US CIBMTR registry[291]	-	✓
		Q-learning: dynamic weighted survival modelling (DWSurv) with doubly robust	Simoneau 2020[275] (borrows from Huang 2014[288] singly robust framework)	Diabetes RWD study using UK CPRD[275]	-	✓

AFT: accelerated failure time; AGIS: The Advanced Glaucoma Intervention Study; AT: as-treated; appl. only: application study only; CIBMTR: Center for International Blood and Marrow Transplant Research; CPRD: Clinical Practice Research Datalink; DTR: dynamic treatment regimens; SSEBCTCG: The Early Breast Cancer Trialists' Collaborative Group; GVHD: Graft-versus-Host Disease; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; ITT: intention-to-treat; PP: per-protocol; RCT: randomised controlled trial; RWD: real-world data; SNFTM: structural nested failure time model; UK: United Kingdom; US: Untied States

* Inverse probability of censoring weight was used in addition to tackle with right-censoring due to lost-to-follow-up

† Examples are all in RCTs, so the method only seem to adjust for non-randomised subsequent treatments in RCT

Table 4.7. Methods used in analysing SMART

Method Groups	Method subcategory	Method	Papers	Case study data details
G-methods	Inverse probability weighting based methods	Marginal mean model-based (MM) estimator (Lunceford, Davidian and Tsiatis (LDT) estimator)	Lunceford 2002[178] (based on Robins 1994[343])	Leukemia: GALGB 8923 data[344]
			Ruppert 2009[309] (only method application study)	Simulated SMART
			Vilakati 2021[305] (only method application study)	Leukemia: GALGB 19808[345]
		Wahed & Tsiatis (WT) estimator	Wahed & Tsiatis 2004[311] (based on semiparametric therapy of Robins 1994[340])	Simulation & Leukemia: GALGB 8923 data[344]
		Weighted risk set estimator (WRSE) (weighted log-rank test)	Guo & Tsiatis 2005[301] (variation of Lunceford 2002 [178] with time-dependent IPW)	Leukemia: GALGB 8923 data[344]
			Vilakati 2021[305] (only method application study)	Leukemia: GALGB 19808[345]
		Supremum weighted log-rank test	Feng & Wahed 2008[300] (an extension of Guo & Tsiatis 2005[301])	Leukemia: GALGB 8923 data[344]
		Weighted log-rank statistic to compare strategies with shared-path that have same initial treatment	Kidwell 2013[312] (an extension of Guo & Tsiatis 2005[301])	Simulated SMART
		Weighted KM estimators (WKM), Weighted KM with dynamic weights (WKM with time-dependent weights)	Miyahara & Wahed 2010[292] (an extension of Lunceford 2002[178]) Vilakati 2021[305] (only method application study)	Leukemia: GALGB 8923 data[344]
		Stratified proportional Cox hazards model to estimate the cumulative hazard ratio estimation for treatment regimes in SMART	Tang & Wahed 2015[304]	Neuroblastoma study 1991-1996 (Matthay 2009[346])
		Cumulative incidence function of DTR for comparing risks	Yavuz 2018[303]	Neuroblastoma study 1991-1996 (Matthay 2009[346])
		Cumulative incidence regression for DTR	Chen 2020[302] (compute CIF using different models than Yavuz 2018[303])	Neuroblastoma study 1991-1996 (Matthay 2009[346])
		IPW for viable DTRs in SMART: IPW is used to evaluate DTR effects while dropouts due to adverse drug events such as toxicities occur. It refines viable drug switch rules, allowing for patients who develop toxicity or progressive disease to switch to non-pre-specified subsequent treatments	Wang 2012[123]	Prostate cancer 1998-2006 (those who randomised poorly were re-randomised)
	G-formula adaption	Joint modelling and multiple comparison with the best of data from SMART survival data (MLE method)	Chao 2022[293] (an adaption of Wahed & Thall 2013 analysing RCT data[287])	Simulation & Leukemia: GALGB 8923 data[344]
Backward induction method	Q-learning	Clinical reinforcement trial with Q-learning	Zhao 2011[308]	Simulated SMART in NSCLC (a virtual clinical reinforcement trial)
		Q-learning with censored data	Goldberg 2012[306]	Simulated three-stage SMART
Others with ambiguous category	Others	Bayesian framework for selecting best strategies	Thall 2007[313]	Metastatic renal cancer trial
		Maximum-likelihood based method: parametric models using a mixture of known probability distributions to model the survival distributions under a specific treatment strategy	Wahed 2010[310] (MLE equations are similar to Wahed & Tsiatis 2004[311], but used with a completely different perspective of IPW)	Leukemia: GALGB 8923 data[344]
		Generalised Cox proportional hazards model that not only applies to comparisons of any combination, of any number of treatment regimes, but also allows the intermediate response to appear as a time-varying covariate	Tang 2011[307]	Neuroblastoma study 1991-1996 (Matthay 2009[346])

DTR: dynamic treatment regimen; IPW: inverse probability weighting; MLE: maximum likelihood estimation; SMART: Sequential Multiple Assignment Randomised Trial

4.6. Narrative synthesis of identified methods

4.6.1. Methods for comparing uniform treatment sequences using sequential treatment RCT and RWD

The methods discussed in the current section (Section 4.6.1) are more readily applicable for HTA compared to those in Section 4.6.2, as they aim to estimate the effects of uniform treatment sequences rather than individualised treatment sequences, as described in the taxonomy Section 4.5.2. This section not only summarises the mechanisms of advanced methods but also outlines those of simpler methods to highlight their common pitfalls. I detail the origins and theoretical characteristics of each method, as well as their applications in the studies reviewed.

4.6.1.1. Simple methods

4.6.1.1.1. ITT in sequential treatment RCTs

- Origin of the method

ITT analysis in sequential treatment randomised controlled trials (RCTs) extends the conventional ITT, which involves analysing RCTs that randomise individuals to single-time treatments[347,348], to analyse RCTs that randomise patients across different treatment sequences. This method does not have a single origin but rather extends ITT principles to analyse RCTs with more complex trial designs.

- Theoretical characteristics

ITT analysis directly compares active treatments and their comparators across all randomised patients in RCTs based on patients' initial group assignments, disregarding the actual treatments received or any post-randomisation events, such as withdrawal from the study or protocol violation. ITT provides an unbiased estimate of the causal effects of treatment strategies followed by each randomised treatment group (as defined in the study protocol), given proper randomisation (i.e., no unmeasured confounders). However, in cases of protocol non-adherence, the estimand addressed by the ITT may not necessarily correspond to the one specified in the study protocol, due to confounding that occurs post-randomisation

In sequential treatment RCTs, ITT analysis assesses the average causal effect of pre-defined treatment sequences, also given perfect adherence to the protocol. To assess the comparative effects of different pairs of treatment sequences, one would theoretically need to conduct a separate RCT for each pair and apply ITT analysis to each, or conduct a single RCT with three or more sequences and analyse them taking into account multi-arm comparisons.

ITT analysis can be compromised by (unintended) post-randomisation treatment switches[43,349], for example in oncology trials where the placebo group may switch to active drugs following disease progression.[43] Conceptually, this introduces an unintended treatment sequence (placebo → active drug), undermining the ability of ITT analysis to accurately estimate the effects of active drugs versus placebo (i.e., a sequence of active drug → active drug versus placebo → placebo). Similarly, in sequential treatment RCTs, non-adherence to the allocated treatment sequence can distort the ITT analysis, leading to inaccuracies in comparing the causal effects of predefined treatment sequences. ITT principles may be applied to RWD analysis to create an ITT Analogue, which requires advanced statistical methods, as specified in Sections 4.6.1.2 and 4.6.2. Theoretically, such an analogue can reflect the causal effects of treatment sequences when emulating a sequential treatment RCT without protocol violations.

- Application in treatment sequences

I identified nine application studies that carried out sequential RCTs, randomising patients to different treatment sequences and employing ITT analysis to determine their causal effects.[173,296,314,316,317,321,324,328] While one study examined advanced glaucoma treatment sequences[295], the rest focused on survival outcomes of early-stage breast cancer treatments.[173,296,314,316,317,321,324,328] The majority of these studies (n = 5) concentrated on the BIG 1-98 study[173,314,317,321,328], which included 8,010 women across a two-arm RCT (comparing letrozole (n = 917) to tamoxifen (n = 911)) and a 4-arm RCT (comparing letrozole (n = 1546), letrozole → tamoxifen (n = 1540), tamoxifen (n = 1548), and tamoxifen → letrozole (n = 1548)). Among these, three employed the ITT Cox regression model to compare HRs for different treatment sequences[173,321,328], drawing on data from the 4-arm RCT in BIG 1-98, assessing the effect of letrozole alone against treatment sequences involving both letrozole and tamoxifen. In their 2009 publication, the BIG 1-98 Collaborative Group reported adjusting their ITT analyses for multiple comparisons (4-arm).[321] Rabaglio et al. furthered this research by combining data from both the two-arm and four-arm RCTs of BIG 1-98, thus enlarging the cohort for the letrozole and tamoxifen alone groups.[317] This integration necessitated adjustments in the ITT Cox model for patient characteristics at randomisation due to data merging. Additionally, the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) aggregated data from nine sequential treatment RCTs, including those in BIG 1-98, for a patient-level meta-analysis on similar treatment sequences. They made adjustments to prevent double-counting events across RCTs but did not specify the methodology.[314]

4.6.1.1.2. PP in sequential treatment RCTs

- Origin of the method

The per-protocol (PP) analysis, similar to ITT, does not have a single origin. This mirrors the "PP censoring—censoring switchers (PPcen)" method described in Latimer et al.'s 2017 study for addressing treatment-switching in RCTs.[350] It diverges from ITT by utilising data only from patient-periods that comply with the study protocol, excluding periods after protocol deviation. In cases of complete protocol compliance, PP analyses should be equal to those of ITT.

- Theoretical characteristics

In RCTs, PPcen (hereafter: PP) analysis benefits from the initial randomisation, and thus does not require any adjustments based on baseline patient characteristics. It operates by censoring patient records following any protocol deviation in the outcome model. It aims to reflect the effects of the intervention unaffected by protocol deviations or non-adherence, which is crucial when protocol violations might significantly affect results inferred from ITT analyses. Nonetheless, PP analysis also carries the risk of introducing bias, especially if the censoring reflects underlying characteristics that are not randomly distributed (i.e., informative censoring) and adjustments are not made accordingly (e.g., using advanced statistical methods, such as inverse probability of censoring weighting (IPCW) in Section 4.6.1.2.1).[43,349]

In the context of sequential treatment RCTs, PP involves censoring patients once they deviate from their assigned treatment sequence, such as receiving an unplanned subsequent treatment, potentially introducing bias without proper adjustments. Applying PP principles to RWD analysis creates PP Analogues. These require the support of statistical methods, such as inverse probability treatment weighting (IPTW) (Section 4.6.1.2.1) to correct biases due to non-randomised baseline treatments in RWD (i.e., baseline confounding). To address remaining biases caused by informative censoring in PP Analogues, advanced methods such as inverse probability of censoring weighting (IPCW) are necessary. These techniques facilitate the accurate estimation of the hypothetical PP effect without treatment deviations, as detailed in Section 4.6.1.2.1.

- Application in treatment sequences

No PPcen analysis cases were identified in my review. Regan et al.[173], however, applied the "hypothetical perfect PP effect" technique in their Cox model analysis of the four-arm sequential treatment RCT survival outcomes in BIG 1-98[173], with IPCW to correct for selective crossover to letrozole in the tamoxifen-only arm estimating the hypothetical PP effect as if no crossover occurred. More details of IPCW can be found in section 4.6.1.2.1.

4.6.1.1.3. AT in RWD sequential treatment studies

- Origin of the method

As-treated (AT) analysis, lacking a specific origin, have been applied in both RCT and RWD analyses.[351-354] In RCTs, it involves comparing patients based on the treatments they actually receive, rather than the treatments to which they were initially assigned. This method excludes or reclassifies patients who deviate from the protocol, potentially introducing bias in estimating the causal effect of treatments by compromising the randomisation—especially if deviations are not randomly distributed across groups.[353-355] This resembles the simple "PP—excluding switchers (PPexc)" method for addressing treatment-switching in RCTs.[350] In real-world studies, the AT analysis involves selecting patients based on the specific treatments they have actually received and comparing their outcomes without adjusting for confounding variables. This is inherently biased in assessing causal comparative effects because treatments in the real world are not randomly assigned.[352] In the context of assessing treatment sequences using RWD, AT analysis is defined as comparing patients who have undergone the entire treatment sequences of interest.

- Theoretical characteristics

In evaluating treatment sequences (e.g., A→B vs. B→A) using RWD, AT analysis is prone to confounding bias due to non-randomised treatments and, more critically, selection bias and immortal time bias.[38,76,356] The latter emerges because AT analysis only considers patients who survive to receive specific subsequent therapies, likely overestimating the absolute effectiveness of any sequence (i.e., the effect of treatment sequences on the "immortal" patients). Specifically, it overlooks those who might not survive after the initial treatment.

Furthermore, including patients who received treatment A but did not survive to receive subsequent treatment in the A→B group cannot resolve estimation inaccuracies of the A→B sequence's effect. Specifically, this approach fails to account for patients who might proceed to treatments other than B after A (e.g., A→C, A→D, A→E) in real world contexts, leading to an overrepresentation of patients not surviving to further treatments. This introduces a selection bias by enrolling patient based on post-treatment eligibility[357], contravening the TTE framework's core principle of adhering to baseline eligibility for patient inclusion.[3] Traditional outcome modelling approaches (Section 4.2.1), which adjust for baseline confounders alone or alongside time-varying confounders in a regression model cannot address the issues inherent in AT analysis. This limitation stems not only from the inadequacy of conventional adjustment methods to handle time-varying confounding, but also largely due to their inability to rectify immortal time bias through mere adjustment. Importantly, even advanced methods discussed later may not fully resolve this issue,

which can only be effectively countered together with careful RWD study design, supplemented by necessary advanced statistical methods. Carefully considering the risks of such biases in RWD study is crucial, particularly in cases where analyses limited to patients receiving second-line treatments omit those who died naturally, failing to reflect the joint effect of the initial treatment within the treatment sequence strategy. As advanced statistical methods cannot rectify this issue; naturally, neither can simple outcome regressions that simply incorporate time-varying confounders, which are inherently flawed in answering causal questions (see the example of simple outcome regression with linear model involving time-varying covariates in Section 1.4.1, Chapter 1, which also extend to explain the flaws in simple time-varying Cox models that simply incorporate time-varying covariates in the model).

- Application in treatment sequences

A significant proportion of retrospective RWD application studies (27%, n = 15) focusing on treatment sequence comparisons, exclusively including patients who completed the specified sequences were identified. Eight of these studies performed AT analyses without any adjustments, including a study comparing letrozole alone to letrozole in sequence with tamoxifen for breast cancer[315], and others exploring various sequences involving vascular endothelial growth factor inhibitors (VEGFis) and mammalian target of rapamycin inhibitors (mTORis) in advanced/metastatic renal cell carcinoma (mRCC), including the following comparisons:

- VEGFi → mTORi, VEGFi → VEGFi, mTORi → VEGFi, and mTORi → mTORi[318]
- Sunitinib → sorafenib versus sorafenib → sunitinib[322,323,326]
- Sunitinib → everolimus versus sunitinib → axitinib[329]
- Sunitinib → mTORi → sorafenib versus sorafenib → mTORi → sunitinib[330]
- Treatment sequences with and without third-line therapy[327]

These studies, while insightful regarding real-world treatment patterns, are limited by their non-randomised design and immortal-time bias for causal interpretation. Although these studies suggest an association between longer treatment durations, improved survival outcomes (e.g., OS), and specific sequences, they do not assert strong causal relationships. Notably, Ishihara et al.[327] compared sequences with and without third-line therapy, inherently favouring better survival in the third-line group as these patients survived long enough to receive it. This mirrors the fairness issues in comparing sequences with different number of LOTs in health economic evaluation discussed in Chapter 3 (Table 3.10, Challenge F18).

The remaining seven studies employed multivariate-adjusted Cox models to estimate OS across different treatment sequences,[294,297-299,319,320,325] with two explicitly incorporating time-dependent covariates (i.e., additive or sequential nucleoside analogue therapy versus continued zidovudine

monotherapy in HIV[294]; and continuous entecavir monotherapy versus entecavir → tenofovir[299]). While the remaining five studies comparing sequences in mRCC did not specify the inclusion of time-dependent variables[297,298,319,320], some of the variables they included suggest otherwise (e.g., primary resistance to first-line treatment[297,298,319]). Despite these efforts to address confounding, challenges such as immortal time bias, and the bias from including time-varying covariates that may act as intermediate outcomes in the outcome model remain (i.e., Figure 1.2 in Chapter 1, where conditioning on time-varying covariate L_1 closes a front door path, but failing to adjust for it does not properly account for time-varying confounding).

4.6.1.2. Generalised methods (g-methods)

4.6.1.2.1. <u>Inverse probability weighting and marginal structural models (MSMs)</u>

- Origin of the method

To tackle missing data from non-random loss to follow-up in HIV clinical trials, Robins and Rotnitzky introduced the IPCW method.[358] Inspired by Jamie Robins' 1986 paper[54], the IPCW method interpreted censoring as a time-varying treatment. Robins acknowledged that adjusting for censoring, in reality, can be interpreted as an estimation of survival under the hypothetical (i.e., counterfactual) scenario where all subjects remain uncensored.[60] Hence, informally, IPCW simulates this scenario by upweighting the remaining uncensored participants based on their inverse probability of staying in the study, at each point when others were censored, compensating the missing records of those who were censored due to lost to follow-up (Figure 4.3).

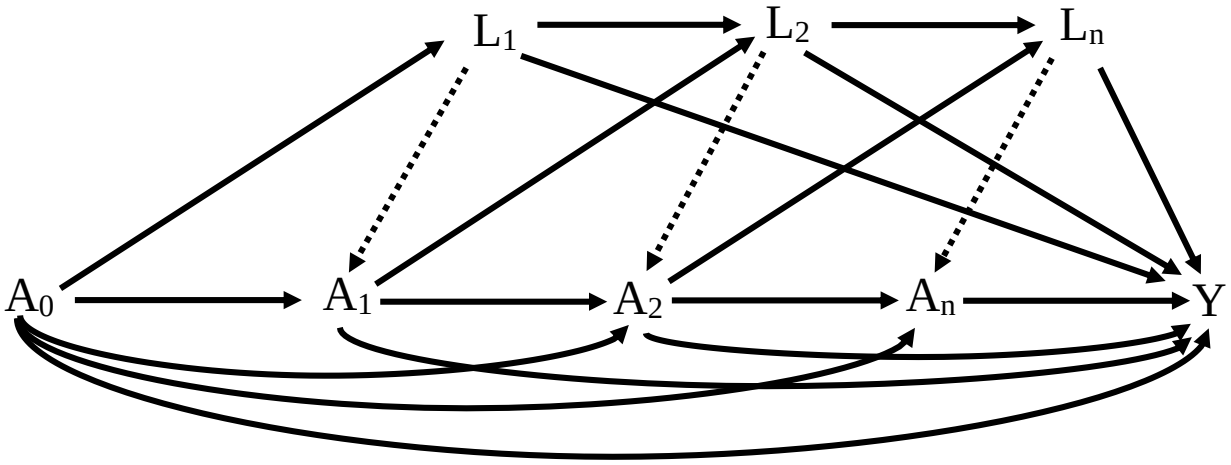


Figure 4.3 Directed acyclic graph (DAG) in RCTs with time-varying treatments and confounders

A_0: treatment at baseline (time 0); A_1, A_2, ..., A_n: treatments at times 1, 2, ..., n; L_1, L_2, ..., L_n: confounders at the start of times 1, 2, ..., n (e.g., performance status, weight, progression); Y: outcome (e.g., death). Dotted lines indicate the effect of time-varying covariates (L_x, where $x > 0$) on subsequent treatment decisions (A_x, where $x > 0$). In RCTs, time-varying treatments may result from protocol violations (e.g. receiving non-randomised subsequent treatment) or loss to follow-up, with censoring being considered a type of treatment change.

Robins highlighted IPCW's capability to assess the effect of any "treatment regimen" *g* by marking the moment a participant diverges from their treatment regimen as their censoring time, and termed it "inverse probability of treatment weighting (IPTW) " for such cases.[359,360] In its simplest application, IPTW can be used to compare the causal effects of two single-time treatments—A versus B (e.g., receiving or not receiving a flu shot)—in observational studies (Figure 1.1 in Chapter 1). The method begins by treating patients who received the opposite treatment as "censored" due to treatment deviation from the start, as though they deviated from the treatment in question. For example, to assess the marginal effectiveness of treatment A, patients who received B are treated as censored from the start (i.e., deviating from treatment A). Then, patients who received treatment A (i.e., the remaining uncensored ones) are weighted based on their inverse probability of receiving A (e.g., inverse probability of receiving a flu shot), given their baseline covariates L. Similarly, to assess the marginal effectiveness of treatment B, those who received A were treated as censored from the start, and recipients of B (i.e., the remaining uncensored ones) were weighted based on their inverse probability of receiving B (e.g., inverse probability of not receiving a flu shot), given their baseline covariates L. Informally, this approach removes the association between baseline confounders and treatment (the arrow of confounders to treatment in Figure 1.1 in Chapter 1), thereby creating a balanced pseudo-population for assessing the marginal (i.e., unconditioned on a specific patient subset) causal effect of treatments (A versus B) on outcome.

However, treatment regimens may vary widely, from sustained treatment strategies, such as "always treat" or "never treat", to more dynamic strategies (i.e., DTR) like "treat if below a certain lab threshold" or "continue treatment A until adverse events occur" (see Section 4.3.2 and Section 4.5.2 for terminology).[5,279] Treatment status in real-world longitudinal studies, measured at multiple discrete points A_k (Figures Figure 1.2 in Chapter 1 & Figure 4.4), often shows limited adherence to a single regimen among participants.[360] Robins, therefore, introduced Marginal Structural Models (MSM) to tackle the challenge of estimating effects from complex regimens and variable treatment regimen adherence.[58,68,69,361] Informally, patients continuing to adhere to a treatment regimen are assigned weights based on the inverse probability of remaining uncensored between one time interval (k) to the next (k+1), given covariates at the beginning of each interval k (e.g., secondary, thirdly, baselines L_1, L_2), treatment status in the previous interval (A_{k-1} for time > 0), and initial baseline covariates (L_0) (i.e., removing arrows from L_k to A_k, A_{k-1} to A_k, and L_0 to A_k in Figure 4.4). A patient's weight at time k is conditioned on their status of being uncensored at time k-1. Consequently, a patient's final weight in the study is the product of all inverse probabilities of remaining uncensored from time 0 to time k. This weighting method can be seen as either IPCW (uncensored vs. censored) or IPTW (adherence to treatment protocol vs. non-adherence), collectively referred to as IPW. Informally, IPW at the outset emulates the random assignment of treatment regimens, and for

subsequent periods, it emulates a hypothetical scenario where all patients in each treatment group consistently follow their assigned treatment regimens. In HTA, this technique has been broadly adopted to tackle unwanted treatment-switching in oncology RCTs, producing hypothetical scenarios where no control group members switch to the active treatment (i.e., commonly referred to as IPCW in HTA).[43,277] Patients in the control group are censored as soon as they diverge from their initial treatment plan, with the remaining uncensored members subsequently upweighted to adjust for the informative censoring.

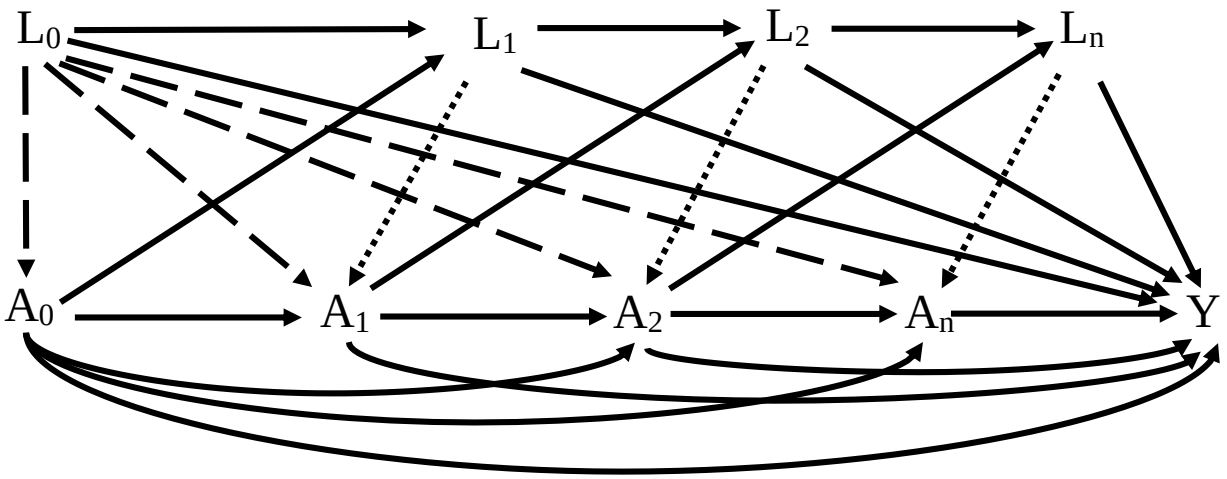


Figure 4.4 Directed acyclic graph (DAG) in observational studies with time-varying treatments and confounders

A_0: treatment at baseline (time 0); A_1, A_2, ..., A_n: treatments at times 1, 2, ..., n; L_1, L_2, ..., L_n: confounders at the start of times 1, 2, ..., n (e.g., performance status, weight, progression); Y: outcome (e.g., death). Dash lines represent the effect of baseline covariates (L_0) on baseline treatment (A_0) and subsequent treatments (A_x, where $x > 0$), while dotted lines indicate the effect of time-varying covariates (L_x, where $x > 0$) on subsequent treatments (A_x, where $x > 0$).

In practice, the IPW method can follow a k-stage framework with fixed intervals (e.g., daily, monthly), or be simplified to k-stage decision points for switching treatments within a regimen, such as decision to modify treatments upon disease progression.[289] The simplification assumes no time-varying confounding within each interval, but acknowledges the potential confounding between decision points and considering time spent in previous intervals as a possible confounding factor. This approach is reflective of the structure found in k-stage SMARTs. In Chapter 2, I discussed Robins' views on the similarities and differences between real-world treatments and SMARTs, noting that SMARTs feature pre-determined randomisation probabilities, whereas real-world scenarios deal with unknown probabilities that could potentially be estimated with sufficient data.[60] For illustration, consider a simple 2-stage SMART trial where patients are initially randomised to receive treatment A or B (1:1), followed by re-randomisation to either treatment (1:1) upon disease progression (i.e., a structure resembling the one in Figure 4.2B or a more complex SMART). The IPW technique can be employed to calculate the marginal effect of each treatment regimen (A→A, A→B, B→B, B→A)

(Note: a simple SMART, as shown in Figure 4.2A, may not require weighting). For instance, for the A→B sequence, patients receiving A→A are considered as diverging from A→B and are censored at progression, then weighted by the inverse probability of receiving B as a second-line treatment given covariates measured at the secondary baseline (i.e. disease progression), including the duration spent in the first-line treatment. Simply censoring without weighting, or limiting the analysis to a specific subset of patients (i.e., observed receiving A→B or who died without a second-line treatment) can introduce biases. Specifically, the first approach lead to informative censoring bias as seen in the PP approach (Section 4.6.1.1.2), while the second approach result in selection bias as noted in the AT analysis (Section 4.6.1.1.3). Informally, both biases result from inaccurately accounting for the proportion of patients/patient-record who survive long enough to receive a second-line treatment after starting with A. In real-world studies that mirror such setup, IPW adjustments—or methods that achieve the same effect—naturally become necessary at every decision point k, including the initial one, for estimating the marginal effects of a specific regimen as discussed earlier in this section (Section 4.6.1.2.1).

- Theoretical characteristics

The IPW method is used to evaluate the marginal effect of a (dynamic) treatment regimen (A) and an outcome (Y) as if confounding (i.e., back door) paths through measured variables (L_k) did not exist (Figure 4.4). It generates a pseudo-population where treatment at each time is independent of confounders by "removing" the influence of confounders on treatment ($L_k \rightarrow A_k$). In the context of adapting the method to estimate the marginal effect of treatment sequences on time-to-event outcomes using RWD, the method involves the following steps:

(1) Censoring observations at time k for individuals not adhering to the treatment sequence under study at time k and before time k+1;
(2) Developing a model to estimate each individual's probability of receiving the initial treatment of the treatment sequence under study (e.g., logistic model) at time 0, factoring in baseline covariates;
(3) Developing a model to estimate the probability of an individual patient being censored (i.e., not adhering to the treatment sequence under study) at each time interval k, accounting for both time-varying covariates and baseline covariates. Though not explicitly mentioned in literature, the censoring probability model can be specific to each treatment sequence, recognising that reasons for non-adherence may vary[282];
(4) Computing the cumulative probability of adhering to a specific treatment sequence up to a certain time t, which involves multiplying the probability of receiving the initial treatment, as calculated in step (2), by the product of the probabilities of remaining uncensored (one minus the probability of being censored) at each subsequent time interval, as calculated in step (3), for each individual.

(5) Conducting a weighted analysis for time-to-event outcomes (e.g., survival analysis using Cox model or pooled logistic regression) using the inverse probabilities of treatment sequence adherence up to each time t for each individual, derived from step (4), to estimate the treatment sequence's marginal effect.

Within this approach, two essential types of weights can be employed: non-stabilised (W) and stabilised weights (SW). Non-stabilised weights adjust for confounding by inversely weighting individuals according to their probability of receiving treatment (estimated in Step (2): W^T, where the superscript T denotes weights for treatment) or remaining uncensored owing to no treatment strategy deviation (estimated in Step (3): W^D, where the superscript D denotes weights for treatment deviation, or more precisely, weights for remaining non-deviated from the assigned treatment) given covariates (i.e., Formula 4.1 and 4.3 for IPTW and IPCW, respectively). To refine these weights, both Step 2 and Step 3 weights have stabilised counter parts: SW^T and SW^D, respectively. SW^T is calculated by multiplying the W^T by the marginal probability of receiving treatment (i.e., the numerator in Formula 4.2). SW^D is calculated by multiplying the W^D by the marginal probability of remaining uncensored due to no treatment strategy deviation (i.e., the numerator in Formula 4.4). The stabilised weights reduce the risk of extreme weights and thereby decreasing the variance of weighted estimates.

The general formula for weights derived from probabilities estimated in Step (2) are as shown in Formulas 4.1 and 4.2. These weights apply to balance the initial baseline treatment allocation:

$$W^T = \begin{cases} \frac{1}{\Pr[A_0 = 1 \mid L_0]} & for\ individuals\ receiving\ A_0 = 1 \\ \frac{1}{1 - \Pr[A_0 = 1 \mid L_0]} & for\ individuals\ receiving\ A_0 = 0 \end{cases}, \text{where } A \in \{0, 1\} \quad (4.1)$$

$$SW^T = \begin{cases} \frac{\Pr[A_0 = 1]}{\Pr[A_0 = 1 \mid L_0]} & for\ individuals\ receiving\ A_0 = 1 \\ \frac{1 - \Pr[A_0 = 1]}{1 - \Pr[A_0 = 1 \mid L_0]} & for\ individuals\ receiving\ A_0 = 0 \end{cases}, \text{where } A \in \{0, 1\} \quad (4.2)$$

For individuals initially receiving treatment $A_0 = 1$, the non-stabilised weight W^T is derived from the inverse probability of receiving $A_0 = 1$ given baseline covariates. For those initially treated with $A_0 = 0$, W^T is based on the inverse probability of not receiving treatment, calculated as one minus the probability of receiving treatment $A_0 = 1$, given baseline characteristics. The numerator of SW^T is the baseline probability of receiving treatment, regardless of baseline covariates. In more complex scenarios, treatment regimens may involve repeated on-and-off treatment A based on lab results. This differs from simpler cases where treatment is consistently applied or withheld until an irreversible event, such as disease progression (which is more likely to be the case of treatment-sequencing). In those more complex scenarios, the calculation of W^T and SW^T at time t may be extended to the cumulative inverse probability of receiving A = 0 or 1 at each time k from 0 to t. To evaluate (treatment) effect modification (in the context of baseline treatment allocation weights), stabilised weights can

use the probability of receiving treatment A based on baseline characteristics as the numerator, instead of the marginal probability.[5] In this case, the outcome model should include both the treatment and baseline covariates, where baseline covariates can be seen as effect modifiers.

The general formula for weights derived from probabilities estimated in Step (3) are as Formula 4.3 and Formula 4.4. These weights are used to address informative censoring resulting from treatment strategy deviation, hence are designated with a "D" to signify deviation (as opposed to administrative censoring due to loss to follow-up). The notation D_k indicates whether patients deviated from a treatment strategy during a specific interval k. For example, when $D_1 = 1$, it indicates that patients deviated from the treatment strategy during interval 1. In these formulas, t refers to the non-stabilised (W_t^D) and stabilised weights (SW_t^D) for a patient at a specific interval t. To calculate these weights, the cumulative probabilities up to time t in the numerator and denominator of the formula need to be computed, with k denoting a vector for each time point from 0 to t.

$$W_t^D = \begin{cases} \prod_{k=0}^{t} \frac{1}{\Pr[D_k = 0 \mid A_{k-1}, L_k]} = \prod_{k=0}^{t} \frac{1}{1 - \Pr[D_k = 1 \mid A_{k-1}, L_k]} & for\ D_k = 0 \\ 0 & for\ D_k = 1 \end{cases} \quad (4.3)$$

$$SW_t^D = \begin{cases} \prod_{k=0}^{t} \frac{\Pr[D_k = 0 \mid A_{k-1}]}{\Pr[D_k = 0 \mid A_{k-1}, L_k]} = \prod_{k=0}^{t} \frac{1 - \Pr[D_k = 1 \mid A_{k-1}]}{1 - \Pr[D_k = 1 \mid A_{k-1}, L_k]} & for\ D_k = 0 \\ 0 & for\ D_k = 1 \end{cases} \quad (4.4)$$

For individuals who remain uncensored (not artificially censored due to remaining not deviating from the treatment regimen, $D_k = 0$) up until time t, their non-stabilised weight W_t^D is the product of the inverse probabilities of not being censored across each time point k from 0 to t. This calculation follows Formula 4.3, where $\Pr[D_k = 0 | A_{k-1}, L_k]$ represents the probability of staying uncensored, given prior treatment and covariates at the beginning of time k. Essentially, it is the complement of the probability of being censored ($D_k = 1$), given prior treatment and covariates at the beginning of time k (i.e., $1 - \Pr[D_k = 1 | A_{k-1}, L_k]$). If an individual becomes censored in the interval from k to k + 1, their weight for the interval starting at k is set to 0. However, their information up to the point of censoring still contributes to estimating the relevant probabilities. The stabilised censoring weight SW_t^D follows similar logic to that of the stabilised initial treatment weight SW^T.

Unlike Step (2), which typically uses the entire population to calculate probabilities, the model for deriving the probability of censoring in Step (3) might utilise subsets of the population if the treatment regimen violation mechanism varies among different groups. For instance, in comparing the effects of receiving no COVID vaccine versus a consecutive 3-dose COVID vaccine regimen, the reason for deviating from the first strategy (e.g., a patient in the no-vaccine group suddenly opting for the vaccine due to risk factors) may differ from the reason for deviating from the second strategy (e.g., stopping further vaccine doses after two, due to severe adverse events). In such cases, models for deriving W_t^D and SW_t^D can be built separately for each treatment group, rather than using a single

deviation model for the entire population. Furthermore, the mechanism may also depend on the specific treatment administered in the previous interval. In such cases, censoring probabilities can be tailored for subgroups based on the treatment received in the prior period (A_{k-1}) rather than their initial treatment assignment at baseline.

Formulas 5 and 6 outline the computation of the final overall weights for each time t, for non-stabilised ($W_t^{T,D}$) and stabilised ($SW_t^{T,D}$) weights in Step (4), which are then used in the outcome model in Step (5), respectively.

$$W_t^{T,D} = W^T \times W_t^D \quad (4.5)$$

$$SW_t^{T,D} = SW^T \times SW_t^D \quad (4.6)$$

An optional step involves multiply the final weight by the censoring weight due to lost-to-follow-up (W_t^C, SW_t^C, where the superscript C denotes administrative censoring) if deemed informative. Given the variations in deriving weights, the adaptation of such method to treatment sequencing is detailed in my case study implementation in Chapter 8, with adapted steps outlined in Figure 8.4. For clarity, I explain how the steps from this section (Steps (1)-(5) in Section 4.6.1.2.1) correspond to those in Chapter 8 (Steps I.1-4 and II.1-3 in Figure 8.4, as explained in Sections 8.3.1 and 8.3.3-8.3.4), despite being arranged slightly differently.

In practice, implementing the method begins with data cleaning, which is divided into two parts in Figure 8.4 of Chapter 8: Part I (Steps I.1-4) focuses on cleaning the data to derive the necessary variables for building the weight models, while Part II (Steps II.1-3) covers building the weight models, calculating the weights, and applying them in outcome analysis. Step (1) in this section corresponds to part of the data cleaning tasks outlined in Steps I.3 and I.4 of Figure 8.4 in Chapter 8. In real-world studies, data cleaning involves more than just determining the timing of censoring due to treatment deviation (hence the additional data cleaning steps in Part I of Figure 8.4 in Chapter 8). However, for the purpose of method demonstration and simplicity, Step (1) in this section focuses specifically on determining censoring timing related to treatment deviation. After data cleaning, the remaining steps of Figure 8.4 in Chapter 8 (Steps II.1-3) correspond to Steps (2)-(5) in this section. Specifically, Step II.1 of Figure 8.4 in Chapter 8 aligns with Steps (2) and (3) in this section (i.e., building weight derivation models), Step (4) maps to Step II.2 (i.e., calculating weights), and Step (5) aligns with Step II.3 (i.e., final outcome analysis using weights).

In the context of sequential treatments (a type of DTR), employing IPW to estimate the marginal effect of a specific treatment regimen requires making the following four fundamental assumptions as shown below[5]:

(1) Positivity (patient overlap): At every decision point throughout the treatment regimen, each individual must have a non-zero probability of receiving any of the potential treatments, given

their past treatment history and covariates. This ensures that both patients who continue following the treatment protocol (uncensored) and those who deviate from it (censored) share similar attributes, allowing uncensored patients to be effectively used as proxies for their censored counterparts in the analysis through weighting.

(2) No unmeasured confounders (sequential ignorability/conditional exchangeability): There should be no unmeasured confounders that affect both the treatment assignment at each decision point and the potential outcomes, given past treatment history and covariates. This assumption ensures the comparability of between treatment regimen groups after implementing IPW.

(3) Consistency: The potential outcomes for each individual are aligned with the observed outcomes along the actual treatment regimen followed. This implies that if an individual's observed treatment regimen coincides with a specific treatment regimen, then the observed outcome should match the outcome that would have been expected had they, hypothetically, been assigned to follow that treatment regimen from the start.

(4) Correct model specification: The models estimating treatment assignment probabilities at each decision point should reflect the actual mechanisms of treatment allocation, including relevant covariates, past treatment history, and their interactions. Given the high-dimensional nature of real-world data (due to sample size constraints, a saturated model that includes all covariates and their interactions is often impractical, thus necessitating an “unsaturated model”), subject knowledge is integral for developing valid models for estimating the probabilities of treatment assignments at each decision time point.

- Application in treatment sequences

In the sequential RCT BIG 1-98[173], researchers supplemented the ITT analysis with an IPCW-adjusted analysis to simulate a hypothetical perfect PP effect. This adjustment reweighted patients in the letrozole alone group who switched to tamoxifen, creating a scenario as if no switch had occurred. Viewed through the lens of treatment sequencing, the ITT analysis captures the combined effect of letrozole → letrozole and letrozole → tamoxifen sequences, whereas the IPCW analysis isolates the marginal effect of the letrozole → letrozole sequence. The randomisation of initial treatments eliminated the need for initial treatment weighting in this study.

Huang et al. revisited the Cormier 2004 observational study[338] with the IPW method with a simplified approach two-stage weighting to estimate counterfactual mean restricted lifetime over 10 years for each treatment sequences[289], instead of incorporating the weights into a Cox model. The Cormier 2004 study followed a group of soft tissue sarcoma patients who underwent different sequences of treatment: initially receiving chemotherapy (A) and then either continuing (AA) or stopping (AB), or initially not receiving chemotherapy (B) and then either starting (BA) or not starting

later (BB). They approached the analysis by simplifying the treatment decisions into a binary choice (receiving chemotherapy versus not receiving chemotherapy) at two points in time (initial treatment and salvage treatment), with the recognition that each patient could have different time intervals between these points. This method contrasts with uniform interval weighting (cumulative weight of each interval k form 0 to t) across all participants as previously described. These methods are fundamentally alike, with Huang et al.'s simplified approach distinguished by having additional implied assumptions that there are no time-varying confounders within each time interval and that treatment decisions coincide with the measurement of covariates at both baseline and at the point of treatment-switching (secondary baseline). It also assumes no gap between measuring covariates and initiating subsequent treatment.

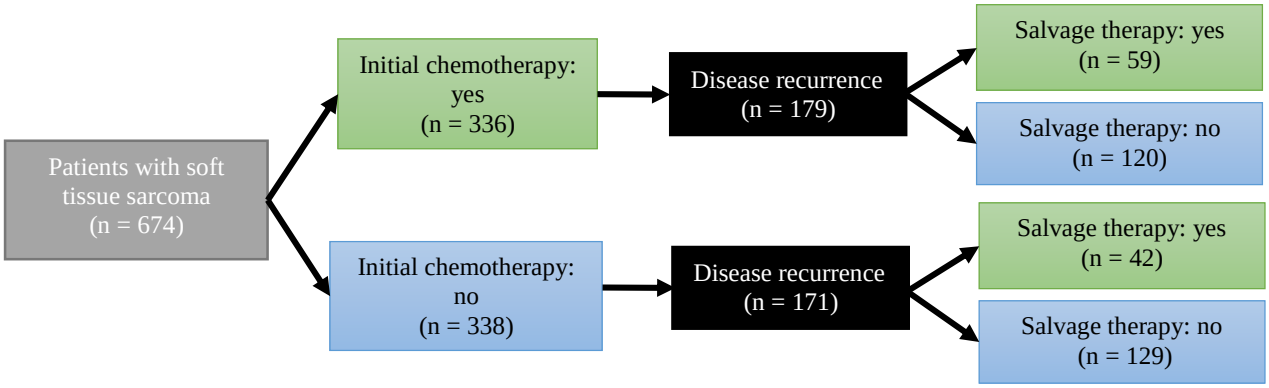


Figure 4.5 Structure of the Cormier 2004[338] observational study

The study by Huang et al. employed the product of two key weights, W_1^{AA} and W_2^{AA} for estimating the effect of a specific sequence on patient outcomes (e.g. the weight for the AA sequence is W^{AA}, as shown in Formula 4.7).[289]

$$W^{AA} = W_1^{AA} \times W_2^{AA} \quad (4.7)$$

W_1^{AA}, is the inverse probability of receiving the initial treatment A given baseline covariates. The second, W_2^{AA}, assesses the inverse probability of receiving the salvage treatment A, incorporating all covariates up to that juncture. While the exact covariates are not specified, it is suggested that they include both baseline characteristics and factors pertinent to the second-line treatment. Besides prognostic factors specified, Huang et al. included three additional variables for estimating the probability of receiving salvage chemotherapy: the time until first recurrence, the type of first recurrence (local or distant metastasis), and whether surgical resection was part of the salvage therapy. This approach suggests an underlying assumption by the authors that the probability of receiving a specific salvage treatment depends on timing of the salvage treatment initiation (i.e., a prognostic factor), which is also hypothesised to have a linear relationship (or another specified functional form)

with the outcome. This, informally, addresses the issue of uneven length intervals in calculating weights within the simplified approach.

Weights W^{AA} and W^{AB} are applied to patients receiving an initial treatment A to estimate the marginal effects of sequences AA and AB, respectively, Similarly, weights W^{BB} and W^{BA} are used for patients starting with treatment B to determine the marginal effects of sequences BB and BA, respectively. To implement their methods, Huang et al. used logistic regression models to estimate the probabilities of patients receiving chemotherapy as initial and salvage treatments. To some extent, Huang et al. implicitly used the "clone-censor-approach" approach, a recent variation of the IPW method detailed in the next Section 4.6.1.2.2, to estimate the marginal effects of AA and AB sequences using data from all patients initially treated with A. Huang et al.'s approach employs the concept of IPTW, treating the decision to receive salvage treatment as a binary choice, rather than focusing on regimen adherence.

In contexts with multiple treatment options at each stage, their approach can be adapted to evaluate specific treatment regimen using the concept of IPCW (e.g. multiple treatment options such as A, B, C, and D can be viewed as binary decisions—either adhering to treatment A or not adhering to treatment A, for example). It is crucial not to confuse the fact that Huang et al. further applied IPCW specifically to address issues of censoring due to administrative loss to follow-up, rather than censoring resulting from deviations from the treatment regimen. Huang et al. conducted a simulation study using Structural Nested Failure Time Models (SNFTM) to generate data for simulating the effects of different treatment sequences.[289] Their findings revealed that the mean estimated standard errors for the treatment effect parameters aligned with their empirical counterparts. However, their simulation implicitly involved an assumption that patients immediately transition to salvage treatment following their secondary baseline (i.e., disease progression) without any gap. The timing of salvage treatment was identified as a prognostic factor influencing both the outcome and the choice of salvage treatment.

In another real-world study by Nishino et al., Huang et al.'s methodology was referenced to compare different DTRs in treating 335 Japanese patients with non-small cell lung cancer (NSCLC).[289,332] They analysed the effects of two treatment sequences starting with gefitinib, including:

- Gefitinib → chemotherapy → gefitinib
- Gefitinib → chemotherapy → ghemotherapy

However, Nishino et al. did not detail how the method was practically implemented, leaving uncertainties about potential selection biases and immortal time biases, as discussed in Section 4.6.1.1.3.[332] Unlike Huang et al.'s paper, where all treatment sequences are binary (treated or untreated with chemotherapy), it was unclear how Nishino et al.'s model accounted for multiple

treatment choices at each decision point (e.g. the initiation of each LOT), making it difficult to assess its accuracy. Specifically, at each LOT, several others treatments may also be used interchangeably, including platinum combination chemotherapy, erlotinib, and single-agent chemotherapy. A valuable aspect of Nishino et al.'s study was the inclusion of a swimmer plot that visually displayed staggered cumulative treatment durations and untreated periods until the end of each patient's last LOT, providing a way to present treatment sequence patterns and durations across the population in RWD.sss[332]

4.6.1.2.2. Clone-censor-weight

- Origin of the method

The clone-censor-weight method[333] is recent advancement variation of the IPW method (Section 4.6.1.2.1), particularly useful in comparing DTRs (not limited to treatment sequences) starting from an identical baseline. In the context of comparing treatment sequences, it is particularly useful in comparing treatment sequences that begin with the same initial treatment without adjusting for baseline characteristics (e.g., A→B vs. A→C) by "cloning" patients starting with A.

In the "What If" book, Robins[5] and Hernan direct readers to Robins et al. (2008)[362] and Cain et al. (2010)[363] for further insights into the clone-censor-weight method, particularly in the supplementary material of Cain et al.'s study, where the procedure of "cloning" and weight derivation practical aspects and data structure was detailed. Although concepts resembling cloning appeared in earlier studies over a decade ago[279,289,362,363], the specific term and its methodology have only been formally coined and gained significant traction in the last five years, possibly primarily through the advocacy of the Harvard CAUSALab group's training courses.

- Theoretical characteristics

The theoretical properties of this method mirror those discussed in Section 4.6.1.2.1, but incorporate a preliminary step known as cloning. This addition eliminates the need of the initial IPTW (Step 2) in scenarios where the cloning results in all treatment strategy groups beginning with the exact same set of patients[75], thereby reducing potential bias from baseline unmeasured confounders. However, when treatment groups have only partially overlapping patients (clones), baseline weighting with IPW remains necessary to correct for confounding at the start of follow-up.

Beyond comparing of DTR involving treatment sequences. The clone-censor-weight method can also address challenges in other DTR observational studies, such as comparing treatment strategies factoring in the grace period of treatment initiation periods (e.g., starting treatment within 6 months of diagnosis represents compliance in the treatment group, while failure to do so leads to censoring at the end of 6 months), and avoiding and mitigating immortal time bias by avoiding patient selection

based on post enrolment assessment period characteristics.[75,334]

It is important to differentiate the cloning process from the concept of reusing patients in sequential trial emulation. Sequential trial emulation involves simulating a series of emulated target trials[364,365], each with a distinct follow-up start time (i.e., time zero), using the same patient data. Here, a patient's start of follow-up time varies, leading to different categorisations across trials (e.g., a patient not vaccinated in June but vaccinated in July might be classified as unvaccinated in an emulated target trial starting in June and as vaccinated in an emulated trial starting in July. In contrast to reusing patients in sequential trial emulations, the cloning procedure within the clone-censor-weight approach involves duplicating the same patients for different treatment strategy groups at the same starting point in a single trial emulation. This method can be used in conjunction with sequential trial emulation when necessary.

- Application in treatment sequences

The adoption of the clone-censor-weight technique for investigating DTRs has seen significant growth in recent years, mostly studies beyond DTR involving sequential treatments.[75,334,335] An illustrative example is provided in Hernan's 2018 study[333], which involves a toy RWD study that evaluates the effects of different durations of aspirin treatment on mortality. The study examined three treatment durations: 0, 1, and 2 years. Initially, patients who had not taken aspirin during the first year were categorised into the 0-year group (patient 1-4). For patients who started aspirin in the first year, the study cloned them into two additional groups, 1-year (patient 5a-12a) and 2-years (patient 5b-12b) of treatment. Any treatment initiated in the second year led to censoring in the 0 and 1-year groups, while discontinuation in the second year led to censoring in the 2-year group. Following the methodology described in Section 4.6.1.2.1, weights were then derived for analysing the weighted outcomes (i.e. marginal effect of receiving different duration of aspirin treatment). While framed as comparing treatment durations, this approach technically evaluates time-related strategies within treatment sequences, namely 1-year of aspirin → 1-year of aspirin, 1-year of aspirin → no aspirin, no aspirin → no aspirin. While the sequence of no aspirin → 1-year of aspirin could technically be also viewed as a 1-year aspirin treatment duration, it may be less clinically relevant (as aspirin is typically started without delay when indicated). Therefore, patients not taking aspirin in the first year were not cloned for estimating the marginal effect of this unlikely treatment sequence.

4.6.1.2.3. Adaption of structural nested failure time models (SNFTM) without g-estimation

- Origin of the method

Initially introduced by Robins and colleagues, Structural Nested Failure Time Models (SNFTMs) are a subset of Structural Nested Models (SNMs) developed for assessing treatment effects through

comparing counterfactual outcomes, factoring in time-varying factors.[67,366,367] Robins and Greenland's 1994 study investigated the survival of acquired immunodeficiency syndrome (AIDS) patients in a RCT comparing high-dose versus low-dose zidovudine treatment.[367] They found better survival in the low-dose group, attributing this to lower treatment toxicity and possibly increased prophylactic measures against pneumocystis pneumonia (PCP), a common co-infection in AIDS patients. To address non-random PCP-prophylaxis rates between groups, they used SNFTMs to estimate survival under a hypothetical scenario where both groups had similar PCP prophylaxis rates. For instance, in their most basic model, a subject's theoretical time to death is when they would have died had prophylaxis therapy been withheld. Their analysis demonstrated that the survival benefit for the low-dose group remained, even if the prophylaxis rates had been equalised between the groups.

I included a study by Huang and Ning, in which they adapted the SNFTM method to compare treatment sequences by re-analysing Cormier et al.'s study on chemotherapy's effects in soft tissue sarcoma (See Figure 4.5 in Section 4.6.1.2.1 IPW method).[290,338] Huang and Ning highlighted the importance to assess the joint effect of initial and subsequent treatments on patient survival in oncology.[290] They pointed out that solely measuring PFS from initial treatment might fail to capture its benefits in extending OS. However, evaluating initial treatment's impact on OS is complex as it can be compounded by the variability in salvage treatments, which could influence outcomes. Huang and Ning highlighted challenges in using conventional Cox proportional hazards models or Accelerated Failure Time (AFT) models for modelling time-to-event outcomes when disease recurrence and salvage treatments are incorporated as time-dependent covariates. This approach complicates the interpretation of the impact of baseline covariates, including initial treatment indicators, on OS, since the estimated effect of initial treatment is not marginal. The problem mainly originates from treating disease recurrence, an intermediate outcome of initial treatments, as an exploratory variable (predictor), which leads to a misrepresentation of the actual influence of initial treatments on OS (as briefly introduced in Chapter 1, see Figure 1.2). Thus, Huang and Ning investigated the joint impact of initial and salvage treatments on patient survival through hypothetical scenarios where salvage treatments were "optimised", assessing the initial treatment's impact in this setup. The definition of "optimisation" and theoretical characteristics of their method are outlined in the following section (Section 4.6.2). This approach differs from another publication introduced by Huang and other colleagues in 2006, who also revisited the Cormier study using the IPW method[289,338], as introduced in Section 4.6.2.1. In Huang et al.'s 2006 paper[338], they used the IPW method to analyse the average effects of specific induction and salvage treatment sequences (A→A, B→B, A→B, B→A). This estimand contrasts with approach detailed in the current section, Section 4.6.1.2.3, which compares the effect of induction treatments with optimised salvage strategies (i.e., A → optimal salvage treatment versus B → optimal salvage treatment). These differences relate to the terminology

introduced in Section 4.5.2: Essentially, Section 4.6.1.2.1 presents a method for understanding an uniform treatment sequence's effect within a population, whereas Section 4.6.1.2.3 discusses a method to identify an optimal version of that effect, one which maximises the population's overall survival.

Although Huang and Ning adapted the SNFTM structure, their practical approach in estimating treatment effect coefficients technically aligns more closely with the simple two-stage estimation (TSEsimple) described in Latimer et al.'s paper that discussed improved two-stage methods for tackling unwanted treatment-switching in RCT for HTA.[152] Specifically, Huang and Ning did not apply g-estimation. In contrast, the improved two-stage estimation with g-estimation (TSEgest) in Latimer et al.'s work accounts for potential gaps between disease progression and the receipt of second-line treatment. This approach utilises g-estimation to estimate the parameters, thereby adding complexity to the statistical model. Although SNFTM is often used with g-estimation, subsequent sections reveal that most methods for identifying optimal individualised treatment sequences have either adapted or paralleled the structure of SNFTM without employing g-estimation (Section 4.6.2). Instead, alternative estimation methods are utilised, leading to a divergence in methodological extensions. Examples include IPW applied in conjunction with the SNFTM structure and Q-learning, illustrating the variety of approaches in this area. The SNFTM structure facilitates backward induction to identify the "optimal strategy", whether individualised or not, with further details provided in the following paragraphs.

- Theoretical characteristics

Huang and Ning addressed leveraged SNFTM along with the ideas of the backward induction method—a method for mathematical dynamic programming optimisation[368]—to identify the optimal treatment sequence for maximising average OS of the population.[290] Their method hinges on three key time intervals: the time from initial treatment to disease recurrence (R), from salvage treatment to death (S), and from initial treatment to death (T), where the observed T equals the sum of R and S (as shown below in Formula 4.8), assuming salvage treatment starts immediately after recurrence.

$$T = R + S \quad (4.8)$$

Their backward induction method "jointly" models the effects of initial and salvage treatments through two sequential AFT models, starting with the last(-line-of) treatment. It first uses an AFT model to evaluate the impact of salvage treatments on survival after recurrence (S), assuming no unmeasured confounders at salvage treatment initiation, as shown in the below Formula 4.9:

$$S_i(P_i, Q_i) = \exp(N_i'\lambda + Q_i'\theta + \epsilon_i) \triangleq \exp(M_i'\beta + \epsilon_i) \quad (4.9)$$

For each subject i, P and Q denote the binary initial and salvage treatments. N represents the covariate values at the time of disease recurrence, including initial treatment P). An apostrophe

denotes the transpose of a vector or matrix, facilitating the appropriate algebraic operations. In the model (Formula 4.9), λ quantifies the effect of covariates N, while θ assesses the impact of salvage treatment Q. M is a combination of N and Q as $M_i = (N_i', Q_i')'$. β is a combined vector that incorporates both λ and θ (i.e., $\beta = (\lambda', \theta)'$), and ϵ_i is assumed to be independent of M_i.

From the above AFT model on S, for each subject i with initial treatment $P_i = p_i$ and salvage treatment $Q_i = q_i$, the potential survival time after recurrence under optimal salvage treatment $Q_i = b$ (S_i^*) can be estimated . When the optimal salvage treatment was administered, the actual survival time S_i aligns with the hypothetical optimised S_i^*. If the given salvage treatment was suboptimal, the actual survival time S_i would be less than the ideal S_i^*. Then the (potential) optimal S_i^* can be calculated as follows:

$$\begin{aligned} S_i^* \triangleq S(p_i, b) &= \exp(N_i'\lambda + b_i'\theta + \epsilon_i) \\ &= \exp(N_i'\lambda + q_i'\theta + \epsilon_i - q_i'\theta + (\theta)_+) \\ &= S_i \exp(-q_i'\theta + (\theta)_+), \end{aligned} \quad (4.10)$$

where (θ)+ equals |θ| when θ is positive, and 0 when θ is non-positive. If $q_i \neq b$, indicating a non-optimal salvage treatment was given, the equation modifies the actual survival time S_i to estimate potential survival S_i^* under optimal salvage treatment by multiplying the S_i to an adjustment factor $\exp(-q_i'\theta + (\theta)_+)$. Huang and Ning simplified the model by assuming a uniform optimal salvage treatment ($Q_i = b$) across all patient groups but mentioned that, for instance, incorporating interaction terms between initial and salvage treatments can address more complex scenarios.[290] They suggested the above formulation of AFT model was adopted from Robins and Greenland's SNFTM[367], but with an adapted formulation that avoids the issue of negative values encountered in Robins and Greedland's equation (1b)[367]. Huang and Ning acknowledged the complexity of estimating S in the presence of censored observations[290,369] and further employed the IPCW method[358] to accommodate these, leveraging approaches similar to those in Section 4.6.2.1.2, a detail I do not delve into here. Hence, informally and conceptually, it represents adding an additional step to the TSEsimp method in Latimer et al.'s paper, mentioned earlier in the current section, with IPCW.[152]

Upon determining the optimal salvage treatment, Huang and Ning built the second AFT model to assess the effect of initial treatment on OS by "plugging-in" the hypothetical optimised S_i^* (i.e., an "expanded" S_i). Specifically, for an individual treated with $P_i=p_i$, their potential OS time T_i^* under optimal salvage treatment can be calculated as follows:

$$T_i^* \triangleq (p_i, b) = \delta_{Ri}\,(R_i + S_i^*) + (1 - \delta_{Ri})T_i, \quad (4.11)$$

where δ_{Ri} indicates if a patient experienced disease recurrence (i.e., $\delta_{Ri} = I(\delta_{Ri} < T_i)$). If a patient experienced recurrence ($\delta_{Ri} = 1$), their T_i^* would be the sum of the observed time to recurrence with the optimised post-recurrence survival:$(R_i + S_i^*)$. For those without recurrence ($\delta_{Ri} = 0$), T_i^* equals

to their observed T_i. This setup leads to the application of a second AFT model expressed as follows:

$$T_i^* = \exp(Z_i'\gamma + \xi_i) \quad (4.12)$$

The parameter γ represents the effect of baseline covariates, including the initial treatment, on OS when optimal salvage treatment is given. The vector Z_i includes covariates observed prior to initial treatment and the initial treatment itself, thus $Z_i = (V_i', P_i')'$, where V represents the vector of all covariates observed prior to the initial treatment P. The component of γ related to P identifies the optimal initial treatment that can maximises T_i^*.

In summary, Huang and Ning's approach leveraged two AFT models to evaluate the joint effect of two-line sequential treatment on survival, specifically in cases where a subsequent treatment is administered upon disease recurrence. The first model evaluates the salvage treatment's effect on post-progression survival (S), identifying the optimal salvage treatment (Qi = b) that maximises the potential post-progression survival (S^*). The second model then identifies the optimal first-line treatment based on S^*. In-depth explanation of the models' estimation processes and the estimators applied can be found in the Huang and Ning's study.[290]

Huang and Ning et al. did not employ g-estimation, and it is unclear how and whether they have addressed potential time-varying confounding biases that could arise between treatment initiation and disease progression—the critical juncture where SNFTM delineates "sections of survival times"— in estimating the coefficients for treatment effect as shown in Formula 4.10. This issue, theoretically may be less of an issue the IPW method described in Sections 4.6.1.2.1 and 4.6.1.2.2, where patients' time-varying characteristics are considered at the point of deviation from the treatment strategy. Notably, treatment deviation in the IPW method does not necessarily have to coincide with the timing of disease progression, although this remains an underlying assumption in Huang et al.'s 2006 study, as introduced in Section 4.6.2.1.1.[289]

Their approach can be extended to identify optimal sequence of multi-line treatments by dividing S into intervals between further salvage treatments and recurrences, using k number of AFT models in reverse order to identify optimal treatments from the last line to the first (i..e, k, k-1, …1). Huang and Ning noted the methodology relies on several assumptions:

(1) No unmeasured confounders: the allocation of treatment, conditioned on observed history (covariates), is independent of potential outcomes (sequential ignorability[370]).
(2) Consistency: The potential outcome of a specific treatment or treatment sequence aligns with the actual outcome observed when that treatment or treatment sequence is given.
(3) The AFT models are correctly specified.
(4) Independence between subjects: each subject's treatment and outcomes do not affect or predict another's.

- Application in treatment sequences

Huang and Ning conducted a simulation study to assess the aforementioned methodology, exploring the impact of both covariate-independent and covariate-dependent censoring.[290] Their simulation results revealed that their estimators performed consistently well across varying levels of censoring, demonstrating low empirical bias. Although their simulation appeared to operate under ideal scenarios with critical assumptions—e.g., no treatment gaps and no unmeasured confounders.

As mentioned earlier, Huang and Ning re-analysed the Cormier study data[338], specifically on the effect of chemotherapy on survival in 674 soft tissue sarcoma patients (Figure 4.5). They first adjusted the observed survival times (T) for 350 patients who had recurrences by applying the estimated effect of salvage chemotherapy on post-recurrence survival from the first AFT model, generating hypothetical survival times (T*) as though all progressed patients received optimal salvage treatment. Salvage chemotherapy's impact on survival was found to be negligible (post-progression survival was 0.99 times the length for those not receiving it, $p = 0.95$). For the 324 patients without recurrence, original survival times were retained.

Utilising data from all 674 patients, Huang and Ning's second AFT model assessed the effect of initial chemotherapy on OS with the premise that all patients who progressed received optimal salvage treatment (i.e., no salvage chemotherapy) and accounted for censored survival times using IPCW and truncated at 15 years.[290] They found that initial chemotherapy increased the average OS times by 32% ($p = 0.04$). Conversely, Cormier et al.'s study[338], using Cox models, reported HRs of 0.37 (favouring chemotherapy) in the first year and 1.36 (favouring no chemotherapy) separately due to survival curves crossing between groups. The study did not separate initial and salvage treatment effects and advised caution in interpreting HRs, as these were based solely on the 591 patients who survived more than one year.

Huang and Ning conducted sensitivity analyses to examine the impact of unmeasured confounders using methods developed by Lin et al.[371] , and found unmeasured confounders can result in both underestimation or over estimation benefits of the initial chemotherapy, provided that the salvage treatment has been optimised. Furthermore, while their method focused on identifying the optimal strategy, by adjusting the definition of "optimised", it could technically compare other treatment sequences (e.g., setting "optimised" to mean the shortest population survival). However, significant adjustments might be necessary to avoid logical inconsistencies.

4.6.2. Methods for identifying optimal individualised DTR involving multiple treatment sequences using RWD or RCTs

As mentioned earlier, the methods discussed in the current section (Section 4.6.2) are deemed less readily applicable in HTA and may need significant modification for adaptation. This is because

these methods aim to identify the optimal individualised treatment sequences and assess the effect treatment strategies that involved multiple treatment sequences, as outlined in Section 4.5.2. This approach contrasts with HTA's focus on evaluating uniform treatment sequences (see Chapter 3). Therefore, I offer a more concise overview of each method in this section, primarily emphasising their conceptual relevance to other methods. Here, I first outline the overarching relevance of the methods in this section, which underpins its layout.

Importantly, as briefly mentioned in Section 4.6.1.2.3, most methods in the current section (Section 4.6.2) were recognised to have either adapted or paralleled the structure of SNFTM and leveraged the backward induction approach (see Section 4.6.1.2.3) to identify the optimal effects of strategies at each LOT.[275,287,288,291,331] The backward induction approach was utilised to optimise strategies, moving from the last section back to the first LOT, while the SNFTM structure established distinct junctions (with model assumptions) to segment the optimisation process. In my view, these methods can be seen, informally and very broadly, as variations and extensions of the TSEsimp method described in Latimer et al.'s paper, although optimising treatment strategy was not Latimer et al.'s primary focus.[152] The similarity lies in how they simplify the question of real-world treatment sequences into multiple stages, analogous to how TSEsimp in Latimer et al.'s paper simplifies the issue of unwanted treatment-switching issue into two stages.[152] However, these methods generally extend beyond two stages in the optimisation along the complex and individualised treatment pathways. Furthermore, these methods employ a variety of parameter estimation techniques instead of relying solely on g-estimation, which is commonly linked with SNFTM. This includes IPW (Section 4.6.2.1.1), adaptation of the g-formula (Section 4.6.2.1.2), and Q-learning (Section 4.6.2.2.1). Notably, the TSEsimp method in Latimer et al.'s paper did not use g-estimation, positioning it as a simpler version of TSE compared to TSEgest, which incorporates g-estimation to account for time-varying biases occurring between disease progression and the initiation of subsequent treatment.[152] Conceptually, Wahed and Thall's approach models counterfactual outcomes at each junction of the SNFTM structure for each patient based on their characteristics (i.e., outcome models).[287] In contrast, IPW weights patients who adhere to a treatment strategy based on their inverse probability of continuing adhered to a specific treatment strategy to represent those who deviated (i.e., treatment models).[287] Given that Latimer et al.'s TSEsimp also modeled outcomes, I view it as having the flavour of the g-formula.[152]

4.6.2.1. G-methods

4.6.2.1.1. Inverse probability weighting

- Origin of the method

The IPW method's origin is the same as in the description in Section 4.6.1.2.1. However, in my review, two studies—by Wahed and Thall, and Xu et al.— further leveraged IPW in identifying optimal DTR involving multiple treatment sequences across a population.[287,331] Both studies adapted IPW for use in conjunction with the SNFTM model. This application of SNFTM highlights the decision points where IPW is used to weight each segment of LOT. Both papers utilised IPW as their reference method for investigation alongside their main investigated method—adaption of g-formula—which is detailed in the following section (Section 4.6.2.1.2).[287,331] In addition to the standard IPW method, Xu et al. also computed augmented IPW (i.e., a version of IPW with doubly robust feature) in their study.

- Theoretical characteristics

The theoretical characteristics of the methods investigated by Wahed and Thall, and Xu et al., generally follow the same principles as those described for the IPW method in Section 4.6.1.2.1. Additionally, their use of the SNFTM structure introduces further assumptions that need to be considered, as those outlined in Section 4.6.1.2.3.

- Application

Both Wahed and Thall, and Xu et al., reanalysed the 4-arm RCT with non-randomised salvage treatments in leukaemia by Estey et al., to demonstrate their methodologies. Figure 4.6 illustrates the trial setting where patients were randomised to receive one of four different induction treatments (decision point A). Following this, patients might either die before receiving further treatment or achieve complete remission and subsequently receive salvage therapy (decision point: B1) or, upon progression, receive an alternative salvage therapy (decision point: B2). These salvage therapies were not randomised.

Figure 4.6 illustrates all possible transitions in the Estey 1999 trial, simplified into a SNFTM format, where each transition represents a segment within the SNFTM (denoted as T with different superscripts). Notably, the model setup makes the decision points for Salvage treatments B1 and B2 mutually exclusive, reflecting patients with different characteristics post first-line treatment (i.e., resistance versus complete remission) might have different theoretically optimal salvage strategies and are thus modelled separately. In the previously introduced method that adapted the SNFTM to identify optimal uniform treatment sequences by Huang and Ning, patients with progression were not

differentiated based on whether the progression followed resistance or a period of complete remission.[290] Consequently, Huang and Ning's study did not identify "individualised" optimal treatment sequences. In contrast, the approaches used by Wahed and Thall, and Xu et al., incorporate more stages in the SNFTM, including all 9 transition times T shown in Figure 4.6, as opposed to the simple two transition times in Huang et al.'s SNFTM outlined in Formula 4.6). This expanded SNFTM enables the identification of optimal, individualised treatment sequences.

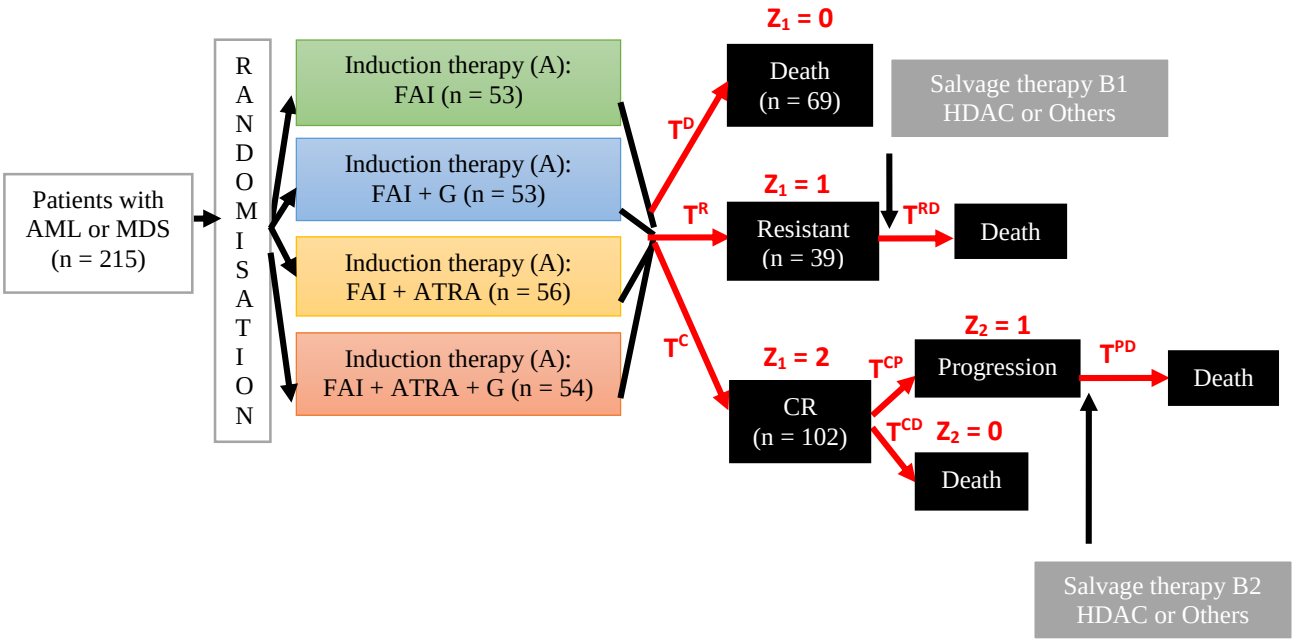


Figure 4.6 Possible patient flow of the Estey 1999 study[339] (reproduced from Wahed and Thall 2013[287])

AML: acute myeloid leukaemia; CR: complete remission; MDS: myelodysplastic syndrome

The equations for deriving IPW weights used by Wahed and Thall, and Xu et al., were relatively complex and are not reproduced here. Conceptually, in Figure 4.6, for a given segment of T, each patient's T is weighted if the treatment received at that junction prior to segment T was not optimised. This weighting reflects the counterfactual T* in which the patient had received the optimised treatment at the beginning of the junction. The same backward rollback technique in Section 4.6.1.2.3 was applied to identify optimal treatment strategies, which were individually optimised in this instance due to the setup of the SNFTM as explained in the previous paragraph. Essentially, in this context, IPW serves as an alternative strategy for computing counterfactual outcomes, contrasting with the methods used by Huang et al. described in Section 4.6.1.2.3. Although both studies by Wahed and Thall, and Xu et al. have used the aforementioned methods in the context of analysing a RCT with non-randomised subsequent treatment, these techniques can theoretically be adapted to identify optimal individualised treatment sequences in RWD. This adaptation theraitlcaly would include an additional IPW weighting at the last stage of the rollback technique to address non-randomised first-

line treatments.

4.6.2.1.2. G-formula

- Origin of the method

The g-formula, also known as g-computation or (g-)standardisation[54], has been used to estimate the causal effect of a treatment in the presence of time-varying confounders that themselves may be affected by prior treatment (e.g. intermediate outcome). The g-formula was first introduced by Robins in his 1986 paper[54,372], predating other g-methods such as g-estimation of SNM[366] (Section 4.6.1.2.3) and IPW of marginal structural models[58,68,69] (Section 4.6.2.1).

The g-formula has been applied in both RWD and RCT analyses, where outcomes may be confounded by subsequent treatment/events.[66,373] The g-formula can be implemented using either non-parametric or parametric approaches based on data complexity. The non-parametric approach does not assume relationships between covariates and outcomes[374], resembling a saturated (outcome) model, incorporating all covariates and interactions in the estimation. Conversely, the parametric approach, more commonly seen in research due to its applicability to high-dimensional data, assumes specific forms (e.g., linear, quadratic) for the relationships between covariates and outcomes, as well as about the interactions among specific covariates, and is thus known as the parametric g-formula.[373,375] In its most basic application, the g-formula can be used to estimate the causal effect of a time-fixed treatment[376], such as the effectiveness of a single-dose of vaccine. It can also be employed in comparing sustained treatment strategies[376], such as "always treat" versus "never treat," or to tackle time-varying treatments can alter based on specific characteristics.[373]

My review included two studies that adapted and extended the g-formula to identify optimal individualised DTR in acute leukaemia that maximise patients' OS in conjunction with SNFTM.[287,331] Specifically, the same two studies that previously used IPW to identify individualised treatment sequences from the Estey 1999 trial (see Section 4.6.2.1.2) also re-analysed same trial using g-formula adaptations with the same objectives.[287,331,339] Specifically, Wahed and Thall applied a likelihood-based frequentist approach, while Xu et al. adopted a Bayesian nonparametric (BNP) approach for the g-computation. Both studies tailored the g-formula equation to estimate counterfactual outcomes under the same adapted SNFTM framework described in Section 4.6.2.1.1 and Figure 4.6.

- Theoretical characteristics & application

The g-formula estimates the average causal effect of a treatment (regimen) by modelling the outcome as a function of both (time-varying) treatment and confounders.[5] It predicts the outcome that would have resulted from different treatment (regimens) by adjusting for confounders at each time

point. This involves estimating the expected outcome for each individual under a hypothetical scenario, then contrasting the outcomes under an alternative scenario (for instance, comparing the outcomes if all individuals received a particular treatment versus if none received it). By aggregating and averaging the expected outcomes across these scenarios, it effectively estimates "standardised outcome distributions for specific patient covariate profiles". The difference in these average outcomes then provide an estimation of the average causal effect of the treatment. Similar to IPW, the g-formula also relies on assumptions regarding unmeasured confounders. However, unlike IPW, which requires a correctly specified model to predict the propensity of receiving treatments or censoring (for IPTW and IPCW, respectively), the g-formula necessitates correctly specified outcome models.

Technically speaking, the g-formula is theoretically more computationally intensive compared to IPW due to its requirement for multiple iterations and simulations of counterfactual outcomes. However, in the studies by Wahed and Thall, and Xu et al., the computational demand may be reduced because they simplified the continuity of time into 9 transition times rather than using uniform time intervals for computing counterfactual outcomes. Further, they incorporate a backward induction technique, which contrasts with the typical implementation of the g-formula that is carried out in a forward fashion. Wahed and Thall, and Xu et al.'s approach simplifies this by focusing on key transition points and applying backward induction. Their adaption, therefore, may not be as computationally demanding as the standard g-formula applications. This clarifies why Latimer et al.'s TSEsimp, although sharing similarities with the g-formula[152], does not exactly follow it due to its SNFTM-like structure. Conversely, the typical application of IPW, does not involve iterative simulations but rather models that predict the probability of treatments or censoring, depending on the context, as discussed in Section 4.6.1.2.1. These models directly factor in time and previous treatments, allowing for calculation of weights for each interval in a single step (or few steps), given patient time-varying characteristics. Thus, this approach is theoretically less computationally demanding than typical g-formula. However, IPW may face challenges with extreme weights, particularly when there is inadequate overlap between groups, such as between different treatment groups or among patients who are censored versus uncensored, especially if the rate of censoring is high.

- Application

Wahed and Thall, and Xu et al., applied their respective adapted g-formula approaches (likelihood-based versus Bayesian) in conjunction with the SNFTM as depicted in Figure 4.6, to identify the optimal individualised treatment sequences from the Estey 1999 study.[287,331,339] Conceptually, this follows the same procedure as outlined in Section 4.6.2.1.1, with the sole difference

being the derivation of counterfactual outcomes for each transition time T using a different method, replacing IPW with their respective adapted g-formula. The authors evaluated their findings with adapted g-formula against those obtained using IPW from Section 4.6.2.1.1. Both the adapted g-formula and IPW methods yielded similar conclusions regarding the efficacy of 16 different treatment strategies (i.e., all possible combinations of A, B1, and B2 in Figure 4.6). Although IPW and g-formula adaptions produced different mean survival estimates, they both identified the same treatment strategies as most effective and the least effective. The 90% confidence intervals of mean survival estimates from two different techniques overlap to some degree, even though their point estimates are notably different. Furthermore, despite aiming to identify the optimal individualised strategy, both studies assessed the marginal effectiveness of each of the 16 possible strategies.

4.6.2.2. Reinforcement learning

4.6.2.2.1. Q-learning

- Origin of the method

Q-learning originated from the fields of dynamic programming and reinforcement learning.[291] It is a model-free algorithm that seeks to find an optimal action-selection policy for any given finite Markov decision process. The approach operates by estimating the value of action-state pairs (i.e., Q-values), and iteratively updates these values based on received rewards and the maximum anticipated future rewards.[377,378] In the context of DTR, Q-learning is used to sequentially optimise decisions, aiming to determine the most effective actions (treatments) based on various patient conditions.

My review identified three studies that leveraged the principles of Q-learning, similar to IPW and adapted g-formula in Section 4.6.2.1.1-4.6.2.1.2, these studies employed a backward induction method to identify the optimal individualised treatment sequences. This includes research by Huang et al., who re-analysed the Estey 1999 study as shown in Figure 4.6.[288] The other two studies are real-world studies: one aimed at understanding treatment strategies for Graft-versus-Host Disease (GVHD) in leukaemia using the Centre for International Blood and Marrow Transplant Research (CIBMTR) registry in the United States (US), and another aim to understand diabetes treatment strategies using the UK Clinical Practice Research Datalink (CPRD) data.[275,291]

- Theoretical characteristics

The actual parameter estimation process of Q-functions and how these relate to the counterfactual outcome estimation were not immediately clear from the reviewed papers. However, if they were applied following the principles of the original version of Q-learning, this could prove problematic as Q-learning does not necessarily adjust for confounding per se. Rather, it focuses more

on predicting and maximising the outcomes for patients with certain characteristics and deciding the appropriate actions for patients carrying specific traits at various stages. For example, in the diabetes study by Simoneau, patients' Hemoglobin A1c (HbA1c) levels and body mass index (BMI) were used as tailoring variables to identify the best add-on treatments at different stages that maximises patient's outcomes.[275] That is, decision rules learned from Q-learning appear to be better tailored to evolving patients' characteristics (in a sense that it maximises predicted outcome). The approach exhibits counterfactual characteristics; however, it remains uncertain whether it aligns fully with the previously discussed methods for estimating causal effects.

- Application

Huang et al. revisited the Estey 1999 study, offering a distinct approach compared to Wahed and Thall who also re-analysed the same RCT as described in Section 4.6.2.1.1 and 4.6.2.1.2.[287,288]

Utilising all available longitudinal patient data, Huang et al. identified optimal treatment strategies for individuals, offering a range of possibilities beyond predefined transition times. In contrast to Wahed and Thall, who estimated discrete treatment strategies from a set of predetermined options (e.g. 16 strategies from all possible combinations of treatment A, B1, and B2 in Figure 4.6), Huang et al.'s method facilitates the creation of non-pre-defined, more detailed treatment strategies that were learned from the process of Q-learning, such as "beginning with treatment A1, and if disease recurrence occurs within six months, switching to treatment B1 for salvage treatment; if not, proceeding with treatment B2 for salvage treatment".[287,288] His approach not only personalises treatments at each "stage" but also integrates additional patient characteristics into decision-making (e.g. time of disease recurrence). According to Huang et al., the selection of individual characteristics for treatment decisions is determined through model selection, with cut-off values for key variables, like the six-month cut-off for disease recurrence, derived from solving statistical models with estimated parameters in Q-learning. This highlights the previously mentioned characteristic of Q-learning, which theoretically does not require a model. However, in practice, it still involves some degree of statistical modelling, although some values can be learned without pre-specification (e.g. the cut-off value of patient characteristics). The other included two real-world studies leveraging Q-learning conceptually followed a similar approach to that presented by Huang et al. Specifically, Simoneau et al. dynamic weighted survival modeling (DWSurv) expanded on Huang's single robust framework with doubly robust features.[275]

4.6.2.3. Other methods

In Table 4.6, I classify the paper by London et al. under the category of g-estimation as it referenced Robins' 2004 paper on optimal structural nested models for optimal sequential decisions, and the optimal DTR paper by Murphy.[286,340,341] The specifics of their method application were not

detailed beyond these references. However, to illustrate its relevance, a brief description of their study is provided.

London et al. analysed a neuroblastoma trial where patients were randomized to begin treatment with either topotecan alone or topotecan plus cyclophosphamide. Using the DTR concept, they explored optimal subsequent treatment decisions based on whether to proceed with autologous stem-cell transplantation, contingent upon the patient's remission status (complete remission versus no remission) or to undergo transplantation irrespective of remission status. This approach parallels the treatment strategy evaluation seen in Robins and Greenland's 1994 study on PCP prevention in HIV (see Section 4.6.1.2.1), where each treatment option (high-dose versus low-dose treatment) potentially influences the likelihood of receiving further interventions across different treatment arms (i.e., PCP prevention) and, consequently, could impact survival outcomes. Similarly, London et al. aimed to evaluate topotecan versus topotecan plus cyclophosphamide, but the comparison was complicated by the fact that these two randomised groups had different probabilities of receiving transplantation, which could potentially affect outcomes. Thus, they explored the DTR embedded within the trial to determine the optimal subsequent treatment for each initial treatment group.

4.6.3. Methods for analysing SMART

For analysing SMART, most methodologies identified (Table 4.6) share foundational elements with those utilised in RCT and RWD that address time-varying confounding (Table 4.7). This aligns with the taxonomy of study types and their parallels discussed in Section 4.5.1.

IPW stands out as the predominant method origin, followed by the adaption of Q-learning and g-formula. Despite these shared roots, SMART analysis has spurred the development of unique extensions that introduce a broader range of estimators, such as the cumulative incidence function (CIF), which are less common in other study types. While these alternative estimators enrich the analytical toolkit available, they play a less role in the context of this review, which aimed to identify methods suitable not only for comparing treatment sequences but also for facilitating proof-of-concept case studies that can be compared against established benchmarks in RCTs for HTA, as detailed in Chapters 6 and 7. This section, therefore, does not detail each specific methodology used in analysing SMART. Instead, it highlights the relevance between different methods in Table 4.7, offering insight (a "map") for future extensions for future research extensions.

4.7. Discussion

According to Mahar et al.'s review on methods used for analysing DTR in observational studies, IPW and g-formula are the most prevalent methods, followed by Q-learning.[274] My review observed similar findings, but extends to cover the examination of RCT featuring non-randomised subsequent treatments and SMART, yielding broader insights. Specifically, IPW have been utilised in real-world

studies to compare uniform treatment sequences, making these insights particularly readily applicable in an HTA context.[289,332,333] While one study did adapt SNFTM (with a "flavour" of g-formula) to compare uniform treatment sequences, it was aiming to identify an optimal uniform treatment sequence.[290] Furthermore, although Q-learning was also observed to estimate the effects of DTR, its use was confined to studies aiming to identify individualised treatment sequences. The causal inference justification for this approach proves questionable, as the counterfactual outcomes calculated by Q-learning are largely informed by the learned Q-function from the data, which focuses on maximising outcomes, including variable cut-offs, to inform treatment strategies.[275,288,291] IPW and the g-formula have also been adapted in conjunction with SNFTM to identify the optimal individualised DTR.[287,331] In contrast to Q-learning, these methods require predefined strategies and cannot "search" for cut-off values to inform new treatment decision rules (i.e., strategies) that were not previously defined. All of these methods share similar roots with those used to analyse SMARTs, with methodological studies on SMARTs being several steps ahead in applying these techniques across a wider range of estimators.[302,303]

Interestingly, regardless of the data types used, the majority of studies reviewed focus on oncology applications. However, this may be due to the fact that my review specifically only includes studies reporting time-to-event related outcomes. Nevertheless, this focus is particularly beneficial because oncology has been identified as one of the fields with the most prevalent discussions around treatment sequences HTA, as well as facing substantial challenges due to data scarcity on treatment sequences (Chapter 2 and Chapter 3). Therefore, it is logical to narrow the focus to designing proof-of-concept case studies that assess the feasibility of using RWD to derive unbiased estimates for comparing treatment sequences (Thesis Aim 5 & 6, Section 1.5, Chapter 1) in oncology. This approach was later further affirmed by the richness of benchmark trials in oncology identified in Chapter 6.

The IPW methods (Table 4.6), particularly the combination of IPTW and IPCW, emerge as the most promising for use in my case studies for a variety of compelling reasons. Firstly, only IPW has been used for comparing uniform treatment sequences without the aim of identifying optimal treatment sequences. This approach aligns closely with what is typically required in HTA, making IPW particularly relevant and readily applicable without substantial adaption. Secondly, studies based on the SNFTM framework typically need extra IPCW to manage survival right censoring.[287,290] In contrast, IPW allows for a more straightforward combination of various IPW types (e.g., IPTW*IPCW) making it more attractive. Thirdly, IPW's extension with the "cloning" strategy further highlights IPW's versatility, especially when combined with the TTE framework as cloning helps mitigate immortal time bias. Although cloning is not a "statistical" method per se, it is powerful in mitigating bias and comparing treatment sequences sharing the same initial path (that bypass any

unmeasured confounding), a capability that other methods do not exhibit. Furthermore, there is considerable experience of IPCW within HTA, particularly for addressing unwanted treatment-switching in trials.[43] If IPCW proves to be effective in providing unbiased estimates when comparing treatment sequences using RWD, this familiarity from broader applications will likely facilitate its wider adoption. Importantly, the conceptual simplicity of IPCW makes it easier to explain, which is particularly beneficial when collaborating with clinicians to identify relevant prognostic factors in establishing IPW weight models. Finally, the availability of IPW code from my attendance in the CAUSALab courses at Harvard University for adaption enhances its practicality within my thesis. These factors together make IPW a compelling option of advanced statistical method for tackling time-varying confounding in comparing treatment sequences for HTA and a logical first step for exploration under time constraints within the PhD, where other methods might necessitate extensive adaptations.

The g-formula could be considered for the next step due to its conceptual parallels to the IPW method (Table 4.6). However, it may require further adaptation since the available examples do not pertain to comparing uniform treatment sequences. Additionally, the g-formula may be computationally substantially more expensive compared to IPW. The implementation of Q-learning could also be explored; however, it should be undertaken with caution due to its ambiguous causal interpretations, and because it has mainly been utilised in studies seeking to determine the best personalised treatment strategies, which are not the central concern of HTA.

The review has several limitations. Firstly, it has only one stage of pearl-growing. However, the pearl-growing procedure primarily identified SMART studies that used similar methodologies to those found in a systematic search using the term “treatment sequences”, so it likely covered the most relevant methods. Further, the targeted maximum likelihood estimation (TMLE) method could be a valuable addition, as indicated in Mahar et al.'s review, but it has not been as widely adopted as other methods for DTR and may therefore require substantial effort to adapt to the context of my focus.[274,379,380] Specifically, no research using TMLE has been seen in explicitly comparing uniform treatment sequences. Despite its limitations, the review supports the exploration of IPW as a logical first step forward, based on the justifications provided earlier.

4.8. Chapter summary

This chapter provided a detailed review of advanced statistical methods that could potentially be useful in providing unbiased estimates of treatment sequences from RWD. It also closely examined the caveats of simple statistical methods. The chapter summarised several key taxonomies essential for understanding the advanced methods identified, including types of clinical studies, variants of treatment sequences, and method categories. IPW was identified as a logical first step forward for

exploration, followed by the g-formula. This groundwork was later applied in designing the case studies and establishing benchmarks for design (Chapters 7 & 6, respectively), and in evaluating suitable RWD for a proof-of-concept study (Chapter 5), particularly in assessing whether a database has the quality and breadth necessary to accommodate these methods. Most importantly, it provides insights for practically implementing these methods to my case studies in Chapter 8.

Chapter 5 Real-world data sources for proof-of-concept benchmarking case studies

5.1. Chapter overview

In Chapters 2 and 3, I highlight the potential of real-world data (RWD) as an alternative source for quantifying the effectiveness of treatment sequences. Chapter 4 then explores causal inference statistical methods for minimising biases in estimating the comparative effectiveness of treatment sequences from RWD. To effectively leverage these methods, it is crucial to have data that contain detailed information regarding treatment decisions and patient characteristics over time. Patient-level data are needed, and ideally the data should be collected over a substantial time period to ensure a sufficient number of outcomes occur within the follow-up timeframe.

Assessing the performance of statistical methods in a given data source against established standards is crucial before these methods and datasets can be reliably used to inform decision-making (the use of benchmarking as a tool for assessing performance is further detailed in Chapter 6, Section 6.2).[381,382] This chapter, in conjunction with Chapter 6, collaboratively identifies suitable clinical trials as benchmarks for treatment sequence comparisons (Chapter 6), and evaluates which RWD sources show promise in providing the basis for replicating these trials' results (Chapter 5). This synergy sets the stage for the development of my proof-of-concept benchmarking case studies in Chapter 7. Successful benchmarking in these case studies would demonstrate that the same statistical approaches can be used to explore similar research questions in comparable contexts, within a specific database.

In this chapter, I first detail the reciprocal refinement process in identifying the benchmark trials and candidate databases for my PhD project. This involves an iterative critical assessment of clinical trials comparing treatment sequences and suitable RWD sources for replicating them. Section 5.3 then provides an overview and comparison of the two data sources chosen for my benchmarking case studies (Chapter 7), highlighting their unique attributes and relevance. Finally, Section 5.4 reflects on the challenges and learning experiences I encountered as an early career researcher dealing with data access complexities, particularly focusing on the de novo independent applications for patient-level health data. This chapter does not cover the process of transforming data from the chosen databases into “research-ready” datasets[383,384] tailored for Chapter 7's case studies, a topic addressed in Chapter 8's discussion on data curation pipelines.

5.2. Reciprocal refinement: identifying benchmark trials and fit-for-purpose data

Figure 5.1 outlines the reciprocal refinement process for identifying candidate benchmark trials (Chapter 6) and fit-for-purpose data databases (Chapter 5) for my proof-of-concept case studies (Chapter 7). This process highlights the tandem considerations in identifying suitable RWD sources

to test the applicability of methods identified in Chapter 4 to inform the English health technology assessment (HTA) decision-making process when treatment sequences are involved, aiming to pinpoint pertinent (ideally English) nationwide databases or disease registries. Although extended trial data containing sequencing information, such as non-randomised subsequent treatments, could be viewed as a form of RWD (see Chapter 2, Section 2.5.2.4), it is beyond the remit of my thesis.

My database selection broadly aligns with the principles set out in the recently published Structured Process to Identify Fit-For-Purpose Data (SPIFD) framework[385], the NICE real-world evidence (RWE) framework[37], and guidelines for good database selection in pharmacoepidemiology studies[386]. These include checking if the database offers variables for selecting relevant patients, defining covariates for statistical analysis, and capturing specific treatment patterns with sufficient sample size (details in Section 5.3, and Chapter 6, Section 6.6). Furthermore, budget constraints (£15,000-£25,000), my PhD timeline, and available connections for database access also influenced the decision.

A pilot search of candidate benchmark trials (Chapter 6, Section 6.4.1.) highlighted a predominance of oncology studies (over 50% of the 20 identified), steering my focus towards oncology-specific databases. This choice also reflects the prevalence of sequencing in oncology HTA and the frequent use of RWD in oncology technology appraisals (TA) (Chapter 3). Additionally, many methods outlined in Chapter 4 were originally applied in oncology studies.[291,338,339] Consequently, the English National Cancer Registration and Analysis Service (NCRAS) database[387,388], particularly its SystemicAnti-Cancer Therapy (SACT) dataset[389,390], stands out for my study due to its recognised role in monitoring the use of drugs placed in the Cancer Drug Fund (CDF) and as a source of data used to address uncertainties in National Institute for Health and Care Excellence (NICE) appraisals of CDF drugs.[391-399] Furthermore, my supervisor Professor Latimer's experience in navigating the application process via the Office for Data Release (ODR) at Public Health England (PHE), custodian of English NCRAS data before the COVID-19 pandemic, strengthened its practicality for my study.

As of November 2020, I identified two routes to access NCRAS data: via the ODR[400] or Clinical Practice Research Datalink (CPRD)[401-403]. However, CPRD's estimated cost of over £25,000 made it unfeasible for my PhD (approx. £7,000 via ODR).[404] Consequently, I concentrated on identifying benchmarks that are replicable using NCRAS data via ODR (Chapter 6, Section 6.4.3), which cover the entire NHS population nationwide. This led to the identification of 12 potential oncology benchmark studies, of which only a single-arm from the RECORD-3 renal cell carcinoma (RCC) trial (sunitinib to everolimus)[86,405,406] were deemed potentially replicable with sufficient sample size using NCRAS data (Chapter 6, Section 6.6.7 includes details on sample size estimation). This shaped the design of a direct benchmarking RCC case study for a single-arm trial with NCRAS data, detailed in Section 7.5.4, Chapter 7.

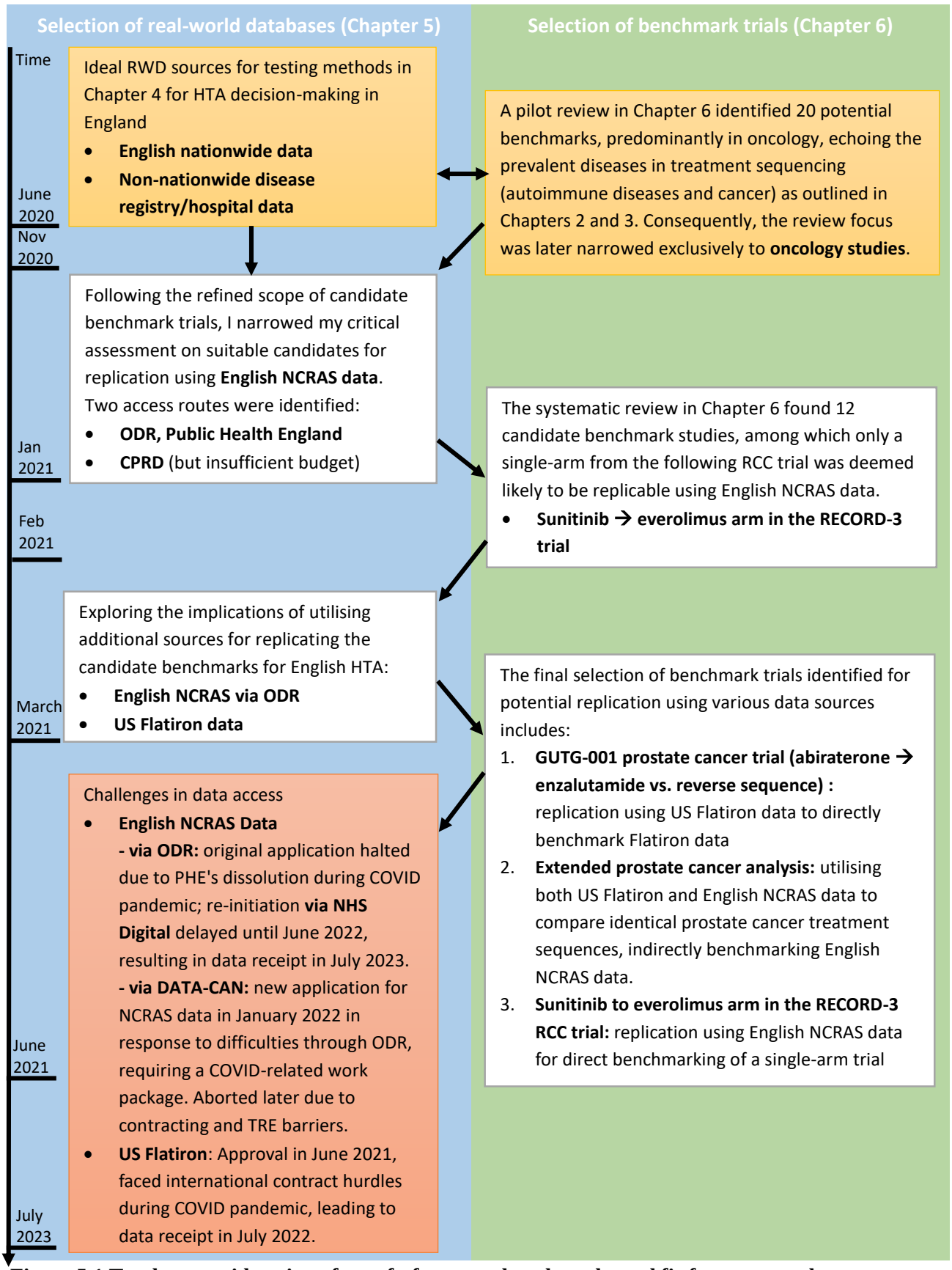


Figure 5.1 Tandem consideration of proof-of-concept benchmarks and fit-for-purpose data

CPRD: Clinical Practice Research Datalink; NCRAS: The National Cancer Registration and Analysis Service; RWD: real-world data; ODR: Office for Data Release

The blue-highlighted section represents considerations of candidate databases, while the green-highlighted section pertains to considerations of candidate benchmark trials. Boxes highlighted in yellow represent the initial broad considerations encompassing both candidate databases and benchmark trials. Boxed highlighted in orange outlines challenges in data access. Arrows in the figure indicate the flow of decision-making processes. Decision timing is outlined on the left of the graph.

Unfortunately, the sudden dissolution of PHE during the COVID-19 pandemic following my submission of the NCRAS data application to the ODR led to severe delays and belated redirection of my application to its new data custodian, National Health Service (NHS) Digital. During the transition period, I investigated emerging post-COVID data sources, specifically the Health Data Research Hub for Cancer (DATA-CAN)[407,408], a newly established NHS affiliated Trusted Research Environment (TRE) (i.e., safe haven) that houses extracts of NCRAS and other NHS Digital datasets. This exploration prompted me to revise the protocol, incorporating additional COVID-19-related study objectives (not covered in this thesis) to align with the consortium's requirement for studies that utilised their data access. As DATA-CAN was a newly formed consortium with internal uncertainties and prioritised projects, I also concurrently reapplied for NCRAS data through NHS Digital as soon as it became feasible. Hurdles of exploring these alternative avenues for accessing NCRAS data are detailed in Sections 5.3.1 and 5.4.

Several disease audits also hold extracts of NCRAS data, tailored to patients with specific cancers and enriched with additional information. For example, the United Kingdom (UK) National Prostate Cancer Audit (NPCA)[409-413] comprises standard NCRAS datasets and enriched with additional laboratory results from the Cancer Outcomes and Services Data (COSD)[414], such as the prostate-sensitive antigen (PSA) test results. The NPCA database has been a key resource in the NPCA's annual reports, tracking treatments and outcomes of prostate cancer, but is inaccessible to the public. The Lung Cancer Data Audit (LUCADA)[415,416] and the National Lung Cancer Audit (NLCA)[416] are the only publicly accessible, disease-specific audits, offering detailed lung cancer patient data that can be linked with standard NCRAS datasets. Nonetheless, they are restricted to particular cohorts of patients and may not provide enough longitudinal information for complex sequencing analysis when used alone.

Apart from NCRAS data, three major UK-based primary health care databases, CPRD[401,402], QResearch[417-419], and The Health Improvement Network (THIN)[420-422], also show potential for oncology research and have been used to generate RWE for NICE TAs.[423] They share common features but vary in coverage of practice groups and external data linkages.[423] Their greatest advantage lies in providing additional information on clinical test data[419,424] and prescription records beyond SACT, such as hormone therapy in treating cancers.[425] The linkage between these databases and registries are growing, notably CPRD and QResearch's links to the English Cancer Registry.[417,426] However, using them in my study poses several limitations. CPRD is publicly accessible but exceeds my budget.[404] QResearch, its use once restricted to certain affiliates, is now open to all UK researchers but unfortunately did not align with my PhD timeline.[427] Moreover, its linked cancer datasets only seem to be available on University of Oxford servers.[417,418] Regarding THIN, although it shows comparable standardised cancer incidence ratios to those in the UK registry for England and Wales[421],

there is no clear documentation on its direct linkage to the English Cancer Registry. Furthermore, these databases only cover a subset of patients in the UK (e.g. CPRD contains < 10% of the English population[403,426]), which reduces the likelihood of achieving an adequate sample size for my research. Therefore, I did not explore these databases further.

Meanwhile, challenges in obtaining timely access to NCRAS data and the absence of a suitable RCT for directly benchmarking using NCRAS data prompted me to explore alternative data source from another country — the US-based Flatiron Health database[85] (hereinafter "Flatiron database"). Consequently, I reassessed the 12 candidate benchmark trials, focusing on their potential replication using Flatiron data (Chapter 6, Section 6.4.3). This led to the identification of an additional prostate cancer trial, GUTG-001[84], as a suitable candidate for direct benchmarking with Flatiron data. This also enabled me to develop case studies jointly using both NCRAS and Flatiron data to indirectly benchmark RWD analyses. Additionally, the recent partnership between NICE and Flatiron Health, focused on investigating the potential of US data (which frequently offers earlier access to medications) to enhance NICE's TAs in oncology, highlights the relevance of Flatiron data.[83] Chapter 7 details the tandem analysis design for my prostate cancer case studies.

In the following section, I provide an overview and comparison of the two chosen databases, the NCRAS and Flatiron databases. Despite potential shortcomings of each selected database, such as the absence of lab results in NCRAS data via NHS Digital and insufficient history of health records prior to patient's initial cancer diagnosis in the Flatiron data, it remains crucial to test the feasibility of these databases. Testing them, rather than dismissing them for their known limitations, can highlight important areas for improvement.[428] This is particularly relevant in English HTA, where NCRAS data is instrumental in informing local drug reimbursements and CDF drug re-appraisals, yet its capability in producing reliable comparative effectiveness estimates of treatment sequences remains largely unexplored. In the next section and further in Chapter 8, I discuss the trade-offs of using these databases for deriving estimates of the treatment sequence effectiveness

5.3. Candidate data sources

5.3.1. English NCRAS database

NCRAS, a part of England's National Disease Registration Service (NDRS), coordinates the publication of aggregated information from cancer registries across England, Northern Ireland, and Scotland at the UK level[387], while regional authorities facilitate access to patient-level data within each country. NCRAS oversees the patient-level cancer registry data in England.[387] Researchers can now access the English Cancer Registry with linkage to a selection of non-cancer specific National Health Service (NHS) England datasets through NHS England's DARS (Data Access Request Service). The application was previously managed by the ODR, PHE.

Within the NCRAS database, the English Cancer Registry[82,387,388] and its linkage to the SACT dataset[389,390] are pivotal for my treatment sequencing case studies (Chapter 7). The English Cancer Registry provides audited data on patients' initial cancer diagnoses, covering primary cancers (i.e., excluding secondary malignancies like metastasis), along with details of basic patient characteristics (age, sex, date of diagnosis, cancer stage, performance status) and tumour prognostic factors (histology, morphology, tumour size, co-morbidities). Each patient may have multiple primary cancer records, where tumours develop at different locations, including bilateral tumours in cases such as kidney, breast, and lung cancer. While the English Cancer Registry itself includes the death date of patients (i.e., for estimating overall survival) and lost-to-follow-up due to relocation, additional information may be needed for other outcomes, such as indicators of disease relapse or treatment discontinuation.

The SACT dataset offers information on patients' anti-cancer treatments[389], detailing specific treatments, combinations, durations, and the prognostic characteristics (i.e., performance status) and outcome at the end of each treatment regimen. Additional datasets such as the National Radiotherapy Dataset (RTDS)[429], NCRAS-linked Hospital Episode Statistics (HES)[430,431] for admitted care (APC), outpatient (OP), and accident & emergency datasets (A&E), and the Cancer Waiting Time (CWT) dataset[432,433] can be linked to the English Cancer Registry[82,387,434], providing further time-varying details on patient cancer treatment journeys. NCRAS data also has the capability to link community-dispensed prescriptions[435] and COSD lab results[414], yet such linkage is not a routine feature of standard applications through the NDRS.[387,434]

The English Cancer Registry typically has a 2- to 3-year delay in data availability; for example, in 2023, data available only cover up to the 2021 cohorts. By the time that data are made available, the English Cancer Registry includes death records from the Office for National Statistics (ONS) and can be linked with the latest updates of SACT, RTDS, HES, and CWT datasets.[388] Consequently, the 2021 cancer registry cohort's records may contain associated treatment, hospital visit, and death data post-2021.

Figure 5.1 illustrates that I applied for NCRAS data using the same case study protocol via three distinct routes. Applications to ODR and NHS Digital, despite targeting the same range of datasets linkable to the English Cancer Registry, differed in their procedures. The latter required refilling submission forms with revised wording, adhering to varied privacy and data protection standards. Both offered one-off extracts of specified datasets (e.g., for particular cancer types as per protocol) and variables, downloadable to the data controller’s (university's) secure local virtual machine. Conversely, DATA-CAN, adopting a new data-sharing model, confines all parent data within a TRE, preventing transfer to local machines. Analysis costs are tied to Data Bricks usage, and Structured Query Language (SQL) is necessary for self-extracting specific study samples from the national

NCRAS data. During the analysis phase, only dummy data is available for visual preview, whereas the actual data analysis proceeds in the background and is not accessible for visual inspection. Initially tailored for COVID-19 research, DATA-CAN applications necessitate additional study aims aligning with the consortium's work packages, which are COVID-19 related. The primary advantage of DATA-CAN is that it includes a blend of audited and more recent data, thus providing updates fresher than traditional data-sharing models with linkages to potential extra datasets, such as community prescriptions. However, some dataset linkages necessitate additional cleaning. Importantly, its cost predictability and computational power flexibility are less clear compared to conventional data-sharing models. Additionally, the University of Sheffield (UoS) was less accustomed to this type of contract review as opposed to standard agreements. Considering these factors and the parallel advancement of different application routes (Section 5.4), I ultimately obtained NCRAS data through NHS Digital.

A full list of NCRAS datasets linkable with the English Cancer Registry and available for application is accessible online.[434] Researchers are required to specify the details of each dataset and variable needed for their study, which undergoes review by NHSD during the data application process. In my case study protocol (Chapter 7), I provide a detailed account of the variables and NCRAS datasets that I applied for in my research.

5.3.2. US-based Flatiron database

The Flatiron database comprises de-identified oncology-focused electronic health records (EHR) from over 1.6 million patients in the US.[85] It includes structured data such as patient demographics, diagnosis, lab results, treatment prescriptions/administration, vital signs, and cancer staging. Additionally, it features enriched data abstracted manually by trained abstractors from unstructured medical notes, making it a comprehensive resource for oncological research. Numerous recent studies have utilised Flatiron data to examine real-world treatment patterns and survival outcomes across various cancers, particularly aiding in the post-marketing surveillance of pharmaceuticals for the US Food and Drug Administration (FDA).[436,437]Moreover, the recent collaboration between NICE and Flatiron Health, aiming to examine how US data could contribute to NICE's HTA processes in oncology, underscores its significance.[83]

Accessing Flatiron's data is more straightforward than NCRAS data due to its established secure data-sharing process. Flatiron data scientists assessed and advised on the feasibility and limitations of using their data based on the protocol submitted, advising applicants on potential limitations and suggesting proxies for certain variables, wherever relevant. This is vital since Flatiron's data dictionaries are not publicly accessible, meaning available variables can only be inferred from previously published studies utilising Flatiron data at the protocol drafting stage.

Contrary to NHS Digital, which only delivers requested NCRAS variables at data transfer, Flatiron provides an entire disease dataset relevant to one's research. For example, the standard Flatiron metastatic prostate cancer (mPC) dataset includes all relevant variables needed for my study. This includes longitudinal data for all mPC diagnosed patients during their follow-up within the Flatiron Health Network of health care providers (community/academic), up to the most recent data cut. I did not request additional manual extractions from unstructured data beyond the standard Flatiron mPC dataset, such as detailed comorbidity data linked to claims information or manual abstraction of comprehensive progression data.[438] Significantly, Flatiron updates its data cut more frequently and with a much shorter lag (i.e., 30-day lag), offering nearly immediate access to current information. For instance, I received an updated data cut at the end of September 2023 with patient records up to the end of August 2023.

5.3.3. Comparison of selected databases

Table 5.1 provides a side-by-side comparison of the contents of the Flatiron mPC dataset (a disease-specific subset of the larger Flatiron database) and the NCRAS database. NCRAS datasets comprise the English Cancer Registry, SACT, RTDS, HES (A&E, APC, and OP), and CWT. Flatiron data features variables that largely overlap with those in NCRAS data, but they are structured differently across datasets and vary in specific details. Table 5.1 outlines the specific locations of each variable within these databases and their notable differences.

Details about the cohort definitions and items requested for my NCRAS data extraction are in my case study protocol (Chapter 7, Section 7.6). The following discussion mainly revolves around the differences between the two databases in prostate cancer research. This is because both databases are used in my prostate cancer case studies, while my RCC case studies rely solely on NCRAS (Chapter 7, Figure 7.1). Nevertheless, most of these insights are applicable to other types of cancer research. See Table 5.1 for details for the following discussions.

Table 5.1 Comparison table: Flatiron metastatic prostate cancer dataset vs. NCRAS database

Content	Item	Location in the Flatiron mPC dataset	Location in NCRAS database	Comparison
Patient demographics	Cohort Identifier	Cohort indicator table	Cancer Registry	The Flatiron dataset I obtained only includes metastatic prostate cancer patients. Conversely, the NCRAS dataset I obtained contains data on both kidney and prostate cancer patients. These groups can be distinguished using SITE_ICD10_O2 codes in the English Cancer Registry. The Flatiron mPC dataset exclusively includes patients diagnosed with metastasis, either initially or through progression. In contrast, the English dataset I obtained encompasses patients at various stages, including those with early diagnoses not yet progressed to metastasis.
	Patient ID	In all tables	In all datasets	Key identifier of individual patients: Patient IDs are pseudonymised in both Flatiron and NCRAS datasets enabling linking of patient-level data across all tables/datasets.
	Tumour ID	N/A	Cancer Registry, SACT	Key identifier of individual primary tumours: Available only in NCRAS data, this identifier distinguishes primary tumour diagnosis records from secondary ones, such as metastatic tumours originating elsewhere, and from bilateral tumours of the same organ. In the Flatiron prostate cancer dataset, Patient ID is equivalent to Tumour ID since prostate is not a bilateral organ. In NCRAS, Tumour ID can be utilised to link with tumour-specific SACT data. This is theoretically beneficial for distinguishing treatments for concurrent cancers occurring simultaneously. However, due to quality concerns, linking SACT with Tumour ID should be done in conjunction with quality variables (that flag whether precise matches between a patient and specific treatment records exist) and with an understanding of its limited availability.[439-441]
	Practice ID	Visits table	Cancer Registry, SACT	For Flatiron data, practice IDs are attached to records that have encounter or prescription records. In NCRAS, information on practices that initiate a patient's primary treatment for a particular primary tumour is available in the Cancer Registry. Additionally, practice information is available for each SACT treatment record. Practice IDs in Flatiron can be directly linked to a table indicating practice types (e.g. medical centre versus community), while NCRAS provide practice ID at varied levels, such as hospitals or trusts.
	Practice Type	Available for each Practice ID in the practice information table	Cancer Registry	Practice IDs can be linked to the Flatiron data's practice information table to deduce the practice type. Information on GP practice that made the referrals (route to diagnosis) for primary tumours is accessible in NCRAS, but I did not request this variable.

Practitioner/Consultant ID	Patient demographic table	Cancer Registry, SACT	Practitioner/consultant IDs are pseudonymised in both Flatiron and NCRAS data. While only the primary practitioner's ID is available in Flatiron, NCRAS separates practitioners for each treatment record and patient's initial cancer diagnosis.
Age at diagnosis	Patient demographic table	Cancer Registry	Both Flatiron and NCRAS data mask detailed age information. Age at diagnosis can be approximated in Flatiron data by subtracting the birth year from the diagnosis year, whereas NCRAS provides age at diagnosis for each primary tumour record.
Gender	Patient demographic table	Cancer Registry	No significant difference.
Race/Ethnicity	Patient demographic table	Cancer Registry, SACT	Ethnicity information in the Flatiron mPC dataset is provided as race (While, Asian and African American, and Others) and ethnicity (Hispanic versus non-Hispanic). In contrast, race/ethnicity in the NCRAS database is a single variable, but offers more nuanced levels compared to the Flatiron dataset (e.g. different subcategories of Asians/Asian British). Although several datasets in NCRAS contain the ethnicity variable, NDRS advises using the ethnicity information from the Cancer Registry (as it is has been audited based on information from a range of datasets).
Geographical region	Patient demographic table	Cancer Registry	Geographical region information is available as states in Flatiron data and as CCG codes, county codes, government office region, and country codes in NCRAS data. More granular details in NCRAS data like postcodes may be available upon special request. CCG codes are typically considered the most relevant indicator for regional health care differences in England.
Socioeconomic status	Socioeconomic status table	Cancer Registry	Both NCRAS and Flatiron datasets provide socioeconomic status/indices of multiple deprivation based on the place of residency/health care facility, categorised into quantiles.
Insurance coverage	Insurance information table	N/A	Flatiron's dataset details patients' insurance coverage, including the start and end dates and the payer category, while all patients in the NCRAS data are under NHS coverage. Despite having various insurance providers, patients in the Flatiron database often continue receiving care from health care providers (community/academic) within the Flatiron Health Network, even if they relocate, for instance, to a different state.
Embarkation flag	N/A	Cancer Registry	NCRAS data includes a marker indicating if and when patients left the UK.
Co-morbidity score	N/A, though it may have been available through Flatiron's add-on service[438]	Cancer Registry	The standard Flatiron mPC dataset I applied for does not include comorbidity scores, but Flatiron's add-on service can provide manual abstraction of comorbidity information, along with additional claims data.[438] Charlson co-morbidity scores for 2-year and 6-year periods are available in the English Cancer Registry, derived using the NCRAS default algorithm and HES data. Each score is accompanied with an indicator noting the availability of HES record linkage used in its derivation.

Cancer Diagnosis	Date of initial cancer diagnosis	Metastatic prostate cancer cohort table	Cancer Registry	In the Flatiron mPC cohort table, the date of patients initial prostate cancer diagnosis is recorded for prostate patients who had metastasis. However, not all patients were followed up at healthcare providers (community/academic) within the Flatiron Health network from the start of their initial diagnosis, and thus some patient's initial date of diagnosis might be unavailable. The English Cancer Registry offers multiple possible initial cancer diagnosis dates for each patient, along with an indicator specifying the best source for diagnosis date.
	Date of metastasis diagnosis	Metastatic prostate cancer cohort table	N/A (CWT)	In the Flatiron mPC dataset, each patient's metastasis date is confirmed using multiple methods (see Appendix 8.1), unlike in NCRAS datasets where these dates are not recorded. However, NCRAS's CWT dataset offers metastasis site data for specific treatment periods, potentially serving as a proxy for the date of progression to metastatic diagnosis in patients initially without metastasis. The reliability of this proxy variable remains uncertain.
	Group and TNM stages at initial diagnosis	Metastatic prostate cancer cohort table	Cancer Registry, SACT	Both NCRAS and Flatiron datasets provide patients' group stage (i.e., the summary of separate TNM stage) and each component of their TNM stages at initial diagnosis. The NCRAS dataset additionally distinguishes between imaging-based and pathology-based staging, specifying which method perform better for each patient, along with the reporting system. Additionally, each SACT record has staging information, though it is uncertain whether the staging is updated or the same as in the English Cancer Registry data.
	Histology at initial diagnosis	Metastatic prostate cancer cohort table	Cancer Registry	Both NCRAS and Flatiron datasets offer histology data for patients, each with varying levels of coding detail.
	Morphology and behaviour at initial diagnosis	N/A	Cancer Registry, SACT	The English Cancer Registry data includes morphology and behaviour codes for tumours using the ICD-O-2 system. Every SACT entry contains a morphology code, but it remains uncertain whether the staging details are updated or match those in the cancer registry data.
	Gleason score at initial diagnosis (prostate cancer)	Metastatic prostate cancer cohort table	Cancer Registry	Both NCRAS and Flatiron datasets provide Gleason score data.
	Additional tumour details at initial diagnosis	N/A	Cancer Registry	In NCRAS data, there are records of the number of tumours, indicator of large tumour, number of lymph nodes excised, size of the lesion at the time of primary treatment, whether the cancer is present in multiple locations, the margin of tissue excised around the tumour, and the grade or aggressiveness of the cancer. These are not available in the standard Flatiron dataset.
	PSA at initial diagnosis and metastatic diagnosis (prostate cancer	Metastatic prostate cancer cohort table	N/A	Flatiron data details PSA levels at initial and metastatic diagnoses when available. These testing results are not available in NCRAS datasets.

	CRPC status, date of castration resistance (prostate cancer)	Metastatic prostate cancer cohort table	N/A	Flatiron data includes a marker indicating whether mPC patients have developed castration resistance, along with the date of castration resistance if applicable. There is no such variable in the NCRAS database.
Cancer treatments	Cancer treatment details	Medication order table, medication administration table, specific treatment tables	SACT, CWT	The standard Flatiron mPC dataset include cancer treatment details from multiple sources, encompassing medication orders, medication administration, and treatment duration abstraction from unstructured data for specific treatments (e.g. abiraterone, enzalutamide). The medication order data contains the date of prescription, expected start date, dosage, planned cycles, and information on any cancellations. Medication administration records the actual date and dosage of received treatments, which is especially pertinent for non-oral treatments administered in healthcare facilities. Flatiron has pre-defined a range of anti-cancer treatments for a specific tumour (e.g. mPC) and maintains detailed records of those specified therapies. SACT documents all anti-cancer treatments, but poor Tumour ID linkage may hinder identifying treatments for specific tumours, especially in concurrent cancers. Diagnosis codes in SACT can aid differentiation. Unlike Flatiron, which draws from multiple sources, SACT's details originate solely from registration. Nevertheless, similar treatment details are provided, such as the decision date to treat, regimen drug/regimen groups, dosage, actual administration date, modifications, treatment intent, treatment cycle and OPCS treatment delivery code. It does not have planned cycle information but has numbered the cycle prescribed/administered. SACT records include the end date (version 2) and outcome of treatments (version 3), as well as information on whether patients experienced toxicity (version 3). However, the completeness of this data is unknown. SACT data can be enhanced with CWT information, which provides cancer treatment waiting times and reasons for delays. Each Flatiron medication administration record is flagged to indicate whether it is considered a maintenance treatment, while each SACT record includes an indicator of the treatment's intent, whether it is adjuvant or palliative.
	Drug episodes	Drug episode table	N/A	Flatiron uses a default algorithm to compile a drug episode table, tracking the start and end date of chemotherapy cycles or consecutive oral prescriptions by cross-referencing various data sources. For instance, the start and end dates of a chemotherapy cycle, initially captured in the medication order, may be augmented with administration data. It is unclear whether each SACT record represents an entire “drug episode” or if the data necessitates further refinement, such as merging records from similar times to construct episodes of combination therapy.

	Line of therapy	Line of therapy table	N/A	Flatiron's default algorithm compiles consecutive drug episodes, such as chemotherapy cycles, into a treatment line table. This table includes the start and end date of each treatment line and their sequence. While this provides a ready-to-use summary of treatment lines, researchers also have the option to use raw data from medication orders, administrations, or drug episodes to create custom definitions of lines of therapy. This flexibility is particularly beneficial when clinicians have varying opinions on what constitutes a line of treatment. There are no default line of therapy table in NCRAS data.
	Androgen deprivation therapy (prostate cancer)	Androgen deprivation therapy table	N/A	The Flatiron androgen deprivation therapy table includes details on whether patients receive androgen deprivation therapy, covering aspects such as the treatment setting (advanced versus curative), start and end dates, with some data retrospectively traced through data abstraction. However, the comprehensiveness of this data table is uncertain, primarily because not every patient received follow-up at health care providers (community/academic) within the Flatiron Health Network before undergoing metastatic treatment, while androgen deprivation therapy often started prior to metastasis. In contrast, the extent to which androgen deprivation therapy is captured in SACT data is uncertain. This therapy might be administered in a primary care setting and not consistently recorded in SACT, with practices potentially varying
	Primary treatment	Primary treatment table	Cancer Registry (treatment details)	In the NCRAS database, a treatment flag indicates whether patients receive radiotherapy, systemic treatment, or surgery, and is accompanied by OPC4 codes, codes for radiotherapy and imaging. Although an event date is provided, it is unclear whether this date represents the primary treatment following a cancer diagnosis. Additionally, this information is not available for all patients. The Flatiron dataset includes information on the primary treatment patients received following their initial diagnosis, detailing types such as radiotherapy and surgery. However, not all patients had this information as not all patients had follow-ups at health care providers (community/academic) within the Flatiron Health Network from the beginning of their initial diagnosis.
	Radiotherapy details	N/A	RTDS	The RDTS provides details on cancer patients' radiotherapy, including primary procedure OPCS and dates. It can only be linked to Cancer Registry data via Patient ID, but includes diagnosis ICD codes, aiding in identifying records for specific cancer treatments. Flatiron data lacks detailed radiotherapy information, except for a singular “primary treatment” table that notes if patients received radiotherapy as their primary treatment upon initial diagnosis of early-stage prostate cancer.

	Cancer waiting time	N/A	CWT	NCRAS includes the CWT dataset, detailing waiting times and possible reasons for delays during cancer treatment periods. Conversely, Flatiron does not provide comparable information on cancer waiting times.
Mortality	Date of death	Patient demographic table	Cancer Registry	Both NCRAS and Flatiron datasets include patients' death dates. Flatiron's dates are verified through multiple sources and may include confirmed deaths after lost follow-up, while NCRAS's death records are audited with ONS data. NCRAS also contain an indicator of whether a patient remain alive by the time of snapshot and provides detailed coding for causes of death.
Health care facility visits (longitudinal data)	Records for medical encounters	Visits table	HES A&E, HES APC, HES OP	Flatiron dataset capture all encounters (including tele visits) at cancer-focused facilities. This means that visit records from other facilities might not be included. Typically, patients begin visits to these specialised facilities following a cancer diagnosis, often resulting in incomplete records of comorbidities prior to their initial cancer diagnosis. NCRAS-linked HES, however, records all non-primary care NHS hospital visits, including A&E, hospitalisations, and outpatient visits, including those prior to their initial cancer diagnosis. Validation studies indicate that HES record completeness has improved in recent years, though completeness of outpatient records still presents challenges.[431,442] In Flatiron data, each visit can be linked to diagnoses, lab results, vital signs, or prescription records depending on the visit type, but lacks information on procedure codes. Conversely, NCRAS-linked HES provides codes of diagnoses and procedures for each visit/hospitalisation, but does not include detailed prescriptions, vital signs, or lab results.
Diagnosis (longitudinal data)	Diagnoses, diagnosis date, diagnosis system	Diagnosis table	Cancer Registry, HES A&E, HES APC, HES OP	See comments in "health care facility visits". Additionally, in both NCRAS and Flatiron datasets, each encounter may result in multiple diagnoses, each accompanied by a diagnosis date and a code indicating the diagnostic system (e.g., ICD-9 versus ICD-10). NCRAS includes an indicator of whether a tumour's registration is based only on a death certificate.
Lab results (longitudinal data)	Test, test date, test results	Lab table, NGS table	N/A	Lab and NGS test results in Flatiron dataset are exclusive to tests conducted at health care providers (community/academic) within the Flatiron Health Network. Test results from other facilities may occasionally transfer to the system, but this is not consistent for all patients. Each test record includes the test name, LOINC code, lab components, lab source, normal range, test dates, test results, and test units. The table contains both raw test results and those cleaned by Flatiron analysts. NCRAS datasets do not include lab test results. Access to PSA information through specialised linkage necessitates a customised request, typically not available to the public, as outlined in Section 5.2.

Vital signs (longitudinal data)	Test, test date, test results	Vitals table	SACT (only height and weight)	Flatiron data contains height and weight details, and a broad range of vital sign measurements, such as blood pressure, whereas SACT records only has weight and height at the start of each treatment cycle/regimen. Vital sign measurements in Flatiron are limited to health care providers (community/academic) within the Flatiron Health Network, with occasional transfers from other facilities, though not consistent for all patients. The Flatiron data includes both raw and analyst-cleaned results.
Performance status (longitudinal data)	ECOG value/performance status	Structured ECOG table, machine-learning augmented enhanced ECOG table	SACT	Flatiron data offers two unique sets of ECOG values. The first set consists of all longitudinal ECOG values derived from structured data (not necessarily on the same date of treatment initiation). The second set combines these structured data values with additional ECOG information abstracted from unstructured data sources (e.g. medical notes). However, this second set specifically includes ECOG values recorded from 30 days before the start of a treatment line to 7 days after the initiation of therapy. The definition of a treatment line is based on Flatiron's standard line-of-therapy algorithm. See the section on cancer treatments for further details. SACT reports patient performance status at the beginning of each regimen/cycle. However, there are discrepancies in how performance status levels are recorded across different versions of SACT, specifically between version 2 and version 3.[434]
Biomarker	Date of biomarker test and results	Biomarker table	N/A	Currently, the Flatiron dataset only contains information on BRCA mutations. This data collection has only started recently, driven by the emergence of new drug targets.

BRCA: The Beast Cancer Gene; CCG: Clinical Commissioning Group ; CWT: Cancer Waiting Time; ECOG The Eastern Cooperative Oncology Group; GP: General Practitioner; HES: Hospital Episode Statistics; HES A&E: Urgent and Emergency Care Activities; HES APC: HES Admitted Patient Care; HES OP: HES Outpatient Attendances; ICD: International Classification of Diseases; mPC: metastatic prostate cancer; N/A: not available; NCRAS: National Cancer Registration and Analysis Service; NGS: Next-generation sequencing; ONS: Office for National Statistics; OPCS: The OPCS Classification of Interventions and Procedures; SACT: Systemic Anti-Cancer Therapy Dataset;

5.3.3.1. Comparing Flatiron and English NCRAS databases for cancer research

Flatiron collects EHR data from health care providers (community/academic) within the Flatiron Health Network, while NCRAS captures nationwide cancer data across in England. Both databases record patient demographics (age, gender, race/ethnicity, geographical region, practice, practitioner, socioeconomic status), initial cancer diagnosis date, staging, tumour characteristics (e.g. histology), risk scores (e.g. Gleason scores), and anti-cancer treatment histories. Flatiron excels in providing detailed post-diagnosis data, including lab results and vital signs, not available in NCRAS. NCRAS, however, offers comprehensive tumour-specific details at diagnosis, including tumour morphology/behaviour, size, lymph node count, lesion size, and excised tissue margins.[434]

Flatiron's unique feature is its ability to transform unstructured medical data into standardised, usable information through validated manual abstraction and/or machine-learning.[443] A key example is its provision of specific data like metastasis and castration-resistance dates — typically unavailable in traditional cancer registries— for an extensive patient cohort.[437] Moreover, Flatiron's method of standardising data curation by pooling EHRs from various facilities is likely effective in reducing biases frequently encountered in RWE from EHRs[444], especially those that are single-facility-based. While crucial dates for identifying metastatic castration-resistant prostate cancer (mCRPC) patients (i.e., the target population in my prostate cancer case studies, see Chapter 7, Figure 7.1) are readily accessible in the Flatiron database, locating similar dates in the NCRAS data may rely on proxy algorithms. For instance, NCRAS's CWT data offers indicators for metastatic sites and treatment wait times, but the precision of these dates aligning with actual metastasis events and their consistent availability across different patient populations remains unclear.

5.3.3.2. Pre-diagnosis information

A limitation of the standard Flatiron database that I applied for is its potential underrepresentation of non-cancer-related patient visit records; such as details of comorbidities or information from before the cancer follow-up at associated facilities. It may be necessary for researchers to assume that if a patient's comorbidity significantly affects treatment decisions and outcomes, the relevant diagnosis codes would be captured during cancer-related visits, or appropriate laboratory tests would be conducted prior to treatment initiation. However, Flatiron's add-on service can provide customised manual abstraction of additional comorbidity information through linkage to claims data, offering a more comprehensive view of the patient's comorbidities—information not included in the standard Flatiron mPC dataset.[438]

In contrast, NCRAS data can be linked with HES data, allowing a retrospective examination of a patient's comorbidities at least six years before their cancer diagnosis (extendable with special data requests). This approach, however, is confined to structured diagnosis and procedural codes and only

accounts for visits within NHS facilities, thereby excluding private healthcare visits. The English Cancer Registry also calculates a Charlson comorbidity score using either 2-year or 6-year HES data, timed in relation to the cancer diagnosis. This score includes an indicator to denote the availability of the necessary HES data for these calculations.

5.3.3.3. Longitudinal treatment details

Both the Flatiron and NCRAS databases provide extensive longitudinal data on anti-cancer treatments with nuanced differences in terms of data sources and details recorded. Flatiron's treatment details include multiple sources, primarily drawn encompassing records of medication orders and administration. This includes prescription dates, expected start dates, dosages, planned cycles, and cancellation details. It also contains data from unstructured sources, mainly for specific treatments lacking administration records, detailing their durations, such as oral treatment abiraterone and enzalutamide. Each of Flatiron's disease-focused database are set up to maintain records of pre-defined anti-cancer treatments relevant to a specific cancer (e.g. mPC). They also provide basic information on primary treatment histories (e.g. types), including surgery and radiotherapy, though availability varies among patients.

Contrasting with Flatiron, the NCRAS database primarily sources its treatment details from registrations, yet it includes similar information. For example, the SACT dataset provide details on systemic treatments, encompassing details like the decision date for treatment, drug regimens, dosages, actual administration dates, modifications, treatment intent, treatment cycles, and OPCS delivery codes. The SACT dataset tracks prescribed or administered cycles but lacks planned cycle information. Its more recent versions further include treatment end dates (version 2), outcomes (version 3), and toxicity details (version 3), though the completeness and consistency of these variables remain uncertain. SACT data is further enriched by information in the CWT dataset, offering insights into cancer treatment waiting times and delay reasons. Additionally, the RTDS offers comprehensive radiotherapy information, a feature not included in the Flatiron mPC database, possibly because radiotherapy is less commonly used in the treatment of metastatic cancer. Furthermore, linked HES data within NCRAS allows for investigations into the dates and techniques of surgical procedures, with the specificity varying by surgery type.

Besides the “raw” treatment data, the Flatiron mPC dataset also features a “drug episode” table. This table integrates data from multiple sources of raw treatment data to better define the start and end dates of treatment episodes, such as augmenting prescription dates with the actual administration dates. Building upon the “drug episode” table, Flatiron further provides a standardised “line of therapy (LOT)” table. This table consolidates consecutive drug episodes, using rules confirmed with clinicians to differentiate treatment lines. For example, treatment changes without a gap may be

considered a new line, while re-initiating the same treatment after a discontinuation of more than 90 days is also classified as a new line. While Flatiron's LOT table comes standard and ready to use, it also provides the flexibility for customisation based on specific research requirements. Researchers can utilise either raw treatment data or the drug episode table to modify definitions of treatment gaps and establish their criteria for what treatments constitute a "line" of treatment. Due to potential incompleteness in Flatiron's pre-diagnosis data as mentioned in section 5.3.3.2, Flatiron analysts recommend that researchers could conduct sensitivity analyses limited to patients with less than a 90-day gap between their advanced diagnosis date (like metastatic diagnosis) and their first structured activity date (e.g. visits, prescription records). This specific sensitivity analysis addresses scenarios where a patient's LOT numbers in the Flatiron LOT table might be inaccurately marked, owing to missing immediate post-diagnosis treatment information (e.g. medication administrations or orders).

In contrast, the NCRAS database lacks a default algorithm for defining LOTs. Researchers are required to develop their own methods to consolidate independent treatment records into LOTs using the SACT dataset. PHE offers guidance for determining the duration of several oral and IV treatments for certain prevalent cancers, as specified in their CDF analytical methods guide.[445,446] Additionally, these working documents highlight the limitations and considerations of utilising SACT data for cancer treatment studies.

The Flatiron database and NCRAS SACT may not consistently provide information on non-anti-cancer treatments, such as therapies for bone metastasis relief (e.g., for hypercalcemia) or treatments for other conditions. This limitation can affect research focused on drug-drug interactions, combinations or employing particular drug use as an indicator of adverse events. For example, while abiraterone is approved and intended to be used in combination with prednisolone[127,447,448], these two databases may lack records of steroids like prednisolone. Furthermore, it is uncertain to what extent both databases cover anti-cancer therapies provided at primary care levels, such as androgen deprivation therapy for prostate cancer patients. While the Flatiron mPC dataset does contain a specific table for androgen deprivation therapy, the extent of its completeness can vary among patients.

5.3.3.4. Treatment outcomes

Both Flatiron and NCRAS datasets validate their death records with Flatiron using three different sources and NCRAS relying chiefly on the ONS records. Neither datasets, however, offers detailed data relating to precise disease progression dates for specific treatments/LOTs. Consequently, an estimated progression date may be proxied using the start date of a subsequent LOT or discontinuation of a LOT. This limitation could likely be improved by opting for add-on variables from Flatiron's service, such as comprehensive manually abstracted information on progression dates.[438] However, these are not included in the standard Flatiron mPC dataset.

The English Cancer Registry records the patient's death date precisely when available. In contrast, owing to confidentiality concerns, patient death dates in the Flatiron database are approximated to the month. In their previous research, Flatiron analysts have carried out sensitivity analyses on the effects of assigning a consistent imputation date (such as the 15th of each month) to all death records, finding that this approach typically does not present significant problems. However, this lack of granularity could potentially still pose challenges in accurately comparing outcomes between groups with very similar, or very short, survival times (e.g., 2 to 3 months).

5.3.3.5. Time-varying covariates

The diagnosis tables from the Flatiron mPC dataset and NCRA-linked HES records both provide structured diagnostic data, enabling the identification of comorbidities and condition-specific covariates through ICD-9/10 codes. Additionally, the Flatiron database provides further time-varying patient information, such as laboratory results (e.g. PSA levels), vital signs (e.g. blood pressure), and biomarker data, including BRCA in metastatic prostate cancer (mPC).

An oncology expert who provided clinical advice to inform my case study analyses, Dr. Carmel Pezaro, highlights the significance of ECOG (Eastern Cooperative Oncology Group) status as a key prognostic factor in patient treatment decisions and outcomes. The Flatiron ECOG tables and the NCRAS SACT dataset both contain time-varying ECOG performance status values. Performance status values in the SACT dataset are associated with the commencement of each anti-cancer treatment record. Notably, there are discrepancies in performance status coding systems across different SACT versions, particularly between versions 2 and 3.[434]

Flatiron, in contrast, offers two unique ECOG tables. The first includes all longitudinal ECOG values from structured data, not necessarily linked to the initiation of a treatment prescription or LOT. The second merges structured data with ECOG information from unstructured sources, such as medical notes, specifically focusing on the period from 30 days before to 7 days after the initiation of a LOT, following the default LOT algorithm in the Flatiron mPC dataset (see section 5.3.3.3 for more details). The incomplete ECOG data for certain patients in the otherwise comprehensive Flatiron ECOG table can present challenges in defining study cohorts. Flatiron analysts suggest that instead of limiting the selection to patients with ECOG scores of 0 and 1, researchers could consider excluding those with documented ECOG scores above 1. This approach operates under the assumption that patients without an ECOG value possess an adequate ECOG score for treatment.

5.3.3.6. Lost-to-follow up

Flatiron data analysts recommend using the patient's last structural activity, such as visits or prescription records, with or without a grace period, as the standard approach for determining the lost-to-follow-up date. This is because Flatiron, as an EHR database, collects data from medical

records, independent of the patient's insurance coverage or billing information. Flatiron Health possesses a broad network of connections. If patients remain with healthcare providers within the Flatiron Health Network, their records are transferable across different US states, regardless of their insurance status, ensuring continuity in their medical documentation.

This contrasts with the NCRAS database, where patient follow-up is maintained as long as they continue receiving care within the English NHS system, which has nationwide coverage. The embarkation (i.e., emigration) date in English Cancer Registry can be utilised to identify patients who are lost-to-follow-up when they leave the country. Yet, a key shortcoming of NCRAS is that it cannot monitor patients who seek treatment at private healthcare facilities. In 2012, 10.9% of patients held some form of private health insurance in England.[449] Data linkage to hospital databases could offer additional insights in to the impact.[450]

5.4. Data access: timeline and challenges

Figure 5.2, a two-page Gantt chart, illustrates the data application process for my PhD from late 2020 to September 2023. It outlines applications to four key data sources, represented by color-coded bars: blue for NCRAS data via ODR, turquoise for NCRAS via NHS Digital (replacing the ODR application), green for Flatiron data, and yellow for NCRAS via DATA-CAN. Bar lengths indicate task durations, and pink bars mark significant obstacles.

Despite having identified suitable databases and developed proof-of-concept benchmarking case studies, my PhD journey was marked by unexpected delays in data acquisition. My original application for NCRAS data via ODR[400] in 2021 was put on hold for over a year and then redirected to NHS Digital in June 2022 following PHE's dissolution (August 2021) during the COVID-19 pandemic. During this period, I initiated alternative NCRAS data access through DATA-CAN[407,408] facilitated through connections established with the University of Leeds by my primary supervisor Professor Latimer. However, this route later proved unfeasible due to technical and contracting barriers and was aborted at the end of 2022. Eventually, I gained NCRAS data access through NHS Digital in July 2023.

The application for the Flatiron mPC dataset was significantly more straightforward. However, access to the actual data extract was delayed by the prolonged review of international contracts by the UoS during the COVID-19 pandemic, leading to the receipt of an initial data-cut in July 2022. In the following sections I detail the major challenges encountered in accessing data, especially with de novo application as an early career researcher and outlines the strategies taken to surmount these issues.

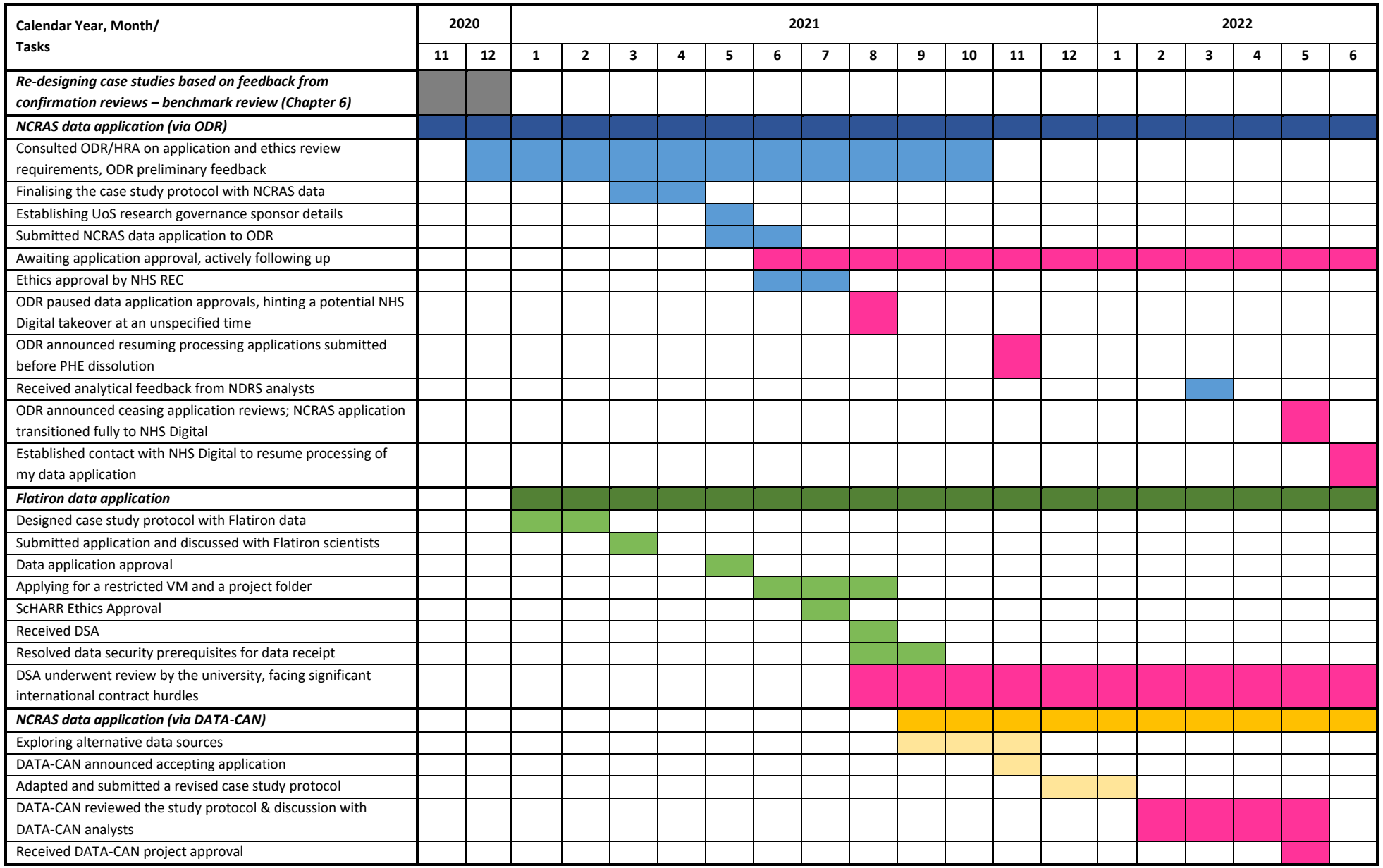


Figure 5.2 Gantt chart - timeline of efforts and hurdles to acquire data

Calendar Year, Month Tasks/Months	2022						2023								
	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9
NCRAS data application (via NHS Digital)															
Prepared and resubmitted NCRAS application to NHS Digital; iterative discussions followed															
Discussion with NDRS analysts post data table review															
Requested an honorary contract from the university as per NDRS's requirement, facing challenges with standard contract wording not accepted by NDRS															
Obtaining a Purchase Order number from the University for DSA generation															
Submitted privacy notice per NDRS request															
Applying for PhD funding extension from the Wellcome Trust															
Passed IGARD review and received DSA approval															
Revised data specifications based on NDRS analysts' requests during data extraction															
University set to sign off DSA; NDRS withheld approval due to email-related hurdles															
Awaiting data receipt, actively following up															
Obtaining a new Purchase Order from the University due to NHSD's changing vendor details and requesting an NHSD invoice															
Received data extract, errors found															
Obtained corrected data extract															
Received an additional corrected data extract following data destruction certificate hurdles															
Flatiron data application															
DSA approved issues escalated															
Received data extract															
Received an updated data-cut															
NCRAS data application (via DATA-CAN)															
DATA-CAN patient involvement meeting															
Received DATA-CAN service level agreement															
Faced University/DATA-CAN contract review hurdles and internal data hub shifting issues within DATA-CAN															
Aborted this line of data application															

Figure 5.2 Gantt chart - timeline of efforts and hurdles to acquire data (continued)

DSA, data sharing agreement; HRA, Health Research Authority; NCRAS, The National Cancer Registration and Analysis Service; NDRS: National Disease Registration Service; NHS, National Health Service; NDRS: National Disease Registry; ODR: Office for Data Release; REC: Research Ethics Committee; VM: virtual machine

Each row in the chart represents tasks required for obtaining data access, and each column corresponds to calendar months. Blue marks the progress in NCRAS data application through ODR, while turquoise shows progress with the same application, re-initiated with NHS Digital, the successor custodian. Green bars represent progress in Flatiron data application, and yellow denotes progress of NCRAS data application via DATA-CAN. Peachy pink marks significant hurdles encountered, and grey indicates the benchmark trial review (Chapter 6) for formulating the study protocol prior to data application.

5.4.1. Impact of the COVID-19 pandemic on data access

The decision to pursue the aforementioned datasets, confirmed upon my confirmation review, remained valid despite the challenges presented by the COVID-19 pandemic. Overall, the resources of ODR, DATA-CAN, and NHS Digital were severely constrained during the COVID-19 pandemic as public sector services. Understandably, research associated with the COVID-19 pandemic has been prioritised in the last few years. Specifically, the redirection of PHE and NHS resources towards addressing COVID-19 resulted in slower communications and unexpected pauses in NCRAS data application approvals, including a notable hold by the ODR in August 2021. Additionally, COVID-related staff absences in various departments led to further staggered delays, notably in processes requiring sequential steps (e.g. securing hierarchical approvals, waiting for regular meetings to schedule decisions, and engaging in iterative discussions between organisations).

Throughout the process, my supervisors and I actively worked to counter the delays in data access by rearranging study orders, adjusting thesis content, and exploring alternative data sources. However, we often encountered unreliable information and unexpected changes from external entities and the University, making planning and adjustments challenging. I took the initiative to escalate the issue and engage in higher-level communication, such as directing discussions to decision-makers within NHSD, particularly after my application shifted from NCRAS to NHS Digital. As an early career researcher, I found that involving my supervisors in email communications was sometimes crucial. Their direct involvement in correspondence often proved to be a decisive factor in receiving responses, particularly with new contacts. Despite these delays, the COVID-19 pandemic also encouraged data sharing within research consortiums, such as COVID-CVD[451], DATA-CAN[408], accelerating COVID-related studies and benefiting related research projects within scope (e.g. studies that are both oncology and COVID related).

5.4.2. Complex data application process

The data application process can be lengthy and complex particular for applications through ODR/NHSD, as highlighted by a recent study by Macnair et al.[452] and my own experience. My NCRAS data application was particularly delayed by the switch from ODR to NCRAS, obstacles in contract signings, the need for additional honorary contracts, and issues with incorrect data extractions. These challenges led to longer durations than the estimates provided in the Macnair et al. study.

Gaining a deeper insight into ODR/NHS internal processes and organisation structures may enable more accurate forecasting of data application timelines. For instance, with NHS Digital applications, while individual steps like the Privacy, Transparency and Ethics (PTE) review[453] and Independent Group Advising on the Release of Data (IGARD) review[454] are outlined online, fully

grasping their actual sequence often demands prior experience or specific inquiries with the application staff. The timeline outlined in Macnair et al.'s study offers a useful framework, yet real-world applications may show significant variation. During my NHSD application, I was informed about the standard NHS application stages: pre-application consultation, application preparation and submission (duration depending on the application), PTE team review (about 2 weeks) with pre-PTE check, IGARD review (2-4 weeks) with a pre-IGARD check, one's organisation (i.e., UoS) reviewing IGARD feedback (1-4 weeks), data sharing agreement (DSA) sign-off (1-2 weeks, varies on the availability of NHSD signatories), and NDRS data release post-DSA (1-4 weeks, depending on the data release team). My application's extended timeline was due to several key factors: (1) Adapting the original application from ODR to NHSD format required preparing additional information, notably publishing an additional layman's privacy notice online; (2) intricate requirements for student applications, even with supervisors as designated information asset owners (See Section 5.4.3.1 for details); (3) a prolonged DSA sign-off period (3 months), resulting from NHSD's unexpected rejection of a previously accepted UoS email address for performing the sign-off (See Section 5.4.3.2 for details); and (4) complications stemming from incorrect data extraction (See Section 5.4.4 for details). As a result, data is received about a year after initiating a new application with NHS Digital, rather than the suggested 4 months.

Repeat applicants, familiar with case-specific timelines from prior references in the application system, generally have an enhanced understanding. This knowledge is beneficial in drafting protocols that meet initial requirements and in identifying effective ways to expedite applications. Any small details in the application process can have substantial impacts on timelines, as these processes are composed of staggered tasks. This is especially relevant for early-career researchers who may be less familiar with university-specific procedures and contacts, impacting their ability to gather details efficiently. From personal experience, I benefitted from extensive support from my supervisors, NHSD application staff, and our Information Governance (IG) Manager, Dr. Amanda Loban, in formulating my application. Yet, increasing transparency in the application process and the application pack, both from the data provider and the applicant's organisation, could expedite data application for early career researchers and facilitate more efficient communication for both data applicants and providers, reducing the trial-and-error in developing applications with specific requirements, like legal compliance with the General Data Protection Regulation (GDPR).

5.4.3. Administrative challenges in contracting

In addition to the benefits of repeated personal or research team experience in data applications, an organisation's expertise in facilitating these applications, particularly contracting, is exceedingly vital. A recent study shows that some organisations access NHS data far more frequently, presumably

benefiting from their collective experience in this area.[455] This section explores my own experience of how the spectrum of an organisation's experience in this domain can markedly affect timely research data access.

5.4.3.1. Honorary contract challenges for student applications

During the pre-PTE check for my NCRAS data application, NHSD requested that I have an honorary contract from the university, providing templates from other universities for reference. This was a first-time request for UoS, which could only provide a standard honorary contract, leading to a pause in our application. To resolve this, our department's IG team suggested revising the application to include wording regarding UoS's data protection policies, demonstrating that these policies apply to regular and honorary staff, as well as research students. NHSD accepted the revision, recognising students using their data are explicitly required to adhere to UoS's universal data protection policies. This was achieved without the need for the University to create a new honorary contract. However, NHSD required an additional "signature block" on my honorary contract, which initially only featured the University's signature, despite including all my details. These additional steps underscored the complexity of the process, and it ultimately took five months to fulfil NHSD's expectations, which were somewhat vague and were not extensively publicly documented online. The substantial time spent navigating these regulations and coordinating with relevant parties significantly limited the time available for my actual research. This experience highlights the need for a University to develop a standard procedure for future student applications involving NHSD data or projects that require student data analysis.

5.4.3.2. Administrative challenges in NHSD DSA sign-off

NHSD rejected the official email address of the UoS research contracts team upon final DSA sign-off, demanding a personal official university email in line with their privacy requirements. This was despite UoS's longstanding practice of using this email address for previous NHSD applications. The UoS's team email system was used to facilitate job coverage within the team. As a result of this impasse, with both parties firmly upholding their positions, the DSA sign-off was put on hold for three months. In my particular instance, UoS eventually agreed to provide an individual's email address for the DSA sign-off, which could potentially indicate a shift in how the UoS might approach similar NHSD requests in the future.

5.4.3.3. Challenges of international contracting for Flatiron data

The Flatiron data application process, in contrast to the ODR/NHSD data application, was initially more straightforward due to fewer procedural steps, leading to quick progress at the outset. However, it soon faced a year-long delay post-approval. The complexity of signing international

contracts for sharing de-identified patient-level data across countries and COVID-related contracting staff absences significantly hindered the progress. The situation reached a resolution only after escalation to the Postgraduate Research (PGR) Committee and intervention by the Dean of ScHARR (now Dean of the School of Medicine and Population Health), Mark Strong.

This experience highlights the intricate challenges researchers face in navigating administrative processes, especially in the context of international collaborations and legal considerations regarding data sharing. The situation underscores the critical importance of robust administrative support within academic institutions to enable timely research.

5.4.3.4. Challenges in Trusted Research Environment data contracts

Much like the Flatiron case, the initial phase of the DATA-CAN application was straightforward, despite a justifiable three-month delay owing to DATA-CAN prioritising other urgent COVID-19 projects. Once our data request passed the DATA-CAN's Patient and Public Involvement (PPI) meeting, we promptly received the data usage contract template. However, DATA-CAN's innovative data sharing approach, deviating from traditional models, required a direct Service-Level Agreement (SLA) with individuals, contrasting with the UoS's more familiar general data-sharing contracts. The UoS contract team noted that this variance required additional time for further investigation. Given my concurrent NHSD NCRAS data application at that time, and the unpredictability of DATA-CAN's contracting process, I chose not to further pursue the DATA-CAN NCRAS data application in Autumn 2022. Meanwhile, my fellow PhD student Saleema Rex chose an alternate approach by navigating the contracting process and successfully signing a SLA with the University of Leeds, which hosts DATA-CAN, for her project. This route was particularly advantageous for her as she did not have a pending NCRAS data application when the ODR was closed, meaning she would not benefit from the prioritisation given to NCRAS data applicants when NHSD resumed operations. This scenario clearly demonstrates the complex and varied nature of data contract negotiations.

5.4.4. Incorrect NCRAS data extraction

Despite an expected 4-6-week timeline for receiving NCRAS data following DSA sign-off, it took over three months to obtain the extraction from NHSD. On receipt, a thorough examination — including comparing calendar years of data, creating Venn diagrams of PatientID overlaps between datasets, and contrasting cancer incidence rates with publicly available ONS statistics — revealed that NHSD had provided an incorrect extract (detailed in the preliminary data checks for the implementation of my case studies in Chapter 8, Section 8.2). NHSD promptly addressed this, citing a system change error during their transition from ODR to NHSD (Oracle to R), which led to the initial release of an incomplete data extract.

However, the revised extract still omitted the TUMOURID variable from the SACT data, a detail

explicitly mentioned in the DSA (see Chapter 7, Appendix 7.1). This omission, attributed to data quality concerns by NHSD analysts, was unexpected, as it had never been raised during the extensive two-year application process and iterative discussions. After detailed discussions and presenting references to acknowledge my understanding of the limitations of using this variable, NHSD agreed to include TUMOURID with appropriate quality indicators in a subsequent data update. The acquisition of an updated extract was further complicated by the need to produce a data destruction certificate, despite the old data essentially being a subset of the newly provided data.

Colleagues at the UoS suggest that this instance of incorrect NHSD data provision is not an isolated event. Consequently, it is imperative to perform an immediate and thorough examination of any data received (regardless of its source). Having a thorough prior knowledge of the data, particularly through comparisons with published statistics using similar datasets, can be helpful. Importantly, prompt communication with both the data provider and the IG Officer upon discovering discrepancies can expedite the resolution of any resulting impact.

5.5. Chapter summary

This chapter, in conjunction with the next Chapter (Chapter 6), investigated suitable RWD sources for conducting proof-of-concept studies aimed at generating unbiased comparative effectiveness estimates of treatment sequences for English HTA. It highlighted the English NCRAS and US Flatiron databases as key data sources, setting the groundwork for the design of direct and indirect benchmarking case studies in Chapter 7. Despite the limitations of each database, notably the absence of lab results and disease progression dates in NCRAS data, and the potential inconsistencies in the availability of health histories among patients within Flatiron data, the utility of these databases—especially the richness of their patient-level oncology longitudinal data—underscores their significance for treatment sequence studies. Case studies are necessary to verify if these databases, combined with the advanced statistical methods from Chapter 4, can be leveraged to provide reliable effectiveness estimates for treatment sequences and identify areas for improvement. Hence, along with Chapters 4 and 6, this chapter set the stage for Chapters 7 and 8, which delve into case studies and their implementation.

Additionally, this chapter outlined the potential challenges in database access from an early career researcher's perspective. It proposed that more transparent, publicly accessible timelines and concise flowchart guides from both data custodians and academic institutions could streamline the approval process and reduce excessive revisions. However, even in cases where publicly accessible timelines are not provided, as with the application to Flatiron Health, the process can be straightforward if all expectations are clearly communicated at the outset. This highlighted the importance of clear, early communication of requirements among all parties, addressing potential

nuances to avoid any hidden or ad-hoc requirements.

I recommend having a tailored data application starter pack, building upon existing procedures, to accommodate customised checklists for different applicant types (substantive staff versus student and others) and data scenarios (e.g., NHSD, domestic vs. international). This would simplify the process for both applicants and contract reviewers. Engaging senior legal experts early in reviewing contracts, especially for complex tasks such as international data transfers or updating Terms and Conditions for standard honorary contracts to incorporate data protection clauses required by custodians like NHSD could minimise delays. Specially, these tasks often require senior approval. Updating the UoS's honorary contract terms is currently not feasible, yet could aid sustainable research, especially if the lead data applicant leaves the university but needs data access for publication rebuttals.

Despite challenges, these experiences have contributed to the collective knowledge, such as the university's experience in handling international contracts for sharing patient-level data, proving advantageous for future similar scenarios. Sharing these collective insights in structured, yet informal settings within the community, could prove beneficial, fostering collective capacity, sustainability, and gains for research within similar contexts.

Chapter 6 Selecting benchmark trials for designing proof-of-concept Target Trial emulation case studies in treatment sequence comparison: A systematic review and feasibility assessment

6.1. Chapter overview

This chapter presents the rationale and methods underpinning the selection of suitable benchmarks for my proof-of-concept Target Trial emulation case studies (Chapter 7). These case studies are an integral part my PhD research, aiming to assess the feasibility of using real-world data (RWD) to determine the comparative effectiveness of treatment sequences.

The chapter begins by reiterating the significance of the Target Trial Emulation (TTE) Framework[3], previously introduced in Section 1.4.2, Chapter 1. This framework, forming the backbone of my case studies, aids in systematically designing real-world studies to reduce biases while tackling causal questions. Despite this, verifying the validity of results from RWD analysis can still prove challenging. Thus, this chapter delves further into how I design my case studies to enable methodological validation via *benchmarking* against well-established standards (i.e., randomised controlled trials (RCTs)).

Section 6.2 reiterates the importance of TTE in heath technology assessment (HTA) and expands on the development of benchmarking within the research sphere.[381,456,457] Section 6.3 underscores the value of benchmarking in my research and outlines the overarching strategies to best actualise it in my case studies. The process begins by selecting benchmark RCTs that exhibit a high(er) possibility to replicate (Chapter 6). Then, Target Trials are designed to mimic these chosen RCTs (Chapter 7). Given the proof-of-concept nature of my case studies, this pre-emptive selection is favoured over a post-hoc search for benchmarks after selecting topics for emulation, due to the uncertainty surrounding the availability of suitable benchmarks. At the end of Section 6.3, I clarify what constitutes "suitable benchmarks" for my research. Put simply, these should be existing trials that compare two or more treatment sequences which are likely to be replicable using the data sources outlined in Chapter 5 (Section 5.3) and the statistical methods identified in Chapter 4.

To identify the ideal benchmarks, I undertook a systematic review and an exhaustive feasibility assessment (Section 6.4 to 6.7). Section 6.4 elaborates the strategies employed to locate the most suitable benchmarks, starting with search methods used and how an oncology focus was determined from these searches (Section 6.4.1). This is followed by an evaluation based on the inclusion and exclusion criteria in Section 6.4.2. Post the initial evaluation, I proceeded to extract key features from the candidate trials for an in-depth assessment. Section 6.4.3 provides a comprehensive account of these feature, focusing on aspects related to the proposed databases from Chapter 5 (Section 5.3) and

local (e.g. within each country) treatment patterns (known as fit-for-purpose data feasibility assessment[385]). In Section 6.5, I present a brief overview of all studies included, while Section 6.6 details my in-depth assessment for each included study. Section 6.7 wraps up with my final choice of benchmark trials, and Section 6.8 delineates the disease demographics of them. Section 6.9 summarises this chapter and its connection to the rest of the thesis.

6.2. Benchmarking of Target Trials

6.2.1. Emulating Target Trials in HTA

Briefly, introduced in Chapter 1 (Section 1.4.2), the TTE Framework is a pragmatic guide enabling researchers to design real-world studies to answer causal questions while minimising biases stemming from study design and analysis.[3] This framework is gaining recognition among HTA agencies and regulatory authorities for its ability to increase the transparency and robustness of using real-world evidence (RWE).[37,78,172] Therefore, I adopted it as the foundation structure of my case study protocol (Chapter 7), aiming to assess the applicability of statistical methods from Chapter 4 in addressing treatment-sequencing questions using RWD.

Alongside the development of this thesis, a recent review by Gomes *et al.* explored the prospects of implementing the TTE framework in HTA[78], revealing parallels with my investigation. Despite the Target Trial concept being implicitly leveraged in fields such as epidemiology for over a decade, its application has been predominantly confined to the analysis of treatment effects to tackle confounding and improve result reporting in observational studies (e.g., The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline[458]). Only recently has a greater emphasis been laid on systematically structuring study designs from the outset to explicitly address causal questions using RWD (i.e., TTE), which is an emerging concept in HTA. Within this context, the design of my TTE case studies (Chapter 7) serve as valuable showcases for this approach in HTA.

Gomes *et al.* also highlighted additional hurdles in implementing TTE in HTA, including integrating causal inference methods, identifying appropriate data sources and their accessibility, and understanding the mechanism of data curation to address inherent limitations.[78] They further promoted the idea of "calibrating" TTE results against established RCTs—a version of "benchmarking" in epidemiology[457]—, to assess the capacity of using a specific RWD source to reproduce trial-mimicking results in pertinent populations. These facets align closely with the methodology investigation in my thesis. Precisely, in preceding chapters (Chapters 4 and 5), I explored relevant statistical methods and issues concerning data sources in the context of comparing treatment sequences in observational research. Moving forward, I will elaborate the significance of benchmarking in validating findings from observational studies, by presenting its background and relevant research (Sections 6.2.2), as well as its relevance to my PhD project (Section 6.2.3).

6.2.2. Benchmarking—assessing validity through comparisons

Benchmarking, a long-established method in epidemiology, involves comparing observational research findings with large, well-executed randomised trials to evaluate the reliability of the observational studies, given both types of evidence are available for a specific clinical question.[459-462] An example is the evaluation of myocardial ischemia risk associated with the diabetes drug, rosiglitazone. The manufacturer initiated the Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (RECORD) trial in 2000 to investigate if rosiglitazone increased cardiovascular events.[463-466] Before the RECORD trial concluded, numerous studies leveraged RWD to explore the same issue, results were, however, inconclusive.[467] The final RECORD results published in 2009, indicated that rosiglitazone did not significantly increase cardiovascular event risks. This finding aligned with those from two large US-based observational studies.[467,468] This suggests that properly conducted RWD studies could provide early safety signals in the absence of trials. Comparing the RECORD trial results with earlier RWD studies serves as a form of benchmarking, where the real-world studies precede trial results. However, the RECORD trial's open-label design, despite randomisation, led to doubts about data quality, potential sponsor bias, and data integrity.[467-469] Consequently, the European Medicines Agency (EMA) still decided to suspend rosiglitazone's market authorisation, while the United States (US) Food and Drug Administration (FDA) only restricted its use.[469] This highlights the importance of discussing the limitations of RCTs benchmarks.

Typically, researchers cross-check findings from completed observational studies through retrospective identification of existing RCTs when available (rather than creating one).[470] When results from observational studies using similar methods align with those of randomised trials across various medical topics, it not only builds additional trust in the methods themselves but also in the datasets we might want to use for addressing other clinical questions. The methods mentioned here broadly refer to strategies to mimic the trial procedure, such as defining the starting point of the study cohort (known as "time-zero")[356] and statistical methods to emulate the randomisation procedure (e.g., propensity score methods).[382] That is, before applying specific methods to unexplored areas, databases, or for generating novel hypotheses[382], it is crucial to confirm their ability to yield trial-like results. This validation process, primarily achieved through benchmarking, is particularly significant when intending to extend the methods' use in observational studies where randomised trials are not feasible.[382] This necessity stems from the fact that post-hoc searches are unlikely to yield available benchmarks in those areas. A recent review examining propensity score methods highlighted that explicitly attempting to emulate Target Trials could improve the benchmarking of observational studies.[382] In other words, by explicitly emulating the conditions of RCTs, observational studies can produce estimates that more closely align with RCT outcomes, thereby determining the ability to

generate unbiased estimates of treatment effectiveness from observational data by using validated statistical methods within a given database.

The RCT-DUPLICATE Initiative in the US is a leading entity focusing on benchmarking Target Trials.[471] They leveraged TTE benchmarking to assess the concordance between real-world effectiveness of medical products and regulatory RCT assessments. In their evaluations, they created multiple trial-mimicking populations by emulating a series of pre-selected RCTs.[171,472] Their findings indicate that RCT and RWE results might not always consistently align, which could be partly attributed to "unavoidably imperfect" emulation process, such as the absence of crucial variables in the database. Another article suggests that a pitfall of benchmarking RCT-mimicking RWE is the existence of an efficacy-effectiveness gap.[473] This gap refers to the difference between a treatment's effect in ideal trial conditions (efficacy) and its effect in a real-world setting (effectiveness). The term "effectiveness" is consistently used throughout the thesis, with the rationale, which is related to the scope of HTA, explained in Section 1.3.4, Chapter 1. In light of these challenges, a recent commentary underscored the merits of understanding and interpreting factors that might contribute to the discrepancies in estimates between RCT and RWD when similar findings cannot be replicated.[474]

It is not necessary for every RWE study to replicate a RCT as the utmost value of RWE lies in its capacity to address inquiries that cannot be sufficiently answered through trials alone.[475] Nevertheless, benchmarking RCT-mimicking RWE serves as a valuable means to validate methods and understand their limitation in relation to the data used before applying them in subsequent studies, especially in cases where there are no existing RCTs to validate the results. Specifically, this approach enables an explicit examination of potential caveats when validated methods are extrapolated to other generalised scenarios (e.g. in a different disease), thereby easing further adaptation.

6.2.3. The significance of benchmarking in my research

Chapters 4 and 5 respectively identified potential statistical methods for using RWD to compare treatment sequences and outlined potential RWD sources. Now, the fundamental question arises—can these methods be feasibly implemented within the identified databases, and can they be validated? This inspired me to have the idea of creating proof-of-concept case studies, employing the promising TTE benchmarking method (see Section 6.2.2). Such studies serve as a powerful tool for evaluating and scrutinising the applicability of these methods within specific databases, conditions and treatment types. At the same time, they provide a structured approach for an examination of the potential limitations. Specifically, the Target Trial design enables a comprehensive comparison that covers not only the results but also the procedures, thereby providing deeper insights into potential causes if the emulation significantly diverges from the benchmark trials—a critical facet of my intended investigation. Further, such efforts may also lead to identifying potential challenges when adapting

validated methodologies to more generalised scenarios, such as different diseases and databases.

In trials, the recorded variables are often crucial predictors of outcomes, with randomisation and blinding used to address the balance of potentially crucial unrecorded predictors. However, these variables may not always be present in RWD, depending on the purpose of the database. For example, the performance status of a patient is often not documented in claims databases as it may not be relevant for reimbursement purposes. However, it is vital in predicting a patient's outcome and thus may be necessary for the implementation of causal inference methods in RWD. In short, it is important to explore whether the existing variables within a chosen observational database are sufficient enough for generating trial-mimicking results. If significant differences arise between an RCT and its emulation, it suggests that additional information might be necessary to enhance the quality of RWE derived from the targeted observational database. By utilising the TTE benchmark design, I can delve deeper into these discrepancies and offer suggestions for potential improvements.

Despite the existence of ongoing large-scale initiatives like RCT DUPLICATE and OPERAND that aim to benchmark their findings with RCTs using a highly structured trial-patient-mimicking-design (i.e. TTE)[172,476], these studies have primarily focused on using US-based databases. Furthermore, as far as I know, these studies have not attempted to assess the feasibility and reliability of estimating the comparative effectiveness of treatment sequences using RWD. Therefore, creating proof-of-concept Target Trial case studies that replicate existing RCTs using UK observational databases could enhance insights from the aforementioned initiatives in two in two key aspects. First, conducting proof-of-concept Target Trial case studies with UK-based observational databases would provide valuable confirmation of their reliability in generating trustworthy RWE. Any significant discrepancies between RCT and RWE could help identify limitations and caveats associated with the databases, especially because there has been limited exploration of UK databases for such purposes. Second, these proof-of-concept studies would expand the knowledge of generating RWE specifically for treatment sequences. Theoretically, if validated, the methodology used in these studies could be adapted to assess the effectiveness of sequential treatments in populations not previously studied in clinical trials. This is particularly relevant when trial evidence is insufficient to compare treatment sequences, a major challenge of the HTA (Chapter 2 and 3). Findings could lead to refind methodology framework for obtaining comparative effectiveness estimates of sequential treatments using UK-based RWD to support health economic decisions.

In summary, there is significant value in designing proof-of-concept studies using the TTE benchmark technique—evaluating the viability of using RWE to inform comparisons of treatment sequences and determining the practicality of utilising UK databases for this purpose. The subsequent section, Section 6.3, will delve into how the benchmarks for my proof-of-concept case studies are identified and how they are employed in the design of the case studies.

6.3. Identifying benchmark sequential treatment trials for proof-of-concept studies

Section 6.2.3 highlights the importance of using existing RCTs as a benchmark assess the feasibility generating reliable effectiveness estimates of treatment sequences from RWD (i.e. proof-of-concept). This requires identifying suitable benchmark trials. Using multiple benchmarks has advantages in methodological development, as different diseases and stages might require different sets of variables for effective emulation. That is, having more benchmarks, when possible, can enhance the generalisability of the methods under investigation.

Hence, I carried out a systematic review in this chapter to identify possible benchmark trials that compared treatment sequences. The key requirement was exhibiting potentiality to be emulated using databases identified in Chapter 5—the English National Cancer Registration and Analysis Service (NCRAS) data and the US Flatiron. Additionally, patients in each trial arm should be randomly assigned to receive a specific treatment sequence, rather than just a single treatment. Despite some trials containing treatment sequence information, they may not suitably serve as benchmarks for proof-of-concept studies due to interruptions in randomisation caused by unplanned treatment switching. This can introduce a post-randomisation confounding bias, which often require additional adjustments to effect estimates when applied for health economics purposes.[43]

The following section, Section 6.4, describes the process and criteria of selecting potential benchmarks for designing the Target Trial emulation case studies for my project, focusing on the use of NCRAS and Flatiron databases. The reasons for this selection are detailed in Chapter 5 and Section 6.4.3. It is worth noting that NCRAS data can be accessed via multiple channels, as noted in Chapter 5. During my assessment, all NCRAS sources, specifically Office for Data Release (ODR), Public Health England (PHE) and Clinical Practice Research Datalink (CPRD), were considered. The feasibility of using NCRAS data from newly emerged post-COVID sources, including NHS Digital and DATA-CAN, which involve different administrative procedures, was evaluated ad hoc and separately in Chapter 5.

My review strategy partially aligns with an emerging concept of fit-for-purpose data identification, as advocated in the Structured Process to Identify Fit-For-Purpose Data (SPIFD) framework in 2022.[385] This framework provides clear strategies on how to determine the most suitable data sources for adequately addressing a clinical question for decision-making. While there are procedural similarities, the motivation behind my assessment diverges from the SPIFD framework. Specifically, I did not aim to identify the most optimal databases for a specific clinical question. Instead, I pre-selected a number of databases to assess their applicability for implementing advanced statistical methods identified in Chapter 4 to address treatment-sequencing questions.

It might seem atypical to identify data sources prior to defining a decision question. However, within the confines of my project, it is fitting, because none of the data sources have been tested to

answer causal treatment sequencing questions. Therefore, the ultimate aim was to explore whether any of the RWD sources can be used to produce rigorous evidence for treatment-sequencing in the future, which was identified as a knowledge gap in the technology appraisal (TA) review in Chapter 3. Therefore, I identify the trials that are mostly likely to be replicated under the selected databases rather than identifying the most suitable databases for emulating all identified trials (however I do compare the fit-for-purpose between the two databases for the same benchmark in my assessment.)

The benchmark RCTs serve as a methodology and database validation tool in this evaluation. Crucially, my aim was not to identify benchmarks that could definitely be replicated, but those with a good chance of replication. That is, I focused on finding benchmark RCTs that are likely replicable with the chosen databases, rather those that seem clearly unfeasible to replicate. If findings suggest that even those RCTs deemed highly replicable cannot be emulated, several implications may arise. These include the potential need for methodological adjustments, issues within the databases that require improvement (such as missing data, unavailable data variables or insufficient data time frame), or a disconnect between efficacy and effectiveness (see Section 6.2.2 efficacy-effectiveness gap). All these scenarios merit further investigation. On the other hand, if findings suggest that methods identified in Chapter 4 coupled with databases identified in Chapter 5 can be used to successfully mimic results of existing trials comparing treatment sequences, it underscores the reliability of RWE derived from the proposed methods. These considerations somewhat mirror a recent commentary by Morris and van Smeden[477], suggesting that as analysable data continue to grow, instead of deciding whether or not to conduct an analysis in a database, one should consider if enough data have been collected before launching an analysis.

6.4. Methods for identifying ideal benchmark sequential treatment trials

6.4.1. Searches and defining therapeutic focus

In November 2020, I conducted a preliminary PubMed search using the following strategy: (*"sequential therap*" OR "sequential treatment*" OR "treatment sequenc*" OR "therapy sequenc*" Filters: Randomised Controlled Trial*). The pilot review identified about 20 trials from the past five years that involved treatment sequences. Over half of these were oncology trials[84,86,174,175,478-483], with the remaining scattering across other therapeutic areas, including rheumatoid arthritis[484], hepatitis B[485], osteoporosis[486,487], weight loss[488], attention-deficit/hyperactivity disorder (ADHD)[489], psychosis[490] and H. pylori infection[491]. These areas broadly align with those identified in my in-depth NICE TA review in Chapter 3. Hence, I made a decision to focus exclusively on oncology examples for the remainder of the review, due to budget constraints and the intricate, time-consuming nature of the data application process as described in Chapter 5. This decision subsequently shaped the scope of my database evaluation within oncology (Chapter 5).

Lastly, the term "reverse sequence" emerged as salient during my pilot review. Therefore, the following search strategy was supplemented to the original one in January 2021, with the goal to capture additional relevant studies for a full systematic review: *("reverse"[Title/Abstract] AND "sequence"[Title/Abstract] Filters: Randomized Controlled Trial)*.

6.4.2. Eligibility assessment

To select suitable benchmarks, I focused on selecting RCTs where patients were randomised to receive two or more sequential treatment strategies (i.e. sequential treatment RCTs). The initial eligibility assessment began with a screening of titles and abstracts of the identified records against the criteria outlined in Table 1, where justifications for each criterion are also provided. This was followed by an assessment of full texts of the remaining articles using the same criteria. This process aimed to shortlist candidates for the next in-depth feasibility assessment, as outlined in Section 6.4.3.

Table 6.1. Inclusion and exclusion criteria of sequential treatment RCTs

	Justification
Inclusion criteria	
Randomised control trials (RCTs) designed to compare two or more sequential treatment strategies	This is the principal criterion for selecting appropriate candidates as benchmarks for designing my Target Trial case studies (details explained in Section 6.3)
Exclusion criteria	
(1) Studies published before 2010	Studies conducted before 2010 may not accurately represent recent NHS treatment patterns, particularly data on English cancer treatments only became accessible for research in the 2010s.[389]
(2) Non-randomised trials	The focus of my study lies in emulating RCTs.
(3) Non-oncology studies	I aim to employ oncology databases for my case studies, as explained in Section 6.4.1. In brief, a large proportion of RCTs comparing treatment sequences fall under oncology, and it emerged as a notable theme in my review of NICE TAs in Chapter 3.
(4) Studies not available in English	Unable to retrieve results for comparison with the simulated Target Trials.
(5) Results of the trial are unavailable	
(6) Trials with fewer than 45 patients in any of the sequential treatment arms	A trial with a small sample size may yield inconclusive results, whereas a Target Trial emulated under comparable conditions might reveal statistical significance due to the larger sample size inherent in observational data. This complexity can make benchmarking more challenging. Thus, trials with small sample sizes are not deemed suitable for my case studies.
(7) Studies that do not contain any sequential treatments	My study plans to focus on comparing two different treatment sequences.
(8) Trials that focus solely on variations of the same treatments across treatment arms, such as two treatments being used in the same order in two trial arms, but with different duration or dosage of each treatment.	
(9) Trials comparing identical or similar treatment combinations in a sequential versus concomitant manner (e.g. Treatment A > Treatment B versus concomitant use of Treatment A + Treatment B, or Treatment A > Treatment B versus concomitant use of Treatment A + Treatment C)	

(10) "Traditional crossover trials" In these trials, patients typically receive two different treatments in reversed sequences. The second treatment is administered with or without a washout period following the first, a design commonly applied to drugs with short-term effects.	These trials are designed such that each patient serves as their own control, thereby reducing residual confounding between two randomisation groups and to achieve statistical efficiency (i.e. fewer subjects are needed in this type of trial setting). These trials aim to assess the efficacy of two individual therapies rather than comparing different sequential treatment strategies, working under the assumption that there are no carry-over effects between sequentially administered treatments.[492] In other words, they do not effectively capture the sequence or carry-over effects between different treatments.
(11) Trials evaluating second, third, or subsequent lines of therapy, where the randomisation occurred in earlier lines of therapy	It is uncertain whether patients may have very different characteristics at the time of receiving a second- and third-, line therapy in an observational database compared to those included in the trials. Particularly if the only reference point provided in a trial is the time of receiving a first-line therapy. Thus, to simplify my methodological study, I have restricted my selection to studies examining newly diagnosed patients or those whose early lines of treatments can be traced (that is, being able to capture all treatments dating back to the point of randomisation).
(12) Trials that compare two different sets (bundles) of regimens, which contains "sequential therapies within each treatment cycle/course", such as bundled chemotherapy regimen or antibiotics that had fixed dosing schedule within a short period (e.g. cycles).	Patient characteristics are generally expected to remain relatively stable within the same cycle of a treatment regimen, and therefore, the switching of treatments within a cycle is not considered as changing lines of treatments. However, it is important to note that this criterion does not exclude trials with sequential regimens that have a fixed duration for each treatment that span over a long period, which could be several years.
(13) Trials without available time-to-event effectiveness estimates	Given the oncology focus of my case studies (section 6.4.1), I have chosen to prioritise outcomes commonly used in oncology trials and NICE TAs, namely time-to-event estimates like progression-free survival and overall survival. Moreover, Chapter 4 of the statistical methods review is specifically tailored to time-to-event outcomes, making it important to identify replicable estimates using the methods outlined therein. Additionally, quality of life measures may not always be available in national registry data.
(14) Trials that exclusively use quality of life measures as study endpoints or trials that do not evaluate treatment effectiveness as study endpoints	
(15) Trials assessing medical devices, surgeries, dental procedures, behavioral therapies, physiotherapies, biological mechanisms, or pharmacokinetics/pharmacodynamics	The main emphasis of this study is not on these interventions. This is primarily due to the uncertainty surrounding their feasibility for capture within the existing databases. For example, trials involving biosimilars typically investigate whether transitioning from a branded drug to a generic result in comparable efficacy. In real-life scenarios, biosimilars often substitute the original patented drug once it goes off-patent and practitioners usually do not use them sequentially with the patented drug unless there is a deliberate switch to a more cost-effective option when available.
(16) Studies of biosimilars and generic drugs, or drugs in different pharmaceutical forms	
(17) Trials incorporate treatments that have not been available within the NHS or through CDF	Trial emulation can only be effectively performed on treatments that can be captured and recorded within the existing databases.
(18) Trials that evaluate outcomes that occur within a day (e.g., hypoglycemia attack after insulin usage)	Assessing extremely short-term outcomes through claims databases or registries can be relatively challenging.

CDF: Cancer Drugs Fund; NHS: National Health Service; NICE: National Institute for Health and Care Excellence; RCT: randomised controlled trial: TA: technology appraisal

6.4.3. Data extraction and in-depth feasibility assessment

After shortlisting a set of trials, I conducted a comprehensive review to evaluate their suitability as benchmarks for my research. This involved extracting their key features, including the trial's objective, disease focus, patient characteristics, treatment sequences, timing of treatment-switching, and the percentage of patients receiving subsequent treatments. Any salient information from studies

cited in these articles, and trial registration details from ClinicalTrials.gov were also extracted wherever applicable, such as details from the same trials with different data cut-off points. These features were compared against local treatment patterns, available data variables, and the operational period of the two selected oncology-centric databases—English NCRAS and US Flatiron databases.

Originally, my objective was to identify benchmarks within the context of English data (i.e. NCRAS data). For this, I appraised each trial's potential by comparing their attributes with relevant NICE clinical guidelines[493-495], TAs, and the NCRAS data dictionary[434] in January 2021. This assessment involved verifying the availability of trial treatments within the NHS, the capacity of the NCRAS data to capture a sufficient sample size of the identical patient population, the presence of sufficient overlaps in patient characteristics between patients who received different treatment sequences (i.e. a principle to apply causal inference methods in Chapter 4), and the sufficiency of trial information to enable a comprehensive emulation. These assessments were primarily conducted through reviewing NICE TAs, clinical guidelines (current and historical ones, wherever relevant), published statistics, and NCRAS data dictionaries.

However, challenges related to data access (Chapter 5) and a lack of trials featuring more than one frequently used treatment sequence in England made it impractical to depend exclusively on English data for emulating "a pair of treatment sequences in an existing trial". This necessitated a shift towards non-English data. Thus, the same assessment was applied to the US-based Flatiron data, leveraging the US National Comprehensive Cancer Network (NCCN) guidelines (current and historical ones, wherever relevant)[496-499] and published studies utilising Flatiron data[85] as of February 2021. Since Flatiron's data dictionary was not accessible prior to the approval of the data application, the availability of data variables was inferred from published studies.

While initial eligibility (Section 6.4.2) did not exclude studies with long follow-up periods (e.g., median follow-up > 5 years), this was scrutinised during the in-depth assessment. This was necessary because certain key variables, such as biomarkers, might not have been included in the databases until recently. The limited timeframe of these variables could pose challenges in replicating longer-term trials accurately, especially if the missing variables are crucial prognostic factors for the disease. I evaluated these factors on a case-by-case basis, acknowledging that each disease has unique prognostic factors and that there may be circumstances where certain variables can act as proxies for others that are unavailable.

Lastly, I consulted with clinical oncologists—Professor Derek Rosario, Dr. Carmel Pezaro, and Professor Janet Brown—to interpret the findings from the most promising examples. They offered insights on how the treatment patterns and survival or other treatment response measures (such as radiology reports and biomarker levels) from these trials compared with their clinical experience.

After these consultations and further discussions with my supervisors, we made the final decision on which sequential treatment trials to use as benchmarks for my case studies.

6.5. Results

Among the 1,213 studies identified, 590 were excluded for being published before 2010. After applying the inclusion/exclusion criteria (Table 6.1), 12 studies were selected for in-depth assessment. Figure 6.1 displays the PRISMA diagram of the literature selection process. The shortlisted studies covered breast (n = 2), colorectal (n = 2), lung (n = 1), ovarian (n = 1), pancreatic (n = 2), prostate cancers (n =1) and renal cell carcinoma (n = 3). Only the breast cancer studies involved early-stage patients, while others focused on advanced or metastatic cancers. Table 6.2 summarises the extacted features from each study, including treatment sequences, patient characteristics, time of switching and cross-over rates (or rate of receiving a second treatment in a treatment sequence). Of these, only the two breast cancer studies and the AIO-PK0104 pancreatic cancer trial did not involve comparing a treatment sequence with its reverse (n = 3). Further analysis of each trial's benchmark potential is discussed in the subsequent section.

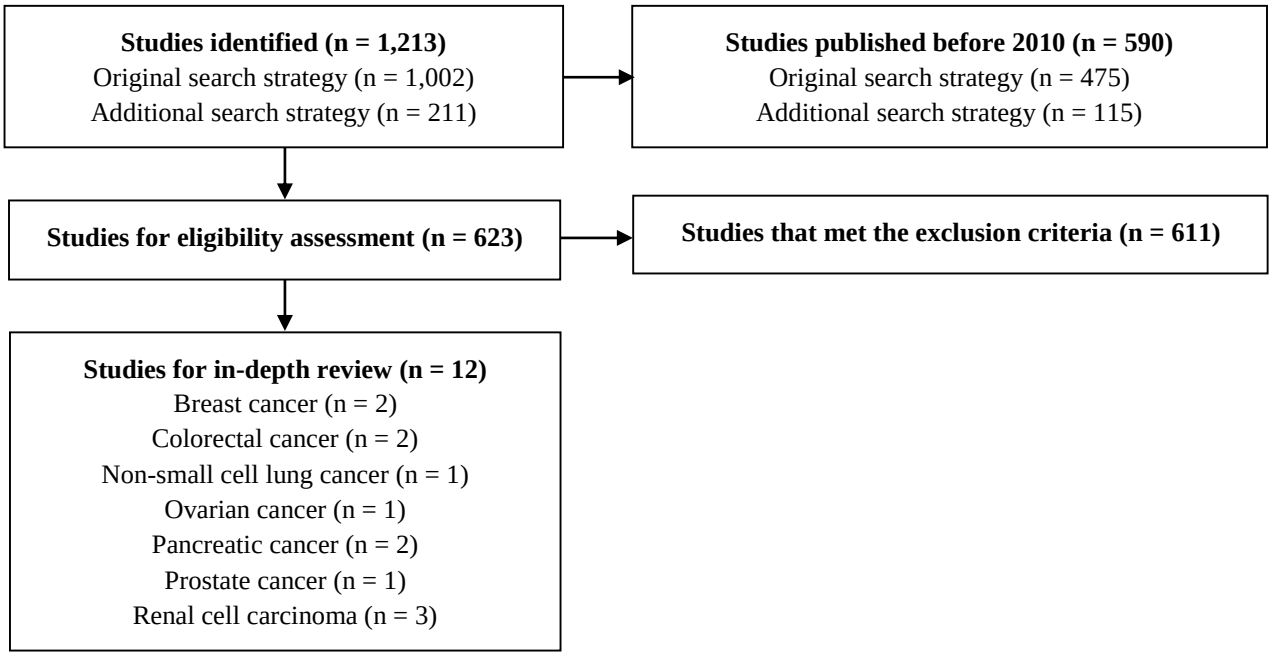


Figure 6.1 PRISMA diagram

Table 6.2 RCTs comparing treatment sequences

RCT identifier	Purpose	Patients	Treatment sequence 1	Treatment sequence 2	Time of treatment switching	Crossover rates*
Breast cancer						
BIG 1-98[173,328]	Compare the efficacy of tamoxifen and letrozole monotherapies or in sequence	Early-stage hormone receptor-positive breast cancer	**Sequence 1: Tamoxifen (5 years) (n=6182)** **Sequence 2: Letrozole (5 years) (n=1546)** **Sequence 3: Letrozole (2 year) → Tamoxifen (3 years) (n=1540)** **Sequence4: Tamoxifen (2 year) → Letrozole (3 years) (n=1548)**		Fixed time-line	25.2% patients receiving tamoxifen monotherapy selectively crossed over to receive letrozole prior to a disease-free survival event
TEAM Trial[324,500]	Compare the efficacy of exemestrane monotherapy or tamoxifen followed by exemestrane	Early-stage hormone receptor-positive breast cancer	**Exemestane (5 years) (n = 6120)**	**Tamoxifen (2-3 years) →exmestane (n = 9776)**	Fixed time-line (but with grace period?)	Not available
Colorectal cancer (CRC)						
REVERCE[479]	Compare the treatment sequence of regorafenib → cetuximab +/- irinotecan versus cetuximab +/- irinotecan → regorafenib	mCRC, KRAS exon 2 wild-type patients after failure of fluoropyrimidine, oxaliplatin and irinotecan	**Regorafenib → cetuximab +/- irinotecan (n = 68)**	Reverse sequence (n = 65)	Time of progression after the first treatment in the treatment sequences	68% versus 65%
COMETS[480]	Explore two different treatment sequences in order to define an optimal therapeutic strategy in KRAS wild-type colorectal cancer patients	Wild-type mCRC patients progressed after FOLFIRI/bevacizumab first-line treatment	**Irinotecan/cetuximab > FOLFOX-4 (as second and third line) (n = 54)**	Reverse sequence (n = 56)	Time of progression after the first treatment in the treatment sequences	54% versus 56%
Non-small cell lung cancer (NSCLC)						
GFPC 0504[501]	Evaluate two different treatment sequences in elderly NSCLC patients	Stage IIIB/IV NSCLC elderly patients (EGFR non-selected) (treatment naïve)	**Docetaxel/gemcitabine → erlotinib (n = 48)**	Reverse sequence (n = 51)	Disease progression during the first-line treatment (treatment failure)	54.2% versus 47.1%
Ovarian Cancer						
MITO-8[502]	Compare the efficacy PBC and NPBCc in two different treatment sequences in a second-line and third-line setting	Ovarian cancer patients who experienced disease recurrence or disease progression in 6-12 months after the last PBC	**PBC → NPBC (n = 108)** PBC: carboplatin + paclitaxel, or carboplatin + gemcitabine NPBC: PLD, topotecan, topotecan + PLD	Reverse sequence (n = 107)	Time of progression after the first treatment of the sequences compared	76.9% versus 69.2%

RCT identifier	Purpose	Patients	Treatment sequence 1	Treatment sequence 2	Time of treatment switching	Crossover rates*
Pancreatic cancer						
FFCD 0301[503]	Compare the efficacy of two different treatment sequences	Metastatic pancreatic adenocarcinoma (treatment naïve)	**LV5FU2-CDDP → gemcitabine (n = 102)**	Reverse sequence (n = 100)	Disease progression during the first-line treatment (treatment failure)	67.6% versus 55.0%
AIO-PK0104[504]	Compare the efficacy of two different treatment sequences	Advanced pancreatic cancer (treatment naïve)	**Gemcitabine + erlotinib → capecitabine (n = 148)**	Capecitabine + erlotinib → gemcitabine (n = 133)	Disease progression during the first-line treatment (treatment failure)	42.6% versus 57.8%
Prostate cancer						
GUTG-001[84] **(Khalaf 2019)**	1. Decide an optimal treatment sequence in prostate cancer 2. Evaluate the efficacy of second-line treatments	Metastatic castration-resistant prostate cancer without neuroendocrine differentiation (treatment naïve)	**Abiraterone + prednisolone → enzalutamide (n = 101)**	Reverse sequence (n = 101)	PSA progression	72% versus 74%
Renal cell carcinoma (RCC)						
RECORD-3[86,406]	Compare the efficacy everolimus > sunitinib and the reverse sequence	mRCC (treatment naïve)	**Everolimus → sunitinib (n = 238)**	Reverse sequence (standard sequence) (n = 233)	Disease progression during the first-line treatment (treatment failure)	55% versus 51%
SWITCH I[174]	Compare the efficacy of sorafenib > sunitinib and the reverse sequence	mRCC (treatment naïve)	**Sorafenib → sunitinib (n = 182)**	Reverse sequence (n = 183)	Disease progression during the first-line treatment (treatment failure)	57% versus 42%
SWITCH II[175]	Compare the efficacy of sorafenib > pazopanib and the reverse sequence	mRCC (treatment naïve)	**Sorafenib → pazopanib (n = 189)**	Reverse sequence (n = 188)	Disease progression during the first-line treatment (treatment failure)	56% versus 46%

CRC: Colorectal cancer; EGFR: Epidermal Growth Factor Receptor; FOLFIRI: folinic acid, fluorouracil, and irinotecan; FOLFOX-4: oxaliplatin plus leucovorin and 5-fluorouracil; -KRAS: Kirsten rat sarcoma viral oncogene homolog; LV5FU2-CDDP: 5-fluorouracil (5FU), folinic acid and cisplatin combination (V5FU2-CDDP); mCRC: metastatic colorectal cancer; mRCC: metastatic renal cell carcinoma; NSCLC: non-small cell lung cancer; NPBC non-platinum-based chemotherapy; PBC: Platinum-based chemotherapy; PLD: pegylated liposomal doxorubicin; PSA: Prostate-specific antigen ; RCC: Renal cell carcinoma; RCT: randomised controlled trial

* For comparisons that do not involve a reverse sequence as a comparator, it refers to the percentage of patients receiving a second treatment within each sequential treatment arm.

6.6. In-depth feasibility assessment

This section critically appraises the feasibility of each shortlisted trial as a benchmark, organised by cancer types. Within each cancer type, I first detail the specifics of the trials, then discuss their benefits and potential issues as benchmarks according to local treatment guidelines. Additionally, I discuss the practicability of replicating these trials using the chosen databases. Since the suitability of a trial as a benchmark can vary depending on the context (see Section 6.2), assessments are conducted separately for each database.

Pragmatically, the primary consideration is whether the treatment sequences of a trial are present in the specified database; absence of these sequences immediately rules out a trial. Subsequently, the focus shifts to assessing the adequacy of the sample size and the capability to capture the same outcome measures in the databases. Additionally, I evaluated the benchmark's quality by examining the detail in the reporting of findings and checking for any severe non-adherence or violations of post-randomisation treatment assignments. These issues may compromise its suitability as a benchmark and complicate the emulation of the trial. These factors do not necessarily disqualify a trial but are crucial in determining its relative suitability.

6.6.1. Breast cancer

Two large RCTs were identified with an aim of comparing multiple treatment sequences in managing early-stage hormone receptor-positive breast cancer patients.[173,500] The BIG 1-98 trial[173,328] compared the survival outcomes of the following four treatment sequences. Noted that the treatment arms with a single treatment, such as using Tamoxifen for 5 years, can be conceptualised as a sequence of Tamoxifen for 2 years followed by Tamoxifen for 3 years (see Section 4.4.1.3, Chapter 4).

- Sequence 1: Tamoxifen (5 years) (n = 6182)
- Sequence 2: Letrozole (5 years) (n = 1546)
- Sequence 3: Letrozole (2 year) → Tamoxifen (3 years) (n = 1540)
- Sequence 4: Tamoxifen (2 year) → Letrozole (3 years) (n = 1548)

The TEAM trial[324,500] compared the survival outcomes of two treatment sequences:

- Exemestane (5 years) (n = 4,868 (5-year analysis), n =3,045 (10-year analysis))
- Tamoxifen (2-3 years) → exemestane (2-3 years) (n = 4,898 (5-year analysis), n = 3,075 (10-year analysis))

Assessment based on NCRAS data

All treatment sequences included in the BIG 1-98 and TEAM trials appeared to be available within the NHS, as confirmed with UK clinical guidelines and clinical experts.[505] Moreover, the substantial sample sizes of both trials make them excellent candidates as benchmarks, as larger

samples typically result in smaller uncertainties in effect estimates. Nevertheless, additional input from clinical oncologists was needed to determine whether there were any significant differences among patients who received different sequences in English clinical practice. This is crucial since the application of causal inference methods relies on a certain degree of overlap between patient groups.

Despite their merits, using these two trials in my project presents unique challenges and limitations due to their focus on early-stage cancers. Specifically, methods validated through emulating these trials have limited generalisability to other cancers or diseases for the following reasons: Firstly, the discrepancy in results between an early-stage benchmark trial and its emulation inherently involve more uncertainties compared to those of late-stage benchmark trial and its emulation. This is because patients diagnosed at early-stage have a longer lifespan, allowing for more treatment lines. However, most treatment-sequencing trials typically compare only two treatment lines. That is, if real-world treatment patterns substantially deviate from those in the benchmark trial beyond the second-line, it could contribute to the discrepancy between the benchmark and its emulation. This obscures other crucial causes of discrepancy and their magnitude, which are vital for my investigation.

Secondly, the longer survival of early-stage cancer patients introduces competing risks from other diseases, such as cardiovascular outcomes, which can complicate the estimation of survival outcomes.[506] This factor is generally less prominent in late-stage cancers. Thirdly, unlike late-stage cancer treatments where the duration of each line of therapy (LOT) is often related to the progression-free time, the treatment duration in both trials is fixed (e.g., 2-3 years). In real-world practice, there may be other factors that influence treatment switching between tamoxifen and aromatase inhibitors (e.g., letrozole/exemestane) rather than switching after a fixed duration, but these require additional confirmation from oncologists. Additionally, the average duration of these treatments in England may differ from the design in the two trials. These factors make it challenging to replicate fixed-duration treatments using RWD, thus necessitating the establishment of a grace period for treatment duration discussed with oncologists.

Fourthly, the varying level of adherence to long-term treatments, as examined in the BIG 1-98 trial, can pose challenges for emulation.[328] In particular, the varied adherence behaviour of routine care patients and trial participants may also contribute to the discrepancies of effect estimates between a benchmark and its emulation, thus introducing additional uncertainties. Furthermore, a significant number of ad-hoc analyses have been carried out for these trials, such as adjustments to account for non-randomised sequential treatment. To ensure a thorough comparison between the benchmark and its emulation, it may be necessary to replicate these ad-hoc analyses as well. This further adds to the complexity and challenge of replicating the findings.

Lastly, considerable concerns arose regarding data availability due to the long duration of these

trials, which have published findings with follow-up ranging from 5 to 15 years. Data from NCRAS, accessible through the ODR, may offer a quality four-year follow-up (2014-2018, the latest available at the time of the assessment) for the patient cohort diagnosed in 2014, given that the SACT dataset for tracking cancer treatments was only established in 2014. On the other hand, the CPRD coupled with NCRAS linkage provides a longer follow-up for patients that are likely to be managed by primary care, starting from 2002. This includes early-stage breast cancer patients and can yield a sample size of approximately 750 patients per year.[506] Moreover, a recent study confirmed that breast cancer patients who received either aromatase inhibitors or tamoxifen from 2002 to 2015 can be identified via CPRD with NCRAS linkage. Specifically, these treatments are recorded in primary care prescription dataset, which, however, is not included in the NCRAS data accessible through the ODR. To summarise, there are substantial challenges and potential issues with generalisability when attempting to replicate these two breast cancer trials. However, it is important not to dismiss their value as benchmarks in situations where better options are not available. For example, if restricted by the limited available period of data, earlier data cuts of these trials (if available) or earlier segments of Kaplan-Meier curves from these trials could be utilised for comparison. Hence, while emulation may be feasible, it could be compromised, making the aforementioned trials not the "first choice".

Assessment based on Flatiron data

No early-stage breast cancer studies by Flatiron were discovered during assessment. However, the NCCN guideline suggest tamoxifen and aromatase inhibitors (letrozole/exemestane/ anastrozole) have been used as adjuvant theory for early-stage breast cancer in the US.[496] These treatments typically span 5 to10 years, depending on the patient's menopausal status at diagnosis and the duration of the first treatment (i.e., tamoxifen or an aromatase inhibitor). The treatment guideline indicated that treatment patterns observed in the BIG 1-98 and TEAM trials may potentially be observed in Flatiron data. Nevertheless, apart from treatment availability, other concerns in conducting a target trial emulation for early breast cancers using NCRAS data are likely to apply when using Flatiron data. Moreover, drawing from other Flatiron publications[85], it appears that Flatiron data is only accessible starting from 2013, which may not offer a sufficient timeframe for emulating these trials.

6.6.2. Colorectal cancer (CRC)

Two trials examining different treatment sequences in metastatic CRC (mCRC) were identified as potential benchmarks.[479,480] The first trial, REVERCE, enrolled Japanese patients who had Kirsten-rat sarcoma virus (K-RAS) exon 2 wild type mCRC and had experienced treatment failure with fluoropyrimidine, oxaliplatin and irinotican (FOLFIRINOX).[479] The patients were randomly assigned to receive either the specified treatment sequence or its reverse, with second-line treatment given upon disease progression.

- Regorafenib → cetuximab +/- irinotican (n = 68)
- Cetuximab +/- irinotican → regorafenib (n = 65)

The other trial, COMETS, is a phase-3 trial conducted to investigate the optimal treatment pathway for K-RAS wild-type mCRC patients.[480] A group of Italian patients who experienced progression after receiving a first-line treatment of folinic acid, fluorouracil (5-FU), and irinotecan (FOLFIRI) + bevacizumab were randomly assigned to two different treatment sequences for their second- and third-line treatments:

- Irinotecan + cetuximab → oxaliplatin plus leucovorin and 5-FU (FOLFOX-4) (n = 54)
- FOLFOX-4 → irinotecan + cetuximab (n = 51)

Assessment based on NCRAS data

The REVERSE and COMETS trials do not appear to be suitable benchmarks within NCRAS data. Firstly, bevacizumab has not been commonly used in the NHS, while cetuximab has been recommended alongside FOLFIRI/FOLFOX instead of irinotecan. Furthermore, regorafenib has not been recommended for use as the TA of regorafenib was terminated without any company submission. Finally, the potentially limited sample size and the inclusion of non-treatment naïve patients in these two trials present further challenges for conducting trial emulations within the scope of my PhD.

Assessment based on Flatiron data

A recent study analysed the treatment patterns of 14,315 CRC patients from 2013 to 2018 using Flatiron data.[507] Patients included in the analysis were identified based on receiving at least one line of systematic anti-cancer treatment. The study indicates that FOLFIRINOX was not used as first-line therapy in the US, making it impossible to emulate the REVERCE trial where patients received FOLFIRINOX as their first-line treatment prior to randomisation. Alternatively, according to the same publication, it is possible to observe treatment sequences from the COMETS trial in the US setting. Among the first-line regimens, FOLFIRI + bevacizumab was the fourth most commonly used (10.9%, n = 1554). Irinotecan + cetuximab (2L, n = 220; 3L, n = 144) and FOLFOX (2L, n = 338, 3L, n = 103) can both be utilised as both second-line and third-line therapies. However, it remains uncertain how many patients experienced failure with FOLFIRI + bevacizumab as a first-line therapy before these treatments. Therefore, there may be a small number of patients who follow the targeted treatment sequences of FOLFIRI + bevacizumab → irinotecan + cetuximab → FOLFOX or FOLFIRI + bevacizumab → FOLFOX → irinotecan + cetuximab.

In summary, emulating the REVERCE trial is impossible using Flatiron data since FOLFIRINOX was not a standard treatment regimen in the US. While it may be possible to replicate the COMETS trial, the diverse range of regimen choices for CRC is likely to result in an insufficient

sample size for the comparison of targeted treatment sequences. Consequently, it is also not an ideal choice as benchmark.

6.6.3. Non-small cell lung cancer (NSCLC)

The GFPC 0504 trial[501], which is a phase-2 trial in elderly patients (aged 65-89) diagnosed with stage IIIB/IV NSCLC, was identified as a potential benchmark. The trial aimed to compare the effectiveness of the following two treatment sequences in treatment naïve patients:

- Docetaxel or gemcitabine → erlotinib (n = 48)
- Erlotinib → docetaxel or gemcitabine (n = 51)

Assessment based on NCRAS data

The GFPC 0504 trial has a relatively small sample size and it is unclear whether docetaxel/gemcitabine (chemo-monotherapy) and erlotinib (an epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor (TKI)) have been used interchangeably as first- and second-line treatments for NSCLC elderly patients in England. This uncertainty stems from the fact that the standard front-line chemotherapy regimen for NSCLC in the NHS previously consisted of platinum doublets (a combination of platinum drugs and other chemotherapy drugs) for wild-type patients, irrespective of age, before the introduction of cancer immunotherapies (e.g. pembrolizumab).[508] Additionally, the study was conducted in the mid-2000s when the treatment paradigm differed from today's standards. As such, the trial did not differentiate between patients with positive and negative EGFR mutations, which is in contrary to the current practice of using EGFR mutation testing to determine the use of an EGFR TKI.[509] As a result, it may be challenging to identify patients who were given erlotinib as a first-line therapy without an EGFR mutation from recently collected data (i.e., after 2015). This limitation makes GFPC 0504 an inappropriate candidate for benchmarking.

Assessment based on Flatiron data

A recent study conducted by Li et al used Flatiron data to investigate how NSCLC patients with EGFR mutations (n = 1,564) were managed in real-world settings.[510] The study revealed that erlotinib (n = 593) and chemotherapy (n = 169) have been used as first-line and second-line treatments for patients with EGFR mutations in the US. Of those who initially received erlotinib, 91 individuals subsequently underwent chemotherapy as a second-line therapy. In contrast, among patients who initially received chemotherapy, 74 individuals later received erlotinib as a second-line therapy. However, it remains uncertain whether the specific chemotherapy regimen utilised in the GFPC 0504 trial (i.e., docetaxel or gemcitabine) was commonly used in the US, as Li et al's analysis combined different chemotherapy regimens for their evaluation. Furthermore, the challenges associated with

non-EGFR stratified patients in the GFPC 0504 trial are likely to persist regardless of the data source used. Furthermore, the GFPC 0504 trial participants were predominantly elderly patients, indicating a highly selective cohort. Consequently, it is unlikely to replicate the conditions of the GFPC 0504 trial using Flatiron data due to the limited sample size, which is expected to be fewer than 91 and 74 patients in each treatment group.

6.6.4. Ovarian cancer

The MITO-8 trial, which was carried out among ovarian cancer patients, was identified as a potential benchmark.[502] The MITO-8 included patients who had experienced a recurrence or progression of the disease within 6-12 months after their last treatment with platinum-based chemotherapy (PBC). These patients were then randomly assigned to two different treatment sequences for their subsequent therapies:

- PBC → non-PBC (NPBC) (n = 108)
- NPBC → PBC (n =107)

Assessment based on NCRAS data

In the MITO-8 trial, the majority of patients were initially given carboplatin + paclitaxel as the PBC, and pegylated liposomal doxorubicin (PLD) as NPBC. However, due to drug shortages, the trial design was amended to include carboplatin + gemcitabine as an additional PBC option, and topotecan or topotecan + PLD as alternative NPBC options. Although NICE did not endorse the use of the following regimen: carboplatin + gemcitabine, topotecan + PLD, and topotecan alone, for treating the initial recurrence of platinum-sensitive ovarian cancer[59], these alternative treatments accounted for a relatively smaller portion (0-25%) within each treatment line in the MITO-8 trial compared to the originally intended treatments (carboplatin + paclitaxel, and PLD alone). Considering the number of treatment-naïve stage III/IV ovarian cancer cases in England (less than 3000 per year) and the specific eligibility criteria of the MITO-8 trial (restricted to platinum-sensitive patients who have failed at least one PBC agent), which further narrows down the sample size, it may not be the most suitable benchmark unless there are no alternative options. Obtaining input from clinical oncologists was crucial to better understand the landscape of ovarian treatment in England, enabling a conclusive decision. However, no further consultations were pursued on this topic, as more suitable benchmarks were identified by the end of the assessment.

Assessment based on Flatiron data

No ovarian cancer studies using Flatiron data were found. The NCCN guideline recommends carboplatin + paclitaxel as first-line therapy, which matches the inclusion criteria of the MITO-8 trial.[497,502] For platinum-sensitive disease upon recurrence, carboplatin + gemcitabine with or without

bevacizumab is recommended, while liposomal doxorubicin (e.g., PLD) or topotecan can be used as monotherapy for platinum-resistant patients, but not in combination. In short, it seems feasible to capture treatment sequences from the MITO-8 trial using Flatiron data. However, there might be limited overlap between patients receiving different second-line treatments (the first treatment position in MITO-8) in the US, as PBC is typically indicated for platinum-sensitive disease and NPBC for platinum-resistant disease. This lack of overlap poses challenges for implementing causal inference methods that rely on sufficient overlap (Chapter 4). Thus, the MITO-8 trial might not be an ideal benchmark within the Flatiron data. Further confirmation from oncologists was not sought, as more suitable benchmarks were identified by the end of the assessment.

6.6.5. Pancreatic cancer

Two pancreatic trials comparing the efficacy of sequential treatments were identified.[503,504] In both trials, patients received a second-line treatment when they experienced disease progression during the first-line treatment. The FFCD 0301 trial assessed the following treatment sequence and its reverse counterpart in treatment naïve patients with metastatic pancreatic adnenocarcinoma[503]:

- 5-FU + folinic acid + cisplatin (V5FU2-CDDP) → gemcitabine (n = 102)
- Gemcitabine → V5FU2-CDDP (n = 100)

The AIO-PK0104 trial investigated the efficacy of the following two treatment sequences for patients with treatment-naïve advanced pancreatic cancer[504]:

- Gemcitabine + erlotinib → capecitabine (n = 148)
- Capecitabin + erlotinib → gemcitabine (n = 133)

Patients in both trials received a second-line treatment at disease progression during the first-line treatment.

Assessment based on NCRAS data

It is unclear whether the chemotherapy regimen LV5FU-CDDP (containing cisplatin) has been used for pancreatic cancer within NHS. Furthermore, the NICE pathway for pancreatic cancer does not include cisplatin-based therapy[511], suggesting that the treatments administered in the FFCD 0301 trial[503] may not be reflected in the NCRAS data. It is also unknown whether erlotinib can be administered in combination with chemotherapy in an English healthcare setting. Considering the evolving treatment landscape in pancreatic cancer and the potential to capture a promising sample size from the AIO-PK0104 population through NCRAS data[504], saw it as a potential option as a benchmark. However, further consultation with clinical oncologists is needed, if preferred alternatives are not identified.

Assessment based on Flatiron data

A recent abstract presented at the American Society of Clinical Oncology (ASCO) conference investigated the management of pancreatic cancer using Flatiron data.[512] O'Reilly *et al.*'s abstract revealed that the treatment regimens employed in the FFCD0301 and AIO-PK0104s trials were not commonly utilised.[503,504] Additionally, the V5FU2-CDDP regimen used in the FFCD0301 trial is not included in the NCCN pancreatic cancer guideline, rendering it impractical to replicate this trial using Flatiron data.[499] Although all treatments in the AIO-PK0104 trial[504] (i.e., gemcitabine + erlotinib, capecitabine) can be used as both first-line or second-line therapies according to the NCCN guideline[499], their use was restricted to specific patient subsets and were not routinely used. For instance, capecitabine monotherapy was reserved for patients with poor performance status, while gemcitabine + erlotinib was used as a second-line therapy only for patients who had previously undergone a fluoropyrimidine-based therapy. In summary, emulating the AIO-PK0104 trial using Flatiron data is challenging due to potential limitations associated with small sample sizes and limited overlap among patients receiving different treatment sequences.

6.6.6. Prostate cancer

Khalaf et al. conducted a phase-2 RCT, GUTG-001, to compare two different treatment sequences in managing treatment-naïve metastatic castration-resistant prostate cancer (mCRPC) patients.[84] They compared the survival outcomes of the following treatment sequences:

- Aabiraterone (+ prednisolone) → enzalutamide (n=101)
- Enzalutamide → abiraterone (+ prednisolone) (n=101).

Assessment based on NCRAS data

Prostate cancer had limited treatment options compared to other types of cancer within the available timeframe of NCRAS data up until 2018 (only 6 TAs available: TA 101, TA259, TA316, TA377, TA391, TA387).[126,127,237,448,513,514] Since 2014-2016, both abiraterone and enzalutamide have become available in the NHS for treating mCRPC, alongside the standard chemotherapy option of docetaxel. Additionally, carbizitaxel has been introduced as a second-line therapy in the treatment pathway since 2016. Despite the ongoing debate about the efficacy of different treatment sequences in the GUTG-001 trial, the NHS explicitly advised against administering abiraterone and enzalutamide in a sequential manner. Even though a recent commentary indicate that these treatment sequences could still be given in specific situations in an English setting.[515] Consultation with oncologists confirmed that these sequences are not standard practice in England. The only likely scenarios for using these treatment sequences are when patients cannot tolerate either abiraterone or enzalutamide, typically resulting in a switch within the first few months of initiating first-line treatment. Although such type of treatment switching was also permitted in the GUTG-001 trial,

receiving subsequent treatments primarily occurred upon disease progression. This suggests that such treatment sequences rare in NCRAS data and unlikely to yield a sufficient sample size for emulation.

The GUTG-001 trial holds considerable merit as a benchmark due to several factors. Firstly, the study presents a comprehensive range of results, including the two primary endpoints: time-to-second-prostate-specific-antigen-(PSA)-progression and PSA response on the second-line therapy. Additionally, it examined secondary outcomes, such as the time-to-progression on the first-line therapy, time-to-progression on the second-line therapy, as well as the composite progression-free-survival (PFS) — time-to-progression through a sequence of treatments. Also, the trial conducted sensitivity analyses on these endpoints, notably, the time-to-second-PSA-progression excluding patients with delayed crossovers. Furthermore, both study arms had crossover rates surpassing 70%, suggesting that more than 70% of patients in both groups received second-line treatment before their death. Thirdly, this study summarises details regarding patient characteristics, including not only those observed at the time of randomisation but also at the point when they transitioned from first-line to second-line treatment. This aspect makes GUTG-001 a compelling benchmark because it offers extra reference points for the emulation process and aids in understanding any potential differences in outcome measures between the benchmark and its emulation.

Fourthly, it is important to highlight that among all the cancer types in the shortlisted trials, only prostate cancer is linked with TAs that explicitly compared treatment sequences (Chapter 3). Beyond proof-of-concept studies, the scope of Target Trial emulations could potentially expand to include comparisons of treatment sequences evaluated in those prostate cancer TAs, given data focused on prostate cancer patients is available and the methodologies have been evaluated and validated as effective. Finally, a recent systematic review suggests that a series of observational studies have been conducted to compare the treatment sequences in GUTG-001, reflecting the ongoing clinical interest and debate.[516] This implies that the results obtained in a Target Trial emulating GUTG-001 can also be contrasted with the findings from these observational studies that may not have explicitly attempted to replicate any trial results. Such a comparison can help provide insights into the strengths and challenges associated with the proposed methods in my thesis.

While adopting GUTG-001 as a benchmark presents compelling advantages, there are also several issues to consider. The main concern is the availability of PSA values, which are crucial as prognostic factors and for assessing the trial's primary outcomes. Although these values seemed to be collected in English primary care records and the Cancer Outcomes and Services Dataset (COSD), there are concerns around accessing them.[414,424] According to a 10-year study of PSA levels in UK men[424], CPRD provides comprehensive PSA screening results for patients in primary healthcare settings. However, it is unclear if it also includes PSA tests conducted in hospitals. Moreover, CPRD records only cover a small percentage (6-13%) of the English population[403,426], which may not provide

a sufficient sample size of prostate cancer patients. Additionally, the budget constraints of my PhD make it challenging to apply for access to CPRD. On the other hand, the COSD dataset, which is part of the National Cancer Registration and Analysis Service (NCRAS) data collection, is not typically accessible through standard data application routes. In theory, one can obtain it by applying for NCRAS linked primary care data through CPRD and requesting a special linkage to COSD, as demonstrated in a recent study by Strongman *et al.*[517] Accessing combined PSA data from both primary healthcare records and COSD would be beneficial, but it possesses the limitation of accessing CPRD data as mentioned earlier.

Alternatively, the National Prostate Cancer Audit (NPCA) dataset for England contains PSA records.[411] This dataset combines data of prostate cancer patients from various sources, including the standard NCRAS datasets (Cancer Registry, National Radiotherapy Dataset (RTDS), and the Systemic Anti-Cancer Dataset (SACT)), the Office for National Statistics (ONS) dataset, and the NPCA Minimum dataset. The NPCA Minimum dataset comprises the subset of COSD variables that are specific to prostate cancer, which have been collected as part of NCRAS routine data collection process since April 2018.[411,413] Unfortunately, the NPCA data is not accessible for public use. Despite the challenges in accessing PSA data, there are ways to reduce bias in emulation when it is unavailable. One approach is to emulate other outcomes in the GUTG-001 trial, such as time-to-progression. Another option is to use surrogate measures, such as treatment discontinuation, as an alternative way to gauge PSA progression.

Finally, the 3-year follow-up duration of the GUTG-001 trial might pose a problem for emulation, given that neither enzalutamide nor abiratearone were endorsed by NICE as first-line treatments until 2016. To ensure a minimum 3-year follow-up for patients diagnosed in 2016, it would be ideal to utilise NCRAS data from 2016-2019. However, the available data only limited up to 2018 at the time of this assessment. Abiraterone and enzalutamide might have been used as first-line treatments through the Cancer Drugs Fund (CDF) before being used in the NHS. Including patients diagnosed in 2014 and 2015 would ensure a minimum 3-year follow-up, but it is difficult to determine the percentage of those patients who received the targeted treatment sequences. Therefore, estimating the sample size becomes challenging.

To summarise, the GUTG-001 trial may not be the best option for benchmarking in an English context due to uncertainties regarding sample size and limited access to PSA data, and crucially these treatment sequences are not recommended by NHS. However, it is highly valued as a benchmark for developing methodologies, due to its comprehensive reports and well-structured sequential trial design without any unexpected protocol amendments.

Assessment based on Flatiron data

A recent study by George *et al.*, utilising Flatiron data from 2013-2017, examined treatment patterns in managing mCRPC patients in the US.[437] The NCCN guidelines suggest enzalutamide and abiraterone as interchangeable for first- or second-line therapy. Moreover, George *et al.*'s research indicates the two treatment sequences in GUTG-001 can be captured within the Flatiron database, with a sample size of approximately 500 patients. PSA values should theoretically be present in Flatiron database as it is EHR-based and have been confirmed with the Flatiron specialists upon data application. In conclusion, employing Flatiron data to emulate the GUTG-001 trial can potentially bypass the major issues associated NCRAS data, making GUTG-001 a compelling candidate for benchmark, provided that Flatiron data is accessible.

6.6.7. Renal cell carcinoma (RCC)

Three trials were identified that assessed the efficacy of different treatment sequences in patients with untreated metastatic renal cell carcinoma (mRCC). RECORD-3, an international phase-2b trial, compared the efficacy of starting treatment with a mechanistic target of rapamycin (mTOR) inhibitor, everolimus, followed by a TKI, sunitinib, and the reverse sequence[86,406]:

- Everolimus → sunitinib (n = 238)
- Sunitinib → everolimus (n = 233)

The SWITCH-I[174] and SWITCH-II[175], both phase-3 RCTs conducted in Germany, each investigated the efficacy of a treatment sequence involving two different TKIs as compared to its reverse sequence, as detailed below.

SWITCH-I:

- Sorafenib → sunitinib (n = 182)
- Sunitinib → sorafenib (n = 183)

SWITCH-II:

- Sorafenib → pazopanib (n = 189)
- Pazopanib → sorafenib (n = 188)

Assessment based on NCRAS data

Sorafenib has not been endorsed for either first- or second-line treatment by NICE[205], unlike everolimus and sunitinib, which have been available within the NHS pre-2011.[233,518,519] Despite their availability in the NHS, everolimus was not recommended as a first-line treatment and sunitinib was not suggested for second-line treatment.[520] Findings of a recent observational study by Hawkins *et al.* involving mRCC patients in two UK hospitals confirmed this.[521] It revealed that among 652 mRCC patients given any first-line treatment between 2008-2015, only 0.6% were administered everolimus.

Moreover, only 0.5% of the patients who received a second-line therapy received sunitinib. That is, most likely, only the treatment sequence of sunitinib followed by everolimus from RECORD-3 can be captured using NCRAS data, while the other treatment sequences from RECORD-3, SWITCH-I, and SWITCH-II are unlikely to be captured.

In conclusion, these three mRCC trials are not optimal choices as benchmarks if one aims to emulate multi-arm sequential treatment trials. Despite their inadequacy in an English setting, akin to the GUTG-001 prostate cancer trial, their structured design and detailed reports render them relatively more appealing examples.

Assessment based on Flatiron data

No studies employing Flatiron data for RCC research were discovered. The latest NCCN guideline does not support everolimus and sorafenib as first-line treatments in the US.[522] While both can be given as subsequent treatments, everolimus is not a preferred regimen and sorafenib is recommended only in certain situations. Pazopanib and sunitinib are both favoured as first-line treatments for patients with a better prognosis and clear cell histology, with sunitinib also suitable for patients with non-clear cell histology. While both can also be considered as subsequent therapies, the NCCN guideline does not specify their optimal sequence in relation to each other and other therapeutic drugs.

The prior NCCN guideline's stance on everolimus, sorafenib, sunitinib, and pazopanib generally aligns with the most recent version.[522,523] In 2015, everolimus was also not endorsed as a first-line treatment, while sorafenib could be a first-line treatment exclusively for a selected patient group. Furthermore, sunitinib and pazopanib could be applied both as first and second-line treatments. Contrary to the clinical practice in England, everolimus was never recommended as a first-line treatment in the US and was not a preferred option as subsequent treatments, making it likely to be unfeasible to emulate the RECORD-3 study using Flatiron data. Despite sorafenib's short-lived endorsement as a first-line treatment in previous NCCN guidelines, it is no longer recommended. Consequently, emulating the SWITCH I and SWITCH II trials using Flatiron data may not be feasible because of potentially limited sample size, given that sorafenib is a part of both trials' treatment sequences. Additionally, the NCCN guidelines confined the use of sorafenib to certain patient groups, irrespective of the treatment lines. This might result in a smaller overlap of patient characteristics between those who had a treatment sequence involving sorafenib and those who received sequences with sunitinib or pazopanib in real-world settings. This situation could pose challenges in employing causal inference methods for Target Trial emulation, as outlined in Chapter 4.

In conclusion, while the three mRCC trials have appealing qualities, emulating their treatment sequences may be challenging due to the limited sample sizes available in Flatiron data.

6.7. Summary of final decision on benchmark trials

For both databases, I concluded that the breast cancer trials (i.e. BIG 1-98[173,328], TEAM[324,500]), CRC trials (i.e. REVERCE[479], COMETS[480]), the NSCLC trial GFPC 0504[501], and the ovarian cancer trial MITO-8[502], are not appropriate benchmarks for my case studies aimed at methodology development. To begin with, the two breast cancer trials are suboptimal because they primarily involve patients with early-stage cancer, necessitating the emulation of an extended follow-up period of exceeding five years. However, the evaluated data sources are unlikely to offer adequate follow-up durations. Utilising data with limited follow-up times can lead to insufficient number of events for survival estimates, resulting in immature estimates during the emulation process. Additionally, these trials had significant issues regarding non-adherence and violations of post-randomisation treatment assignment. These issues further complicated the statistical analysis and rendered them unsuitable as a benchmark for assessing methodologies.

The CRC trials[479,480] are unsuitable as they specifically enrolled patients who had previous treatment experience prior to receiving the treatment sequences in the trials. Establishing consistent definitions for non-treatment-naïve patients in both the benchmark and its emulation is challenging, the pre-treatment regimens in these trials may not match those used in the NHS. This situation creates unresolvable disparities between the benchmark and its intended emulation, which should ideally be minimised when selecting a benchmark. Even if the earlier treatments were available within the NHS, emulating trials focusing on non-treatment-naïve patients requires additional considerations such as incorporating prevalent new-user design[524], which may further complicate the emulation process.

The NSCLC trial GFPC 0504[501], which took place in the mid-2000s, did not stratify patients based on their EGFR mutation status, which diverges from the established norm over the last decade, where NSCLC treatments are tailored to align with patients' EGFR status. Consequently, characteristics of patients subjected to the same treatment sequence in the GFPC 0504 trial and recent clinical practice are anticipated to exhibit considerable differences. Therefore, attempting to replicate the GFPC 0504 trial may unavoidably yield incomparable results, rendering it less ideal as a benchmark. Emulating MITO-8 trial[502] in ovarian cancer has unique challenges due to the interim trial-design amendments made to address drug shortages. The impact of these amendments remains uncertain, making MITO-8 unsuitable as a benchmark.

Of the remaining studies, the emulation of the pancreatic cancer trials (i.e., FFCD 0301[503], AIO-PK0104[504]) and two mRCC trials (i.e., SWITCH-I[174] and SWITC-II[175]) were expected to have very small sample sizes, regardless of the database used , as the corresponding treatment patterns are absent in real-world clinical practice. Conversely, GUTG-001 trial[84] for prostate cancer and the RECORD-3 trial[86,406] for mRCC stoodd out as appealing benchmarks as they feature treatment sequences that align with those utilised in real-world clinical practice. Additionally, they provided extensive reports,

well-defined sequential treatment design, and did not encounter any unexpected protocol amendments.

Emulating the GUTG-001 trial in a US context was deemed highly feasible, but not in England due to the NHS's restriction on the sequential use of abiraterone and enzalutamide. In the case of RECORD-3, emulating both study arms is unattainable, regardless of data sources, as everolimus has not been recommended as a first-line treatment in either England or the US (i.e. impossible to emulate the sequence of everolimus → sunitinib). Nonetheless, it might be feasible to capture the alternate treatment sequence in RECORD-3, where everolimus was used as a second-line treatment (i.e., sunitinib → everolimus), through the use of the NCRAS database. The Flatiron database, on the other hand, may have limited sample size for the same sequence due to everolimus not being a preferred treatment in the US.

Given the English emphasis of my PhD, even though the original goal of this review was to pinpoint potential benchmarks that allow both arms of the trials to be emulated, I also ended up considering the prospect of single-arm trial emulation case studies. This shift was prompted by the likelihood that the sequence of sunitinib → everolimus in RECORD-3 might be the sole sequence among all potential benchmarks that can be emulated using English data and still generate a sufficiently large sample size. The implications of emulating single-arm trials, specifically regarding TTE approaches and statistical methods, are detailed in my study protocol (Chapter 7).

In conclusion, the GUTG-001 and RECORD-3 trial have proven to be pertinent benchmarks for my case studies.[84,86] Even though the emulation of the GUTG-001 trial through Flatiron data disqualifies it as a "direct benchmark" in an English context (but it does qualify in an US context), it nonetheless retains value as an "indirect benchmark". This is made possible by exploiting shared treatment sequences for mCRPC in both the US and England, leading to the innovative design of a series of correlated Target Trial emulations. For details of this design, see Section 7.5.3, Chapters 7. On the other hand, the RECORD-3 trial has proven to be a valuable "direct benchmark" in the English context, despite the limitation of only being able to emulate a single-arm of the trial using NCRAS. For a comprehensive understanding of how the single-arm emulation study was conceptualised, please refer to Section 7.5.4, Chapter 7. The subsequent concluding section of this chapter provides a concise overview of the disease demographics in the two chosen benchmark trials, as well as the clinical insights provided by the oncologists.

6.8. Summary of disease demographics of the chosen benchmark trials

This section provides a concise review of the disease demographics for the benchmarks identified in Section 6.7. Specifically, Section 6.8.1 details the characteristics of mCRPC in both England and the US (Flatiron population). In Section 6.8.2, I outline the disease demographics for mRCC in England. Additionally, I have included valuable insights from clinical experts.

6.8.1. mCRPC

England

Every year in England, there is an estimated annual incidence of approximately 6,000 newly diagnosed cases of metastatic prostate cancer.[525] According to Dr. Carmel Pezaro, prostate cancer patients commonly undergo castration therapy upon diagnosis, which can be achieved through surgical castration (i.e., bilateral orchiectomy) or medical castration (i.e., lifelong androgen deprivation therapy (ADT)). Professor Derek Rosario further suggested that surgical castration represents a very small fraction of patients who require castration therapy. Life-long androgen ADT involves the use of luteinising hormone-releasing hormone (LHRH) agonists or antagonists, also known as gonadotrophin-releasing hormone (GnRH) agonists or antagonists, such as padeliporfin. In the event that a patient's prostate cancer continues to grow despite undergoing castration therapy, they will be categorised as having CRPC. The duration it takes for the transition from castration-sensitive (hormone-sensitive) to castration-resistant (hormone-relapse) status varies among individuals. Around 28% of prostate cancer patients may develop castration-resistant prostate cancer (CRPC), as suggested by an English epidemiology study.[526] Prostate cancer patients are predominantly under the care of oncologists. Nevertheless, patients who exclusively undergo ADT may primarily receive management from GPs following their initial treatment.

Appendix 6.1 presents a summary of all the published TAs associated with the management of mCRPC between 2006-2020. Since 2006, Docetaxel has been accessible as a first-line treatment for treatment-naïve patients with mCRPC within the NHS, whereas abiraterone and enzalutamide, two androgen receptor-targeted agents (ARTA), have been available for the same indication since 2016.[126,127,513] Since 2012 and 2014, respectively, abiraterone and enzalutamide have been introduced as second-line therapy options in the NHS for the treatment of patients with mCRPC who have previously undergone docetaxel treatment.[448,514,527] Prior to the introduction of abiraterone and enzalutamide, docetaxel therapy served as the sole standard treatment for patients with mCRPC. Since 2016, Cabazitaxel has been included as an alternative second-line treatment option in the treatment pathway, exclusively for patients who have previously received docetaxel therapy.[237] Additional treatment options for patients at a later phase of mCRPC are available. Radium-223 has been approved since 2016 for patients with bone metastases[528], while olaparib remained under review by NICE as of June 2021.

Importantly, the English standard practice does not allow for the sequential use of abiraterone and enzalutamide, as mentioned earlier.[527] However, Dr. Carmel Pezaro confirms that if patients experience severe adverse events (such as toxicity) with either drug, they can switch to the other agent without it being considered as disease progression at that time. Such switching typically occurs within

three months of treatment initiation, while switching after three months may indicate disease progression.

Additionally, the prostate cancer treatment landscape has evolved, integrating additional systematic treatments alongside ADT at earlier stages before developing metastasis or castration-resistance. Treatments beyond the scope of mCRPC now include enzalutamide, daralutamide, and apalutamide in combination with ADT for non-metastatic castration-resistant prostate cancer (nmCRPC) (approved in 2019, 2020, and 2021, respectively).[249,529-531] For newly diagnosed metastatic hormone-sensitive prostate cancer (mHSPC), treatment options have expanded to include docetaxel[495,532], enzalutamide (available through CDF since 2021)[533], apalutamide (available through CDF since 2022)[534], and darolutamide (available through CDF since 2023 for patients ineligible for chemotherapy with docetaxel).[531,535] Key clinical trials comparing ADT alone to ADT combined with docetaxel in treating mHSPC were conducted from 2004 to 2013, including the GETUG-AFU15 (France)[536,537], the CHAARTED (US)[538], and the STAMPEDE (UK).[539,540] The results of these studies, published between 2013 and 2019, were inconsistent regarding the benefits of upfront docetaxel and the specific mHSPC patient subgroups that might benefit from it, leading to varied adoption timelines across different medical practices for the use of docetaxel in treating mHSPC.

Reflecting on these recent treatment advancements, our oncologist Dr. Pezaro noted that patients now receiving abiraterone or enzalutamide as first-line treatments for mCRPC might have previously undergone other treatments like docetaxel when their cancer was castration-sensitive. Dr. Pezaro suggested that this trend in treatment strategy likely became more apparent after 2016-2017, following publication of UK-based STAMPEDE trial results.[539,540] The widespread use of medications in England depends not only on the UK Medicines and Healthcare Products Regulatory Agency (MHRA) approval but also on recommendations from NICE. Consequently, treatments other than docetaxel as upfront treatment prior to the development of mCRPC did not become widely adopted until after 2021. In rare cases, docetaxel may be re-administered upon disease relapse while receiving abiraterone or enzalutamide for mCRPC (i.e., docetaxel for hormone-sensitive prostate cancer → abiraterone/enzalutamide for mCRPC → re-administration of docetaxel for mCRPC).

Identifying castration-resistant patients in NCRAS data can be challenging due to the absence of a variable documenting the date of patients becoming castration-resistant. Dr. Pezaro suggested that the emergency use of certain drugs in England upon the onset of castration resistance could serve as a potential proxy indicator, although it is uncommon. These drugs include non-standard treatment options like maximal ADT[541], which involves adding an additional androgen receptor (e.g., bicalutamide) to the standard ADT (e.g., leuprorelin, cetrorelix). Another option is adding low-dose dexamethasone. Professor Derek Rosario seconded these statements but highlighted the potential uncertainty surrounding the full documentation of emergency drug usage in NCRAS data. Inclusion

of such information, even if available, may result in the identification of a distinct group of patients, given the variability in treatment preferences among different physicians without a defined standard practice.

Given that my case study designs were formulated during the COVID-19 pandemic; it is important to acknowledge certain limitations associated with utilising retrospective data. Professor Rosario noted that docetaxel was not recommended for prostate cancer patients since the COVID-19 outbreak around mid-2020, while abiraterone emerged as a preferred treatment option. Hence, my case studies designed to utilise NCRAS data from 2012-2018/2019 may not accurately depict the current or future treatment patterns, specifically those influenced by the COVID-19 situation. However, this does not hinder the trial emulation for benchmarking, as once methodologies are validated, they can be applied to answer questions in a more generalised population, such as those involving post-COVID clinical practice.

US (Flatiron population)

In the US, the treatment options for prostate cancer are generally similar to those in England, with one significant difference being the ability to use abiraterone and enzalutamide in a sequential manner. A publication based on Flatiron data indicated that the primary treatment sequences for prostate cancer in the US during 2013-2017 were abiraterone followed by enzalutamide or the reverse sequence.[437]

Additionally, while abiraterone and enzalutamide have been approved by the US FDA for hormone-sensitive prostate cancer[447,542], NICE was still evaluating this indication when I submitted my case study protocol to ODR for applying NCRAS data in June 2021. As mentioned earlier, NICE later approved enzalutamide for mHSPC in 2021, but not abiraterone.

Flatiron data experts were consulted during the data application process. Their preliminary analysis, indicated that their database could capture approximately 600 mCRPC patients who were treated with abiraterone followed by enzalutamide, as well as around 400 mCRPC patients who received enzalutamide followed by abiraterone. As of the data cut-off on March 31, 2019, a total of 4,000 metastatic prostate cancer patients were identified as having received a first-line treatment, and among them, 1,700 patients had undergone a second-line therapy. Based on this information, I concluded that the sample size for my case study using Flatiron data is likely to be sufficient. In particular, one aspect of my emulation involves a scenario analysis that would increase the number of participants in each group, which involves including patients who were initially treated with either abiraterone or enzalutamide as their first-line therapy, but unfortunately, they passed away without undergoing any further treatments. For further insight into the rationale behind this particular design, please refer to Chapter 7. Consequently, I proceeded with the data application process.

6.8.2. mRCC

Each year in the UK, an estimated 13,000 new instances of kidney cancer are reported based on data from 2015 to 2017.[543] The majority of these patients (exceeding 80-90%) are diagnosed as RCC.[525] In England, approximately one out of every four RCC patients is initially diagnosed at an advanced stage, and at the time of diagnosis, 25-34% of RCC patients already have metastases (i.e. mRCC).[543,544] Approximately 75% of advanced-stage RCC patients are considered eligible to commence a first-line systematic cancer treatment. [545]

Appendix 6.2 presents a summary of all the published TAs associated with the management of mRCC between 2009-2020. Currently, the primary treatment options for patients with mRCC are targeted therapy, such as TKIs (drugs ending with –nib), and immunotherapy (drugs ending with –mab), including Programmed Cell Death Protein 1 (PD-1) inhibitors and Programmed Cell Death Ligand 1 (PD-L1) inhibitors. Since 2009 and 2011, Sunitinib and pazopanib have been recommended as the first-line treatment options for patients with mRCC in the NHS, respectively.[498] Tivozanib and cabozantinib were introduced as additional first-line treatment options for mRCC starting in 2018. In contrast, the regimens of nivolumab plus ipilimumab (since 2019), and avelumab coupled with axitinib (since 2020), have been exclusively accessible as first-line treatments via the CDF. [520] The NHS has adopted the following treatments for mRCC patients who are in need of second-line or later-line therapies: axitinib (from 2015), nivolumab (from 2016), everolimus (from 2017, previously only available through CDF), cabozantinib (from 2017), and the levatinib-everolimus combination (from 2018).[520]

Professor Janet Brown stated that in England, before targeted therapy and immunotherapy became widespread, approximately 10-15% of patients with less advanced cancer were treated with interferon-alpha before receiving any other systemic anti-cancer therapies. Additionally, everolimus has been shifted from second-line treatment to later stages of treatment (third or fourth-line treatment) in recent years for patients with an Eastern Cooperative Oncology Group (ECOG) score below 2. This adjustment reflects the lack of substantial survival benefits of everolimus and instead emphasises its role in improving patients' quality-of-life.

Professor Brown clarified that there is no standardised adjuvant therapy for mRCC patients, indicating that systematic therapy may or may not accompany surgical interventions before metastases appear.[546] She added that patients who are suitable for operations like nephrectomy, and those who do not rapidly develop metastases post-nephrectomy, usually have a more favourable prognosis. Specifically, these patients are typically deemed healthy enough to undergo surgical interventions at the first place. Conversely, elevated calcium levels in patients typically indicates a poorer prognosis.

Professor Brown further suggested that patients with mRCC who remain in good health

conditions (i.e. desirable performance status) may receive up to 5 or 6 lines of treatment with each subsequent relapse. However, a significant proportion (roughly 40-50%) of mRCC patients receive only three lines of treatment before death. This can be attributed to either their frailty preventing further treatment upon disease progression or mortality occurring before subsequent treatments. The introduction of newer treatment options, particularly cancer immunotherapies since 2016, has significantly improved the overall survival (OS) of mRCC patients.[547] Previously, patients who receiving only TKIs had an OS range of 1-2 years up to 4-5 years after being diagnosed.

6.9. Chapter overview

This chapter systematically reviewed and assessed existing oncology trials that randomised patients to different treatment sequences, evaluating their potential as benchmarks for conducting proof-of-concept studies that aim to assess the feasibility using RWD to derive unbiased effectiveness estimates for comparing treatment sequences. Two trials, GUTG-001[84] and RECORD-3[86], were identified as pertinent, each with its own limitations. Specifically, the treatment sequences in GUTG-001 are not standard practice in England and thus can serve only as an indirect benchmark for validating in English NCRAS database, supported by proof-of-concept studies conducted with Flatiron data. Further, only one sequence from the RECORD-3 trial can be captured in English NCRAS, making it a direct benchmark only as a single-arm trial. These findings led to the design of a series of interconnected case studies in Chapter 7.

Chapter 7 Leveraging real-world data to assess treatment sequences in HTA: A Target Trial Emulation protocol using the English Cancer Registry and US Flatiron Health Database

7.1. Chapter overview (Chang et al. (2024) in SCHARR HEDS Discussion Paper Series)

This chapter presents a protocol of a series of case studies that were designed to assess the feasibility of leveraging real-world data (RWD) as an alternative means for generating reliable, unbiased effectiveness estimates for treatment sequences. The protocol was published as a SCHARR HEDS Discussion Paper (24.01) under a Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 License.[170] This allows the following sections (Section 7.2 to 7.6) to present the published version, with text format and article numbering adjusted to fit the thesis format. Additionally, the article's bibliography has been integrated into the thesis bibliography. Minor changes were made to the published protocol for inclusion in the thesis chapter to improve clarity based on PhD examiners' comments, but no major alterations to the protocol were made.

Section 7.2 provides an overview of the structure of the publication. For publication readiness, the protocol offers a concise overview of treatment sequence evaluation in health technology assessments (HTA) and highlights the role of the Target Trial Emulation (TTE) approach[3] in addressing the challenge of scarce clinical evidence for comparing treatment sequences. Specifically, Section 7.3 serves as a concise overview of the background and significance of the evidence scarcity issue, revisiting themes initially introduced in Chapters 1 to 3. It also recaps the role of benchmarking in designing proof-of-concept case studies and the selection of candidate databases, which were discussed in more detail in Chapters 6 and 5, respectively. Section 7.4 further highlights the planned case studies' significance, not only in tackling the evidence scarcity in comparing treatment sequences in HTA but also in their broader relevance to TTE research. Section 7.5 delves into the actual case study design, incorporating key concepts of causal inference and the TTE framework introduced in Chapter 1 (Section 1.4), alongside the role of advanced statistical methods from Chapter 4. Section 7.6 discusses data acquisition considerations. Sections 7.5 and 7.6 also provide the rationale behind these decisions. Finally, the chapter concludes with a summary of the protocol's contribution to the thesis in Section 7.7.

University of Sheffield | Sheffield Centre for Health & Related Research

HEALTH ECONOMICS & DECISION SCIENCE

Discussion Paper Series

HEDS Discussion Paper 24.01

Title: Leveraging Real-World Data to Assess Treatment Sequences in Health Economic Evaluations: A Study Protocol for Emulating Target Trials Using the English Cancer Registry and US Electronic Health Records-Derived Database

Authors:

Jen-Yu Amy Chang ORCID ID: 0000-0001-6660-5246

James B. Chilcott ORCID ID: 0000-0003-1231-7817

Nicholas R. Latimer ORCID ID: 0000-0001-5304-5585

Corresponding author: Jen-Yu Amy Chang, Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK

Tel: +44 (0) 114 222 5454

Email: jy.a.chang@sheffield.ac.uk

Leveraging Real-World Data to Assess Treatment Sequences in Health Economic Evaluations: A Study Protocol for Emulating Target Trials Using the English Cancer Registry and US Electronic Health Records-Derived Database

Jen-Yu Amy Chang[1,*], James B. Chilcott[1], Nicholas R. Latimer[1,2]

[1]Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield
[2]Delta Hat Limited, United Kingdom

*Corresponding author: Jen-Yu Amy Chang
Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield, Sheffield, UK
Address: Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK
Tel: +44 (0) 114 222 5454
Email: jy.a.chang@sheffield.ac.uk

Running title: Target Trial Emulation for comparing treatment sequences

Key words: Treatment sequence, Health economics, Economic evaluation, Real-world data, Target Trial Emulation, English Cancer Registry, Electronic health records, Causal inference, Benchmarking

Funding: This study was supported by the Wellcome Trust [108903/B/15/Z]. The sponsor had no role in the study design, data collection, analysis, and decision to publish or preparation of the manuscript.

Conflicts of interest: All authors have no conflicts of interest that are directly relevant to the content of this article.

Ethics Approval: This project has been approved by the Ethics Committee of Sheffield Centre for Health and Related Research (SCHARR; formerly known as the School of Health and Related Research (ScHARR)) at the University of Sheffield (Ref. 041331) and the NHS Research Ethics Committee (Ref. 21/PR/0908).

Acknowledgement: The authors greatly appreciate Dr. Carmel Pezaro (Singleton Hospital, Swansea Bay University Health Board, NHS Wales), Professor Derek Rosario (Sheffield Teaching Hospitals NHS Foundation Trust), and Professor Janet Brown (Sheffield Teaching Hospital NHS Foundation Trust) for their invaluable clinical expertise and insights, which have been instrumental in enriching this project. We extend a heartfelt thanks to the team at Flatiron Health, especially Dr. Blythe Adamson and Dr. Philani Mpofu for their expert insights on the Flatiron database and assistance in accessing data. Lastly, we are thankful to Dr. Amanda Loban (University of Sheffield) for her expert guidance with information governance and assistance in fulfilling data management standards.

Project Overview	
Project Title	Leveraging Real-World Data to Assess Treatment Sequences in Health Economic Evaluations: A Study Protocol for Emulating Target Trials Using the English Cancer Registry and US Electronic Health Records-Derived Database • Sub-project 1: Investigating the Application of Causal Inference Methods for Modelling the Impact of Treatment Sequences in Health Economic Evaluations: Utilising Real-world Evidence from the Flatiron Database • Sub-project 2: Investigating the Application of Causal Inference Methods for Modelling the Impact of Treatment Sequences in Health Economic Evaluations: Utilising Real-world Evidence from the English Cancer Registry
Project Objective	This project aims to assess the feasibility of harnessing data from the English Cancer Registry and Flatiron electronic health record (EHR)-derived database to derive unbiased effect estimates for comparing oncology treatment sequences. A key aspect of this investigation is the application of advanced causal inference methods, with a particular focus on Target Trial Emulation.
Principal Investigator (PI)	Jen-Yu Amy Chang (PhD Candidate, Sheffield Centre for Health and Related Research (SCHARR), Division of Population Health, School of Medicine and Population Health, University of Sheffield)
Project Team Members	<u>PhD Supervisors:</u> Professor Nicholas Latimer (Professor of Health Economics, Sheffield Centre for Health and Related Research (SCHARR), School of Medicine and Population Health, University of Sheffield) Professor Jim Chilcott (Professor of Healthcare Decision Modelling, Sheffield Centre for Health and Related Research (SCHARR), School of Medicine and Population Health, University of Sheffield) <u>Clinical Experts:</u> Dr. Carmel Pezaro (Consultant in Oncology, Singleton Hospital, Swansea Bay University Health Board, NHS Wales for Swansea and Neath)

	Professor Derek Rosario (Consultant Urological Surgeon, Sheffield Teaching Hospitals NHS Foundation Trust) Professor Janet Brown (Medical Oncologist/Professor of Translational Medical Oncology, Sheffield Teaching Hospital NHS Foundation Trust)
PI's institution	University of Sheffield
Project Funder	Wellcome Trust [108903/B/15/Z]
Protocol version	2.0

Protocol Version History		
Version	**Date**	**Description**
Sub-project 1. Version 1.3	June 2021	**Protocol for Sub-project 1:** "Investigating the Application of Causal Inference Methods for Modelling the Impact of Treatment Sequences in Health Economic Evaluations: Utilising Real-world Evidence from the Flatiron Database" • Content: This protocol focuses on case studies using the Flatiron Database. • Ethics Approval: Approved by the SCHARR ethics committee at the University of Sheffield. • Database Application: Utilised for applying to access the Flatiron database. • Scientific Review: The project underwent a review by Flatiron's data analytics team to assess the theoretical feasibility for conducting the planned analysis using Flatiron data prior to the data application.
Sub-project 2. Version 1.0	June 2021	**Protocol for Sub-project 2**: "Investigating the Application of Causal Inference Methods for Modelling the Impact of Treatment Sequences in Health Economic Evaluations: Utilising Real-world Evidence from the English Cancer Registry" • Content: This protocol details case studies using the National Cancer Registration and Analysis Service (NCRAS) database. It also includes an overview and key elements of its sister

		project (Sub-project 1), presenting a comprehensive view of their interconnected objectives and methodologies. • Ethics Approval: Granted by the NHS Ethics Committee. • Data Application: Initially submitted in June 2021 to the Office for Data Release, Public Health England (PHE), for access to the NCRAS database. The application was later transferred and re-initiated with NHS Digital in June 2022 due to the dissolution of Public Health England (October 2021). • Scientific Review: Underwent peer-review within the institution for feasibility and scientific merit, and also reviewed by the National Disease Registration Service (NDRS) data analytical team during data application to determine data release for the planned analysis.
Combined protocol. Version 2.0	January 2024	**Combined protocol for Sub-project 1 and 2**: Key Updates: • Content: This protocol merged and refined wording of two sub-project protocols for publication readiness. • Literature update: Incorporated additional references from recent studies. • Enhanced details: Augmented the methods section with more details, including the criteria for assessing the agreement of benchmark trials and real-world evidence. • Team update: updated the project team members.

Abstract

Background

Considering the sequence of treatments is vital for optimising healthcare resource allocation, especially in cancer care, where sequence changes can affect patients' overall survival and associated costs. A key challenge in evaluating treatment sequences in health technology assessments (HTA) is the scarce evidence on effectiveness, leading to uncertainties in decision making. While randomised controlled trials (RCTs) and meta-analyses are viewed as the gold standards for evidence, applying them to determine the effectiveness of treatment sequences in economic models often necessitates making arbitrary assumptions due to insufficient information on patients' treatment histories and subsequent therapies. In contrast, real-world data (RWD) presents a promising alternative source of evidence, often encompassing details across treatment lines. However, due to its non-randomised nature, estimates of the treatment effectiveness based on RWD analyses can be susceptible to biases if not properly adjusted for confounding factors.

To date, several international initiatives have been investigating methods to derive reliable treatment effects from RWD — by emulating Target Trials that replicate existing RCTs (i.e. benchmarks) and comparing the emulated results against the benchmarks. These studies primarily seek to determine the viability of obtaining trial-equivalent results through deploying specific analytical methodologies and study designs within the Target Trial emulation framework, using a given database. Adopting the Target Trial emulation framework facilitates the analyses to be operated under causal inference principles. Upon validation in a particular database, these techniques can be applied to address similar questions (e.g., same disease area, same outcome type), but in populations lacking clinical trial evidence, leveraging the same RWD source.

Studies to date, however, have predominantly focused on the comparison of individual treatments rather than treatment sequences. Moreover, the majority of these investigations have been undertaken in non-English contexts. Consequently, the use of RWD in evaluating treatment sequences for HTA, especially in an English setting, remains largely unexplored.

Objectives

The goal of this project is to investigate the feasibility of leveraging RWD to produce reliable, trial-like effectiveness estimates for treatment sequences. We aim to assess the capability of two oncology databases: the US-based Flatiron electronic health record and the National Cancer Registration and Analysis Service (NCRAS) database of England. To achieve this, we plan to harness the Target Trial Emulation (TTE) framework for replicating two existing oncology RCTs that compared treatment sequences, with the intent of benchmarking our results against the original

studies. Further, we aim to detail the practicalities involved with implementing TTE in diverse databases and outline the challenges encountered.

Methods

1. We aim to emulate existing RCTs that compare the effect of different treatment sequences by constructing the study design and analysis plan following the TTE framework. Specifically, the following case studies are planned:
   (1) Prostate cancer case study 1 (PC1) - US direct proof-of-concept study (method direct validation): replicating the GUTG-001 trial using Flatiron data
   (2) Prostate cancer case study 2 (PC2) - US-England bridging study (method extension): emulating Target Trials that compare treatment sequences that have been common in England using Flatiron data
   (3) Prostate cancer case study 3 (PC3) - English indirect proof-of-concept study (method indirect validation): emulating the same Target Trial in PC2 using English NCRAS data
   (4) Renal cell carcinoma case study (RCC) - method direct validation in a single-arm setting: emulating the sunitinib followed by everolimus arm in the RECORD-3 trial using English NCRAS data
2. We will compare results of the emulated Target Trials with those from the benchmark trials.
3. We plan to compare different advanced causal inference methods (e.g. marginal structural models using IPW and other g-methods) in estimating the effect of treatment sequences in RWD.

Expected results

This study will provide evidence on whether it is feasible to obtain reliable estimates of the (comparative) effectiveness of treatment sequences using Flatiron data and English NCRAS data. If applicable, we intend to develop a framework that provides a systematic way of obtaining the (comparative) effectiveness of treatment sequences using RWD. It is possible that the data quality is insufficient to emulate the planned Target Trials. In this case, we will report reasons for the implausibility of data analysis. If applicable, we will make suggestions to whether the national health data collection may be enhanced to make the analyses possible. The results of this study will be submitted to peer-reviewed journals and international conferences.

7.2. Overview

This protocol outlines a series of proof-of-concept case studies focusing on evaluating the use of real-world data (RWD) for making informed decisions in health technology assessment (HTA), particularly in the context of treatment sequences. The protocol begins in Section 7.3 by underscoring the importance of evaluating treatment sequences in HTA and discussing the opportunities and challenges in leveraging RWD. We describe the Target Trial Emulation (TTE) approach with benchmarking as a means to assess the feasibility of deriving reliable estimates from RWD, and specify RWD sources for investigation. We then lay out the primary objectives of the project.

In Section 7.4, we elaborate on the project's significance and relevance to pertinent research. Section 7.5 presents a detailed Analysis Plan for the Target Trial Emulation case studies. Finally, Section 7.6 details the data requirements necessary for the study.

7.3. Background

7.3.1. Evaluating treatment sequences in health technology assessments

With an ever-increasing number of treatment options, the significance of evaluating treatment sequences within HTA has become apparent. Specifically, changing the order of treatments can introduce variability in the overall effectiveness and costs associated with managing a disease, making the assessment of treatment sequences—as opposed to a single line of therapy—vital in HTA wherever relevant.[1,2,8]

Despite established frameworks for modelling treatment sequences in health economic evaluations, challenges persist, particularly the scarcity of data on the effectiveness of treatment sequences.[1,9,34,35] Although clinical trials are considered the “gold standard” in evidence, they rarely assess the impact of sequences, focusing instead on the efficacy/effectiveness of a single line of treatment (LOT). Given the scarcity of trials comparing treatment sequences, analysing RWD offers a promising avenue to determine the (comparative) effectiveness of sequential treatment strategies, thereby supporting more informed clinical and economic decision-making.

7.3.2. Assessing the sequencing effect using rea-world data

The use of RWD is advantageous not only for its capability to capture sequencing information, but also for offering larger, more generalisable sample sizes compared to clinical trials. However, the lack of random treatment allocation in routine practice necessitates careful study design and statistical analysis to avoid biased results[548], notably due to confounding from factors that affect both treatment choices and outcomes, such as disease severity. While existing guidelines discuss RWD's utility for HTA and methods for estimating treatment effectiveness from RWD, none of them explored methods for comparing treatment sequences.[39,40]

To effectively harness RWD for evaluating the causal effects of different treatment sequences—a form of time-related static treatment strategies or dynamic treatment strategies[5]—, it is crucial to employ advanced causal inference methods like marginal structural models with inverse probability weighting (IPW) and other G-methods for ensuring a "fair comparison" across patients receiving different treatment sequences (i.e., achieving balanced patient characteristics between treatment groups, and addressing time-varying confounding).[5,58,279] Moreover, the successful application of these statistical methods often hinges on the availability of adequate data and relevant variable information.

Our project, in response to the absence of established guidelines, aims to determine the feasibility of applying the aforementioned methods to real-world datasets, especially local ones, to reliably estimate the effectiveness of treatment sequences in the context of supporting decision-making by the National Institute for Health and Care Excellence (NICE). Specifically, The National Cancer Registration and Analysis Service (NCRAS) database for England[82] and the US Flatiron Electronic Health Records (EHR)-Derived database[85] were identified as promising data sources for our initial investigations. More detailed information and the rationale behind choosing these databases are provided in Section 7.3.4.

7.3.3. Target Trial Emulation and benchmarking

In addition to advanced statistical methods, the TTE framework, proposed by Hernan et al.[3], emerged as valuable tool for structuring observational studies aimed at answering causal questions. The framework's significance lies in facilitating adequate designs of observational studies, enhancing transparency, thereby further mitigating biases inherent in study designs (which may not be fully rectifiable through statistical methods alone), such as selection bias and immortal time bias. The framework is based on the idea of designing an observational study as a hypothetical Target Trial[3], had such a trial been implementable, and then explicitly emulating this Target Trial using RWD. A standard Target Trial protocol consists of seven key components to resemble the setting of a randomised controlled trial (RCT), including eligibility criteria, intervention strategies being compared, intervention assignment, follow-up period, outcomes of interest, causal contrasts of interest, and analysis plan.

Several initiatives have launched benchmarking studies to determine how effectively the TTE approach can be used within specific real-world datasets to answer causal questions, especially before applying it to other questions in comparable settings.[172,476] These benchmarking studies attempted to replicate the designs and results of existing clinical trials (i.e., benchmark trials) through emulating Target Trials using RWD, including applying the same (or as far as possible) patient inclusion/exclusion criteria and analytical methods to achieve the emulation. Theoretically, if a Target

Trial is correctly specified, estimates derived from RWD may be comparable to those from a benchmark trial, providing a validated means to derive reliable real-world evidence (RWE). However, a review highlights that disagreements between observational studies and RCTs can arise for various reasons, with the specific causes often being indeterminable.[382] Nevertheless, improved benchmarking can be achieved, had an observational study explicitly aimed to emulate a Target Trial.[382]

Building on this concept, our study will employ a similar strategy to assess the applicability of advanced statistical methods in generating reliable RWE for HTA decisions involving treatment sequences. Specifically, in the current study protocol, we outline the design of a series of benchmarking case studies following the TTE framework aiming to replicate the effectiveness estimates from several RCTs comparing treatment sequences[84,86,324,328] (Section 7.5 Analysis Plan). We will compare our findings with those from benchmark RCTs, assessing the potential of RWD to successfully mimic their results. The design of our TTE analyses (Table 7.1-7.3) references the published protocol structures of the RCT DUPLICATE case studies.[471,472]

In developing the protocol of our study, we conducted a systematic review to identify candidate benchmark trials, focusing on RCTs that explicitly randomised patients to receive different predetermined treatment sequences. Our study settled on two oncology trials: the prostate cancer trial, GUTG-001, and the renal cell carcinoma (RCC) trial[86,406] RECORD-3[84]. Due to the extensive nature of the review, we will provide the detailed rationale in a separate publication.

7.3.4. The English NCRAS database and the Flatiron database

Our study focuses on implementing the benchmarking studies using two oncology databases: the NCRAS database for England and the Flatiron database. NCRAS, a part of England's National Disease Registration Service (NDRS), coordinates aggregated information from cancer registries across England, Northern Ireland, and Scotland at the UK level, with regional authorities facilitating access to patient-level data. Additionally, NCRAS oversees the patient-level cancer registry data in England.[387] Researchers can access the English Cancer Registry with linkage to a selection of non-cancer specific National Health Service (NHS) England datasets through NHS England's DARS (Data Access Request Service). The application was previously managed by the Office for Data Release, Public Heath England. The English Cancer Registry provides detailed data on NHS England's cancer patients, including important prognostic factors, such as tumour stages, sizes, and patient performance status at the time of diagnosis. Enhancing this, the registry can be linked with other NCRAS datasets, such as the Systemic Anti-Cancer Therapy (SACT) dataset[390], which provides extensive information on cancer treatments, including those under the Cancer Drugs Fund (CDF). Additionally, its viable linkage with NHS hospital records (i.e., Hospital Episode Statistics (HES)), allows for a thorough

understanding of patients' medical histories.

The Flatiron database, a US-based EHR-derived database focused on oncology care, provides detailed diagnostic and treatment records of patient visits as well as laboratory results.[85] It combines structured and (machine learning assisted) manually abstracted unstructured data, making it a comprehensive resource for oncological research. Despite being US-based, the Flatiron database was included in our study for its potential of more timely data access, diverse patient demographics (across the US) potentially overlapping with the English population, and its capability to capture treatment sequences relevant to the GUTG-001 trial, which are less common in the UK. This choice facilitates the design of our benchmarking study protocol, tethered to the identified benchmark trials, as detailed in Section 7.3.5 and Section 7.5. Furthermore, NICE has partnered with Flatiron Health to explore the use of RWE in improving the assessment of health technologies' clinical and cost effectiveness.[83,549,550] An example of this is the use of Flatiron's data to supplement clinical trial information in a recent NICE technology appraisal (TA).[551]

7.3.5. Summary of project aims

In summary, this project aims to examine the feasibility of using the English NCRAS data and the Flatiron data to obtain reliable effectiveness estimates of treatment sequences in prostate cancer and RCC. The aim of this project will be achieved through completing the following objectives.

1) To emulate existing RCTs that compare the effect of different treatment sequences by constructing the analyses plan following the Target Trial framework.
2) To compare results of the emulated Target Trials with those from the benchmark trials.
3) To compare different advanced causal inference methods (e.g. marginal structural models with IPW and other g-methods) in estimating the effect of treatment sequences in RWD.
4) To detail the practical aspects of implementing TTE across different databases and to describe challenges encountered.

Ultimately, our goal is to leverage the insights from this project to create a systematic framework for generating evidence on the effectiveness of treatment sequences using RWD, particularly in the context of health economic evaluations. In the event that benchmarking proves to be infeasible, our focus will shift to providing detailed insights into the practicality of implementing TTE with these databases. This includes an exploration of the challenges encountered, the reasons behind any limitations, and potential areas for improvement and future research.

7.4. Significance of this study

The significance of this study lies in its potential to expand upon relevant existing research conducted by several initiatives, including the US Food and Drug Administration (FDA) funded RCT DUPLICATE and the pharmaceutical industry-sponsored OPERAND.[172,471,474,476,552] These

initiatives have been focusing on replicating clinical trial results using RWD within highly structured frameworks designed to mimic clinical trials.[172,476] RCT DUPLICATE, initiated in 2018 under the 21st Century Cures Act, seeks to inform the use of RWE studies in regulatory decisions. It leverages the Target Trial Emulation approach to assess the real-world effectiveness of medical products and benchmark them against a large number of RCTs.[471,553] OPERAND, in contrast, explores how treatment effect estimates might differ when the stringent eligibility criteria of RCTs are relaxed. To our knowledge, these studies have primarily focused on cardiovascular diseases and utilised US claims databases (e.g. Medicare), without attempting to assess the feasibility of estimating the comparative effectiveness of treatment sequences.

Our project stands out by seeking to replicate results from sequential treatment trials using data from the English Cancer Registry and the Flatiron EHR-derived database. This approach is expected to enhance the findings from previous large-scale initiatives in several ways. Firstly, it will evaluate different causal inference methods specifically for emulating sequential treatment trials using RWD and benchmark them against existing trials. Secondly, the project will demonstrate the utility of observational data as an alternative source of evidence in health economic evaluations, particularly for modelling treatment sequences. Specifically, it could pave the way for establishing a systematic framework for deriving reliable (i.e. trial-mimicking) comparative effectiveness estimates for sequential treatments, addressing a key challenge in sequence evaluation in HTA. Thirdly, the project aims to improve user experience and enhance the use of English NCRAS data in future HTA.

Furthermore, part of the project (i.e. prostate cancer benchmarking studies) will compare the emulation of the same Target Trials using the US-based Flatiron database versus the English NCRAS database, as outlined in Section 7.5 Analysis Plan). This comparison will offer insights into the differing treatment patterns for prostate cancer in the US and England, and how they may affect the emulation. For example, sequential use of abiraterone and enzalutamide is not permitted in England[527], but they are key first and second-line treatments in the US[437], as a recently study using Flatiron data shows.[437] The global debate over the benefits of this sequence is ongoing.[105,515,516,554-557]

Based on a recent systematic review focusing on treatment sequences in prostate cancer conducted by the Canadian Agency for Drugs and Technologies in Health (CADTH)[555], it seems that docetaxel-containing treatment sequences with androgen receptor-targeted agents (ATRA) (i.e. abiraterone, enzalutamide) may improve progression-free survival (PFS) compared to sequential therapy with ATRA alone in castration-resistant prostate cancer (CRPC) patients. However, none of the studies included evidence from England and studies included were all retrospective, and therefore, should be interpreted with caution. In addition, no published cost-effectiveness studies were found explicitly comparing different treatment sequences in prostate cancer, despite the interest of decision makers in this. The results of our study will supplement the understanding of these topics using the

English cancer registry data to provide English-based effectiveness estimates. Additionally, our analyses will provide insights on whether conducting similar Target Trial analyses using different observational data sources (i.e. English NCRAS data and US Flatiron data) may require modifications in defining important variables (e.g. definition of progression using retrospective data) and/or result in contrasting final results. We will also explore the strengths of each database in informing HTA treatment sequencing decisions in England, identifying potential areas for improvement, especially within the local database (NCRAS).

7.5. Analysis Plan

7.5.1. Overview

This section is structured into four parts. Firstly, we introduce the scope of the section and provide a brief overview of the structure for each set of Target Trial case studies. This is followed by exploring considerations associated with applying causal inference methods for Target Trial analyses using English NCRAS and Flatiron data in Section 7.5.2. Subsequently, we present detailed plans for our two sets of sequential treatment Target Trial case studies in prostate cancer (Section 7.5.3: case studies PC1, PC2, and PC3) and RCC (Section 7.5.4: case studies RCC1 and RCC2).

Each case study set begins with an introduction to the benchmark trials (GUTG-001[84] and RECORD-3[86,406]) and summarises the demographics of cancer patients in the UK/England and the US, assessing the treatment sequences used in NHS and US clinical practice. This is followed by a detailed presentation of the planned Target Trials, featuring a table summarising their seven key components. Primary outcomes in all case studies focus on time-to-event outcomes, particularly the overall survival (OS) of patients receiving specific treatment sequences. At the end of each case study set, we assess whether the NCRAS and Flatiron data offer a sufficient sample size for our planned Target Trials. We will compare the outcomes of our emulated Target Trials with their corresponding counterparts in the benchmark trials. Finally, Section 7.5.5 details the criteria for determining the agreement between our emulated Target Trials and their corresponding benchmark trials, and Section 7.5.6 lists the software that will be used for the analyses.

We acknowledge the scarcity of clinical trials comparing treatment sequences that have been conducted in the England, leading to the absence of fully suitable benchmark trials for a proof-of-concept study with “direct benchmarking” using NCRAS data. Nevertheless, we are confident that our systematic review (to be detailed in a forthcoming publication) has identified the best suited benchmark trials for our project, despite their limitations. Specifically, PC1 will emulate an Analogue Target Trial of the GUTG-001 trial using Flatiron data. This serves as a direct proof-of-concept study for comparing treatment sequences using Flatiron data. Depending on PC1's success, PC2 will expand the same approach to a broader population and comparison of alternative treatment sequences in

prostate cancer that are prevalent in both the US and England. PC2 will continue to use Flatiron data, functioning as a method extension study. PC3 will then follow, replicating PC2's design (i.e., Analogue Target Trial of PC2) and analysis but using NCRAS data, thus acting as an indirect proof-of-concept study for sequence comparison analysis with NCRAS data. Additionally, the RCC case studies, constrained to replicating a single treatment sequence from RECORD-3 due to the unavailability of the other sequence in England, will nonetheless function as a direct proof-of-concept for single-arm studies using NCRAS data. The interconnections between each planned Target Trial analyses are illustrated in Figure 7.1 and further elaborated in Section 7.5.3 and 7.5.4.

In summary, our case study designs creatively overcome the scarcity of direct benchmark trials for examining the feasibility of sequencing comparisons with NCRAS data. By leveraging and optimising existing, albeit imperfect, benchmarks, we aim to evaluate the feasibility of using NCRAS data to support local HTA sequencing decisions in England.

7.5.2. Considerations of applying causal inference methods in the NCRAS and Flatiron data

Advanced causal inference methods (i.e. marginal structural models with IPW and other g-methods) will be applied to mimic the effect of randomisation in analysing RWD through the principle of "no unmeasured confounders". Therefore, it is important to understand if all important prognostic factors affecting treatment decisions and outcomes can be well captured in the NCRAS and Flatiron data. For selecting patients, we require basic characteristics (e.g. age, sex) and key tumour prognosis factors (e.g. tumour size, tumour histology, tumour stage) and any factors that might influence survival at diagnosis and the time of treatment switching. For outcome measures, we need the death date of patients (i.e. estimating OS), indication of treatment relapse (i.e. estimating progression free survival (PFS)), and factors influencing patients being lost to follow-up (e.g. moving out of the country). Details of variables required will be described for each case study. We recognise that some desired variables might be unavailable. Thus, a key aspect of the study involves exploring the possibility of extracting necessary information from a blend of other related variables, in cases where direct mapping is absent. The process of variable selection and operational definition, and the use of proxy variables, will be subject to further discussions with clinical experts.

The principal investigator, JYAC, has experience in analysing disease registry and EHR data. Her PhD supervisors, NL and JBC, have extensive experience in research with cancer trial and registry data, oncology HTA and prostate cancer screening programs in the UK. They will help facilitate project collaborations with clinical experts, including Dr Carmel Pezaro, Professor Derek Rosario, and Professor Janet Brown, all part of the project team specialising in prostate and kidney cancer treatments. This project may include other statistical experts specialised in g-methods in the future in the research team or in post-hoc consultation, if necessary.

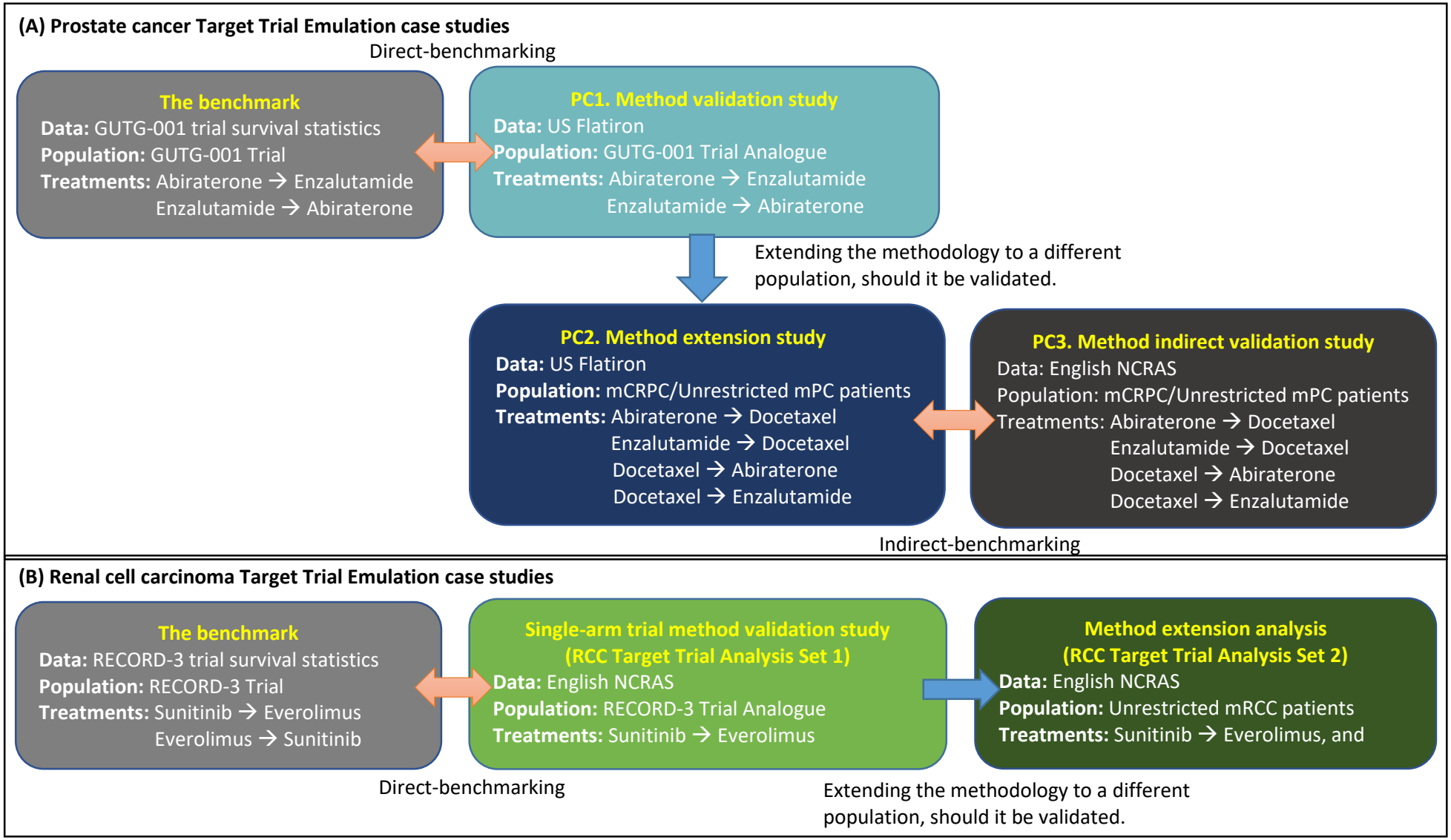


Figure 7.1 Schematic overview of interrelationships and purposes of each proposed Target Trial Emulation case study

mPC: metastatic prostate cancer; mRCC: metastatic renal cell carcinoma, TTE: Target Trial Emulation.

The orange arrows represent the comparison between benchmarks and their emulated counterparts, while the blue arrows indicate the application of validated emulation methods to a new population.

Furthermore, the application of causal inference methods also relies on the comparison of counterfactual pairs. That is, a certain degree of overlap with respect to patient characteristics between study groups are required to create a reasonable comparison. Thus, we will assess the overlap of patients receiving different treatment sequences and summarise in descriptive statistics. This project is also designed to understand the extent to which the completeness of the data may have an impact on the expected study results (i.e. results deviate from the benchmark trial). It may be possible that we are unable to fully replicate results from the chosen benchmarks. In this case, we will document the potential reasons (e.g. insufficient sample size, incomplete data on time-varying prognostic factors) and provide discussions around how NCRAS and Flatiron data may be enhanced to enable the similar analyses in the future.

7.5.3. Sequential treatment Target Trial: prostate cancer case studies

7.5.3.1. Benchmark RCT

In our prostate cancer case study, we identified a phase-2 RCT by Khalaf et al., the GUTG-001 trial[84], as a valuable benchmark. This trial compared the following two treatment sequences in treating treatment-naïve metastatic CRPC (mCRPC) patients:

- abiraterone (plus prednisolone) followed by enzalutamide (n = 101)
- enzalutamide followed by abiraterone (plus prednisolone) (n =101)

Although a recent commentary from some oncologists suggest the use of these two treatment sequences under specific circumstances in the UK, it seems unlikely that we can obtain an adequate sample size (n > 100) for a TTE study involving both sequences using NCRAS data.[515] The reason being these two drugs cannot be used directly after one another within the NHS. Had these treatment sequences been more prevalent in the UK, the GUTG-001 trial would have been a "perfect" benchmark for evaluating the feasibility of comparing treatment sequences with NCRAS data. On the other hand, these sequences have been more commonly used in the US[437], rendering the GUTG-001 trial an effective benchmark when using Flatiron data.

Given the absence of an ideal benchmark trial for a direct proof-of-concept study with NCRAS data, it is challenging to determine if NCRAS data can be used to reliably estimate the effectiveness of treatment sequences (in prostate cancer). To address this, we propose a novel strategy that involves jointly utilising the NCRAS and US Flatiron databases to —indirectly — assess the feasibility of deriving reliable effectiveness estimates from English data for treatment sequence comparisons. Our approach hinges on leveraging analyses with Flatiron data as a benchmark or "bridge" for assessing the analyses performed with NCRAS data. It involves a comparison of emulating identical Target Trials using both NCRAS and Flatiron data. Specifics of this design are elaborated in Section 7.5.3.3.

7.5.3.2. Demographics of metastatic prostate cancer

England

In England, around 6,000 new cases of metastatic prostate cancer are diagnosed annually.[525] Our oncology expert Dr. Carmel Pezaro noted that prostate cancer patients often start receiving castration therapy post-prostate cancer diagnosis, encompassing either surgical castration (i.e., bilateral orchiectomy) or medical castration (i.e., life-long androgen deprivation therapy, ADT). Professor Derek Rosario added that surgical castration accounts for only a very small fraction of these cases. Life-long ADT involves luteinising hormone-releasing hormone (LHRH, also known as gonadotrophin-releasing hormone (GnRH)) agonists or antagonists, such as padeliporfin and degarelix.[558] If the cancer progresses despite castration, the condition is termed castration-resistant prostate cancer (CRPC). The time of developing castration-resistance (i.e. hormone-relapse) from being castration-sensitive (i.e. hormone-sensitive) varies among patients, with an English study indicating that about 28% of prostate cancer patients may develop castration-resistance.[526] Patients with prostate cancer are typically managed by oncologists, while those on long-term ADT alone may be overseen by GPs post-initial treatment.

Docetaxel has been accessible as a first-line treatment for treatment-naïve patients with mCRPC within the NHS since 2006, whereas abiraterone and enzalutamide, two ARTAs, have been available for the same indication through The Cancer Drugs Fund (CDF) since 2016.[126,127,513] Since 2012 and 2014, respectively, abiraterone and enzalutamide have been introduced as second-line therapy options in the NHS (through CDF) for the treatment of patients with mCRPC who have previously undergone docetaxel treatment.[448,514,527] Prior to the introduction of abiraterone and enzalutamide, docetaxel therapy served as the sole standard treatment for patients with mCRPC. Since 2016, Cabazitaxel has been included as an alternative second-line treatment option in the treatment pathway (through CDF), exclusively for patients who have previously received docetaxel therapy.[237] Additional treatment options for patients at a later phase of mCRPC are available, including Radium-223 being approved since 2016 for patients with bone metastases (through CDF).[528] In May 2023, Olaparib was approved for mCRPC patients with breast cancer gene (BRCA) mutations through CDF[559], while the use of Lutetium-177 Vipivotide Tetraxetan for PSMA-positive patients after two or more prior treatments was not recommended in a recent NICE TA.[560]

Importantly, the English standard practice does not allow for the sequential use of abiraterone and enzalutamide, as mentioned earlier.[527] However, Dr. Carmel Pezaro confirms that if patients experience severe adverse events (such as toxicity) with either drug, they can switch to the other agent without it being considered as disease progression at that time. Such switching typically occurs within three months of treatment initiation, while switching after three months may indicate disease

progression.

Prostate cancer treatment has shifted, now integrating additional systematic treatments alongside ADT at earlier stages before developing metastasis or castration-resistance. Notably, darolutamide[529] and apalutamide[530] have been approved for high-risk non-metastatic castration-resistant prostate cancer (nmCRPC) patients, available through the CDF since 2021 and 2022[531], respectively. For newly diagnosed metastatic hormone-sensitive prostate cancer (mHSPC), the treatment options have expanded to include docetaxel[532], enzalutamide (available through CDF since 2021)[533], apalutamide (available through CDF since 2022)[534], and darolutamide (available through CDF since 2023 for patients ineligible for chemotherapy with docetaxel).[531,535] Key clinical trials comparing ADT alone to ADT combined with docetaxel in treating mHSPC were conducted from 2004 to 2013, including the GETUG-AFU15 (France)[536,537], the CHAARTED (US)[538], and the STAMPEDE (UK).[539,540] The results of these studies, published between 2013 and 2019, were inconsistent regarding the benefits of upfront docetaxel and the specific mHSPC patient subgroups that might benefit from it, leading to varied adoption timelines across different medical practices for the use of docetaxel in treating mHSPC.

Reflecting on these recent treatment advancements, our oncologist Dr. Pezaro noted that patients now receiving abiraterone or enzalutamide as first-line treatments for mCRPC might have previously undergone other treatments like docetaxel when their cancer was castration-sensitive. Dr. Pezaro suggested that this trend in treatment strategy likely became more apparent after 2016-2017, following publication of UK-based STAMPEDE trial results. The widespread use of medications in England depends not only on the UK Medicines and Healthcare Products Regulatory Agency (MHRA) approval but also on recommendations from NICE. Consequently, treatments other than docetaxel as upfront treatment prior to the development of mCRPC did not become widely adopted until after 2021. In rare cases, docetaxel may be re-administered upon disease relapse while receiving abiraterone or enzalutamide (i.e. docetaxel (castration-sensitive prostate cancer) → abiraterone/enzalutamide (mCRPC) → docetaxel (mCRPC)).

Identifying castration-resistant patients in NCRAS data can be challenging due to the absence of a variable documenting the date of patients becoming castration-resistant. Dr. Pezaro suggested that the emergency use of certain drugs in England upon the onset of castration resistance could serve as a potential proxy indicator, although it is uncommon. These drugs include non-standard treatment options like maximal ADT[541], which involves adding an additional androgen receptor (e.g., bicalutamide) to the standard ADT (e.g., leuprorelin, cetrorelix). Another option is adding low-dose dexamethasone. Professor Derek Rosario seconded these statements but noted the uncertainty surrounding the complete capture of emergency drug usage in NCRAS data. Inclusion of such information, even if available, may result in the identification of a distinct group of patients, given

the variation in treatment preferences among different physicians without a defined standard practice.

Given this protocol was developed after the onset of Covid-19 pandemic, it is important to acknowledge certain limitations associated with utilising retrospective data from the relevant period. Professor Rosario noted that docetaxel was not recommended for prostate cancer patients during the Covid-19 outbreak since around mid-2020, while abiraterone emerged as a preferred treatment option. This shift implies that treatment patterns during the pandemic might differ from other periods. Therefore, sensitivity analyses may be needed to investigate the adequacy of overlaps between patients receiving comparator treatment sequences, factoring the impact of including patients from different periods. This is crucial because the propensity of a patient receiving a certain treatment sequence may be influenced not only by their personal characteristics but also by the changing nature of treatment paradigms over time.

In summary, directly replicating the benchmark trial GUTG-001 using English NCRAS data is unfeasible due to the specific treatment patterns of mPC patients in England. Particularly, GUTG-001 investigated the effects of sequential treatments with abiraterone followed by enzalutamide, and the reverse sequence, but such sequential use is not permitted in England.

US (Flatiron population)

In the US, the treatment options for prostate cancer are generally similar to those in England, with one significant difference being the ability to use abiraterone and enzalutamide in a sequential manner. A publication based on Flatiron data indicated that the primary treatment sequences for prostate cancer in the US during 2013-2017 were abiraterone followed by enzalutamide or the reverse sequence.[437] In contrast to England, the US FDA granted approval for abiraterone and enzalutamide in castration-sensitive prostate cancer treatments, in 2018 and 2019, respectively[447,542], with adoption influenced by the US National Comprehensive Cancer Network (NCCN) guidelines[561], local practices, and individual insurance coverage.

Flatiron data experts were consulted during the data application process. Their preliminary analysis indicated that the Flatiron database could capture approximately 600 mCRPC patients who were treated with abiraterone followed by enzalutamide, as well as around 400 mCRPC patients who received enzalutamide followed by abiraterone. As of the data cut-off on March 31, 2019, a total of 4,000 metastatic prostate cancer patients were identified as having received a first-line treatment, and among them, 1,700 patients had undergone a second-line therapy. These statistics suggest the potential feasibility of replicating the GUTG-001 trial with Flatiron data.

7.5.3.3. Target Trial Emulation (TTE)

Our planned TTE analyses in prostate cancer involves a series of interconnected and progressive components. The specifics of these steps are outlined in Figure 7.2 and further explained in the

following paragraphs.

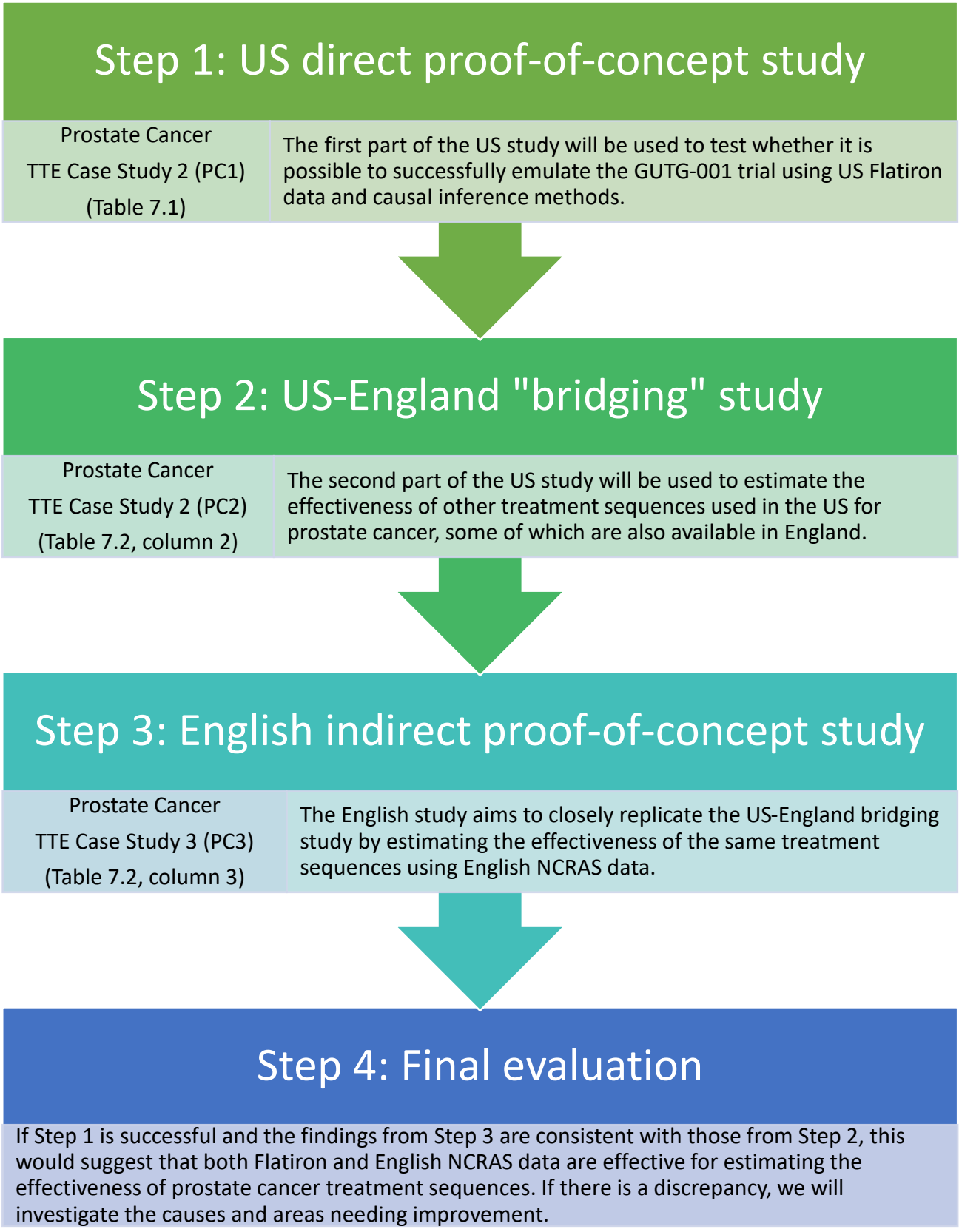


Figure 7.2 Flow chart of the US-England coupled sequential treatment Target Trial emulation studies in prostate cancer

National Cancer Registration and Analysis Service (NCRAS); PC: prostate cancer; TTE: Target Trial Emulation; US, United States

7.5.3.3.1. Prostate cancer TTE 1 (PC1)

The first step involves conducting a TTE (PC1) (see Figure 7.2) to assess the feasibility of emulating the GUTG-001 trial (i.e., GUTG-001 Analogue) with US Flatiron data, leveraging causal inference methods. This serves as a direct critical proof-of-concept study to identify appropriate statistical methods for obtaining reliable effectiveness estimates of treatment sequences from RWD. Success in replicating the GUTG-001 results using US Flatiron data will imply the potential for the same methods to be applied in comparable scenarios, including comparisons of other treatment sequences across broader populations. Table 7.1 outlines the specifics of the PC1 TTE analyses.

7.5.3.3.2. Prostate cancer TTE 2 (PC2)

If the benchmarking in the US proof-of-concept study (PC1) prove successful, Step 2 (PC2) (see Figure 7.2) will expand these methods for TTE in broader populations, focusing on comparing alternative treatment sequences prevalent in both the US and England using Flatiron data. These first- and second-line treatment sequences for treating treatment-naïve mCRPC include:

- docetaxel followed by enzalutamide upon disease progression
- docetaxel followed by abiraterone upon disease progression
- docetaxel followed by cabazitxel upon disease progression
- enzalutamide followed by docetaxel upon disease progression
- abiraterone followed by docetaxel upon disease progression

Subsequently, results from PC2 could serve as an "emulated benchmark trial" (i.e., a bridge), providing a basis for comparison with the Target Trial analyses in PC3 (Section 7.5.3.3.3). Such comparisons aim to indirectly validate the applicability of the same methods for comparing treatment sequences in the NCRAS database. In addition to functioning as an emulated benchmark, PC2 also aims to examine how the estimates of PC1 will change when the restrictions on the patient population are relaxed. PC2's detailed Target Trial design is presented in Table 7.2, column 2.

Dr. Pezaro notes that the exact date of developing castration-resistance may be unavailable in the database, particularly the NCRAS data, given its nature as a disease registry. If NCRAS data lacks specific timing for mCRPC diagnosis an alternative could involve including all newly diagnosed metastatic patients, irrespective of their hormone status. In such cases, treatment sequences for metastatic prostate cancer (mPC) patients could be relevant, which include:

- docetaxel (plus ADT) followed by enzalutamide upon disease progression
- docetaxel (plus ADT) followed by abiraterone upon disease progression
- ADT alone followed by abiraterone upon disease progression
- ADT alone followed by enzalutamide upon disease progression
- ADT alone followed by docetaxel upon disease progression

Table 7.1 Prostate cancer case study 1: Target Trial using US Flatiron Data to replicate the GUTG-001 Trial

	Original benchmark RCT (GUTG-001)[84]	PC1: GUTG-001 Analogue Target Trial
Purpose of the Target Trial	N/A	A direct proof-of-concept study: replicating the GUTG-001 trial using Flatiron data
Eligibility criteria	The eligibility criteria presented here are an abridged version from Khalat et al.'s 2019 publication, with further details in GUTG-001's protocol (Version 6.0) on ClinicalTrials.gov.[562] Patients who were aged 18 years or older and had newly diagnosed prostate adenocarcinoma without evidence of neuroendocrine differentiation, with metastatic disease on CT scan, MRI, or bone scan, and a rising PSA (PSA progression per PCWG2 criteria) with castrate concentrations of testosterone (≤ 1·7 nmol/L) with ongoing medical castration or previous bilateral orchiectomy. Patients were required to receive LHRH agonist or antagonist therapy for the duration of study treatment if not surgically castrated. Eligible patients were required to have adequate organ function, defined as absolute neutrophil count $1{\cdot}5\times10^9$ cells/L or higher, platelet count 100×10^9/L or higher, haemoglobin 80 g/L or higher, creatinine clearance 30 mL/min or higher, serum potassium higher than lower limit of normal range, total bilirubin 1.5 times upper limit of normal or less, and alanine aminotransferase and aspartate aminotransferase five times upper limit of normal or less. Patients who were previously treated with any CYP17A1 inhibitors (e.g. abiraterone, enzalutamide or experimental androgen receptor inhibitors) were excluded, while previous use of docetaxel for castration-sensitive disease was allowed. Patients who had contraindications to abiraterone and enzalutamide were excluded per manufacturer's label. Other exclusion criteria were ECOG performance status more than 2, brain metastases, active epidural disease, severe concurrent illness or comorbid disease, active concurrent malignancy, history of seizures or cerebrovascular events, major surgery within 4 weeks of starting study treatment, gastrointestinal disorders affecting absorption, and life expectancy of less than 6 months. The presence of visceral metastasis and pain requiring opioid analgesia were allowed.	Matching the eligibility criteria of the GUTG-001 trial as far as possible, following GUTG-001's protocol version 6.0.[562]

Treatment strategies	• Group A: patients received abiraterone 1000 mg orally once daily plus prednisone 5 mg orally twice daily as first study treatment until confirmed PSA progression, wide-field radiotherapy of symptomatic bone metastases, unacceptable treatment-related toxicity or withdrawal of consent. They then crossed over to receive enzalutamide 160 mg orally once daily until symptomatic or clinical progression, unacceptable treatment-related toxicity, or withdrawal of consent. • Group B: patients received enzalutamide and abiraterone plus prednisone in a reverse sequence until confirmed PSA progression, wide-field radiotherapy of symptomatic bone metastases, unacceptable treatment-related toxicity, or withdrawal of consent. Patients then crossed over to receive enzalutamide 160 mg orally once daily until symptomatic or clinical progression, unacceptable treatment-related toxicity, or withdrawal of consent. Dose modification for treatment-related adverse events was allowed.	• Group A: patients receiving abiraterone plus prednisolone followed by enzalutamide • Group B: patients receiving enzalutamide followed by abiraterone plus prednisolone For all treatment sequences, patients may switch to second-line treatment in cases of disease relapse* or unacceptable treatment-related toxicity. Additionally, patients may discontinue their first-line treatment without proceeding to subsequent treatment, based on clinical/patient decisions.
Assignment procedures	Eligible patients were randomly assigned (1:1) to receive abiraterone + prednisolone followed by enzalutamide or the reverse. Investigators and participants were not masked to treatment assignment.	Same as in GUTG-001 To effectively emulate the randomisation, we need to adjust for all measurable confounding factors to ensure the comparability of two treatment arms (counterfactual) at baseline (i.e., screening visit prior to randomisation). To align with GUTG-001, we plan to use the initiation date of first-line treatment as the reference point for assessing patients' baseline characteristics (i.e., time zero). This aligns with GUTG-001's tracking of time-to-event outcomes from first-line treatment commencement, which was within five days post-randomisation. The randomisation emulation will be performed using inverse probability weighting or other g-methods (e.g. standardisation). Important prognostic factors will be used to derive propensity score using a multivariable regression model. These important prognostic factors include age, tumour status, ECOG performance status, prior treatments, comorbidities, and PSA level. The final covariate selection will be based upon discussion with

		clinicians, and will be based upon attempting to satisfy the "no unmeasured confounding" assumption.
Follow-up period	Patients were followed up since the initiation of their first-line treatment, which began within five days of randomisation, until either death, data-cut off or lost-to-follow-up, whichever occurred first. The median duration of follow-up in the GUTG-001 trial was 30.7 months (IQR 25.1-36.2) of the data cut-off (May 31, 2018). Given that the final enrolment in the GUTG-001 trial occurred on December 13, 2016, the minimum follow-up period would have been approximately 17 months, had no patients been lost to follow-up.	The follow up begins with the initiation of the first-line therapy until the occurrence of death, loss to follow-up, or data cut-off, whichever occurs first. Our analysis will target patients who could have a theoretical minimum follow-up of 17 months, matching the GUTG-001 trial's follow-up duration as closely as possible. For example, for a data cut-off date of May 31, 2018, we will include all patients who were eligible for enrolment before the end of 2016, regardless of their actual follow-up period. Patients with less than 17 months of actual follow-up will be marked as lost to follow-up (censored), ensuring alignment with the GUTG-001 trial without introducing selection bias. The criterion of 17 months may be relaxed if the sample size is insufficient.
Outcomes**	<u>Primary endpoints:</u> 1. Time to second PSA progression: time from the start of first-line therapy to PSA progression on second-line therapy, or death from prostate cancer before crossover, whichever occurred first. 2. The proportion of patients with PSA response on second-line therapy. <u>Secondary endpoints:</u> 1. The proportion of patients with PSA response on first-line therapy. 2. Time to PSA progression on first-line therapy: time from the start of first-line therapy to confirmed PSA progression on first-line therapy (Preliminary results of this endpoint were reported in Annala et al. 2019[563]) 3. Time to PSA progression on second-line therapy: time from crossover to confirmed PSA progression 4. Overall survival: time from the start of first-line therapy to time of death from any cause, or last follow-up (censored); 5. Time on treatment for second-line therapy: time from crossover to end of second-line treatment or death 6. Time to clinical progression on second-line therapy: time from crossover	<u>Primary endpoints:</u> 1. OS, measured as the time from the start of first-line therapy until death from any cause, or last follow-up (censored). <u>Secondary endpoints:</u> 1. Time to second progression, defined as time from the start of first-line therapy to progression* on second-line therapy, or death from prostate cancer before crossover, whichever occurred first. 2. Time to progression on first-line therapy, defined as the time from the start of first-line therapy to any type of progression*, including death from prostate cancer. <u>Exploratory endpoints (</u>PSA-related endpoints are contingent on the availability and quality of PSA levels in the Flatiron database): 1. Time to progression on second-line therapy, defined as the time from crossover to any type of progression*, including death from prostate cancer. Since both the GUTG-001 trial and the TTE are designed to evaluate the effectiveness of treatment sequences from the initial baseline (i.e., the start of first-line treatment, analyses that use the time

to clinical progression on second-line therapy, including death from prostate cancer (This endpoint was not analysed because the endpoint was subject to variability in individual physician decision making of local study investigators).

7. Safety of second-line abiraterone and enzalutamide
8. Change in Montreal Cognitive Assessment score on first-line and second-line therapy (Results of this endpoint was reported elsewhere).[564]
9. Correlation of cell-free DNA biomarkers with PSA response after first-line and second-line treatment.

Post-hoc analysis:

1. Time to progression on first-line therapy: time from treatment initiation to confirmed PSA progression, radiographic progression (PCWG2 criteria), clinical progression, or prostate cancer-related death, whichever occurred first (preliminary results was reported in Annala et al. 2019[563]).
2. Time to progression on second-line therapy: time from crossover to confirmed PSA progression, radiographic progression (PCWG2 criteria), clinical progression, or prostate cancer-related death, whichever occurred first.
3. Time to second progression: time from treatment initiation to confirmed PSA progression, radiographic progression (PCWG2 criteria), clinical progression on second-line therapy, or or prostate cancer-related death, whichever occurred first.
4. Comparison of second-line PSA responses between groups using Pearson's chi-square test
5. Clinical correlates of time to PSA progression and PSA response in patients receiving second-line enzalutamide
6. Comparison of crossover clinical characteristics between groups
7. Sensitivity analysis of time to second PSA progression (primary endpoint), excluding patients with delayed crossovers, > 2 weeks
8. Comparison between groups of time from first progression of any kind to crossover
9. Subgroup analysis to determine whether second-line enzalutamide was better than second-line abiraterone in all patient subgroups

of treatment crossover as a secondary baseline are prone to bias, and such analyses should be adjusted for prognostic characteristics. Adjustments were not made in the published analyses of the GUTG-001 trial and this is not the focus of our analysis. Therefore, we regard this as an exploratory endpoint and will only present naïve exploratory analyses for this endpoint.

2. Time to second PSA progression: time from the start of first-line therapy to PSA progression on second-line therapy, or death from prostate cancer before crossover, whichever occurs first.
3. Time to PSA progression on first-line therapy: time from the start of first-line therapy to PSA progression, including death from prostate cancer.
4. Time to PSA progression on second-line therapy: time from crossover to PSA progression, including death from prostate cancer.
5. The proportion of patients with a PSA response on first-line therapy.
6. The proportion of patients with a PSA response on second-line therapy.

Causal contrasts of interest	All endpoints were analysed using the intention-to-treat principle, with first-line or combined treatment endpoints evaluated in all randomised patients, and second-line treatment endpoints assessed in those who switched treatments.	Main analysis: Analogue of per-protocol effect: estimating the hypothetical effect had all patients adhered to the treatment strategy to which they are assigned in our analyses Exploratory: 1. Analogue of intention-to-treat effect: estimating the effect according to the first-line therapy. 2. Analogue of as-treated effect: estimating the effect restricted to those who received the specific treatments sequences outlined in our analysis. First-line or combined treatment endpoints will be evaluated in all randomised patients, and endpoints for second-line treatments will be evaluated specifically in those patients who crossover.
Analysis plan**	• Time-to-event outcomes: KM survival curves and log-rank tests were used. Hazard ratios and 95% CI were estimated from Cox proportional hazard models for PFS, OS and combined PFS, stratified by the MSKCC risk criteria. • Proportion of PSA response: compared between groups using Pearson's chi-square test. • Comparison of crossover clinical characteristics between groups: Continuous-valued characteristics were compared using the rank-sum test, and Boolean characteristics were compared using Fisher's exact test. All Cox regression analyses, associated confidence intervals, and Kaplan-Meier curves were calculated using R (version 3.6.0) with the survival package (version 2.44.1.1). Confidence intervals for PSA response, Pearson's chi-square tests, rank-sum tests and Fisher's exact tests were calculate using Julia (version 1.1.0) with the HypothesisTests package (version 0.8.0).	• Descriptive analyses will be conducted to understand the overall treatment pattern in the data and to estimate the sample size. • T-tests and chi-square tests will be conducted to compare patient characteristics at treatment initiation and at cross-over, and compared with those in the GUTG-001 trial • KM survival curves, survival probability (e.g. median survival)/event incidence, and cox proportional hazard ratios (and/or risk ratios using pooled logistic regression), will be conducted for all time-to-event outcomes. Time-to-event outcome analyses will be performed for the per-(TTE)-protocol analogue, intention-to-treat analogue, and as-treated analogue effect. Marginal structural models with inverse probability weighting and other G-methods will be used to emulate the randomisation process of the Target Trial and account for time-varying confounders (see Section 4.7 in Chapter 4 for the discussion of advanced methods, and Chapter 8 for the implementation of inverse probability weighting method in this case study).[3,5,289]

ALT, alanine aminotransferase; AST, aspartate aminotransferase; CI: confidence interval; CYP17A1, Cytochrome P450 Family 17 Subfamily A Member Enzyme; ECOG, Eastern Cooperative Oncology Group performance status; HTA: health technology assessment; KM, Kaplan-Meier; IQR, interquartile range; LHRH, maintain luteinising hormone-releasing hormone, mCRPC, metastatic

castration-resistant prostate cancer; MSKCC, Memorial Sloan Kettering Cancer Center; NCRAS, National Cancer Registration and Analysis Service; PCWG2, Prostate Cancer Working Group 2; PFS, progression-free survival; PSA, prostate-specific antigen test; OS, overall survival; RCT, randomised controlled trials; TTE: target trial emulation; US, United States

* The final operational definition of disease progression can vary across databases and will be determined through discussions with clinical experts and data experts, due to the uncertainty in data quality. If a specific progression date is unavailable, we might use a composite of variables to obtain a proxy date of disease progression, such as treatment discontinuation.
**Colour coding: The colour-coding scheme highlights the comparable outcomes between the GUTG-001 trial and the planned Target Trial Emulation Study PC1, with comparable outcomes marked in the same colour.
- Red: OS
- Pink: Time to second progression.
- Purple: Time to progression on first-line therapy.
- Blue: Time to progression on second-line therapy.
- Green represents comparable endpoints related to PSA assessment.
- Orange represents the comparison of patient characteristics at crossover. Although not explicitly stated as an outcome in the PC1 study, the comparison of patient characteristics is mentioned in the analysis plan as part of the descriptive analysis.
- Black text indicates outcomes in the GUTG-001 trial that are not explored in the PC1 study, including safety endpoints and those unlikely to have data in both the Flatiron and English Cancer Registry datasets, as well as those without full results reported in the GUTG-001 publication for enabling RCT versus real-world evidence (RWE) agreement assessment in Section 7.5.5.
The outcomes in the PC1 study are selected based on their projected feasibility to enable RCT-RWE agreement assessment (i.e., from high to low: primary endpoints > secondary endpoints > exploratory endpoints) as planned in Section 7.5.5. These selections also consider their potential to inform bridging studies PC2 and PC3 (Sections 7.5.3.3.2-7.5.3.3.3). The primary endpoint in the PC1 study is OS, as mortality data is a relatively reliable endpoint available in both Flatiron data and the English Cancer Registry. In contrast, OS is a secondary endpoint in the GUTG-001 trial, potentially due to typically shorter follow-up periods in trials and likely immature OS data. Secondary endpoints in the PC1 study involve progression assessments, which may require proxy definitions using other available information in real-world data (RWD), making them secondary rather than primary. Furthermore, even if progression information is available in RWD, it could be an indicator resulting from a combination of assessments (progression based on clinical, PSA, or radiology; time-to-treatment-discontinuation) rather than a single measure. Hence, the comparable progression-related endpoints between the GUTG-001 and PC1 case studies are those that contain comprehensive assessments of progression. Finally, all PSA-related endpoints are listed as exploratory in the PC1 study, unlike in the GUTG-001 trial, which lists PSA assessments as primary and secondary endpoints. This classification is due to the uncertain completeness and longitudinal consistency of PSA information for each patient in RWD.

Table 7.2 Prostate cancer case study 2 & 3: US-England bridging study and English indirect proof-of-concept study

	PC2: An emulated benchmark using Flatiron data	PC3: PC2 Analogue Target Trial using NCRAS data
Purpose of the Target Trial analysis	• US-England bridging study: comparing prostate cancer treatment sequences common in England using Flatiron data to serve as an emulated benchmark for PC3 • Investigate how estimates from PC1 might differ when patient population restrictions are relaxed.	An indirect proof-of concept study: replicating the "emulated benchmark trial(s)" in PC2 using English NCRAS data.
Eligibility criteria	Analysis Set 1: All patients who were aged 18 years or older with mCRPC will be included. Analysis Set 2: According to Dr. Pezaro, the exact date of a patient becoming castration-resistant prostate cancer may be unknown or only available as a proxy in the database. Should the data quality of NCRAS be inadequate for determining the timing of mCRPC diagnosis, the inclusion of all newly diagnosed metastatic patients who were aged 18 years or older will be considered as an alternative (i.e., including mHSPC patients). For a detailed justification, please refer to the "patient inclusion/exclusion criteria" in Section 7.5.3.4.	Matching PC2 as far as possible.
Treatment strategies	• Group A: docetaxel followed by enzalutamide if disease relapse* or unacceptable treatment-related toxicity. • Group B: enzalutamide followed by docetaxel Alternative treatment strategies may be chosen for comparison based on the sample size of each treatment sequence in Flatiron data and NCRAS data: • docetaxel followed by abiraterone • abieraterone followed by docetaxel • abieraterone followed by cabazitaxel • abiraterone followed by enzalutamide • enzalutamide followed by abiraterone For Analysis Set 2, we will consider comparison of treatment sequences that either start with a systematic treatment (e.g. docetaxel) or no treatment (i.e., with only baseline ADT).	Same as in PC2

	For all treatment sequences, patients may switch to second-line treatment in cases of disease relapse* or unacceptable treatment-related toxicity. Additionally, patients may discontinue their first-line treatment without proceeding to subsequent treatment, based on clinical/patient decisions.	
Assignment procedures	Participants are randomly assigned to one of two strategies at baseline. To effectively emulate the randomisation, we need to adjust for all measurable confounding factors to ensure the comparability of two treatment arms (counterfactual) at baseline. The randomisation emulation will be performed using propensity score matching, inverse probability weighting or other g-methods (e.g. standardisation). Important prognostic factors will be used to derive propensity score using a multivariable regression model. These important prognostic factors include age, tumour status, ECOG performance status, prior treatments, comorbidities, and PSA level[†]. The final covariate selection will be based upon discussion with clinicians, and will be based upon attempting to satisfy the "no unmeasured confounding" assumption.	Matching PC2 as far as possible.
Follow-up period	The follow-up period starts from the time of treatment initiation (sensitivity analysis: starts from the time of diagnosis (mCRPC for Analysis Set 1, mPC for Analysis Set 2)) and continues until the event of death, loss to follow-up, or the data cut-off date, whichever comes first. In PC 1, follow-up begins on the date of first-line mCRPC treatment initiation (baseline), aligned with the GUTG-001 trial, implying that all patients included in the trial have survived to receive their first-line mCRPC treatment. However, this could potentially lead to immortal time bias in RWD analysis if the time from mCRPC diagnosis to first-line treatment initiation differs significantly between the two study groups (which might not be fully adjustable with statistical methods). While this concern might be less significant in PC1, as abiraterone and enzalutamide are often interchangeable according to our oncology experts, we would like to examine such design's impact on estimating the comparative treatment effectiveness from RWD in the sensitivity analyses for PC2. Specifically, we will begin follow-up at diagnosis instead, and use techniques like cloning[75,334] and/or TTE with sequential eligibility criteria[76] to assist in assigning treatment groups, contrasting it with setting the first-line treatment initiation as baseline (time zero).	Matching PC2 as far as possible. However, it is crucial to acknowledge that for sensitive analysis, the dates of castration-resistance and metastasis might be unavailable in NCRAS data.

Outcomes	Primary endpoints: 1. Overall survival: the time from treatment initiation (for primary analysis)/diagnosis (for sensitivity analysis) until death from any cause, or last follow-up (censored). Secondary endpoints 1. Time to second progression: time from diagnosis to progression* on second-line therapy, or death from prostate cancer before crossover, whichever occurred first. 2. Time to progression on a first-line therapy, defined as the time from diagnosis to any type of progression*, including death from prostate cancer. Exploratory endpoints: 1. Time to progression on second-line therapy, defined as the time from crossover to any type of progression*, including death from prostate cancer. Since the GUTG-001 trial and the TTE are designed to evaluate the effectiveness of treatment sequences from the initial baseline (i.e., the start of first-line treatment, analyses that use the time of treatment crossover as a secondary baseline are prone to bias, and such analyses should be adjusted for prognostic characteristics. Adjustments were not made in the published analyses of the GUTG-001 trial and this is not the focus of our analysis. Therefore, we regard this as an exploratory endpoint and will only present naïve exploratory analyses for this endpoint.	Matching PC2 as far as possible.
Causal contrasts of interest	Main analysis: Analogue of per-protocol effect: estimating the hypothetical effect had all patients adhered to the treatment strategy to which they are assigned in our analyses Exploratory: 1. Analogue of intention-to-treat effect: estimating the effect according to the first-line therapy. 2. Analogue of as-treated effect: estimating the effect restricted to those who received the specific treatments sequences outlined in our analysis. First-line or combined treatment endpoints will be evaluated in all randomised patients, and endpoints for second-line treatments will be evaluated specifically in those patients who crossover.	Matching PC2 as far as possible.

Analysis plan	• Descriptive analyses will be conducted to understand the overall treatment pattern in the data and to estimate the sample size. • T-tests and chi-square tests will be conducted to compare patient characteristics at treatment initiation and at cross-over. • KM survival curves, survival probability (e.g. median survival)/event incidence, and cox proportional hazard ratios (and/or risk ratios using pooled logistic regression), will be conducted for all time-to-event outcomes. Time-to-event outcome analyses will be performed for the intention-to-treat analogue, as-treated analogue, and per-(TTE)-protocol analogue effect. Marginal structural models using inverse probability weight and other G-methods will be used to emulate the randomisation process of the Target Trial and account for time-varying confounders (see Section 4.7 in Chapter 4 for the discussion of advanced methods).[3,5,289]	Matching PC2 as far as possible.

ECOG, Eastern Cooperative Oncology Group performance status; HTA: health technology assessment; KM, Kaplan-Meier; mCRPC, metastatic castration-resistant prostate cancer; mHSPC, metastatic hormone-sensitive prostate cancer; mPC: metastatic prostate cancer; NCRAS, National Cancer Registration and Analysis Service; PSA, prostate-specific antigen test; RCT, randomised controlled trials; TTE: target trial emulation; US, United States
N/A: not applicable

* The final operational definition of disease progression can vary across databases and will be determined through discussions with clinical experts and data experts, due to the uncertainty in data quality. If a specific progression date is unavailable, we might use a composite of variables to obtain a proxy date of disease progression, such as treatment discontinuation.
† PSA level is unavailable in the NCRAS data, and therefore will not be included in the English NCRAS analysis.

7.5.3.3.3. Prostate cancer TTE 3 (PC3)

Step 3 (PC3) (see Figure 7.2): The design of PC2 allows us to conduct an indirect proof-of-concept study assessing if the quality of NCRAS data is adequate for reliably assessing the effectiveness of treatment sequences using causal inference methods. PC3 aims to emulate the same Target Trial as PC2, but using English NCRAS data. Differing from the PC1, PC3 will only indirectly reference information from the original benchmark RCT (i.e., GUTG-001 trial). PC3's detailed Target Trial design is presented in Table 7.2, column 3.

Comparison between similar analyses in NCRAS data (PC3) and Flatiron data (PC2) relies on the existence of common treatment sequences in both datasets. There are several common treatment sequences in treating prostate cancer in the US and the UK as described in 4.3.3.2. Docetaxel followed by enzalutamide and its reverse sequence represent one of the most prevalent treatment sequences available in both the US and the UK,[437,527] and therefore these sequences are chosen as the main comparators. However, alternative sequences like docetaxel-abiraterone (and vice versa) or other non-symmetric pairs remain viable (e.g. abiraterone-docetaxel versus enzalutamide-docetaxel, abiraterone-docetaxel versus abiraterone-enzalutamide, enzalutamide-docetaxel versus enzautamide-abiraterone), subject to the sample sizes in both the Flatiron and the NCRAS data.

Additionally, Professor Rosario highlighted the possibility of assuming no systematic difference between abiraterone and enzalutamide, given their similar mechanisms of action. Consequently, for PC2 and PC3, we can also consider including the following two sequences to increase the sample size if necessary:

- docetaxel followed by abiraterone or enzalutamide upon disease progression
- abiraterone or enzalutamide followed by docetaxel upon disease progression

The final selection of sequence pairs for PC2 & PC3 will be based on ensuring larger sample sizes than in the GUTG-001 trial for both Flatiron and English NCRAS data (guided by analyses of treatment patterns analyses from the actual data), and finalised after discussions with clinicians regarding their clinical interests. In the exploratory analysis, sequence pairs with varying rates of crossover/treatment regimen violations may be compared. This may be used to assess the performance of methods potentially sensitive to a small percentage of patients adhering to the assigned treatment strategy.

7.5.3.3.4. Final evaluation

If Step 1 is successful and the findings from Step 3 are consistent with those from Step 2 (see Figure 7.2), this would suggest that both Flatiron and English NCRAS data are effective for estimating the effectiveness of prostate cancer treatment sequences. If there is a discrepancy, we will investigate the causes and areas needing improvement. We intend to assess the extent of discrepancy

using the matrix outlined in Section 7.5.5

7.5.3.4. Patient inclusion/exclusion criteria

The study targets patients aged 18 and over with mCRPC. Depending on the quality of Flatiron and NCRAS data, as well as clinical implications, this focus may broaden to include all metastatic prostate cancer (mPC) or all advanced prostate cancer cases. Specifically, Prof Rosario and Dr. Pezaro are concerned about the databases' ability to accurately identify when patients become castration-resistant, a status typically determined by prostate-specific antigen (PSA) levels. For the English population, this information is unavailable in English NCRAS data and currently only accessible in CPRD data.[403,424] Additionally, as docetaxel is increasingly used in mHSPC patients prior to castration-resistance, distinguishing between patients who received docetaxel post-castration resistance and those treated while hormone-sensitive in the databases poses a significant challenge.

The 2020 UK National Prostate Cancer Audit Annual Report indicates an increase in the use of docetaxel with standard ADT for new metastatic prostate cancer cases, from 27% in 2019 to 36% in 2020.[495] This usage varied widely across different NHS providers in England, ranging from 0% to 47%.[495] Nevertheless, it may still be possible to capture mCRPC patients among all metastatic patients using NCRAS data if we have patients' full treatment trajectory and time-varying prognostic factors not only limited to those upon the emergence of metastases. A crucial part of the study process is to investigate whether a proxy of patient's disease status (i.e. castration-resistant) can be defined by specifying an algorithm when a direct variable is lacking. This rationale supports our requests for specific data periods and relevant variable information outlined in Section 7.6 Data Requirements.

In summary, identifying mCRPC patients in RWD, especially within the English NCRAS, may be challenging. We aim to collaborate with clinicians to identify these patients using a combination of variables if necessary. If we cannot develop an ideal algorithm for this purpose, we will relax the restriction regarding mCRPC patients. Specifically, we will analyse newly diagnosed metastatic patients regardless of their castration-status. Most importantly, the same criteria will be applied to both PC2 & PC3 to enable a fair comparison between these "bridging-studies". Although the study population may slightly differ from what was initially planned, it remains clinically relevant. Notably, Professor Rosario highlighted a shift in the treatment approach for metastatic prostate cancer, increasingly favouring the early use of systematic anti-cancer therapy alongside ADT, even before patients develop castration-resistance.

7.5.3.5. Sample size estimation

The Flatiron dataset is expected to provide a sufficient sample size for our case studies (PC1 & PC2), as indicated by a US treatment pattern study from 2013-2017. This study showed 227 patients receiving docetaxel followed by enzalutamide (including docetaxel-only patients) and 414 receiving

enzalutamide followed by docetaxel (including enzalutamide-only patients).[437] Furthermore, preliminary data evaluations by Flatiron specialists have confirmed the number of prostate cancer patients receiving relevant treatments appears to be sufficient for our study (i.e., preferably exceeding the sizes in the GUTG-001 trial[84] and previous observational studies of prostate cancer treatment sequences, aiming for at least 100-250 patients[516,565]).

On the other hand, while precise estimation of patients receiving docetaxel and/or abiraterone/enzalutamide annually in NCRAS data is challenging, it is likely to be sufficient. This assessment is based on the published UK epidemiology data described Section 7.5.3.2.[525] Specifically, according to a recent NICE TA, there are approximately 5,500-5,800 CRPC patients who may be eligible for a first-line treatment in England and Wales every year.[126] Further, abiraterone, enzalutamide, and docetaxel are among the most frequently administered therapies.[126,127,448,514]

7.5.4. Sequential treatment Target Trial: renal cell carcinoma case studies

7.5.4.1. Benchmark RCT

RECORD-3, an international phase-2b trial, has been identified as a valuable benchmark in evaluating treatment sequences for metastatic treatment-naïve metastatic RCC (mRCC) patients. This trial compares the efficacy of the two following treatment sequences in treating metastatic treatment-naïve mRCC patients: [86,406]

- everolimus as first-line therapy followed by second-line sunitinib (n = 238)
- sunitinib as first-line therapy followed by second-line everolimus (n = 233)

Despite being a potential benchmark, only one of the treatment sequences in RECORD-3 has been available in the NHS. Everolimus was never recommended as first-line therapy and sunitinib was not recommended as second-line therapy in the NHS.[520] This statement has been confirmed with our medical oncologist, Professor Janet Brown and further supported by recent observational data from three UK hospitals.[521] Everolimus accounts for only 0.6% of all treatments among 652 mRCC patients in 2008-2015.[521] Further, only 0.5% of all patients who ever received a second-line treatment received sunitinib as a second-line therapy.[521] On the contrary, sunitinib accounts for 60.7% of first-line therapies, and everolimus accounts for 41.9% of second-line therapies. However, it is likely that the percentage of patients using sunitinib followed by everolimus as first- and second-line therapy has been decreasing because other newer agents for treating mRCC have been available since 2015. Particularly, everolimus has gradually been shifted to be used as a later-line therapy (e.g. 3rd or 4th line) in recent years, while sunitinib remains as a common first-line therapy.[566] Despite these changes, historical data from the NCRAS, especially for the RCC incident cohort in 2015 and 2016 (at the beginning of acquisition of newer treatments in the NHS), may still provide insights into the sequence of sunitinib followed by everolimus.

Similar to the prostate cancer study series (Section 7.5.3), there is no "perfect" benchmark trial for the RCC. However, at least one of the treatment sequences in RECORD-3 is likely to exist in the NCRAS data. Thus, RECORD-3 remains a valuable reference, enabling a direct comparison between the benchmark trial results and the Target Trial analysis conducted using NCRAS data. In summary, the primary aim of this RCC case study is to explore the feasibility and reliability of using NCRAS data to replicate results of a single arm in the RECORD-3 trial (i.e. sunitinib → everolimus).

7.5.4.2. Demographics of renal cell carcinoma (RCC) in England

In England, approximately 13,000 new kidney cancer cases are reported annually (2015-2017)[543], with more than 80-90% being RCC.[525] Approximately one-fourth of these patients present with advanced-stage cancer and 25-34% have metastases at diagnosis.[543,544] About 75% of advanced-stage RCC patients are eligible for a first-line systematic therapy.[545]

Current treatment for mRCC primarily includes targeted therapy (e.g., tyrosine kinase inhibitors, TKIs) and immunotherapy (e.g., PD-1 and PD-L1 inhibitors).[521,567] In the NHS, sunitinib and pazopanib have been first-line therapies for mRCC since 2009 and 2011, respectively.[498] Tivozanib and cabozantinib (both through CDF) were introduced as additional first-line treatment options for mRCC starting in 2018.[531] Subsequently, more treatments became exclusively available as first-line therapies for mRCC through the CDF[531], including nivolumab plus ipilimumab (since 2019 and further passed NICE CDF review in 2022), avelumab-axitinib combination (since 2020), lenvatinib-pembrolizumab combination (since 2023). The NHS has adopted the following treatments for mRCC patients who are in need of second-line or later-line therapies: axitinib[568] (from 2015), nivolumab[545] (from 2016), everolimus (from 2017 through CDF, previously through CDF for more restricted indication), cabozantinib (from 2017 through CDF), and the lenvatinib-everolimus combination (from 2018 through CDF).[531]

Professor Janet Brown stated that in England, before targeted therapy and immunotherapy became widespread, approximately 10-15% of patients with less advanced cancer were treated with interferon-alpha before receiving any other systemic anti-cancer therapies. Additionally, everolimus has been shifted from second-line treatment to later stages of treatment (third or fourth-line treatment) in recent years for patients with an Eastern Cooperative Oncology Group (ECOG) score below 2. This adjustment reflects the lack of substantial survival benefits of everolimus and instead emphasises its role in improving patients' quality-of-life.

Professor Brown clarified that there is no standardised adjuvant therapy for mRCC patients, indicating that systematic therapy may or may not accompany surgical interventions before metastases appear.[546] She added that patients who are suitable for operations like nephrectomy, and those who do not rapidly develop metastases post-nephrectomy, usually have a more favourable

prognosis. Specifically, these patients are typically deemed healthy enough to undergo surgical interventions at the first place. Conversely, elevated calcium levels in patients typically indicates a poorer prognosis.

Professor Brown further suggested that patients with mRCC who remain in good health conditions (i.e. desirable performance status) may receive up to 5 or 6 lines of treatment with each subsequent relapse. However, a significant proportion (roughly 40-50%) of mRCC patients receive only three lines of treatment before death. This can be attributed to either their frailty preventing further treatment upon disease progression or mortality occurring before subsequent treatments. The introduction of newer treatment options, particularly cancer immunotherapies since 2016, has significantly improved the overall survival (OS) of mRCC patients.[547] Previously, patients who received solely TKIs had an OS range of 1-2 years up to 4-5 years after being diagnosed.

7.5.4.3. Target Trial Emulation

This section presents the design of a single-arm Target Trial to assess the viability of using NCRAS data to replicate the sunitinib to everolimus sequence results from the RECORD-3 trial. Table 7.3 specifies these details.

In routine clinical practice, unlike controlled sequential treatment trials, patients may receive various second-line treatments based on the outcomes of their first-line therapy. Additionally, some patients might not receive any second-line treatment, and the reasons for this can differ from those seen in clinical trials. For instance, in the NCRAS data, patients initially treated with sunitinib might have different second-line treatments, though everolimus could have been an option. In our emulation of the RECORD-3 trial's sunitinib to everolimus arm, solely analysing NCRAS data patients who completed this treatment sequence could be problematic. Such an approach implies selecting patients based on a non-random future decision: the progression to second-line everolimus is contingent on the outcome of the first-line sunitinib treatment, rather than a predetermined treatment plan. Selection conditioning on a post-treatment variable (i.e., as-treated effect) could lead to immortal-time bias. Nevertheless, such an approach is not uncommon in observational studies and one should be cautious in interpreting the results of these analyses.[569,570]

In our RCC study, we aim to highlight the strengths of causal inference in estimating treatment sequence effectiveness in a single-arm Target Trial by contrasting the following three effects:

- As-treated effect: include only patients who received sunitinib as their first-line treatment and proceeded to everolimus as their second-line therapy for time-to-event outcome assessment
- Standard per-protocol effect: include all patients who received sunitinib as their first-line therapy for time-to-event outcome assessment and censor those who did not proceed to

everolimus as their second-line therapy by the time of treatment-switching

- Hypothetical per-protocol effect assuming complete adherence to treatment assignment: include all patients who received sunitinib as a first-line therapy for time-to-event outcome assessment and adjust for treatment-switching with causal inference methods if they did not receive everolimus as a second-line therapy

While the as-treated effect clearly faces the risk of immortal time bias, the standard per-protocol effect could encounter issues if censoring is informative. On the other hand, the hypothetical per-protocol effect, assuming complete adherence to the assigned treatment, somewhat mirrors the approach used in RCTs to address treatment-switching.[277] This involves adjustments for unintended switches, such as when patients initially assigned to standard therapy subsequently move to a new drug. Implementing this approach could help mitigate biases in estimating the effects of treatment sequences. While there are similarities in the causal inference methods applicable, our single-arm Target Trial analysis, which uses RWD, presents additional complexities. Specifically, it lacks a randomisation baseline for reference and patients in real-world settings often switch between multiple drugs. Part of our study involves assessing if varying methods for the hypothetical per-protocol effect yield discrepancies, and how to interpret these differences.

The challenges in replicating the RECORD-3 trial single-arm results could stem from several factors. A primary concern is confounding by subsequent treatments beyond second-line. About half of patients in the RCORD-3 trial received additional treatments after the second-line, but detailed information on these is unavailable. It may be possible that the options of subsequent treatments are different or the rate of receiving a subsequent treatment is incomparable to the clinical practice in England. Another potential issue is the difference in patient demographics, including age and adherence to treatment, despite matching the same inclusion/exclusion criteria of the trial as far as possible. If feasible within our project's limited timeline, we might consider undertaking exploratory analyses to see whether additional causal inference methods can aid in resolving these issues. These methods may include adjustments for variations in treatment exposure due to non-adherence (e.g. delay of treatment), or methods of extending inferences and formulating external control arms using RWD (i.e., emulating the sunitinib → everolimus arm as an external control for the trial's everolimus → sunitinib arm).[381,571]

7.5.4.4. Patient inclusion/exclusion criteria

The RCC case study mainly focuses on patients with mRCC. In this study, Target Trial Analysis Set 1 strives to closely match the eligibility criteria of the RECORD-3 trial, essentially creating a RECORD-3 trial-mimicking population (Table 7.3: Eligibility criteria). Conversely, the Target Trial Analysis Set 2 aims to extrapolate the findings in a more generalised population, including all mRCC

patients aged 18 or older.

7.5.4.5. Sample size estimation

In the UK, it is estimated that each year, 2,500 to 3,000 new cases of mRCC are diagnosed. Approximately three-quarters of these patients (est.n = 1,875-2,250), might be eligible for first-line therapy. According to a recent UK study, prior to 2015, sunitinib may account for roughly 61% of first-line therapies before 2015.[521] This suggests that around 500 to 600 patients in the NCRAS data (2012-2018) could be eligible for our study. Given these figures, pooling mRCC incident cohorts from 2012 to 2018 (or later) should provide us with a sufficiently large sample size for the single arm Target Trial analyses.

Table 7.3 A single-arm Target Trial replicating sunitinib-everolimus arm in the RECORD-3 trial using NCRAS data

	Benchmark RCT: RECORD-3 trial[86,406]	A single-arm Target Trial: Analogue of the RECORD-3 sunitinib-everolimus arm using NCRAS data
Eligibility criteria	The eligibility criteria presented here are an abridged version from Knox et al.'s 2017 publication, with further details in RECORD-3's protocol on ClinicalTrials.gov.[405] Patients aged 18 years or older with measurable mRCC as per RECIST v1.0 were included. Prior nephrectomy was not a prerequisite. Key eligibility criteria included no previous systemic therapy, a KPS score of 70% or higher, adequate hematologic, liver, and kidney function, and a normal left ventricular ejection fraction. Patients with brain metastases were excluded.	Target Trial Analysis Set 1: Matching the eligibility criteria in the RECORD-3 as far as possible. The KPS score information is not available in the NCRAS data; however, it may be converted into an ECOG score.[572] Target Trial Analysis Set 2: All patients who were aged 18 years or older with mRCC will be included.
Treatment strategies	• Group A: first-line everolimus 10 mg/day until PD followed by sunitinib (4 weeks on, 2 weeks off as second-line therapy (n = 238) • Group B: first-line sunitinib 50 mg/day (4 weeks on, 2 weeks off) until PD followed by everolimus as second-line therapy (n = 233) The crossover period is defined as the interval between the end of the first-line treatment and the start of the second. The crossover (initiation of the second-line treatment) should occur within 35 days of disease progression. Patients had a minimum 2-week period after discontinuation of the first-line drug because of progression before beginning the second-line drug. Dose modifications were permitted for adverse events.	Single group: first-line sunitinib until disease progression* followed by everolimus as second-line therapy. Treatment crossover (the initiation of second-line therapy) should occur within 35 days of progression, and patients should have a minimum 2-week period after discontinuation of the first-line drug because of progression before beginning the second-line drug. Dose modifications were permitted for adverse events. Professor Brown noted that patients switching treatments due to toxicity, rather than disease progression (thus not advancing in treatment lines), typically did so within 3-6 months of starting therapy. In such cases, patients often moved to another drug within the same class, like from one TKI to another TKI. For those who started with everolimus before 2015, a reduced dosage sequence (e.g., 50 mg/day → 37.5 mg/day → 25 mg/day) upon disease progression was common due to the lack of alternative later-line therapies at that time.
Assignment procedures	Eligible patients were randomly assigned (1:1) to receive either everolimus followed by sunitinib or the reverse sequence. The random assignment was stratified by MSKCC risk criteria (favorable, intermediate, or poor risk). Patients received the first-line drug until disease progression (according to RECIST v1.0), discontinuation due	All patients will be receiving sunitinib as first-line therapy and followed with a second-line everolimus upon disease progression*. In this single-arm Target Trial, randomisation is not required. Our focus is on replicating the condition where patients were assigned a treatment sequence of

	to unacceptable toxicity, or for any other reason. Upon disease progression, patients were eligible to switch to the second-line drug until further progression.	sunitinib followed by everolimus. Although the RECOTD-3 trial stratified patient recruitment using the MSKCC risk criteria, this information is unavailable in NCRAS data. Professor Brown indicated that ECOG might not be an ideal substitute for MSKCC criteria, and therefore, we will conduct further investigations into alternative variable combinations that could serve as better proxies.
Follow-up period	Patients were followed up since the initiation of their first-line treatment, until either death, data-cut off, lost-to-follow-up, whichever occurred first. Since the RECORD-3 trial's final enrolment was in June 2011 and the data cutoff for the final analysis was in June 2014, the minimum follow-up period would have been approximately 3 years, had no patients been lost to follow-up and all remain alive.	The follow up begins with the initiation of the first-line therapy until the occurrence of death, loss to follow-up, or data cut-off, whichever occurs first. Our analysis will target patients who could have a theoretical minimum follow-up of 3 years, matching the RECORD-3 trial's follow-up duration as closely as possible. For example, for a data cut-off date of December 2021, we will include all patients who were eligible for enrolment before the end of 2018, regardless of their actual follow-up period. Patients with less than 3 years of actual follow-up will be marked as lost to follow-up (censored), ensuring alignment with the RECORD-3 trial without introducing selection bias. The criterion of 3 years may be relaxed if the sample size is insufficient.
Outcome**	<u>Primary endpoints:</u> 1. PFS of first-line therapy: time from the first date of first-line treatment to progression during first-line treatment or death from any cause. Patients without progression* or death at data cut-off for the analysis or at the time of receiving additional anticancer therapy, including the second-line drug were censored at their last date of adequate tumour evaluation. <u>Secondary endpoints:</u> 1. Combined first- and second-line PFS: the time from randomisation to progression after second-line treatment or death from any cause. Patients who did not crossover to second-line therapy or who did not experience progression after the start of second-line treatment or who were alive at data cut-off for the analysis or at the time of receiving an additional anticancer therapy, were censored at last date of tumour evaluation.	<u>Primary endpoints:</u> 1. OS: time from the first date of first-line treatment to death <u>Secondary endpoints:</u> 1. Combined first- and second-line PFS: the time from the first date of first-line treatment to progression* after second-line treatment or death from any cause. Patients who did not crossover to second-line therapy or who did not experience progression* after the start of second-line treatment or who were alive at data cut-off for the analysis or at the time of receiving an additional anticancer therapy were censored at their last date of structural activity plus grace period. 2. PFS of first-line therapy: the time from the first date of first-line treatment to progression* during first-line treatment or death from any cause. Patients without progression* or death at data cut-off for the analysis or at the time of receiving additional anticancer therapy, including the second-line drug, were censored at their last date of structural activity plus grace period.

	2. OS: time from randomisation to death (no formal power calculation was made, and the expected number of deaths was 300) Exploratory endpoints: 1. Second-line PFS: time from the start of second-line treatment to progression or death	Professor Janet Brown suggested that the judgement of disease progression in routine practice (i.e. a combination of clinical and radiology progression assessment) may vary between clinicians and may differ from clinical trials, which often involve independent central using RECIST. Patients remain on a treatment or radiology showing < 20% increase may be an indicator of disease remains stable without progression.
Causal contrasts of interest	Per-protocol effect was estimated for first-line PFS, OS and combined PFS. A notable proportion of patients were censored in estimating the combined PFS and OS in RECORD-3, owing to delayed crossover to second-line therapy, with 57% patients in the sunitinib-everolimus arm and 56% patients in the everolimus-sunitinib arm. The crossover, which is the start of receiving second-line treatment, should occur within 35 days of progression. The period between the end of first-line treatment and the beginning of second-line therapy is the crossover period.[86]	Main analysis: 1. Hypothetical per-protocol effect: estimating the hypothetical effect had all patients adhered to the treatment strategy to which they are assigned in our analyses. This approach includes all patients who received sunitinib as a first-line therapy for time-to-event outcome assessment and adjust for treatment-switching with causal inference methods if they did not receive everolimus as a second-line therapy or did not have a timely cross-over. Exploratory: 1. Analogue of intention-to-treat effect: estimating the effect according to the first-line therapy. 2. As-treated effect: estimating the effect according to the actual treatment sequences. This approach only includes patients who received sunitinib as their first-line treatment and proceeded to everolimus as their second-line therapy for time-to-event outcome assessment. 3. Standard per-protocol effect: This approach includes all patients who received sunitinib as their first-line therapy for time-to-event outcome assessment and censor those who did not proceed to everolimus as their second-line therapy by the time of treatment-switching or did not have a timely cross-over without any adjustment for informative censoring
Analysis plan	Kaplan-Meier survival curves and log-rank tests. Hazard ratios and 95% CI were estimated from a Cox proportional hazard models for PFS, OS and combined PFS, stratified by the MSKCC risk criteria.	• Descriptive analyses will be conducted to understand the overall treatment pattern in the data and to estimate the sample size. • Patient characteristics at treatment initiation and at cross-over will be examined and compared with those in the RECORD-3 trial • KM survival curves, survival probability (e.g. median survival)/event incidence, and cox proportional hazard ratios (and/or risk ratios using pooled logistic regression), will be conducted for all time-to-event outcomes.

	Time-to-event outcome analyses will be performed for the as-treated, and standard per-protocol effect, the hypothetical per-(TTE)-protocol effect assuming complete adherence to treatment assignment. Marginal structural models using inverse probability weight and other G-methods will be used to facilitate the emulation of the hypothetical per-protocol effect assuming complete adherence to treatment assignment and account for time-varying confounders (see Section 4.7 in Chapter 4 for the discussion of advanced methods).[3,5,289]

PD, progressive disease; ECOG, Eastern Cooperative Oncology Group performance status; KM: Kaplan-Meier; KPS, Karnofsky performance status; mRCC, metastatic renal cell carcinoma; MSKCC, Memorial Sloan-Kettering Cancer Center; NCRAS, National Cancer Registration and Analysis Service; OS: overall survival; PFS: progression-free survival; RCT, randomised controlled trials; RECIST, Response Evaluation Criteria in Solid Tumors; TKI, tyrosine kinase inhibitors; TTE: target trial emulation

* The final operational definition of disease progression can vary across databases and will be determined through discussions with clinical experts and data experts, due to the uncertainty in data quality. If a specific progression date is unavailable, we might use a composite of variables to obtain a proxy date of disease progression, such as treatment discontinuation.
† PSA level is unavailable in the NCRAS data, and therefore will not be included in the English NCRAS analysis.

**Colour coding: The colour-coding scheme highlights the comparable outcomes between the RECORD-3 trial and the planned RCC case study, with comparable outcomes marked in the same colour.
- Red: OS
- Pink: Combined first- and second-line PFS
- Purple: PFS of first-line therapy
- Black text indicates outcomes in the GUTG-001 trial that are not explored in the RECORD-3 study

The outcomes in the RCC case study are selected based on their projected feasibility to enable RCT versus real-world evidence (RWE) agreement assessment (i.e., from high to low: primary endpoints > secondary endpoints > exploratory endpoints) as planned in Section 7.5.5. The primary endpoint in the RCC case study is OS, as mortality data is a relatively reliable endpoint available in the English Cancer Registry. In contrast, OS is a secondary endpoint in the RECORD-3 trial, potentially due to typically shorter follow-up periods in trials and likely immature OS data. Secondary endpoints in the RCC case study involve progression assessments, which may require proxy definitions using other available information in real-world data (RWD), making them secondary rather than primary.

7.5.5. Benchmark trial-RWE agreement assessment

This section delves into the methods to assess the alignment between benchmark clinical trials and their emulated counterparts, which are crucial for evaluating the extent of and discussing potential reasons for any discrepancies observed. The matrix primarily comprises three components, with the first three assessment criteria adapted from those used in the RCT DUPLICATE studies.[471,553,573]

(1) Regulatory agreement: This component assesses whether the RWE replicates its benchmark's results (such as hazard ratio (HR) and risk ratio (RR)) in terms of both direction and statistical significance observed in the benchmark trials. While RCT DUPLICATE established criteria for evaluating regulatory agreement in both superiority and non-inferiority trials[553], our chosen benchmarks (such as GUTG-001 and RECORD-3) were not intended for regulatory use and hence did not specify any non-inferiority margin. Therefore, our primary focus is to compare RWE and RCT data and treated them as superiority trials, evaluating whether the direction and significance of RWE's estimates matched those of the benchmarks. Endpoints with non-significant effects in RCTs should also show no significant effect in RWE.

(2) Estimate agreement: Considering the potentially disproportionately large sample size of RWE, achieving statistical significance might be easier compared to the benchmarks (and thus not easier to meet the first criteria). Therefore, this step examines whether the point estimate of RWE's effect sizes falls within the 95% confidence intervals (CIs) of the benchmark trial. Furthermore, we added an extra procedure to include the comparison of non-relative effect estimates for time-to-event outcomes. For example, it examines whether the point effect of median survival estimates falls within the 95% confidence interval of the trial.

(3) Exploratory - standardised differences: The third criterion is exploratory in nature. It involves computing the standardised difference to compare the relative effect estimates from the benchmark and the RWE, to determine whether there is a statistically significant difference in the estimated effects, as indicated in the RCT DUPLICATE study, using the formula below.[553]

$Z = \frac{\hat{\theta}_{RWE} - \hat{\theta}_{RCT}}{\sqrt{\sigma^2{}_{RWE} + \sigma^2{}_{RCT}}}$, where an absolute Z-value less than 1.96 indicates no significant difference between the estimates from RWE and RCT.

(4) Exploratory - survival curve comparison: This additional criterion, specific to our study, involves comparing RWE survival curves with those from the benchmark trial for time-to-event outcomes. The key aspect here is assessing whether the point estimates of the RWE survival curve for each treatment-sequence group fall within the 95% CI of the benchmark trial. Since RCTs typically present Kaplan-Meier (KM) curves, without patient-level data, we aim to reconstruct patient-level survival data from benchmark RCTs using Guyot et al.'s digitisation method for extracting information from published KM curves[574], wherever possible. We introduced this extra criterion

beyond what was included in RCT DUPLICATE, emphasising the importance of verifying whether absolute outcomes (like survival times) in our emulation match those in the RCT, and not solely focusing on relative effect estimates. This is crucial because there is a possibility that, even with similar relative effect estimates, the absolute effect may significantly vary, indicating a less ideal emulation. This holds particular importance in scenarios where RWE is employed to form external control arms, notably in our RCC case study (Section 7.5.4) which focuses on the emulation of a single arm.

This assessment matrix is specifically designed for comparing the benchmark-RWE pair in the PC1 case study (GUTG-001 versus its emulation using Flatiron data). However, its application to the benchmark-RWE pairs in the PC2, PC3, and RCC case studies can be limited in terms of interpretation. For example, PC2 utilises an emulated benchmark instead of a traditional RCT. Additionally, in PC3's single-arm study, only the fourth criterion and certain aspects of the second criterion, particularly the comparison of non-relative effects, are pertinent.

7.5.6. Software for analysis

The data will be analysed in R and STATA to conduct Target Trial emulations in different cancer types to explore the feasibility of applying causal inference methods for estimating the effectiveness of different treatment sequences using RWD. All analyses will be documented in R script and Stata. Do files.

7.6. Data acquisition

7.6.1. Overview

This section offers an overview of the specification of datasets and variables we originally planned to request from the NCRAS and Flatiron databases for the analyses detailed in Section 7.5 (Analysis Plan). For transparency, we reported our original specifications and included feedback received from analysts of both databases during the data application process. Our main objective is to gather detailed patient information to identify eligible patients who fit the inclusion/exclusion criteria of our Target Trial case studies. This includes two specific patient groups: 1) individuals with prostate cancer (ICD-10: C61x) in both the Flatiron and NCRAS databases, and 2) those with RCC (ICD-10: C64x) in the NCRAS database. Importantly, we also need longitudinal patient data for measuring time-varying treatments, covariates and outcomes, such as overall survival.

Patient selection and emulation taking into account time-varying treatment exposures will be based on various factors including basic characteristics (age, sex, date of diagnosis, cancer stage, performance status), tumour prognostic factors (histology, morphology, tumour size, co-morbidities), and treatment details (anti-cancer treatments, pre-diagnosis treatment history, treatment duration).

This information is needed at diagnosis, first-line treatment, and, if available, at the time of treatment switching. For evaluating outcomes, we require information on the dates of patients' deaths (to estimate overall survival), signs of treatment relapse (for assessing progression-free survival), and factors leading to follow-up loss (e.g., relocating out of the country), wherever available.

For the Flatiron data application, we submitted our study protocol and variable requirements, as outlined above and in Section 7.6.3, specifically for prostate cancer patients, to the Flatiron scientists for evaluation. This approach was necessary because Flatiron's data dictionary is not publicly available. In response, Flatiron informed us of potential limitations in using their database for our planned analysis. Contrary to NCRAS, which supplies only the variables specifically requested, Flatiron grants access to their entire standard disease-specific database, such as the metastatic prostate cancer database, upon approval of the data application.

For the NCRAS database application, we employed the most recent ODR NCRAS data dictionary template (v4.4)[434] available at the time when we initially applied. This helped us identify relevant variables and customise our data request, including justifications for each variable we requested. A complete list of datasets, variables, and cohort definitions for the NCRAS data application can be found in the Appendix 7.1. This list, along with the protocol, has been under review by the NHS ethics committee and the ODR/NHSD, and has been amended based on their feedback.

The subsequent paragraphs (Section 7.6.2-7.6.5) provide specifics on the datasets and variables we requested from Flatiron and ODR/NHSD, including detailed original content specifications.

7.6.2. Relevant datasets required

7.6.2.1. Flatiron datasets/tables required

All relevant tables in the standard Flatiron mPC dataset, containing necessary variable information as outlined in Sections 7.6.1 and 7.6.3, will be used wherever available.

7.6.2.2. NCRAS datasets required

- English Cancer Registry
- SACT dataset
- Radiotherapy dataset (RTDS)
- HES admitted care
- HES outpatient
- HES accident and emergency
- Cancer Waiting Time

7.6.3. Relevant variable information required

Our primary analyses (Figure 7.1) only planned to include patients that match with the inclusion/exclusion criteria in the benchmark RCTs (case study PC1 and RCC Analysis Set 1). However, we will conduct further "real-world effect" analyses on a wider population compared to the restricted trial population (case study PC2, PC3 and RCC Analysis Set 2). Consequently, our data application requests records of all patients diagnosed with prostate cancer or RCC, not just a specific subset. We requested all available pre- and post-diagnosis patient records, wherever available. Comprehensive data collection is crucial for our sequential treatment Target Trial analyses, especially for integrating time-varying exposures in assessing treatment sequence effectiveness. Table 7.4 presents the operational definitions of the required variable/information, with detailed justifications for each specific requested variable provided separately to ODR/NHS Digital using the ODR NCRAS data dictionary version 4.4[434].

Time-to-event outcomes like overall survival (OS) will be calculated from treatment initiation to the patient's death. In contrast, defining PFS requires careful consideration due to potential unavailability of exact progression dates. Our approach involves integrating proxy variables to estimate these dates, such as instances of treatment discontinuation. Patients who encounter adverse events or dropouts, when identifiable and pertinent to the protocol, may be censored as appropriate.

Operational definitions defined in Table 7.4 may be refined following discussions with clinical experts and database analysts, and upon discovering more relevant algorithms in the literature, such as those for comorbidity definition.

Table 7.4 Summary of variable requirements and operational definitions

	Prostate cancer case studies	**Renal cell carcinoma case studies**
Cancer type	ICD-10: C61x	ICD-10: C64x
Basic characteristics	Age, sex, date of diagnosis, date of first-line treatment, date of second-line treatment, cancer stage, ECOG performance status, tumour size, tumour histology, tumour morphology	
Comorbidities considered for replicating the inclusion and exclusion criteria of the benchmark trials in PC1 and RCC Analysis Set 2	• Brain metastases (ICD-10: C79.3) • Contraindications of abiraterone and enzalutamide - pregnancy (ICD-10: O00x, O01x, O02x, O03x, O04x, O05x, O06x, O07x, O08x)[575] (It is unlikely that men have any diagnosis of pregnancy, but this can be used to test the quality of the NCRAS/HES data) • Active epidural disease (G95x other and unspecified diseases of spinal cord) • Active concurrent malignancy (ICD-10: C00x-C43x, C45x-C96x, D00x-D05x, D07x-49x) • History of seizures or cerebrovascular events (ICD-10: G40x, I60x-69x)	• Metastases (ICD-10: C79x) (of other cancers) • Brain metastasises (ICD-10: C79.3): Professor Janet Brown suggested that roughly 1 in 20 mRCC patients may have de novo brain metastatses. • End-stage renal disease (ICD-10: N18.6)

- Gastrointestinal disorders affecting absorption (ICD-10: K90x)
- End-stage renal disease (ICD-10: N18.6)
- Major surgery within 4 weeks of starting study treatment

Comorbidities considered for adjusting prognostic factors

We planned to use the Charlson Comorbidity Index (CCI) from NCRAS data as an indicator of disease severity for adjusting baseline characteristics. Given the CCI's nature as a summary score with various algorithm versions, we may perform sensitivity analyses with alternative algorithms for defining comorbidities. A study utilising HES data validated a CCI ICD-10 translation (based on the Deyo and Dartmouth-Manitoba ICD-9 adaptations) for predicting in-hospital mortality in urological cancer surgery patients.[576] Although not specific to prostate cancer or RCC, this algorithm developed by Sundararajan et al. (outlined below) may be relevant as it has been validated in NHS cancer patients. Other validated tools, such as Elixhauser scores or individual comorbidity variables may also be considered subject to discussion with clinicians.[577] For example, for the US population, the Combined Comorbidity Score developed by the Harvard group may be particularly relevant.[578,579]

Condition	Weights	ICD-10-AM
Acute myocardial infarction	1	I21, I22, I252
Congestive heart failure	1	I50
Peripheral vascular disease	1	I71, I790, I739, R02, Z958, Z959
Cerebral vascular accident	1	I60, I61, I62, I63, I65, I66,G450, G451, G452, G458, G459, G46, I64, G454, I670, I671, I672, I674, I675, I676, I677 I678, I679, I681, I682, I688, I69
Dementia	1	F00, F01, F02, F051
Pulmonary disease	1	J40, J41, J42, J44, J43, J45, J46, J47, J67, J44, J60, J61, J62, J63, J66, J64, J65
Connective tissue disorder	1	M32, M34, M332, M053, M058, M059, M060, M063, M069, M050, M052, M051, M353
Peptic ulcer	1	K25, K26, K27, K28
Liver disease	1	K702, K703, K73, K717, K740, K742, K746, K743, K744, K745
Diabetes	1	E109, E119, E139, E149, E101, E111, E131, E141, E105, E115, E135, E145
Diabetes complications	2	E102, E112, E132, E142 E103, E113, E133, E143 E104, E114, E134, E144
Paraplegia	2	G81 G041, G820, G821, G822
Renal disease	2	N03, N052, N053, N054, N055, N056, N072, N073, N074, N01, N18, N19, N25
Cancer	2	C0, C1, C2, C3, C40, C41, C43, C45, C46, C47, C48, C49, C5, C6, C70, C71, C72, C73, C74, C75, C76, C80, C81, C82, C83, C84, C85, C883, C887, C889, C900, C901, C91, C92, C93, C940, C941, C942, C943, C9451, C947, C95, C96
Metastatic cancer	3	C77, C78, C79, C80
Severe liver disease	3	K729, K766, K767, K721
HIV	6	B20, B21, B22, B23, B24

Treatments	• All patients whoever received any systematic anti-cancer therapy for their prostate cancer. • Common treatments: abiraterone, enzalutamide, docetaxel, carbizitaxel, sipuleucel-T, radium-223	• All patients whoever received any systematic anti-cancer therapy for their RCC. • Common treatments: sunitinib, everolimus, pazopanib, sorafenib, temsirolimus, axitinib, nivolumab, cabozantinib, tivozanib, lenvatinib with everolimus, nivolumab with ipilimumab, avelumab with axitinib
Outcome measurements	Death date, disease status (i.e., disease progression, treatment discontinuation, metastasis (e.g. ICD-10: C79x), wherever available)	

7.6.4. Data time period

The full duration of the Flatiron mPC dataset is not publicly available, but the data curation lag is typically minimal. As of the March 31, 2019 cut-off, Flatiron's analysts provided initial sample size estimates for our protocol population. They identified about 4,000 mCRPC patients who underwent first-line treatment, with 1,700 advancing to second-line therapy. This includes about 600 patients treated first with abiraterone then enzalutamide, and 400 with the reverse sequence, suggesting a sufficient sample size for the study, especially with further inclusion of patients who received abiraterone or enzalutamide as first-line therapy without subsequent treatments. Based on this confirmation, we were able to confidently proceed with the data application for Flatiron's mPC dataset.

Patients diagnosed before April 2012 might lack comprehensive treatment data, essential for our treatment sequence analysis. The NDRS analysts further recommended focusing on patients treated from 2014 onwards due to the improved completeness of SACT data. Moreover, explicitly emulating the theoretical minimum follow-up periods of approximately 2 and 3 years, as observed in the GUTG-001[84] and RECORD-3[86] trials (i.e., from the last patient's enrolment to the data cut-off date), promises fairer comparisons. With data access expected to cover up to the end of 2019/2020 by the time of application, we aimed to include patients diagnosed or starting their first-line treatment between 2014-2017 for prostate cancer, and 2014-2016 for RCC case studies, taking into account both the respective minimum follow-up periods and SACT data maturity. However, to accommodate sensitivity analyses assessing the impact of treatment data and the inclusion of patients not meeting the theoretical minimum follow-up periods, we requested data for all patients diagnosed with prostate cancer (C61) or renal cell carcinoma (C64) between 01/01/2011 and 31/12/2020 (See Appendix 7.1). For these patients, we also requested an extended period of records—six years prior to their diagnosis and all records post-diagnosis. This is necessary to accurately define time-varying covariates, treatments, and outcomes as specified in Section 7.6.3 (See Appendix 7.1).

7.6.5. Geography criteria

For Flatiron data, we requested records from the entire population associated with US health

care providers (community/academic) within the Flatiron Health Network. Our NCRAS data request was exclusively for patients in England.

7.7. Chapter summary

This chapter revisited the significance of leveraging RWD as an alternative means to derive unbiased estimates of treatment sequence effectiveness in HTA, along with the associated challenges. It highlighted the appropriateness of designing proof-of-concept studies that borrow the strengths of the TTE framework and benchmarking techniques. The case studies in this chapter's protocol were designed to evaluate if databases identified in Chapter 5—NCRAS and Flatiron—are of sufficient quality to provide reliable estimates of treatment sequence effectiveness.

This chapter set forth specific aims for the aforementioned quest and introduced an innovative strategy for utilising the NCRAS and Flatiron databases in tandem for a series of direct and indirect benchmarking studies. It highlighted the novelty, uniqueness, and potential of indirect benchmarking, especially because an ideal benchmark RCT for direct benchmarking was unavailable for calibrating the utility of NCRAS data. Specifically, assessing the feasibility of using NCRAS data to provide unbiased comparative effectiveness estimates of treatment sequences—requiring at least two sequences within one case study—can only hinge on the support from Flatiron data (see Chapter 6). The design of my interconnected prostate cancer studies bridged the gap caused by the lack of benchmark RCTs featuring treatment sequences common in English clinical practice. However, it is essential to acknowledge that the ability to proceed with the indirect proof-of-concept studies (PC2 & PC3, Section 7.5.3.3.2 & 7.5.3.3.3), which aim to assess the utility of NCRAS data in comparing treatment sequences, is dependent upon the success of the direct proof-of-concept study utilising the Flatiron database (PC1, 7.5.3.3.1). Hence, the RCC case study (Section 7.5.4) was designed as a direct proof-of-concept study to assess the utility of NCRAS data, albeit with the limitation of being single-arm validations. However, it lays initial steps to examine the leverage of TTE and advanced statistical methods for creating external control reference treatment sequence arms.

The significance of this study goes beyond assessing the use of RWD in informing the effectiveness of treatment sequences for HTA. It also holds importance in the broader context of TTE studies. To date, this project serves as the first attempt within TTE research to compare treatment sequences, making it an initial foray into applying these methods in this specific area. While similarly rooted statistical techniques have been employed to compare other types of dynamic treatment regimen (e.g., determining the timing for initiating treatments based on lab test thresholds), they have not been adopted to compare treatment sequences for HTA.

Furthermore, in evaluating the agreement between RCT and RWE, my protocol has advanced beyond and complemented the status quo established by the RCT DUPLICATE research initiative[553]: The predefined assessment matrix introduced an additional criterion to assess the visual concordance in survival curves between the emulated Target Trial and the benchmark trial (Section 7.5.5). This

addition addressed a caveat in the established criteria that focus on examining the agreement of the relative risk between treatment groups. For instance, even when HRs in the emulated Target Trial appear similar to those of the benchmark RCT, the absolute outcomes (e.g., survival times) across groups may still be biased. It is, however, important to evaluate the absolute outcomes in HTA as they have implications for extrapolating patients' quality-adjusted life years and the cost of treatments in economic evaluations. Thus, the extra criterion aimed to ensures a more comprehensive assessment.

The protocol has been published online for transparency.[170] Chapter 8 dives into the complexities of adapting these outlined TTE into practical operations. It touches upon data quality concerns and provides a tutorial-style breakdown of the protocol's implementation process. Due to constraints on time and data access described in Chapter 5, the thesis presents only the implementation of the first prostate cancer case study (PC1) as a representative example. More information can be found in Chapter 8.

Chapter 8 Implementing Target Trial Emulation for Comparing Treatment Sequences Using the Flatiron database and NCRAS database

8.1. Chapter overview

This chapter dives into the execution of the Target Trial Emulation (TTE) case studies outlined in Chapter 7. Given the constrains of time, scale of the project, and challenges of data access (Chapter 5), my focus narrowed to implementing the first prostate cancer study (PC1) using Flatiron data (Chapter 7, Section 7.5.3.3.1). This chapter details the TTE implementation for comparing treatment sequences using RWD in a step-by-step manner, detailing the challenges encountered and the necessary compromises. It establishes the groundwork for further exploration of subsequent case studies in a Wellcome Trust transition fund-funded research position that I plan to undertake post-PhD, and aims to serve as a tutorial for similar future research.

Despite the primary focus on the PC1 case study, the initial part of this chapter (Section 8.2) gives a thorough overview of the importance of preliminary data checks for sequencing analysis, relevant across all case studies, including both prostate cancer (PC1 to PC3) and renal cell carcinoma (RCC) case studies. These were performed upon receipt of data from all candidate databases specified in my case study protocol (Chapter 7)—the Flatiron and the English National Cancer Registration and Analysis Service (NCRAS) databases. I explain how to validate received data against published data, perform consistency checks, and preliminary treatment pattern checks to refine sample size estimation prior to initiating TTE. I conclude by discussing the implications of these preliminary checks.

Section 8.3 transitions into the execution of the PC1 TTE case study, beginning with a 15-person toy example, illustrating the key elements of an analysis-ready dataset (Section 8.3.2). This toy example sets the stage for implementing the advanced inverse probability weighting (IPW) method, an advanced statistical technique identified in the methods review in Chapter 4, for comparing treatment sequences. It demonstrates how to operationalise the method at a conceptual level. The section then moves on to methods for identifying the target cohort and preparing fit-for-purpose analytic datasets using RWD of my case study (Section 8.3.3). I then outline the practical application of the IPW method to adjust for baseline and time-varying confounding factors. This theoretically enables the derivation of unbiased estimates for comparing two treatment sequences using RWD. I explain how I adapted the IPW method specifically for the context of PC1's research and the data available in the Flatiron database (Section 8.3.4). I then report the study findings, compare the results obtained with the IPW method with those using simple methods, and assess how well the real-world evidence (RWE) from the emulated TTE aligns with the benchmark RCT (Section 8.4). The assessment follows the predefined criteria set out in my study protocol (Chapter 7, Section 7.5.5).

Finally, in Section 8.5, I delve into the implications of these findings and steps to advance this line of research further.

8.2. Preliminary data checks for sequencing analyses

Upon receiving the Flatiron and English NCRAS data extracts, I conducted preliminary checks to assess the practicality of implementing the planned TTE analyses in Chapter 7. These checks were vital to ensure that the sample size was sufficient, either comparable to or exceeding the patient populations specified in the benchmark trials (GUTG-001[84] and RECORD-3[86]). No specific minimum sample size for TTE aimed at benchmarking is recommended in the literature.[3,37,77] Less strict inclusion and exclusion criteria can increase sample sizes, while matching the strict criteria of benchmark RCTs theoretically tends to result in smaller sample sizes.[580] A review indicates that sample sizes for TTE studies can greatly vary, depending on the research topic and data source.[581] Although clinical trial sample size estimation methods, designed to detect a specific level of treatment efficacy, may be applied[37,582], their relevance become less clear where assessing real-world impacts or safety signals in a generalised population where effect magnitude and direction are uncertain. Nonetheless, this highlights the reasoning for match or exceeding benchmark trial sample sizes in my analysis to enhance likelihood of identifying comparable differences between treatment groups and thus facilitating effective benchmarking.

The data checking process begins with logic checks to ensure that the received data matched what I had requested and aligned with the data dictionaries. This includes comparing the cancer incidence in my data with published statistics, checking the received data variables against the requested parameters, and ensuring all datasets within each database could be linked through key parameters. Then, I analysed all patients in each database to understand the treatment patterns, assessing the potential sample size for the planned TTE analyses. In the following sections, I describe the implications of these checks in detail.

8.2.1. Cancer incidence, variable completeness, and dataset linkage

The annual incidence of metastatic castration-resistant prostate cancer (mCRPC) patients in the Flatiron metastatic prostate cancer (mPC) dataset aligned with recent publications using the same database.[437] Moreover, all tables within this mPC dataset could be linked using patient unique ID without logical inconsistencies.

For the English NCRAS data, metastatic dates are only available for patients initially diagnosed with metastatic cancer, unlike the Flatiron database where the progression date to metastatic cancer can be determined even for those initially diagnosed at an early stage (see Section 5.3.3, Chapter 5). Consequently, I focused on comparing the incidence of all prostate and kidney cancer cases in England with published statistics for logic checks. The initial data extract provided by the NHS

Digital fell short of expectations, showing only about 4000 patients in total, markedly below the expected 30,000-45,000 new cases of prostate and kidney cancer annually according to published NHS statistics.[583,584] This shortfall was particularly notable since I requested data for patient cohorts from 2011 to 2020, but found that most patients in the dataset were diagnosed in 2011. Additionally, a Venn diagram analysis of unique patient IDs in the received NCRAS datasets, revealed patient ID mismatches, indicating issues with data linkage. This prompted me to request a data re-extraction, which resolved these inconsistencies upon receipt of the revised data.

Lastly, minor discrepancies between requested and provided variables were found in both databases and resolved in the updated data-cut for both Flatiron and NCRAS data.

8.2.2. Treatment pattern and sample size check

In evaluating the sample size, I analysed treatment patterns for mCRPC patients from the Flatiron database for the PC1/PC2 case studies (Chapter 7, Section 7.5.3.3.1 and 7.5.3.3.2), and prostate and kidney cancer patients from the English NCRAS database for the PC3 and RCC case studies (Chapter 7, Section 7.5.3.3.3 and 7.5.4.3). In analysing the Flatiron database, I used its default line-of-therapy (LOT) table to define the duration, start, and end dates of each LOT (Chapter 5, Section 5.3.3.3, with raw prescription data also available for creating user-defined LOT). In contrast, user-defined LOT was necessary for analysing the English NCRAS Systemic Anti-Cancer Treatment (SACT) dataset. For comparative purposes, I applied similar criteria across both datasets. In all cancer treatments, the introduction of any new treatment marks the beginning of the next LOT, except when the new treatment is part of a combination regimen. Table 8.1 defines the treatments that qualify as a LOT for advanced prostate cancer, and Table 8.2 outlines common treatments for mRCC. In the preliminary check of treatment patterns across two databases, a subtle difference should be noted: in the Flatiron mPC dataset, by default, restarting the same treatment more than an extended period (i.e., 90 days, is also classified as a new line of therapy. Nevertheless, our clinical expert, Dr. Carmel Pezaro, highlights that it is rare for the same medication to be used again in subsequent treatment lines. For example, the consecutive use of abiraterone or enzalutamide as first and second lines is rare (i.e. abiretarone → abiretarone, enzalutamide → enzalutamide), as shown in Figure 8.1. This underscores the negligible effect of the default LOT definition in Flatiron data (i.e., one of the criteria being that re-initiating the same treatment after a discontinuation of more than 90 days is considered a new line of therapy, whether it was a pause or a true re-challenge) on my treatment pattern checks for sample size estimation. That is, merging LOTs involving repeats of the same treatment—such as categorising the sequence abiretarone → abiretarone → docetaxel as abiretarone → docetaxel—would not significantly change the landscape of treatment sequence patterns.

Table 8.1 Criteria for defining line-of-therapy in advanced prostate cancer

Treatment types	Examples of specific treatments (non-exhaustive)
Typical therapies included in LOT duration calculations	
Androgen receptor-targeted agents (ARTAs)	Abiraterone, Enzalutamide, Apalutamide, Darolutamide
Taxanes (chemotherapy)	Docetaxel, Cabazitaxel
Internal radiotherapy	Radium-223
Cellular immunotherapy	Sipuleucel-T
Atypical therapies included in LOT duration calculations	
Combinations of any typical therapies with other agents	-
Other chemotherapies	Cisplatin, Carboplatin, Vinorelbine
Immunotherapies	Durvalumab
Antifungal agents being used as treatment for prostate cancer	Ketoconazole
Hormone	Medroxyprogesterone, Diethylstilbestrol
Drugs being used in clinical trials	In all Flatiron tables, medications used in clinical trials are labelled as "clinical trial agents" instead of the specific treatment received. Some records specify the trial name, such as the "STAMPEDE trial", but due to blinded randomisation, it is generally unclear if a specific treatment, such as docetaxel, was administered.
Therapies excluded in LOT duration calculations	
Radiation and surgery	-
GnRH, LHRH agonists/antagonists	Cyproterone, Goserelin, Triptorelin, Leuprorelin, Degarelix Note: These agents are not counted as a LOT for treatmenting advanced prostate cancer as they are androgen deprivation therapy, often initiated in earlier stages of prostate cancer to reduce testosterone levels and may continue alongside advanced therapies for metastatic prostate cancer.
Nonsteroidal antiandrogen	Bicalutamide, Nilutamide, Flutamide Note: According to Dr. Pezaro, these agents should not be counted as a LOT for treating advanced prostate cancer as they do not offer a survival benefit in treating mCRPC. Instead, they are sometimes used emergently to alleviate symptoms related to metastasis.
Treatments for cancer-related hypercalcemia (e.g. bisphosphonates)	Denosumab, Zoledronic Acid Note: These agents are not classified as a LOT because they primarily serve to manage symptoms of bone metastasis, supplementing the main treatment.
Appetite-stimulating agents used as atypical treatment for prostate cancer	Megestrol, Megestrol Acetate Note: These agents are not counted as a LOT because they are considered atypical hormone therapies.
Combinations of the above	Bicalutamide + Goserelin, Bicalutamide + Leuprorelin, Cyproterone + Goserelin, Cyproterone + Leuprorelin Note: These combinations are not classified as a LOT since these combinations solely consist of agents that should not be counted as a LOT.

GnRH: Gonadotropin hormone-releasing hormone; LHRH: Luteinizing hormone-releasing hormone; LOT: line-of-therapy

Table 8.2 Common treatments in metastatic kidney cancer

Treatment types	Examples of specific treatments (non-exhaustive)
Targeted therapies	Pazopanib, Sunitinib, Cabozantinib, Axitinib, Levatinib
Immunotherapies	Nivolumab, Ipilimumab + nivolumab, Ipilimumab
mTOR Inhibitors	Everolimus, Tesirolimus
Combinations	Everolimus + Levatinib, Avelumab + Axitinib

mTOR: mammalian target of rapamycin

When evaluating NCRAS data sample sizes, I focused on general prostate and kidney cancer cases, diverging from the specific target populations of mCRPC and metastatic RCC (mRCC) outlined in the PC3 and RCC studies, respectively. These simplifications are solely for obtaining an indicative sample size estimates. For the actual TTE analysis in Section 8.3, I followed the pre-defined eligibility criteria in my study protocol (Chapter 7) only relevant patients were included. I deemed this approach reasonable as RCC accounts for most kidney cancer cases.[585] Additionally, the systemic treatments of interest are primarily only used for metastatic diseases (Tables 8.1 and 8.2). Hence, including early-stage patients in the analysis is unlikely to lead to misinterpretation of the treatment sequences for patients in their mCRPC and mRCC phases, since these treatments are typically initiated upon progression to metastatic disease. This method is preferable for sample size estimation because it does not exclude early-stage diagnosed patients who later develop metastasis, a detail that is not included in the English NCRAS data (Section 8.2.1). Acknowledging recent treatment trends is vital in interpreting treatment sequences for mCRPC and mRCC, especially the increasing early use of docetaxel and enzalutamide in non-metastatic prostate cancer[249,540], and the potential overlap of systemic treatments as adjuvant or neoadjuvant therapy in early-stage RCC[586].

Figures 8.1 to 8.3 present Sankey diagrams showing treatment patterns for three groups: treatment-naïve mCRPC patients in Flatiron data, prostate cancer patients in English NCRAS data, and kidney cancer patients in English NCRAS data, respectively. The patterns broadly align with the disease demographics outlined in Chapter 6, Section 6.8, and correspond with sample size estimates from Chapter 7, Sections 7.5.3.5 and 7.5.4.5.

In the US, the primary treatment sequences for mCRPC patients include abiraterone followed by enzalutamide, enzalutamide followed by abiraterone, abiraterone or enzalutamide followed by docetaxel, and no further treatment after the initial therapy (Figure 8.1). Conversely, in England, the prevalent treatments for mCRPC starts with docetaxel, followed by abiraterone or enzalutamide, or in the reverse order (Figure 8.2). Both countries often use sequences involving docetaxel, enzalutamide or abiraterone. Dr. Pezaro highlighted that in England, using abiraterone and enzalutamide sequentially is uncommon, generally limited to those who develop intolerable toxicity to one of these agents.

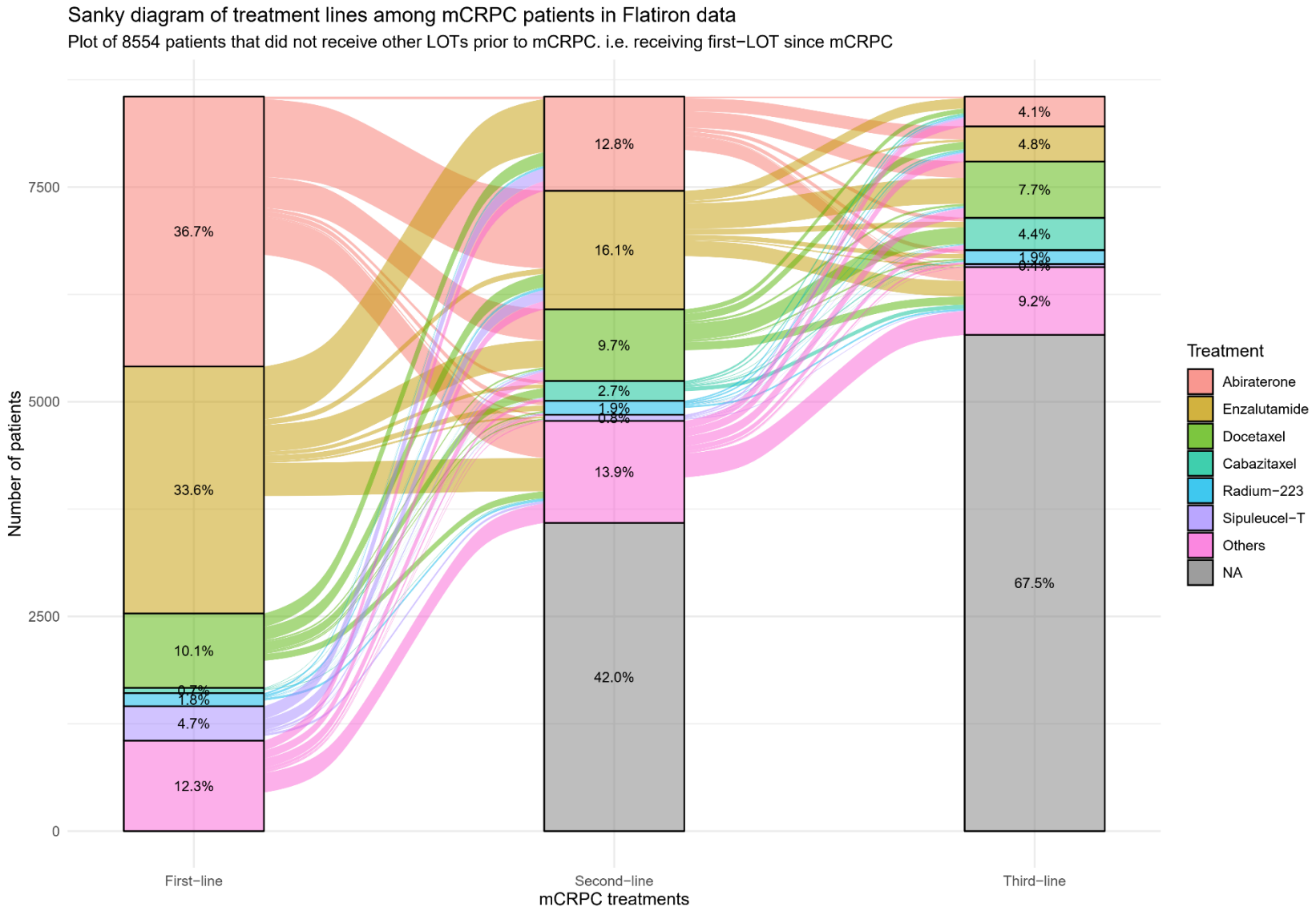


Figure 8.1 Sankey diagram of treatment patterns in treatment-naïve mCRPC patients in Flatiron data (data-cut: April, 2022)

Patients were diagnosed with metastatic castration-resistant prostate cancer (mCRPC) between 2013 and 2022.

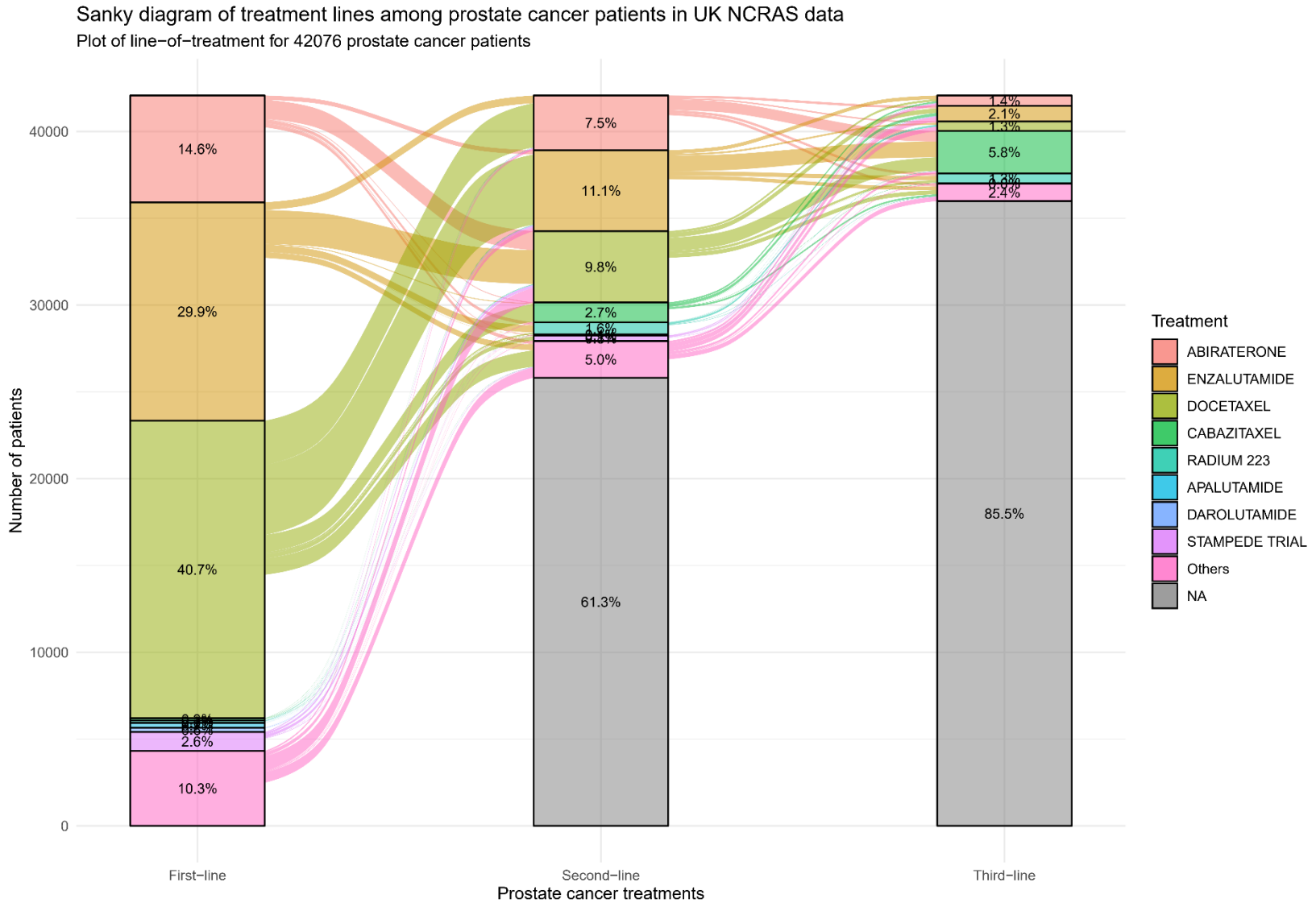


Figure 8.2 Sankey diagram of treatment patterns in treatment-naïve prostate cancer patients in English NCRAS data (data-cut: Dec 2021)

Patients were diagnosed with prostates cancer between 2011 and 2021.

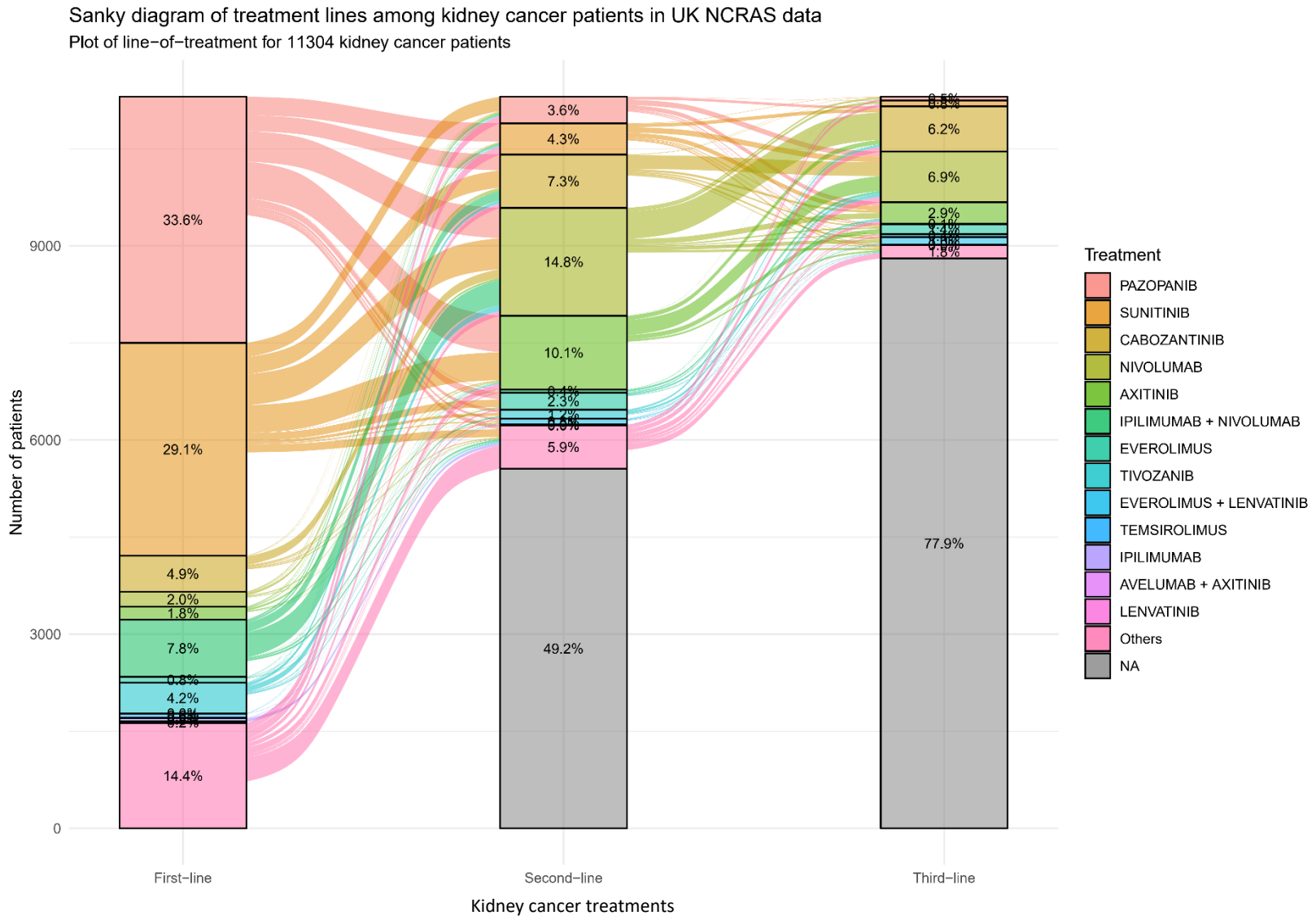


Figure 8.3 Sankey diagram of treatment patterns in treatment-naïve kidney cancer patients in English NCRAS data (data-cut: Dec 2021)

Patients were diagnosed with kidney cancer between 2011 and 2021.

These findings indicate a sufficient sample size for PC1 analyses with Flatiron data (Chapter 7, Table 1), with over 1,000 patients in each group (abiraterone → enzalutamide versus its reverse sequence), and over 2,000 when factoring in those who did not receive any second-line treatment. Common treatment sequences in Figure 8.1 & 8.2 also help identify common sequences for PC2 and PC3 case studies (Chapter 7, Table 2), such as docetaxel → enzalutamide and its reverse sequence. However, it is crucial to consider refining the sequences for prostate cancer analyses using English NCRAS data, factoring in proxies for progression dates and sequences starting from non-metastatic stages. Accurately differentiating certain scenarios—such as docetaxel for non-metastatic phases followed by enzalutamide for metastasis versus both used following metastases —is key to avoid immortal time bias when comparing treatments potentially across different disease phases.

For patients with kidney cancers in England, the most common treatment sequences begin with either pazopanib or sunitinib (Figure 8.3). In my RCC Analysis Set 1 (Chapter 7, Table 7.3), the target sequence of interest is sunitinib followed by everolimus. This particular sequence was observed in only 109 patients, representing a smaller subset than anticipated. Despite the small number, the analysis may remain feasible, especially factoring in 1517 patients who were treated exclusively with first-line sunitinib and did not receive any further treatment. It is also clear, and unsurprising, that replicating other treatment sequences from the RECORD-3 study[86] (everolimus → sunitinib) is impossible due to the scarce use of everolimus as a first-line treatment in England.

8.2.3. Implications of preliminary data checks

The Sankey diagrams (Figure 8.1-8.3) align well with the disease demographics identified for selecting benchmark trials and databases, indicating that sample size is likely be sufficient for the planned TTE analyses. This reinforces the need for a thorough fit-for-purpose assessment in the design phase, particularly for estimating potential sample sizes in targeted treatment sequences and factoring in treatment patterns based on published data and clinical expert insights. A key question remains: Can smaller sample sizes in certain treatment sequences (e.g. suinitinib → everolimus for RCC Analysis Set 1) be offset by including all patients who started with the same first-line treatment (i.e., sunitinib), as required for the advanced IPW method (see Section 8.3.4)?

Since only a small proportion of patients follow the sunitinib → everolimus sequence among those starting with sunitinib, applying the advanced IPW method requires censoring patients who deviate from this sequence when they receive a non-everolimus second-line treatment (see Section 8.3.4). This would result in a censoring rate of over 50% among patients who received a second-line treatment, unlike in the PC1 case study, where the rate would be below 50% due to the prevalence of the abiraterone → enzalutamide sequence and its reverse. Consequently, the challenge of smaller sizes is closely tied to the performance of the advanced IPW method under high levels of censoring

due to treatment deviations, an issue not covered in the thesis (i.e., not applicable to the PC1 case study). This aspect, however, is especially relevant in the PC2/PC3 and RCC analyses and should be investigated in the respective actual TTE analyses to assess its impact in the future.

Additional research, as suggested in Section 9.5(4) in Chapter 9, could use simulation studies to explore this in more detail. Two key aspects need further examination: the small proportion of patients following a specific treatment sequence (resulting in a high censoring rate among those receiving the same first-line treatment) and the varying levels of patients receiving only the first-line treatment (still considered as adhering to the treatment sequence strategy). These factors could impact the performance of the advanced IPW method and affect the feasibility of the analysis. A high censoring rate can lead to extreme IPW weights, causing analysis instability and potentially violating the positivity assumption (i.e., too few patients with similar characteristics remain adherent to the treatment strategy to represent those censored due to treatment deviation; see Section 4.6.1.2.1 in Chapter 4 and Section 8.3.2). This issue is notable, for example, in the use of inverse probability of censoring weights (IPCW) to adjust for unwanted treatment switching in health technology assessments (HTA), where the method becomes error-prone when over 90% of control patients in a sample of 500 switch to the experimental treatment.[43]

Finally, ensuring data consistency with published summary statistics is key to confirming the extracted data meets the study's needs and enables prompt communication with data providers for corrections in case of any errors.

8.3. Implementation of sequential treatment Target Trial Emulation: the GUTG-001 Analogue case study using Flatiron data (PC1)

8.3.1. Overview

Due to COVID-19-related delays and challenges in data access highlighted in Chapter 5 (Section 5.4), this thesis solely focuses on the implementation of the PC1 case study (Chapter 7, Section 7.5.3.3.1).

This section offers a step-by-step tutorial on executing the PC1 case study using Flatiron data, harnessing the inverse probability weighting (IPW) method, as illustrated in Figure 8.4. This begins with curating (survival-)analysis-ready datasets for the patient cohort of PC1's Target Trial (i.e., GUTG-001 Analogue), along with defining their baseline and time-varying characteristics, and outcomes, as detailed in Section 8.3.3.

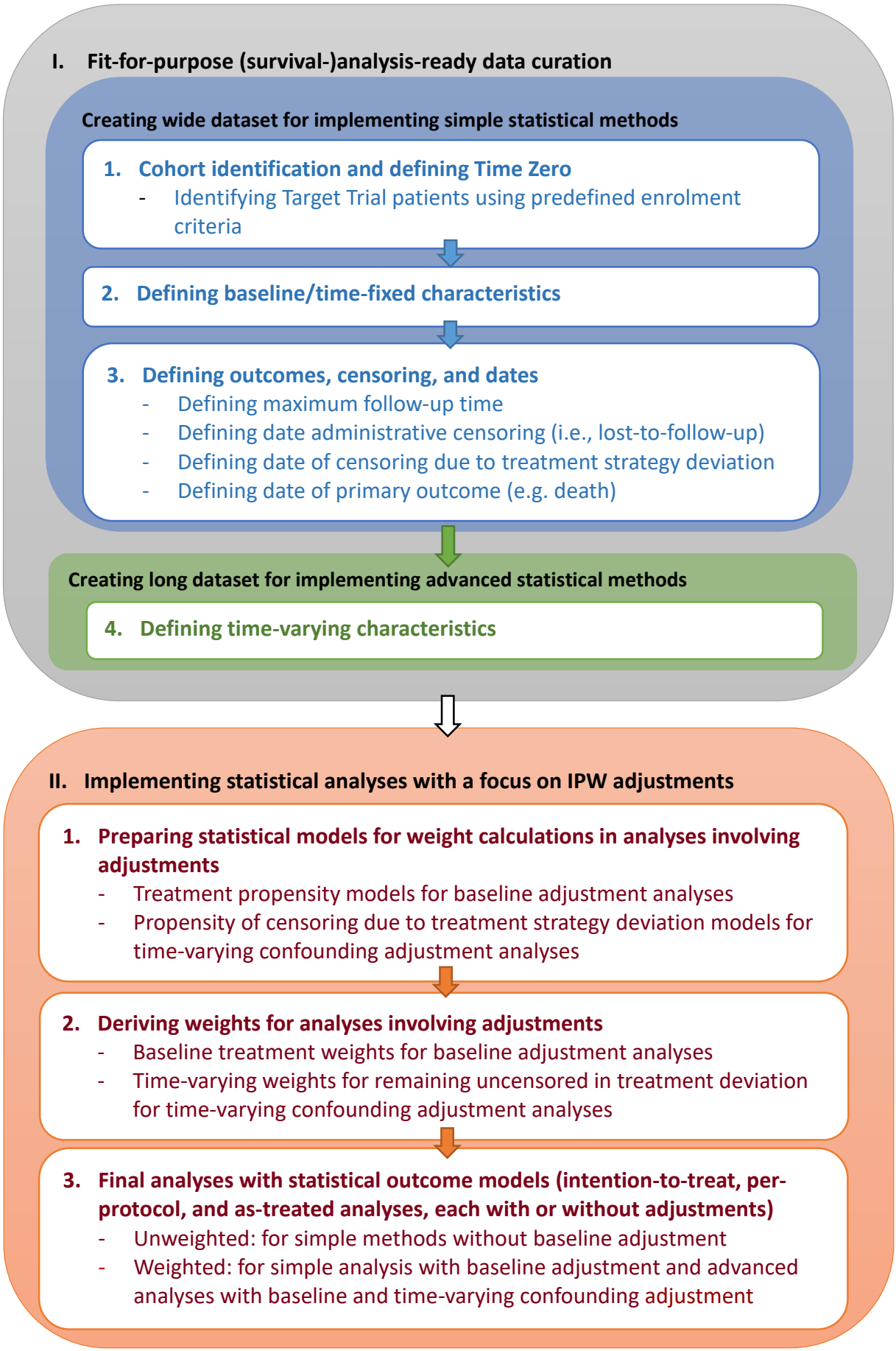


Figure 8.4 Emulating a Target Trial comparing treatment sequences

IPW: inverse probability weighting; Time Zero: the start of follow-up

Following dataset preparation, statistical models are applied to analyse the data, as outlined in Section 8.3.4. The analyses cover both simple and advanced methods identified in Chapter 4. The simple methods include intention to treat (ITT), per-protocol (PP), and as-treated (AT) analyses (Section 4.6.1.1, Chapter 4), both with and without baseline covariate adjustments using inverse probability treatment weighting (IPTW) (Sections 8.3.4.2 and 8.3.4.4). The advanced method covered in the thesis to adjust for both baseline and time-varying confounding is inverse probability weighting (IPW) incorporates IPTW and inverse probability of censoring weighting due to treatment deviation ($IPCW_{txdev}$) (see Section 4.6.1.2.1 in Chapter 4, and Section 8.3.4.5).

A total of nine analyses were conducted to compare the survival outcomes between the treatment sequences abiraterone → enzalutamide and its reverse sequence, including eight simple and one advanced analyses:

- 1. Simple analyses
  - 1a. ITT analysis
  - 1b. PP analysis
  - 1c. AT analysis
  - 1d. AT analysis, limited to patients receiving at least two lines of treatment
- 2. Simple analyses with baseline confounding adjustment using IPTW
  - 2a. ITT analysis with IPTW
  - 2b. PP analysis with IPTW
  - 2c. AT analysis with IPTW
  - 2d. AT analysis, limited to patients receiving at least two lines of treatment, with IPTW
- 3. Advanced analysis with baseline and time-varying confounding adjustment: PP analysis using $IPTW*IPCW_{txdev}$

The series of analyses was set up to contrast simple methods, frequently found in existing literature comparing the effectiveness of treatment sequences (Section 4.6.1.1, Chapter 4), with the chosen advanced method $IPTW*IPCW_{txdev}$ (Section 4.6.1.2.1, Chapter 4), showcasing the potential bias that may arise from not using a theoretically more appropriate advanced technique.

While Chapter 4 identified several candidate advanced methods, this thesis prioritised the advanced IPW method due to time constraints and considerations discussed in Chapter 4, Section 4.7. This approach also acknowledged the challenges and time required to curate RWD analytical datasets for a single method's implementation, discussed further in Section 8.5. PhD transition funding will be utilised to further advance this line of research, such as exploration of alternative advanced methods (e.g. g-formula).The methods implementation steps in Figure 8.4 were tailored to accommodate the advanced IPW method (i.e., $IPTW*IPCW_{txdev}$) described in Section 4.6.1.2.1 of Chapter 4. However, the steps described in Chapter 4's methods review (Steps (1)-(5) in Section

4.6.1.2.1) simplified these steps by omitting the detailed data curation procedures outlined in Part I, Steps I.1-4 of Figure 8.4. Additionally, Figure 8.4 also covers procedures for simple analyses, either without weights or using only baseline weights.

For clarity, I outline below how the steps in Figure 8.4 align with those in Chapter 4, despite their slightly different arrangement. For further details, please refer to the last few paragraphs in Section 4.6.1.2.1 in Chapter 4.

- Determining censoring timing related to treatment deviation: Step (1) in Section 4.6.1.2.1 of Chapter 4 corresponds to part of the data cleaning tasks in Steps I.3 and I.4 in Figure 8.4.
- Building weight derivation models: Steps (2) and (3) in Section 4.6.1.2.1 of Chapter 4 align with Step II.1 in Figure 8.4.
- Calculating weights: Step (4) in Section 4.6.1.2.1 of Chapter 4 maps to Step II.2 in Figure 8.4.
- Final outcome analysis using weights: Step (5) in Section 4.6.1.2.1 of Chapter 4 aligns with the advanced analysis using both baseline and time-varying weights in Step II.3 of Figure 8.4.

8.3.2. A toy example of an analysis-ready dataset for implementing inverse probability weighting

Before moving on to fit-for-purpose data curation (Section 8.3.3) and the application of advanced IPW methods (Section 8.3.4) using real datasets in my case study (illustrated in Figure 8.4), this section provides a toy example. The toy example aims to demonstrate the structure and components of an analysis-ready dataset for implementing the advanced IPW method to compare two treatment sequences with IPTW and $IPCW_{txdev}$ weights. It outlines the essential elements required in the dataset before the final stage of statistical outcome modelling (Step II.3 in Figure 8.4) and explains the underlying principles of the advanced IPW method ($IPTW*IPCW_{txdev}$) conceptually.

The toy dataset, shown in in Table 8.3, is constructed to evaluate the effectiveness of two treatment sequences: A → B and B → A (as denoted in the second column from the left). It includes 15 patients: 10 patients (Patients 1-10) received treatment A as their initial therapy (hypothetically assigned to the A → B strategy), while the remaining 5 patients (Patients 11-15) began with treatment B (hypothetically assigned to the B → A strategy). Table 8.3 outlines the key components for each patient required to apply the $IPTW*IPCW_{txdev}$ method, including:

- Patient ID
- (Hypothetically) assigned treatment sequences
- Baseline characteristics at the start of the first-line treatment
- Time-varying (dependent) characteristics at the start of the second-line treatment
- Observed first-line treatment
- Observed second-line treatment
- Day of starting second-line treatment

- Original observed day of outcome (e.g. death)
- Censoring date/status due to treatment deviation
- IPTW
- $IPCW_{txdev}$
- Total weight (IPTW * $IPCW_{txdev}$)

This toy example provides a simplified overview of variables typically found in both "wide" datasets (variable with fixed values, such as baseline characteristics and IPTW, see Section 8.3.3.2 and Section 8.3.4.3.2) and "long" datasets (variable with time-varying values, such as time-varying (dependent) characteristics and $IPCW_{txdev}$, see Section 8.3.3.3, and Section 8.3.4.3.3). In this simplified example, all patients were observed for the entire period (Day 0 to Day 8) without any loss to follow-up. However, in real-world scenarios, patients may have varying follow-up durations, and additional information regarding administrative censoring should be considered (see Section 8.3.3.3 for details).

Among these components, IPTW, $IPCW_{txdev}$, and the total weight (IPTW*$IPCW_{txdev}$) (i.e., the last three columns on the right in Table 8.3) are derived from the other components, while the rest should be directly cleaned from the raw patient-level data. To clean RWD for analysis in my case study to the extent shown in the toy example, Steps I.1-4 in Figure 8.4 should be followed to generate both wide and long datasets (Section 8.3.3). These datasets can then be used to derive weights (IPTW and $IPCW_{txdev}$) following Steps II.1-2 in Figure 8.4 (Section 8.3.4.3). Below, I explain how IPTW and $IPCW_{txdev}$ weights are derived for each patient using baseline and time-varying characteristics in the toy example, and how they conceptually help address confounding in comparing treatment sequences.

The aim of IPTW is to address potential confounding due to non-random allocation of first line treatments in the real world. IPTW weights ensure comparability between patients receiving first-line treatments A and B by effectively removing the influence of baseline confounders on baseline treatment assignment (i.e., eliminating the association between baseline confounders and baseline treatment, equivalent to removing the arrow from L_0 to A_0 in Figure 1.2 in Chapter 1).

Table 8.3 A toy example of an analysis-ready dataset for implementing inverse probability weighting for comparing treatment sequences

Patient ID	**Assigned treatment sequences**	**Baseline characteristics when starting 1st-line treatment** (L_0 in Figure 1.2, Chapter 1)	**Time-varying (dependent) characteristics when starting 2nd-line treatment** (L_1 in Figure 1.2, Chapter 1)	**Observed 1st-line treatment (T1)** (A_0 in Figure 1.2, Chapter 1)	**Observed 2nd-line treatment (T_2)** (A_1 in Figure 1.2, Chapter 1)	**Day of starting 2nd-line treatment**	**Original observed day of death**	**Censored due to treatment deviation (C)**	**IPTW for each time interval** For A-first group, △: 1/P(A|△) For A-first group, O: 1/P(A|O) For B-first group, △: 1/P(B|△) For B-first group, O: 1/P(B| O)	**$IPCW_{txdev}$ for each time interval** For A-first group: 1/P(2nd-line B|☺) For B-first group: 1/P(2nd-line A|☺)	**Total weight for each time interval (IPTW*$IPCW_{txdev}$)**
1	A → B	△	-	A	-	-	-	No	1/(4/5) = 1.25	1	1.25*1 = 1.25
2	A → B	△	-	A	-	-	8	No	1/(4/5) = 1.25	1	1.25*1 = 1.25
3	A → B	△	☺	A	B	2	-	No	1/(4/5) = 1.25	1	1.25*1 = 1.25
4	A → B	△	☺	A	B	3	8	No	1/(4/5) = 1.25	1	1.25*1 = 1.25
5	A → B	O	☺	A	B	4	8	No	1/(6/10) = 1.67	Day 1 to day 3: 1 From day 4 onwards: 1/(1/2) = 2	Day 1 to day 3: 1.67*1 = 1.67 From day 4 onwards: 1.67*2 = 3.34
6	A → B	O	☺	A	C	4	6	Yes, on day 4	1/(6/10) = 1.67	Day 1 to day 3: 1 From day 4 onwards: 0	Day 1 to day 3: 1.67*1 = 1.67 From day 4 onwards: 1.67*0 = 0
7	A → B	O	☺	A	B	6	-	No	1/(6/10) = 1.67	Day 1 to day 5: 1 From day 6 onwards: 1/(1/2) = 2	Day 1 to day 5: 1.67*1 = 1.67 From day 6 onwards: 1.67*2 = 3.34
8	A → B	O	☺	A	B	6	-	No	1/(6/10) = 1.67	Day 1 to day 5: 1 From day 6 onwards: 1/(1/2) = 2	Day 1 to day 5: 1.67*1 = 1.67 From day 6 onwards: 1.67*2 = 3.34
9	A → B	O	☺	A	D	6	7	Yes, on day 6	1/(6/10) = 1.67	Day 1 to day 5: 1 From day 6 onwards: 1/(1/2) = 0	Day 1 to day 5: 1.67*1 = 1.67 From day 6 onwards: 1.67*0 = 0
10	A → B	O	☺	A	E	6	8	Yes, on day 6	1/(6/10) = 1.67	Day 1 to day 5: 1 From day 6 onwards: 1/(1/2) = 0	Day 1 to day 5: 1.67*1 = 1.67 From day 6 onwards: 1.67*0 = 0
11	B → A	△	-	B	-	-	5	No	1/(1/5) = 5	1	5*1 = 5
12	B → A	O	☺	B	A	4	6	No	1/(4/10) = 2.5	Day 1 to day 3: 1 From day 4 onwards: 1/(1/4) = 4	Day 1 to day 3: 2.5*1 = 2.5 From day 4 onwards: 2.5*4 = 10
13	B → A	O	☺	B	C	4	-	Yes, on day 4	1/(4/10) = 2.5	Day 1 to day 3: 1 From day 4 onwards: 0	Day 1 to day 3: 2.5*1 = 2.5 From day 4 onwards: 2.5*0 = 0
14	B → A	O	☺	B	D	4	-	Yes, on day 4	1/(4/10) = 2.5	Day 1 to day 3: 1 From day 4 onwards: 0	Day 1 to day 3: 2.5*1 = 2.5 From day 4 onwards: 2.5*0 = 0
15	B → A	O	☺	B	E	4	5	Yes, on day 4	1/(4/10) = 2.5	Day 1 to day 3: 1 From day 4 onwards: 0	Day 1 to day 3: 2.5*1 = 2.5 From day 4 onwards: 2.5*0 = 0

In addition to IPTW, $IPCW_{txdev}$ weights are necessary because, although patients were initially (hypothetically) assigned specific treatment sequences based on their first-line treatment (as shown in the second column in Table 8.3), some did not adhere to their assigned second-line treatments. For example, Patients 6, 9, 10, 13, 14, and 15 received different second-line treatments (C, D, E, C, D, and E, respectively) instead of following the A → B or B → A sequences. These patients are censored at the point when they deviated from their assigned treatment sequences to remove the effect of subsequent treatments that are not of interest. Since second-line treatments are also not randomised, this type of censoring is also potentially informative. Hence, $IPCW_{txdev}$ is needed to estimate the effectiveness of A → B versus B → A in a scenario where all patients adhered to their assigned treatments sequences without any deviations (i.e., addressing time-varying confounding). Finally, in the final survival outcome analysis, each patient receives a total weight ($IPTW*IPCW_{txdev}$) at each time interval, calculated as the product of (fixed) IPTW and (time-varying) $IPCW_{txdev}$ (Step II.3 in Figure 8.4).

In this toy example (Table 8.3), the probability of receiving a specific first-line treatment is associated with a baseline characteristic (denoted by △ or O in the third column in Table 8.3). This baseline characteristic also affects the outcome (i.e., a prognostic characteristic), making it a baseline confounder. The probability of receiving a second-line treatment is associated with another factor, denoted by the colour of a smiley face at the initiation of second-line treatment (blue, green, yellow, red, purple, or pink), as shown in the fourth column in Table 8.3. This factor also affects the outcome, making it a time-varying (dependent) confounder. This toy example is built in such a way that patients with the same time-varying characteristics (i.e., denoted by having the same colour of smileys at the initiation of the second-line treatment) share the same baseline characteristic (denoted by △ or O). This simplification allows baseline characteristics to be ignored when considering the prognostic factors influencing the choice of second-line treatment in this example. In real-world scenarios, these relationships are likely more complex, with multiple factors influencing treatment choices cross and outcomes across different lines of therapy. Specifically, $IPCW_{txdev}$ weights are typically calculated based on the history of variables, including both their baseline and time-varying values. For simplicity, this toy example uses only a single variable for each calculation: a baseline characteristic for IPTW (denoted by △ or O) and a time-varying characteristic for $IPCW_{txdev}$ (denoted by the colour of smileys) at the start of each line of therapy.

To account for baseline confounding, the IPTW calculation incorporates the baseline characteristic, as shown in the third column from the right in Table 8.3. For example, in the A-first group, patients with characteristic △ (Patients 1-4) are assigned a weight based on their inverse probability of receiving treatment A given their baseline characteristic (i.e., $1/P(A|△)$), which is 1.25. Here, $P(A|△)$ is calculated as the proportion of patients receiving treatment A among all patients with

the characteristic △ (Patients 1-4 and 11), resulting in 4 out of 5, or 0.8. Therefore, the inverse of P(A|△) is 1.25. Similarly, patients with characteristic O (Patients 5-10) in the same group are given a weight of 1/P(A|O), which is 1.67. In the B-first group, the only patient with characteristic △ (Patient 11) receives a weight of 1/P(B|△), which is 5, while four patients with characteristic O (Patients 12-15) receive a weight of 1/P(B|O), which is 2.5. It is important to note that, for illustrative purposes in this toy example, the probabilities of receiving first-line treatments A and B are directly calculated by finding the proportion of patients receiving each treatment given their baseline characteristic (i.e., a non-parametric approach), under the assumption of only one baseline confounder. However, in most real-world cases, a parametric multivariate logistic model (with treatment as the dependent variable and relevant baseline confounders as independent variables) is required to estimate these probabilities due to the presence of multiple baseline confounders (see Section 8.3.4.3.2 for details).

$IPCW_{txdev}$ calculations are provided in the second column from the right in Table 8.3. For $IPCW_{txdev}$, all patients receive a weight of 1 until they receive a second-line treatment (i.e., each patient represents only themselves until that point), provided they adhere to their assigned treatment strategy. Patients who received only one line of treatment were considered to be adhering to the treatment strategy for the entire duration of the study. As a result, they were assigned an $IPCW_{txdev}$ weight of 1 for all time intervals in the study (e.g., Patients 1, 2, and 11). To address time-varying confounding caused by the artificial censoring of patients who deviate from their assigned treatment sequences, the remaining patients who adhere to the correct second-line treatment, start it at the same time, and have similar time-varying characteristics (i.e., the same colour of smileys) are upweighted to compensate for those who were censored. For example, Patient 6 received second-line treatment C on Day 4, deviating from the assigned treatment strategy, and was therefore censored from Day 4 onward (resulting in an $IPCW_{txdev}$ weight of 0 for all intervals from Day 4 onward). In contrast, Patient 5 also started a second-line treatment on Day 4 but received treatment B, which adhered to the assigned strategy. Patient 5 is the only individual in A-first group who began a second-line treatment on Day 4 and had the same characteristic at the initiation of second-line treatment—a yellow smiley face—as Patient 6. Consequently, Patient 5 is upweighted to represent both themselves and Patient 6, receiving an $IPCW_{txdev}$ weight of 2 for all intervals from Day 4 onwards. The upweighting is based on the inverse probability of receiving the appropriate second-line treatment B on Day 4, given their at the initiation of second-line treatment (i.e., 1/P(2nd-line B|☺)), which is 2. This is because the probability of receiving a second-line treatment B among all patients with a yellow smiley at the initiation of second-line treatment on Day 4 (Patient 5 and 6) is 1 out of 2 (P(2nd-line B|☺) = 0.5). Conceptually, Patient 5 counts for double to maintain the overall effective sample size, representing both themselves and the censored Patient 6 from Day 4 onwards. If Patient 5 did not exist, there would

be no one to compensate for Patient 6's absence from Day 4, reducing the effective sample size. This relates to a positivity assumption violation (see Section 4.6.1.2.1 in Chapter 4), which can result in analysis instability.

Similarly, Patient 7 is upweighted ($IPCW_{txdev}$ weight of 2 for each interval from Day 6 onwards) to represent both themselves and Patient 9, as both started a second-line treatment on Day 6 with a red smiley face. Patient 8 is upweighted ($IPCW_{txdev}$ weight of 2 for each interval from Day 6 onwards) to represent both themselves and Patient 10, as both began a second-line treatment on Day 6 with a purple smiley face. Patient 12 is upweighted by a $IPCW_{txdev}$ weight of 4 from Day 4 onwards to represent themselves and Patients 13-15, who all received a non-A second-line treatment on Day 4 and had a pink smiley face on that day. While both Patients 3 and 4 received second-line treatment and adhered to their assigned treatment sequence, they were not upweighted with a $IPCW_{txdev}$ weight throughout the study. This was because no other patients who deviated to the assigned second-line treatment had similar characteristics at the time they started (i.e., blue and green smileys, respectively). As a result, they did not share comparable characteristics with others at the start of second-line treatment and were not upweighted. It is important to note that, similar to IPTW, in this simplified toy example, $IPCW_{txdev}$ are derived from the proportion of patients receiving a specific second-line treatment given their characteristics at the initiation of second-line treatment. In a more realistic setting, however, estimating $IPCW_{txdev}$ would often require a parametric multivariate logistic regression model that accounts for multiple time-varying confounders and time as a continuous variable, especially since patients are unlikely to switch treatments on the exact same days but rather within similar timeframes.

Finally, the total weight of each patient at each interval can be calculated by multiplying the fixed IPTW by the time-varying $IPCW_{txdev}$ (IPTW* $IPCW_{txdev}$) weight, as shown in the rightmost column of Table 8.3. For example, for Patient 5, the weight is 1.67 for each interval from Day 1 to Day 3 (1.67*1 = 1.67) and 3.34 for each interval from Day 4 onward (1.67*2 = 3.34). The total weight can then be used in the survival outcome analysis to estimate the effectiveness of the A → B versus B → A treatment sequences, as if all patients had been randomised to receive one of these two sequences at the start of the study.

In summary, the toy example in Table 8.3 provides an overview of the data structure needed for implementing advanced IPTW* $IPCW_{txdev}$ method to compare treatment sequences. The dataset curation steps (Steps I.1-4 of Figure 8.4) are crucial for identifying baseline and time-varying prognostic factors related to patients' treatment probabilities, outcome date and status, and censoring dates (due to treatment deviation and administrative censoring). While the toy example does not detail the dataset curation process from raw data, this is covered in the following section (Section 8.3.3). Once cleaned, the dataset can be used to derive IPTW and $IPCW_{txdev}$ weights, as outlined in the toy

example (corresponding to Steps II.2-3 of Figure 8.4). These weights are then applied to the final survival outcome model (corresponding to Step II.3 in Figure 8.4) for comparing treatment sequences. The following sections detail how I completed these steps (i.e., all steps in Figure 8.4) using real-world data in my case study (Sections 8.3.3–8.3.4).

8.3.3. Preparing fit-for-purpose analysis-ready datasets: cohort selection & data curation pipeline

My approach for preparing the final analytical datasets broadly aligns with insights from a British Health Foundation Data Science Centre webinar on constructing research-ready datasets from NHS health records.[383] A key simplification was the utilisation of a pre-curated metastatic prostate cancer cohort dataset from the Flatiron database (see Chapter 5, Table 5.1). This dataset conveniently records critical dates, such as the initial prostate cancer diagnosis, metastasis, and castration resistance, for each patient who developed metastasis. This markedly simplified the identification of the Time Zero (i.e., the start of patient follow-up) in my PC1 study—the initiation of first-line metastatic castration resistant prostate cancer (mCRPC) treatment, minimising the need to sift through longitudinal patient records to pinpoint critical dates and distinguish metastatic cases from the broader prostate cancer patient population.

Section 8.3.3.1 explains how I finalised the identification of the target cohort by meeting predefined eligibility criteria. Section 8.3.3.2 describe how I created a final "wide" dataset for the analysis, capturing baseline characteristic, outcomes, and censoring for each included patient, with each row summarising each patient's information. Subsequently, Section 8.3.3.3 outlines how I curated an additional "long" dataset for tackling time-varying confounding, capturing daily updates on each patient's characteristics and outcomes in a row-wise manner. These procedures have adapted and extended insights from the CAUSALab courses at Harvard University, particularly in Advanced Confounding Adjustment, where the structure of analytical datasets for implementing advanced confounding adjustments was introduced.

8.3.3.1. Cohort identification and defining Time Zero

The procedure for identifying the cohort and defining Time Zero is depicted as Step I.1 in the implementation process shown in Figure 8.4. Table 8.4's first two columns detail the original inclusion and exclusion criteria from the benchmark GUTG-001 trial[562] and the adapted criteria for identifying the Target Trial cohort intended to replicate it (i.e., GUTG-001 Analogue). The adaption represents a refined version of the idealised Target Trial protocol specified in Chapter 7's Table 1, addressing the practicability of executing the TTE given constraints imposed by data limitations. Constraints, such as imperfectly recorded variables, could emerge during data cleaning, leading to necessary iterative protocol reformulation to align as closely as possible with the original ideal protocol while accounting for data limitations.[587] Consulting clinicians is crucial for assessing the

clinical relevance and to determine appropriate modifications, alongside considering potential sensitivity analyses for assessing the impact. Hence, the middle column of Table 8.4 outlines the reasons for each adaption for transparency, incorporating clinician insights.

I adopted the color-coded scheme from the RCT DUPLICATE study protocols[171,588] to classify the match level between the benchmark trial's eligibility criteria and the adapted ones in Table 8.4: green for adequate, yellow/amber for intermediate, orange/coral for poor, grey for unmeasurable but non-critical criteria for the analysis, and white for additional criteria in the PC1 case study that were not in the GUTG-001 trial. The eligibility assessment comprises 12 steps, detailed in Table 8.4's second right column: Steps 1-2 & 7-12 align with the GUTG-001 trial[562] and Steps 3-6 are extra criteria for quality assurance and to better mimic the GUTG-001 trial. The step order does not follow the protocol's criteria numbering, with explanations detailed in the adjacent right column. Figure 8.5 outlines the relevant data timeframe used to identify the study population and covariates assessment in the PC1 study. The timeframe for each criterion in Table 8.4 is explained in detail in the following paragraphs.

Table 8.4 Identification procedures for the PC1 Case Study's Target Trial patient cohort using Flatiron Data: GUTG-001 Analogue

Colour coding scheme for inclusion and exclusion criteria mapping	Original criteria in the GUTG-001 trial	Adequate mapping in RWD	Intermediate mapping in RWD	Poor mapping or cannot be measured in RWD	Cannot be measured in RWD assessed as not critical for the analysis	Additional criteria in the PC1 case study that were not in the GUTG-001 trial

Cohort identification procedures				
Original criteria: Inclusion and exclusion criteria from GUTG-001 trial	**Adapted Criteria:** Modified criteria for identifying targeted patients in Flatiron's metastatic prostate cancer database*, with operational definition	**Rationale behind adapted criteria and caveats, incorporating clinical insights**	**Cohort identification steps**	**Cohort identification step justification:** Logic underpinning the order of cohort identification steps
Inclusion 1. Willing and able to provide informed consent	NA	Patients treated in real-world practice presumably all gave their consent. Major caveat: While GUTG-001 trial participants were informed of receiving either the abiraterone →enzalutamide sequence or the reverse, those in the real-world might have received a specific first-line and subsequent treatments influenced by time-varying patient preferences, disease condition, and physician preference.	NA	NA
Inclusion 5. Evidence of metastatic disease on bone scan or CT scan	Include all patients in the Flatiron metastatic prostate cancer database as they all experienced metastasis. Further eligibility assessments for enrolment follow only after this confirmation, in line with GUTG-001 trial criteria. The *"MetDiagnosisDate"* from the *"MetProstate"* table was used to confirm patients having metastasis and to determine their metastatic diagnosis date. No patients lacked a metastatic diagnosis date in the database, as it was carefully reviewed and determined by trained medical record abstractors.	Patients with a metastatic diagnosis date should largely resemble those with evidence of metastasis in the GUTG-001 trial. Major caveats: The hierarchy for defining metastatic disease in the Flatiron data followed a sequence of biopsy date, radiology report date, and physician-reported date (See Appendix 8.1). This may not necessarily align with the trial's criteria, which seem to primarily rely on radiology reports for establishing evidence of metastasis. Additionally, the "screening date" for GUTG-001 trial eligibility may not always coincide with the date of metastatic prostate cancer diagnosis (see Figure 8.5). Nonetheless, it serves as a valuable reference point for applying further inclusion and exclusion criteria for identifying the Target Trial population in the database.	Step 1	The GUTG-001 trial required a "screening date" to assess patient eligibility, with all the inclusion and exclusion criteria being applied simultaneously. Eligible patients were randomised into two study groups within 28 days, proceeding to treatment initiation within 5 days of randomisation. To best replicate this process using RWD, it is vital to establish a "hypothetical screening date" for each patient (Figure 8.5) in the Target Trial cohort, ensuring a uniform assessment of all

Inclusion 6. Evidence of biochemical or imaging progression in the setting of surgical or medical castration. Progressive disease for study entry is defined by one of the following three criteria: A. PSA progression: minimum of two rising PSA values from a baseline measurement with an interval of ≥ 1 week between each measurement. Minimum PSA at screening visit is > 2.0 ug/L B. Soft tissue or visceral disease progression (see the GUTG-001 protocol's Appendix B for definition of measurable disease as per RECIST 1.1 criteria) C. Bone progression: ≥ 2 new lesions on bone scan	Include patients in the Flatiron database that has "IsCRPC" (in the *"MetProstate"* table) = "Yes". Among these patients, exclude any patient who does not have a defined "CRPCDate" (in the *"MetProstate"* table).	The dataset's CRPC date is derived from a combination of PSA levels and clinical judgment (see Appendix 8.2), which resembles the multifaceted criteria used for defining CRPC in the GUTG-001 trial. Patients without a CRPC date will be excluded from the Target Trial emulation, since they could potentially compromise the accuracy of establishing "Time Zero" (the starting point for follow-up) in the Target Trial designed to replicate the GUTG-001 trial. Particularly, in the GUTG-001 trial, treatments were specifically given as the initial treatment for mCRPC. As a result, it is crucial to identify the mCRPC date (i.e. the metastasis date or the CRPC date, whichever is later) in the GUTG-001 Analogue Target Trial. Caveats: • It's assumed that all patients have undergone either surgical or medical castration, which was deemed reasonable by our oncology specialist, Dr. Pezaro. • Although a physician's clinical assessment of CRPC (Appendix 8.2) in real-world practice is expected to correspond with the GUTG-001 trial's specified criteria (I6.A, I6.B, I6.C), slight variations may exist. However, Dr. Pezaro indicated that such discrepancies are expected to be negligible. • Excluding patients without a CRPC date, despite their documentation confirming CRPC, may limit the generalisability of the study (and resulting in potential selection bias).	Step 2	eligibility criteria, such as age assessment at comparable points in the disease timeline. The mCRPC diagnosis date serves as a pragmatic reference for defining a "hypothetical screening date", aligning with the trial's requirement for the eligibility to be determined post-mCRPC diagnosis. That is, the hypothetical screening date is likely to be after the mCRPC diagnosis for eligible patients, leading to the application of Inclusion Criteria 5 & 6 at the outset of cohort identification using RWD. Specifically, cohort identification Steps 1 & 2 are key to establishing the mCRPC diagnosis date (see "Adapted criteria" and rationale in Steps 1 & 2).
None (the corresponding adapted criterion in this step serves as a quality control measure, i.e., eligibility criteria enforced by RWD analysis)	Exclude any patient whose line-of-therapy involves "LineName" (in the *"LineofTherapy"* table) = "Line Zero"	In the Flatiron database, patients' "structured" electronic health records begin when they start prostate cancer follow-up at oncology health care providers (community/academic) within the Flatiron Health Network, potentially leading to incomplete medical history prior to that time. Flatiron addresses this by abstracting from the unstructured records, such as medical notes, to identify any relevant treatments that began prior to a patient's structured records. These identified treatments were then labelled as "Line Zero," a label that serves as an acknowledgment of the potential uncertainties in the completeness of the patient's	Step 3	Steps 3 to 6, while not directly correspond to any of the inclusion or exclusion criteria in the original GUTG-001 trial, are crucial in identifying "Time Zero" (i.e., the initiation of first-line mCRPC treatment) and the "hypothetical screening date" (i.e., occurring within 28 days prior to Time Zero) for quality control.

		treatment details. The “Line Zero” label is applied to any treatment histories for advanced prostate cancer (specifically nmCRPC or mCRPC, excluding mHSPC or nmHSPC) initiated post-CRPC date but absent from the patient’s structured records. Excluding patients with a “Line Zero” record aligns with Flatiron’s established data analysis procedures using their database. Crucially, in replicating the GUTG-001 trial, the presence of incomplete data suggests that a patient’s treatment history prior to mCRPC may not be thorough enough for proper assessment against the trial’s inclusion and exclusion criteria, particularly E2, E3, E6, and E7. Major caveat: Excluding patients lacking complete treatment data could constrain the study's generalisability. Nevertheless, including patients with incomplete treatment histories prevents a full replication of the GUTG-001 criteria and hinders proper benchmark comparisons. To assess the impact of potential selection bias, this criterion could be relaxed in the Target Trial Emulation for unrestricted mCRPC patients (i.e., PC2 case study).		This is indispensable for establishing an emulated randomisation in the GUTG-001 Analogue Target Trial, ensuring precise correspondence with the randomisation in the GUTG-001 trial (i.e. first-line mCRPC treatment began within 5 days following randomisation). For example, including patients flagged for potentially incomplete pre-mCRPC treatment data could mistakenly identify some as eligible for first-line mCRPC treatment without prior therapies, when in fact they may have received treatments not captured in the data. Therefore, this step is to ensure the inclusion of patients whose records are complete for fair assessment of the following eligibility criteria (e.g. previous treatments). For practicality, both “Time Zero” and the emulated randomisation in the GUTG-001 Analogue Target Trial are defined as the commencement of first-line mCRPC treatment. Caveats of such approach will be discussed in 8..4 Discussion)
None (the corresponding adapted criterion in this step serves as a quality control measure, i.e., eligibility criteria	Exclude patients who did not initiate treatment after being diagnosed with mCRPC	The GUTG-001 trial exclusively focuses on patients capable of surviving until they can commence first-line mCRPC treatment, which occurs within five days after randomisation. Therefore, this criterion aims to identify	Step 4	Same as Step 3. Importantly, the initiation of the first treatment post-mCRPC is defined as “Time Zero” for

enforced by RWD analysis)		those who have survived up to the "emulated randomisation" point. For practical purposes, I have defined this point to be the same date of the initiation of first-line mCRPC treatment. Major caveat: ● This approach aims to closely mimic the GUTG-001 trial but may be susceptible to potential immortal time bias (i.e. patients surviving long enough to begin the first-line mCRPC treatment). Specifically, there is uncertainty regarding whether the duration between patients mCRPC diagnosis and the initiation of their first mCRPC treatment aligns closely between RWD and the GUTG-001 trial. If patients typically commence their mCRPC treatment later in a real-world setting than in the trial, the survival duration in the emulated Target Trial might appear artificially extended when compared to the benchmark GUTG-001 trial. Nevertheless, if the interval from patients' mCRPC diagnosis to the commencement of their first-line treatment remains consistent across both groups in the emulated trial, the influence of immortal time bias might be less pronounced on comparative effectiveness, though it could still affect each group's absolute effectiveness. This criterion may be relaxed in an unrestricted mCRPC Target Trial emulation (i.e., PC2 case study) to test the impact of immortal time bias, wherein patients could be monitored from the point of, for example, their mCRPC diagnosis, rather than from the initiation of first-line treatment. The "cloning" method from Chapter 4, Section 4.6.1.2.2, can be applied to accommodate the grace period of mCRPC treatment initiation. Specifically, for assessing the treatment sequence from abiraterone → enzalutamide, this approach deems all patients with mCRPC diagnosis eligible to begin follow-up. Patients starting with treatments other than abiraterone are censored on the date of such initiation. Then, patients not starting treatment within the specified grace period, for instance, 30		emulating randomisation in the Target Trial and sets the potential timeframe for the "hypothetical screening visit" for assessing the rest of the inclusion and exclusion criteria. (see Figure 8.5).

		days after mCRPC diagnosis, are censored on the 30th day following their mCRPC diagnosis. The implications of this approach are discussed in Section 8.5. ● In practice, some patients may begin mCRPC treatment slightly before their documented mCRPC date and continue past their diagnosis. This can be attributed to Flatiron's mCRPC definition (see Appendix 8.1 and 8.2) (i.e. the real mCRPC definition can be earlier) or emergency procedures necessitating an early start to mCRPC treatment prior to mCRPC diagnosis. This situation is further complicated in scenarios where pre-mCRPC treatments initiated pre-diagnosis are halted only after a progressive diagnosis is confirmed (i.e. concludes after mCRPC diagnosis). Consequently, "mCRPC treatment lines" can become ambiguous when involving treatments starting before and concluding after mCRPC diagnosis, due to the aforementioned various scenarios. Flatiron defaults these as first-line mCRPC treatments. However, in my GUTG-001 Analogue Target Trial analysis, these are classified as pre-mCRPC treatments and evaluated as previous treatments in Step 12 for consistency. Despite this caveat, this ensures the defined treatment lines are consistent with the GUTG-001 trial, where patients were randomised into different treatment groups only after a confirmed mCRPC diagnosis. To assess the impact of potentially incorrectly defined treatment line on survival, sensitivity analyses considering treatments initiated within a specific timeframe (e.g., a week) of the mCRPC diagnosis as part of the mCRPC treatment and redefine the Time Zero for patients with these treatments could be considered if deemed necessary.		
None (the corresponding adapted criterion in this step serves as a quality control measure, i.e., eligibility criteria enforced by RWD analysis)	Exclude patients whose first treatment after mCRPC diagnosis was neither abiraterone nor enzalutamide. Treatment lines initiated prior to the mCRPC	The GUTG-001 trial only had two arms, one started with abiraterone, and the other with enzalutamide. However, in the unrestricted mCRPC Target Trial (i.e., PC2 case study), this can be relaxed, depending on what treatment sequences are compared.	Step 5	same as Step 4

	diagnosis, which continued through and concluded after mCRPC, are not considered as the first treatment post-mCRPC, but pre-mCRPC treatments.			
None (the corresponding adapted criterion in this step serves as a quality control measure, i.e., eligibility criteria enforced by RWD analysis)	Excluded patients who initiated mCRPC treatment (referred to as "Time Zero") after March 31, 2022.	The GUTG-001 trial enrolled participants from October 21, 2014, to December 13, 2016, and the results published in Khalaf 2019 had a data cutoff on May 31, 2013, ensuring a minimum 17-month follow-up. As this study's Flatiron database has a data cutoff on August 31, 2023, patients whose Time Zero fell after March 31, 2022 were excluded, to maintain a comparable follow-up period. To prevent post-randomisation selection bias, our patient selection was not based on the availability of structured data of individual patients. This implies that patients might be lost to follow-up within a shorter timeframe, aligning with the circumstances in the GUTG-001 trial.	Step 6	same as Step 4
Inclusion 2. Adult males ≥ 18 years age	Exclude patients who were < 18 at Time Zero	Same as the original criterion.	Step 7	Having defined Time Zero in Step 2, records from the potential "hypothetical screening visit" period (see Figure 8.5) for each patient can now be used to evaluate their eligibility against the remaining inclusion and exclusion criteria.
Inclusion 3. History of adenocarcinoma of the prostate diagnosed histologically without evidence of neuroendocrine or small cell differentiation, or if patient does not have pathology of adenocarcinoma of the prostate, patient has metastatic disease typical of prostate cancer (i.e., involving bone or pelvic lymph nodes or para-aortic lymph nodes) AND a serum concentration of PSA that is rising and > 20ng/mL at the	Include all patients regardless of their histology	In the Flatiron database, patients' histology is categorised solely as "Adenocarcinoma" or "Prostate cancer, not otherwise specified". Furthermore, the database does not include information to ascertain the type of metastatic disease (e.g., bone metastasis, lymph node metastasis, and brain metastasis). Dr. Pezaro recommended including patients irrespective of their histology, as it is not expected to substantially impact the choice of treatment at the defined "Time Zero" (Step 4, Figure 1). Furthermore, PSA levels at the time of initial prostate cancer diagnosis may have become irrelevant since PSA levels are subject to change and can vary significantly during the time around the diagnosis.	Step 8	Same as Step 7

time of when the patient was clinically diagnosed with prostate cancer.		Caveat: We assume that the selected patients' histology profiles are comparable to those in the GUTG-001 trial. Many patients lack detailed histology and PSA levels at initial diagnosis, leading to challenges in accurate mapping to the original criteria.		
Inclusion 7. ECOG performance status 0-2 (see the GUTG-001 protocol's Appendix C)	Exclude patients with a most recent ECOG score exceeding 2 recorded within 30 days prior to their Time Zero	Maintain the original criterion but make an exception to assume that patients who did not undergo any ECOG assessment within the 30 days prior to Time Zero have an ECOG score of 2 or lower. A 30-day timeframe prior to or at Time Zero was utilised for assessment to reflect potential timeframe of the hypothetical screening visit (Figure 1). Major caveat: There is a possibility that patients lacking an explicitly documented ECOG assessment may have an ECOG score greater than 2. However, Dr. Pezaro indicated that this is theoretically less likely to occur among patients initiating treatments (Step 4). Nevertheless, the adapted criteria are colour-coded as amber to indicate intermediate mapping to the original GUTG-001 trial criteria, as there may be a possibility that the undocumented ECOG values are not missing at random. Furthermore, some patients with an ECOG score greater than two from more than 30 days before Time Zero (e.g., 35 days prior) were not excluded despite the proximity.	Step 9	Same as Step 7
Inclusion 4. Prior surgical orchiectomy or if on LHRH agonist/antagonist then testosterone < 1.7 nmol/L at screening visit (patients must maintain LHRH agonist/antagonist therapy for duration of study treatment if not surgically castrated)	Exclude patients with a most recent testosterone level > 49.03 ng/dL (1.7 nmol/L) within 30 days preceding their Time Zero	Dr. Pezaro highlighted that in standard clinical practice all patients diagnosed with prostate cancer are typically subjected to surgical or chemical castration. However, due to insufficient information in the database, it is unclear if and when patients received surgical, chemical interventions or both. Therefore, our focus became verifying the effectiveness of the castration prior to Time Zero. Therefore, patients with testosterone level exceeding the threshold should be excluded. The threshold 1.7 nmol/L was converted to 49.03 ng/dL to align with the test units used in the Flatiron database. Additionally, it is assumed that patients who did not have any testosterone tests within the 30 days prior to Time Zero have remained within a reasonable range. A 30-day timeframe prior to or at Time Zero was utilised for	Step 10	Same as Step 7

		assessment to reflect the potential timeframe of the hypothetical screening visit (Figure 1). Major caveat: There is a risk of not excluding patients with extremely high testosterone levels if they were not tested; however, this scenario should theoretically be relatively rare given the rarity of such extreme levels per Dr. Pezaro. Nevertheless, the adapted criteria are colour-coded as amber to indicate intermediate mapping to the original GUTG-001 trial criteria, as there may be a possibility that the undocumented testosterone values are not missing at random.		
Inclusion 9. Adequate organ function defined as: A. Absolute neutrophil count ≥ 1.5 x 10^9/L, platelet count ≥ 100 x 10^9/L and hemoglobin ≥ 80 g/L B. Creatinine clearance ≥ 30 ml/min (calculated by Cockcroft-Gault formula, see the GUTG-001 protocol's Appendix D) C. Serum potassium > than lower limit of normal range D. Total bilirubin ≤ 1.5 x ULN except for patients with known Gilbert's syndrome (direct bilirubin ≤ 1.5 x ULN) E. ALT and AST ≤ 5 x ULN	• Exclude patients if their most recent lab test results within the 30 days prior to their Time Zero meet any of the following criteria: o Absolute neutrophil count < 1.5 x 10^9/L o Platelet count < 100 x 10^9/L o Hemoglobin < 8 g/dL o Serum potassium < 3.5 mmol/L o Total bilirubin > 1.8 mg/dL. For patients with Gilbert's syndrome diagnosis within the 30 days prior to their Time Zero, direct bilirubin > 0.45 mg/dL o ALT > 275 IU/L o AST > 240 IU/L • Exclude patients who had a diagnosis of chronic kidney disease with a creatinine clearance of <30 ml/min (Appendix x.3) within 182 days prior to Time Zero.	A 30-day window before or at Time Zero was mainly used to conduct organ function assessment based on lab tests, in line with the timing of the hypothetical screening visit (Figure 1). Relevant reference normal levels for each test were obtained from U.S. guidance, and the test units were converted to match those used in the Flatiron data. For kidney function assessment, I replaced the trial's original criteria, which relied on the assessment of age, weight, and creatinine clearance at similar timing (calculating using Cockcroft-Gault formula), as such data alignment was not always available. Therefore, I adopted a criterion of excluding patients with a diagnosis of chronic kidney disease indicating a creatinine clearance of <30 ml/min within half-year before Time Zero. I used a half-year window to assess kidney conditions, as I interpreted the criteria to exclude patient with severe chronic kidney disease (creatinine clearance <30 ml/min). I assumed that patients typically have at least two annual follow-up visits for chronic conditions, and a shorter timeframe may not capture their long-term health status adequately. Major caveat: There is a risk of not excluding patients with insufficient organ function if they were not tested or given relevant diagnosis within the assessment timeframe. The adapted criteria are colour-coded as amber to indicate intermediate mapping to the original GUTG-001 trial criteria, as there may be a possibility	Step 11	Same as Step 7

		that the undocumented lab values are not missing at random.		
Inclusion 8. Eligible for treatment with either abiraterone acetate or enzalutamide as per standard of care guidelines	NA	I only included patients who initiated abiraterone or enzalutamide as first-line treatment at Time Zero (Step 5). Therefore, it is assumed that these patients were all considered "eligible" for these treatments. Moreover, since the GUTG-001 trial was conducted post approval of these drugs, it is reasonable to anticipate that the physicians' considerations in the trial closely mirrored those in real-world clinical practice.	NA	NA
Inclusion 10. Able to swallow study drug and comply with study requirements including provision of peripheral blood samples at specified time points for correlative studies	NA	It is assumed that all patients are capable of swallowing abiraterone and prednisolone or enzalutamide. Particularly, only patients who initiated either abiraterone or enzalutamide after their mCRPC diagnosis were included (Step 4 and Step 5).	NA	NA
Inclusion 11. Recovery from all prior treatment-related toxicity to grade ≤ 2 (as per CTCAE 4.0)	NA	There were no diagnosis codes specific enough to identify extreme treatment-related toxicity event. It is assumed that all patients have recovered from any extreme prior treatment-related toxicity. Dr. Pezaro suggested this assumption is reasonable, given that patients with an ECOG score greater than 2 were excluded from this study (Step 9).	NA	NA
Exclusion 1. Severe concurrent illness or co-morbid disease that would make the subject unsuitable for enrolment	Exclude patients with any diagnosis of non-prostate primary malignancies (except non-melanomatous skin cancer) or brain metastasis (see Appendix 8.3) within 182 days prior to Time Zero.	It is assumed that patients with comorbid diseases, except concurrent malignancies, are acceptable if their most recent ECOG score is < 2 (Step 9). This stems from the original criteria for severe concurrent illness and comorbid diseases being unclearly defined. Dr. Pezaro suggested that excluding only patients with concurrent primary malignancies (other than prostate cancer) and brain or spinal metastasis is a reasonable approach and should closely align with exclusion criteria E1, E5, and E7. Diagnoses within half-year prior to Time Zero were considered for assessment, for the same reasons as described in I9. Caveat: Patients having ICD codes for brain metastasis and spinal/nerve diseases (Appendix 8.3) may not accurately reflect all cases specified in the GUTG-001 trial due to limited detail. Treatment status of patients having epidural disease is also unknown. Since it is	Step 12	Same as Step 7
Exclusion 5. Active concurrent malignancy (with the exception of non-melanomatous skin cancer)				
Exclusion 7. Brain metastases or active epidural disease (treated epidural disease is permitted)	Excluded patients with a diagnosis of secondary malignancy of the brain or spinal cord (Appendix x.3) within 182 days prior to Time Zero.			

		unclear whether a diagnosis indicates concurrent active cancer or a historical one, the adapted criteria for excluding patients with active concurrent malignancy are colour-coded as amber, indicating intermediate mapping to the original GUTG-001 trial criteria. In contrast, the adapted criteria for excluding patients with brain metastasis are colour-coded green, as it is generally a late-stage condition that persists.		
Exclusion 2. Prior therapy with CYP17 inhibitors (including abiraterone acetate, TAK-700, TOK-001 and ketoconazole), enzalutamide or other experimental antiandrogens (e.g. ARN-509, TOK-001)	Excluded patients if they had previously received treatments, starting before mCRPC diagnosis, as monotherapy or in combination with the therapies listed in the exclusion criteria, clinical study drug, or treatments that became available to the market after the initiation of GUTG-001 trial (see Appendix 8.4). Patients receiving any other systemic therapy not listed prior to the diagnosis of mCRPC are not excluded.	The adapted definition closely aligns with the original exclusion criteria, with the exception of additionally excluding patients having therapies that became available only after the initiation of the GUTG-001 trial. It is anticipated that these treatments would have been prohibited as previous therapies had they been accessible to patients at the enrolment period of GUTG-001 trial. Caveat: It is assumed that all systemic therapy initiated prior to the mCRPC diagnosis was not intended for mCRPC treatment. This may lead to potential incorrect specification of mCRPC treatment lines as described in Step 4. Furthermore, prior systemic therapies for non-mCRPC conditions in real-world practice is assumed to closely resemble those in the GUTG-001 trial.		
Exclusion 6. Wide-field radiotherapy or radioisotopes such as Strontium-89 or Radium-223 ≤ 28 days prior to starting study drug (limited-field palliative radiotherapy for 1-5 fractions is permitted)				
Exclusion 3. Prior systemic chemotherapy for mCRPC				
Exclusion 10. History of seizure or seizure disorder, or history of any cerebrovascular event within 6 months of study entry.	Excluded patients with a diagnosis of seizure or cerebral event (other than transient ischemic attack) (Appendix 8.3) within 182 days prior to Time Zero.	The adapted definition closely aligns with the original exclusion criteria. Caveat: Patients having diagnosis of a seizure or cerebrovascular event within 6 months prior to Time Zero does not conclusively indicate that the event occurred within this timeframe. Nevertheless, for practicality, all such patients were excluded. Dr. Pezaro pointed out that not all diagnoses of seizures or cerebrovascular events imply a clinically significant event necessitating exclusion; the decision should also take into account the severity of the condition. Nevertheless, given the lack of detailed information on the severity of the patients' conditions, all patients having diagnoses with these conditions were excluded for simplicity.		

Exclusion 4. Life expectancy < 6 months	NA	Choosing patients based on "observed survival" post Time Zero is inappropriate and would undoubtedly lead to selection bias. Similarly, in the GUTG-001 trial, patients were evaluated based on their "expected survival" during the screening visit, rather than their actual survival times. Dr. Pezaro suggests that assuming patients with an ECOG performance status of 2 or below are in a satisfactory condition is a reasonable approach to resemble this criterion.	NA	NA
Exclusion 8. Use of herbal products that may lower PSA level (e.g. saw palmetto)	NA	Consumption of herbal products that can interfere with PSA levels cannot be identified from the information in the database. Caveat: It is uncertain whether patients who consume herbal products in actual practice might experience drug-drug interactions and variations in PSA results. Survival outcomes could potentially be confounded by this unmeasured confounder.	NA	NA
Exclusion 9. Contraindication to prednisone therapy including poorly controlled diabetes mellitus	NA	It is challenging to precisely define "poorly controlled" diabetes as the trial did not provide any specific criteria. Dr. Pezaro suggested that, clinically, when patients exhibit poorly controlled diabetes, clinicians tend to prefer enzalutamide as it does not require concurrent use with prednisolone. Therefore, rather than arbitrarily including patients diagnosed with diabetes, diabetes is addressed as a potential confounding variable in the analysis.	NA	NA
Exclusion 11. Gastrointestinal disorder affecting absorption	NA	Defining the extent of a patient's gastrointestinal disorder solely based on diagnosis codes is challenging. Dr. Pezaro points out that if patients had absorption issues, they likely would not have been prescribed oral medications initially. Consequently, it can be assumed that all patients who commenced treatment with abiraterone and enzalutamide have adequate absorption function. Caveat: patients who develop significant gastrointestinal disorders that affect absorption post Time Zero may be a potential unmeasured time-varying confounder, possibly influencing the selection of subsequent treatments.	NA	NA

Exclusion 5. Major surgery within 4 weeks of starting study treatment	NA	The Flatiron metastatic prostate cancer database metastatic prostate cancer is an electronic health record database curated for research purpose (see Sections 5.3.2 and 5.3.3.1 in Chapter 5). Hence, procedure codes are not available, while diagnosis codes of surgeries might not necessarily reflect a recent surgical procedure but history of surgery. Information on treatment before metastasis is limited, as described in Section 5.3.3.3; no raw surgery codes are available in the curated database. Early-stage cancer treatments are recorded as primary treatments with a simple indicator flag showing whether patients received radiotherapy, systemic treatment, or surgery (see Table 5.1 in Chapter 5). Dr. Pezaro posited that patients fit to initiate treatment (Step 4) and with an ECOG score of 2 or lower (Step 9) are unlikely to have undergone any major surgery within the four weeks preceding treatment initiation. Therefore, it is assumed that none of the included patients received major surgery within the four weeks prior to the initiation of the study treatment.	NA	NA

* In the adaptive operational definition, variables used from the Flatiron data were labelled in italics, and data tables were labeled in italics and underlined.
ALT: Alanine aminotransferase; AST: aspartate aminotransferase; CTCAE: Common Terminology Criteria for Adverse Events; mCRPC: metastatic castrate-resistant prostate cancer; NA: not applicable; PSA: prostate-specific antigen; Time Zero: the initiation of the first treatment post-CRPC (i.e., the first-line mCRPC treatment); ULN: upper limit of normal

Steps 1-2 in Table 8.4 identify individuals with mCRPC and the date of their mCRPC diagnosis, as patients undergo further eligibility assessment (i.e., baseline characteristics assessment) only after a confirmed mCRPC diagnosis. The data period for Steps 1-2 in Table 8.4 (i.e., enrolment assessment for mCRPC) is shown by the large dark blue arrow in Figure 8.5. After identifying mCRPC patients in Steps 1-2 of Table 8.4, the date of first-line treatment initiation (i.e., Time Zero) was determined for each patient (i.e., the solid vertical line in the middle of Figure 8.5)). Baseline assessments were then conducted using the relevant data period (i.e., the three light grey horizontal bars in Figure 8.5) to assess patient eligibility in PC1 based on baseline characteristics from Step 3 onwards in Table 8.4.

Here, I describe the baseline eligibility assessment period in the GUTG-001 trial and how I replicated it in the PC1 case study as closely as possible. In the GUTG-001 trial, all eligibility criteria were assessed on a specific screening date prior to randomisation. Eligible patients were randomised into two study groups (abiraterone → enzalutamide versus enzalutamide versus abiaraterone) within 28 days of screening, with treatment initiation within 5 days after randomisation. Hence, in GUTG-001, the maximum possible timeframe between the screening date and the start of first-line treatment ranges from 28 (if randomisation and treatment initiation occur on the same day) to 33 days (if treatment starts 5 days after randomisation). The dark grey area to the left of treatment initiation (i.e., Time Zero) in Figure 8.5 highlights this 28-33 day maximum timeframe between screening and treatment initiation. Since there is no predefined screening or randomisation date in the RWD, a 30-day period before treatment initiation (Time Zero) was defined for baseline assessments in the PC1 study, including assessing patients' latest lab results and ECOG scores (i.e., represented by the bottom grey horizontal bar in Figure 8.5). This 30-day baseline assessment period is designed to replicate the timeframe in which a hypothetical screening visit may occur for each patient in RWD (i.e., 28 to 33 days between screening and treatment initiation, as shown by the dark grey shaded area to the left of Time Zero in Figure 8.5). Records from this period were to define patients' baseline characteristics, ensuring eligibility assessment at a comparable disease stage.

It is important to note that this approach (i.e., first identifying the time of first-line treatment initiation and then defining the baseline assessment period accordingly) relies on the premise that patients are equally likely to initiate first-line treatments in both study groups, namely abiraterone and enzalutamide, at comparable times following mCRPC diagnosis. If this assumption does not hold, there may be a risk of immortal time bias when comparing survival times between the two treatment groups. However, clinicians confirmed that patients are starting abiraterone versus enzalutamide at relatively similar stages of the disease. Despite nuanced differences in how metastasis and castration resistance are defined between the GUTG-001 trial and the Flatiron database (Steps 1 & 2 in Table 8.4), the adapted criteria for mCRPC patient identification are considered to adequately match the original criteria, following consultation with clinicians.

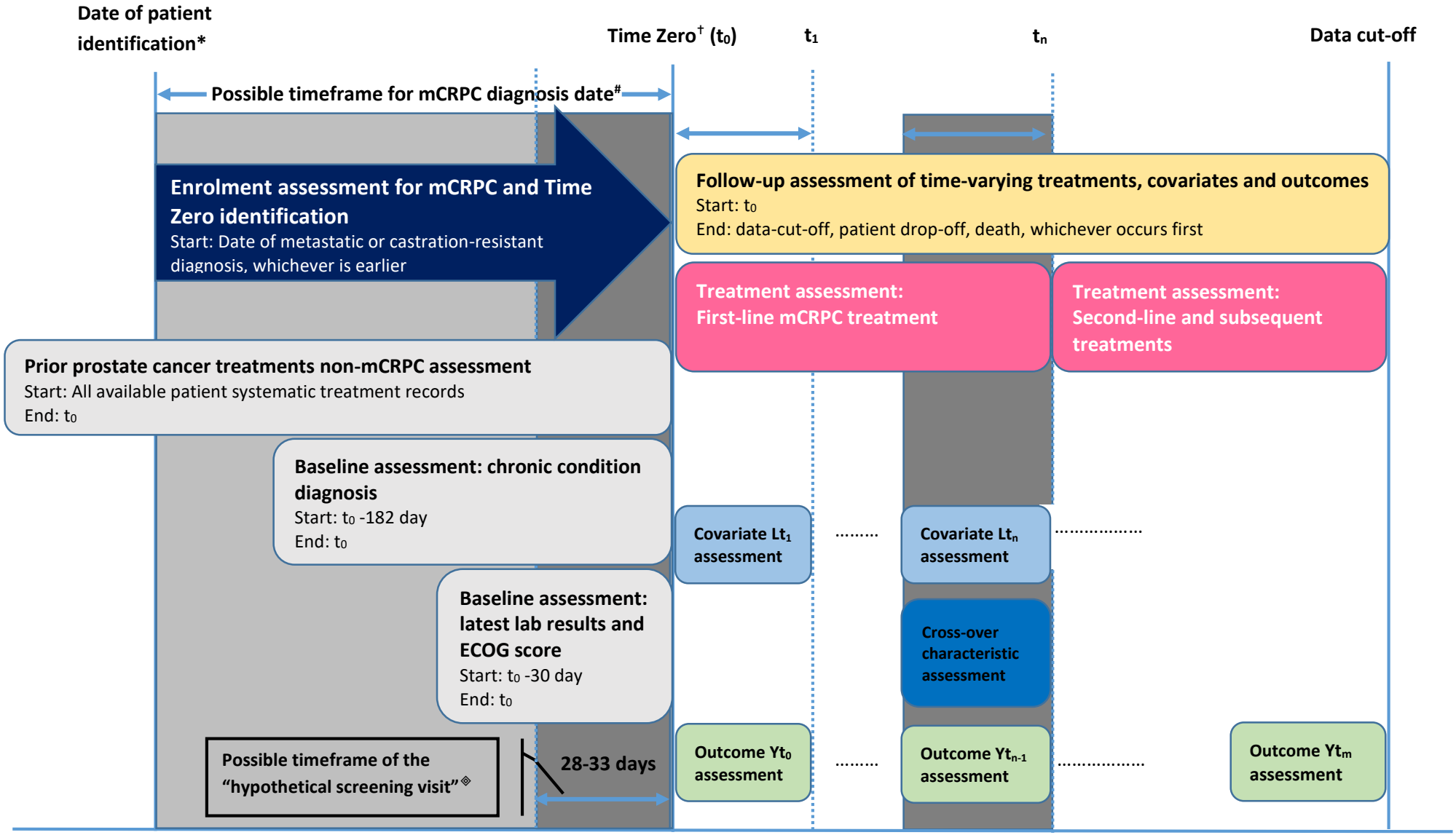


Figure 8.5 Schematic timeframe for data utilised in the Flatiron GUTG-001 Analogue Target Trial population identification and covariate assessment

The x-axis represents the span of time, while each coloured square signifies the start and end date of each assessment, either for target population identification or covariate assessment.
t_0, t_1, …t_x: represent times of covariate and outcome assessment. The interval between t_x and t_{x-1} spans one day; for example, t_1 is t_0 + 1 day, and t_2 is t_1 + 1 day.
Lt_0: represents the covariate assessment for patient characteristics at the Time Zero (i.e. baseline).
Lt_x: represents the covariate assessment for patient characteristics on day x, where x is greater than 0. For each covariate assessment at time t_x, if no new records (lab results, vital sign records and diagnosis) exists for day t_x, the covariate at t_x assumes the last value from t_{x-1} (i.e., last observation carried forward). Lt_n is a special type of Lt_x, specified as follows:
Lt_n: represents the covariate assessment for patient characteristics on the day the n^{th}-line treatment is initiated.
Crossover characteristics assessment (i.e., secondary baseline): represents the assessment of patient characteristics at the initiation of the second-line treatment and is equal to the assessment at $t_{second\text{-}line}$.
Yt_x: represents outcome assessment for patient characteristics in the interval starting on day t_x, where x is greater than 0. For each t_x, outcome assessments record status from events on happened on day t_{x+1}.
Yt_m: represents the final date of outcome assessment, determined by either administrative censoring rom loss to follow-up, data cut-off, or the maximum observation period as defined in Section 8.3.3.3.
* Date of patient identification: defined as the either date of either metastatic diagnosis or castration-resistant diagnosis, whichever is earlier.
# mCRPC Diagnosis Date: defined as either the date of metastatic diagnosis or castration-resistant diagnosis, whichever comes later.
† Time Zero: defined as the date of the initiation date of the first line-of-treatment following an mCRPC diagnosis. Patients with a Time Zero after March 31, 2022 are excluded from the analysis.
◈ In the GUTG-001 trial, patients were randomised to one of the study groups within 28 days following their "screening date", and began first-line mCRPC treatment within 5 days of randomisation. Consequently, for the Target Trials in this study, the "theoretical screening visit" might fall between 28 to 33 days prior to the start of the first-line mCRPC treatment (Time Zero). Therefore, baseline assessments relying on indicators that might change over a brief period (such as lab results and ECOG scores) were drawn from the most recent record within 30 days of Time Zero. Importantly, while trials typically use the date of randomisation for assessing patient's baseline characteristics, the GUTG-001 trial defined the time-to-event outcome from the treatment initiation date due to its proximity to the date of randomisation (difference ≤ 5 days). For practicality in this study, the date of first-line mCRPC treatment initiation (Time Zero) captured in the data was set as both the baseline and the emulated randomisation date. That is, this date served as the point for assessing patient baseline characteristics and performing statistical analyses for emulating the randomisation procedure at baseline.

For the subsequent steps in Table 8.4, patient baseline records up to 30 days prior to and on Time Zero were evaluated for eligibility, using the records closest to Time Zero for each indicator (represented by the bottom light grey horizontal bar in Figure 8.5). For chronic conditions, especially comorbidities defined using diagnosis records, records up to 180 days before Time Zero were reviewed (represented by the second light grey horizontal bar from the bottom in Figure 8.5). This longer period for capturing chronic conditions reflects the nature of RWD collection, specifically the Flatiron database's focus on oncology clinic records, which might not fully cover non-oncology diagnoses. Significant comorbidities affecting treatment decisions were assumed to be documented, though updates may be infrequent, such as semi-annual check-ups (hence the 180-day period).

Due to Flatiron's data collection limitations, which begins with a patient's first visit to an oncology healthcare providers (community or academic) within the Flatiron Health network. (see Chapter 5, Table 5.1, "Cancer Diagnosis"), patients flagged with potentially incomplete prior treatment records were excluded for quality assurance (Step 3, Table 8.4). This step ensured that the criteria for excluding prior disallowed treatments (Step 12, Table 8.4) can be applied without bias, which relies on the assumption that the distribution of patients with incomplete records of prior treatment is randomly distributed. Furthermore, patients who did not initiate treatment with either abiraterone or enzalutamide following their mCRPC diagnosis were excluded (Steps 4 & 5, Table 8.4) as patients could only start with either of these treatments in the GUTG-001 trial. Treatments starting before but with a duration ending after mCRPC diagnosis were classified as pre-mCRPC treatments, with justifications detailed in Steps 5 & 6, Table 8.4. To maintain a follow-up duration comparable to the GUTG-001 trial, patients initiating first-line mCRPC treatment post-March 31, 2022, were excluded (Step 6, Table 8.4).

Steps 7-12 in Table 8.4 present varying degrees of criteria matching. Criteria such as age (Step 7, Table 8.4) and the exclusion of patients receiving disallowed pre-mCRPC treatments or those with brain metastasis (parts of Step 12, Table 8.4) can be adequately matched. Most criteria exhibit an intermediate match due to vague original criteria in the GUTG-001 trial or inconsistent RWD collection for assessments. For example, including patients based on lab tests that confirm sufficient organ function proves challenging (Step 11, Table 8.4), as not all patients undergo consistent testing in real-world settings—often, patients deemed healthy may not necessarily be tested before staring new treatments. Therefore, the adapted criteria exclude patients with inadequate lab test results instead. This assumes those without any test results had acceptable organ function. However, there are likely patients with borderline organ function (per trial definition) not identified due to missing tests in the real-world setting.

Examples of poor matching include histology details, which, while available in the Flatiron database, are not detailed enough for applying relevant eligibility criterion (Step 8, Table 8.4). Further,

the absence of surgery codes and over-the-counter medication records, such as herbal products, complicates the application of several criteria (original GUTG-001 exclusion criteria 5 & 8), potentially leading to the inclusion of patients with poorer prognosis than those in the GUTG-001 trial. The GUTG-001 trial excluded patients with poorly controlled diabetes, a criterion difficult to apply in RWD analysis due to varying definitions. Dr. Pezaro notes that in practice, patients with this condition often receive enzalutamide (or other treatments) instead of abiraterone, which is combined with prednisolone and could worsen diabetes control. This differs from the GUTG-001 trial, which excluded these patients. However, Dr. Pezaro highlighted that unless diabetes control is extremely poor, it typically would not prevent patients from trial enrolment, as the benefits of receiving cancer treatments often outweigh concerns over blood sugar management. Consequently, few patients are likely to have been excluded from the GUTG-001 trial for this reason (original exclusion criteria 9).

The R codes for identifying the GUTG-001 Analogue Target Trial cohort are detailed in Appendix 8.5, ordered by Steps listed in Table 8.4

8.3.3.2. Defining baseline covariates, censoring, and outcomes within the "wide" dataset

After identifying the target cohort and establishing Time Zero for each patient, the first step in forming the final analytical dataset is to define baseline characteristics (Step I.2 and I.3 in Figure 8.4). This involves linking patients to their longitudinal health records up to Time Zero to determine patients' characteristics, using the nearest available value measured before or on the date of Time Zero. Figure 8.5's light grey bars denote the baseline assessment period for different types of characteristics. For this procedure's R code, see Appendix 8.6 (which identifies cohort baseline characteristics) and Appendix 8.8 (which summarises the final "wide" dataset for TTE). Table 8.5 outlines the variables in the final "wide" dataset, which includes patient baseline characteristics, with each row providing information for an individual patient.

Table 8.5 GUTG-001 Analogue Target Trial data codebook – final cleaned "wide" baselined dataset (TTE_GUTG001_base)

Each row in the dataset represents records for one patient (wide format)			
Variable	**Variable Name**	**Variable Type**	**Content/Coding**
1. Participant identifier			
Patient ID	PatientID	Character	Unique identifier for each patient.
2. Treatment at baseline			
Type of first-line treatment	FirstLine	Character	Abiraterone, Enzalutamide
First-line treatment coding	FirstLine_num	Binary	0: Enzalutamide 1: Abiraterone
3. Patient characteristics at baseline			
Time Zero	TimeZero	Date	The initiation date of the first-line treatment for mCRPC
Calendar year of first-line treatment initiation	cal_time	Numerical	Year extracted from the Time Zero date
Year of birth	BirthYear	Numerical	Patient's birth year, ranging from 1928 to 1975

Age at first-line treatment Initiation	Age_TimeZero	Numerical	Calculated by subtracting the birth year from the year of first-line treatment initiation
Group stage at initial prostate cancer diagnosis	GroupStage	Categorical	I, II, III, IV, or Unknown/Not documented
Tumour stage at initial prostate cancer diagnosis	TStage	Categorical	T0/T1, T2, T3, T4, or Unknown/Not documented
Lymph node involvement at initial prostate cancer diagnosis	NStage	Categorical	N0, N1, NX (not evaluable), or Unknown/Not documented
Metastasis presence at initial prostate cancer diagnosis	MStage	Categorical	M0, M1, or Unknown/Not documented
Gleason score at initial prostate cancer diagnosis	GleasonScore	Categorical	Low/Very low risk: ≤ 6, Intermediate risk: 7, High/Very high risk: 8-10, Unknown/Not documented
Previously treated with docetaxel	f_pretreat_treat_allowed_docetaxel	Binary	Indicates if the patient was treated with docetaxel before Time Zero: Yes or No
Previously treated with other systematic treatments	f_pretreat_treat_other_systemic	Binary	Indicates if the patient was treated with any non-docetaxel systematic treatment before Time Zero: Yes or No
Number of previous treatment lines before Time Zero	f_pretreat_lines	Categorical	0, 1, or >1
Race	Race	Categorical	Asian, Black or African American, White, Other Race, Unknown/Not documented
Days since mCRPC diagnosis to Time Zero	Gap_TimeZero_mCRPC	Numerical	Time span from mCRPC diagnosis to first-line mCRPC treatment initiation, measured in days
Days since mCRPC diagnosis to Time Zero (Categorical)	Gap_TimeZero_mCRPC_cat	Categorical	< 30 days, 31-90 days, 91-182 days, 182-365 days, > 365 days
Days since mPC diagnosis to Time Zero	Gap_TimeZero_mPC	Numerical	Time span from mPC diagnosis to first-line mCRPC treatment initiation, measured in days
Days since mPC diagnosis to Time Zero (Categorical)	Gap_TimeZero_mPC_cat	Categorical	≤ 30 days, 31-90 days, 91-182 days, 182-365 days, 366-730 days, > 730 days
Days since initial prostate cancer diagnosis to TimeZero	Gap_TimeZero_PC	Numerical	Time span from initial prostate cancer diagnosis to first-line mCRPC treatment initiation, measured in days. Some patients have unknown values due to unknown date of initial prostate cancer diagnosis.
PSA at Time Zero (ng/mL)	PSA_TimeZero	Numerical	The latest PSA lab test value within 30 days prior to or on the date of Time Zero, some patients may have unknown values
PSA at Time Zero (ng/mL, Categorical)	PSA_TimeZero_cat	Categorical	< 4, 4 ≤ PSA < 10, 10 ≤ PSA < 20, 20 ≤ PSA < 100, > 100, Unknown
Haemoglobin at Time Zero (g/L)	Hb_TimeZero	Numerical	The latest Haemoglobin lab test value within 30 days prior to or on the date of Time Zero, some patients may have unknown values
Haemoglobin at Time Zero (g/L, Categorical)	Hb_TimeZero_cat	Categorical	Hb < 120, 120 ≤ Hb < 140, Hb ≥ 140, Unknown
ALP, relative to ULN	ALP_RULN_TimeZero	Numerical	The latest ALP lab test value, measured within 30 days before or on the date of Time Zero, expressed as a ratio relative to the ULN of 130 IU/L. Some patients may have missing or unknown values.
LDH, relative to ULN	LDH_RULN_TimeZero	Numerical	The latest LDH lab test value, measured within 30 days before or on the date of Time Zero, expressed as a ratio relative to the ULN of 225

			IU/L. Some patients may have missing or unknown values.
ECOG performance status at time Time Zero	ECOG_TimeZero	Categorical	0, 1, 2, Unknown
Diabetes at Time Zero	DM_TimeZero	Binary	Whether patients have any diabetes diagnosis within half year prior to Time Zero or on the date of Time Zero 0: No 1: Yes
SES	f_SES	Categorical	Flatiron Health SES data, assumed to represent patient's SES at Time Zero. 1 represents the lowest SES, while 5 is the highest. Categories: 1, 2, 3, 4, 5, Unknown/Not documented
Patient's combined comorbidity score at Time Zero	Comorbid_Score_TimeZero	Numerical	Calculated using the Harvard algorithm of Combined comorbidity score for claims data[578,579,589] (i.e. a weighted score based on 20 groups of comorbidities) with adapted R codes specified in the Appendix 8.6. Any relevant diagnoses within half a year prior to Time Zero or on the date of Time Zero were used to classify each specific category of comorbidity as defined in the algorithm.
4. Outcome			
Date of death	DateOfDeath	Date	Date when the patient was recorded as deceased in the Flatiron database. NA if there is no death record by the data cut-off (August 31st, 2023).
Date of data cut-off	DateAdminCen_dcutoff	Date	Fixed date for all patients: August 31st, 2023
Date of the patient's last recorded structural activity in the Flatiron database	DateAdminCen_lastactive	Date	The maximum date of the patient's last recorded activity in the Flatiron database, including date of diagnosis, clinic/telemedicine visits, lab tests, vital sign measurements, medication orders/administrations, ECOG performance status records, treatment dates, NGS test dates, and biomarker test dates. This reflects the most recent structural activity for each patient (i.e., date of last follow-up).
Date of the patient's last recorded structural activity in the Flatiron database plus a grace period	DateAdminCen_lastactive_grace182	Date	Extends the last structural activity date by 182 days (grace period) as the date of administrative censoring due to lost to follow-up.
Date used for outcome valuation relative to Time Zero	DateAdminCen_maxK	Date	K days after Time Zero (default K=1440) serves as a maximum unified last follow-up date for all patients.
Date of treatment strategy deviation	DateCen_DTRdev2	Date	This variable represents the date on which patients started second-line mCRPC treatment if they deviated from their initially assigned treatment strategy. In the context of comparing abiraterone → enzalutamide versus enzalutamide → abiraterone, it specifically notes when patients, starting on abiraterone, switch to a second-line treatment other than enzalutamide, and similarly, those starting on enzalutamide but switching to treatments other than abiraterone. NA if the second-line treatment follows the expected sequence or is not initiated. For patients who resumed or re-challenged abiraterone or enzalutamide after more than 90 days (detailed in Section 8.2.2), their treatment re-

			initiation date is marked as the start of second-line treatment, indicating as the date of treatment.
Date of outcome for ITT analysis	DateY_FirstLine	Date	For ITT analysis, the outcome date is the earliest among DateOfDeath, DateAdminCen_maxK, DateAdminCen_dcutoff, and DateAdminCen_lastactive_grace182, prioritised by this order in case of a tie.
Type of outcome for ITT analysis	TypeY_FirstLine	Categorical	For ITT analysis, type of outcome on the outcome date depends on what defines DateY_FirstLine: - DateOfDeath - DateofAdminCensor (if DateY_FirstLine comes from DateAdminCen_dcutoff, DateAdminCen_lastactive_grace182, or DateAdminCen_maxK)
Death status on ITT analysis outcome date	Y_FirstLine	Binary	1: Death (if TypeY_FirstLine equals DateOfDeath) 0: Alive
Status of administrative censoring for ITT analysis	C_Admin_FirstLine	Binary	1: Censored due to lost to follow up (if TypeY_FirstLine equals DateofAdminCensor) 0: Not censored
Overall survival time in ITT analysis	SurvTime_FirstLine	Numerical	Calculated as the difference between DateY_FirstLine and Time Zero
Date of outcome for PP analysis	DateY_DTR	Date	For PP analysis, the outcome date is the earliest among DateOfDeath, DateCen_DTRdev2, DateAdminCen_maxK, DateAdminCen_dcutoff, and DateAdminCen_lastactive_grace182, prioritised by this order in case of a tie.
Type of outcome for PP analysis	TypeY_DTR	Categorical	For PP analysis, type of outcome on the outcome date depends on what defines DateY_DTR: - DateOfDeath - DateofAdminCensor (if DateY_DTR comes from DateAdminCen_dcutoff, DateAdminCen_lastactive_grace182, or DateAdminCen_maxK) - DateOf2LTreatDeviation (if DateY_DTR comes from DateCen_DTRdev2)
Outcome status for PP analysis	Y_DTR	Binary	1: Death (if TypeY_DTR equals DateOfDeath) 0: Alive
Status of administrative censoring for PP analysis	C_Admin_DTR	Binary	1: Censored due to lost to follow up (if Y_DTR equals DateofAdminCensor) 0: Not censored
Status of censoring due to treatment strategy deviation for PP analysis	C_DTRdevL2	Binary	1: Censored due to treatment strategy deviation (if Y_DTR equals DateOf2LTreatDeviation) 0: Not censored
Overall survival time in PP analysis	SurvTime_DTR	Numerical	Calculated as the difference between DateY_DTR and Time Zero

ALP: alkaline phosphatase; Diabetes mellitus; ECOG: Eastern Cooperative Oncology Group; Hb: haemoglobin; ITT: intention to treat; LDH: lactate dehydrogenase; mCRPC: metastatic castration-resistant prostate cancer; mPC: metastatic prostate cancer; NA: not available; PP: per-protocol; PSA: prostate-specific antigen; ULN: the upper limit of normal; SES: socioeconomic status

To facilitate the comparison with Khalaf et al.'s GUTG-001 trial findings[84], I derived all listed characteristics from their publication wherever possible. This includes age, prostate specific antigen (PSA), alkaline phosphatase (ALP), lactate dehydrogenase (LDH), haemoglobin (Hb), Eastern Cooperative Oncology Group (ECOG) performance score, and history of docetaxel treatment for castration-sensitive conditions at Time Zero. I adjusted the reporting units of these variables to align with those used in Khalaf et al.'s publication for consistency.[84] Further, patients' ALP and LDH levels

were categorised relative to their respective upper normal limits (130 IU/L[590] and 225 IU/L[591], respectively) for the same reason. Mapping the locations of metastases (bone, lung, liver) was found to be inadequate due to the non-differential ICD diagnosis and inconsistent coding practices, which did not provide enough detail to clearly distinguish each location. Consequently, this information was not included. However, Dr. Pezaro believes that, at this disease stage, the location of metastases is not as an influential prognostic factor for treatment decisions when compared to others, such as the ECOG score. Following discussions with clinical experts, Professor Rosario and Dr. Pezaro, I also recorded additional patient characteristics, including the number of prior treatment lines for non-mCRPC status, indicator of prior non-docetaxel systematic treatments, Gleason score, TNM staging (both overall and specific subcategories), a combined comorbidity score, diabetes status, race, socioeconomic status (SES), duration since mCRPC diagnosis, confirmation of metastasis, and initial prostate diagnosis, and the calendar year of initiation for first-line mCRPC treatment.

Most characteristics were derived from records close to Time Zero, as shown in Figure 8.5. However, the Gleason score and TNM staging were only available from the initial prostate cancer diagnosis—not necessarily at the mCRPC stage. In clinical practice, updating patient stages at progression is uncommon, it is simply noted that progression has occurred. Nevertheless, according to Dr. Pezaro, the ECOG score at Time Zero, time since mCRPC diagnosis, and whether the patient had metastasis at initial diagnosis or progressed to develop metastases, are considered more critical factors in guiding treatment decisions. Race and socioeconomic status, recorded once per patient in the database, are expected to remain unchanged at treatment time. Patients for whom no records for specific indicators were available during the baseline assessment period were marked as having "Missing" or "Unknown/Not Documented" characteristics. Patients missing an ECOG score at baseline were assumed to have a score of less than 3, as justified in Table 8.4, Step 9, during cohort identification. Not all listed patient characteristics influence treatment decisions were considered to be confounders and so not all were included in statistical models for adjustment. These determinations, made in consultation with clinicians, are elaborated in Section 8.3.4.3.

The combined comorbidity scores were calculated using an algorithm published by the Harvard group[589], which includes 20 comorbidities with specific weights assigned to each. For a comorbidity to be considered present in my study, patients need to have relevant ICD-9-CM or ICD-10-CM diagnosis codes recorded within six months leading up to Time Zero. Originally coded in SAS[589], I converted the algorithm to R, with the details provided in Appendix 8.6. The development of this algorithm leveraged data from the Clinformatics Data Mart (OptumInsight) claims (including a subset of 10% of patients with a Medicare Supplement Plan) and Pennsylvania Medicare enrollees.[578,579] Its weights were established through correlation with 30-day hospital readmission and 1-year mortality, respectively. The algorithm's population theoretically overlaps with the Flatiron population, as both

are based on American healthcare data and likely include Medicare beneficiaries.[592] This is because individuals with prostate cancer are often aged 65 or older. Further, the algorithm accommodates comorbidity definitions through both ICD-9 and ICD-10 diagnosis codes, making it ideal for analysing Flatiron data that spans the transition between these coding systems.

Subsequently, the "wide" dataset in Table 8.5 was enriched by linking it to patients' mortality records, second-line treatment information, and dates related to administrative censoring, such as the last structural activity, database cut-off, and the maximum observational period (set at 1440 days). This facilitated the determination of date of death, administrative censoring, and treatment deviation. The date of administrative censoring is defined as the earliest date among the following: the last structured activity plus a grace period of 182 days, the database cut-off (August 31st, 2023), and the maximum follow-up time (by default as each patient's Time Zero plus 1440 days, aligning with the follow-up period in Khalaf et al.'s publication[84]). The date of the last structured activity is defined as the latest date when patients have records (e.g., diagnosis, testing dates, see R code in Appendix 8.8 for details) in the database, aligning with standard censoring practices in Flatiron's publications. In my analysis, I applied an additional 182-day grace period beyond the last structural activity date to define the date of administrative censoring, based on the assumption that patients generally may have at least check-ups twice a year. This modification aimed to capture any death records that may occur shortly after the last structured activity, thereby avoiding a substantial underestimation of mortality that would result from censoring precisely on that last structural activity date. Date of treatment deviations are defined as receiving second-line treatment that deviates from the assigned protocol, specifically, transitions to second-line therapies other than enzalutamide following first-line abiraterone, and transitions to second-line therapies other than abiraterone following first-line enzalutamide. Patients receiving second-line treatments as per protocol were not assigned a non-deviating treatment date.

For determining the outcome date and type (death vs. administrative censoring) in both ITT and AT analyses (see Section 8.3.4 for all analyses undertaken), only the death date and administrative censoring date were considered. If a patient's administrative censoring date precedes their death date, they are censored at the administrative censoring date (i.e. last structural activity date + 182-day grace period), and their death outcome is marked as 0. If not, death is marked as 1, with their death date as the outcome date. This approach is adopted because Flatiron's death data, which comes from various sources including hospital death records, obituaries, and others, may feature patients who are lost to clinic follow-ups (e.g., potentially missing subsequent treatment information) but are later reported deceased after a period of no follow-up. Consequently, setting a uniform administrative censoring date as the data cut-off (August 31, 2023) could introduce bias. This is particularly pertinent in treatment sequence analyses, where patients lost to follow-up could theoretically deviate from their

assigned treatment strategy only after transferring to be managed at a healthcare provider (community or academic) outside the Flatiron Health Network. Implicitly including a period of lost follow-up in patient outcome calculations can lead to incorrect patient classification, especially in AT and PP analyses.

For PP analyses, outcomes and their dates are separately defined by incorporating an additional date: the date of treatment deviation. If a patient's treatment deviation date is earlier than either their death or administrative censoring date (as previously defined for ITT and AT analyses), they are censored on the treatment deviation date, with their death outcome marked as 0.

8.3.3.3. Defining time-varying covariates and finalising the "long" dataset

To accommodate the use of advanced statistical methods for adjusting for time-varying confounding (Section 8.3.4.5), the "wide" dataset capturing only patients' baseline characteristics and outcomes needs to be expanded into a "long" format that can incorporate patients' time-varying characteristics. In my case study, I expanded the wide dataset from Table 8.5 into the long dataset of Table 8.6 (Step I.4 in Figure 8.4). In this long format, each row represents a patient's status at a specific interval k (each interval k spans from day k to day k+1).

Table 8.6 GUTG-001 Analogue Target Trial data codebook – final cleaned "long" time-varying dataset (TTE_GUTG001_DTR_IPW)

Each row in the dataset represents a patient's status at interval k (between day k and k+1)			
Variable	**Variable Name**	**Variable Type**	**Content/Coding**
1. Time interval			
Start time of an interval	time	Numerical	The dataset captures time-varying covariate information, starting from Time Zero (time = 0) and tracking each patient through daily intervals until an outcome is recorded (death, censoring due to treatment strategy deviation, or administrative censoring). Outcomes at the end of each day, or interval k, are marked: y = 1 indicates a death on the subsequent day (k+1). Observation for each patient continues up to a maximum of K days after Time Zero (with a default K = 1440), unless they developed an outcome before this time. Patients with TypeY_DTR equal to DateOfDeath have number of observations matching their SurvTime_DTR (as death on day k+1 are captured within interval k). Those censored due to any reasons receive an extra observation (SurvTime_DTR + 1), with censoring noted at an interval's start—censoring on day k (i.e., last follow-up date, or the date of receiving a non-protocol treatment) falls within interval k. The only exception applies to patients remaining alive at day 1440, whose observation concludes at interval k = SurvTime_DTR.
Start day of the an interval	tstart	Numerical	Indicates the beginning of an interval, equals to the "time" variable in this study.

Stop day of an interval	tstop	Numerical	Indicates the end of an interval, equals to the tstart + 1.
Second-line treatment initiation indicator	TimeCrossover	Binary	0: No initiation of second-line treatment within an interval k (i.e., on day k) 1: Second-line treatment was initiated within an interval k (i.e., on day k)
2. Participant identifier			
Patient ID	PatientID	Character	Unique identifier for each patient.
3. Treatments			
First-line treatment coding	FirstLine_num	Binary	0: Enzalutamide 1: Abiraterone
4. Patient characteristics at baseline			
Age at first-line treatment Initiation	Age_TimeZero	Numerical	Calculated by subtracting the birth year from the year of first-line treatment initiation
Calendar year of first-line treatment initiation	cal_time	Numerical	Year extracted from the Time Zero date
Group stage at initial prostate cancer diagnosis	GroupStage	Categorical	I, II, III, IV, or Unknown/Not documented
Metastasis presence at initial prostate cancer diagnosis	MStage	Categorical	M0, M1, or Unknown/Not documented
Gleason score at initial prostate cancer diagnosis	GleasonScore	Categorical	Low/Very low risk: ≤ 6, Intermediate risk: 7, High/Very high risk: 8-10, Unknown/Not documented
Previously treated with docetaxel	f_pretreat_treat_allowed_docetaxel	Binary	Indicates if the patient was treated with docetaxel before Time Zero: Yes or No
Previously treated with other systematic treatments	f_pretreat_treat_other_systemic	Binary	Indicates if the patient was treated with any non-docetaxel systematic treatment before Time Zero: Yes or No
Race	Race	Categorical	Asian, Black or African American, White, Other Race, Unknown/Not documented
SES	f_SES	Categorical	Flatiron Health SES data, assumed to represent patient's SES at Time Zero. 1 represents the lowest SES, while 5 is the highest. Categories: 1, 2, 3, 4, 5, Unknown/Not documented
Days since mCRPC diagnosis to Time Zero (Categorical)	Gap_TimeZero_mCRPC_cat	Categorical	< 30 days, 31-90 days, 91-182 days, 182-365 days, > 365 days
Days since mPC diagnosis to Time Zero (Categorical)	Gap_TimeZero_mPC_cat	Categorical	≤ 30 days, 31-90 days, 91-182 days, 182-365 days, 366-730 days, > 730 days
PSA at Time Zero (ng/mL)	PSA_TimeZero	Numerical	The latest PSA lab test value within 30 days prior to or on the date of Time Zero, some patients may have unknown values
PSA at Time Zero (ng/mL, Categorical)	PSA_TimeZero_cat	Categorical	Categories based on the PSA_TimeZero value: < 4, 4 ≤ PSA < 10, 10 ≤ PSA < 20, 20 ≤ PSA < 100, > 100, Unknown
Haemoglobin at Time Zero (g/L)	Hb_TimeZero	Numerical	The latest Hemoglobin lab test value within 30 days prior to or on the date of Time Zero, some patients may have unknown values
Haemoglobin at Time Zero (g/L, Categorical)	Hb_TimeZero_cat	Categorical	Categories based on the Hb_TimeZero value: Hb < 120, 120 ≤ Hb < 140, Hb ≥ 140, Unknown
ECOG performance status at time Time Zero	ECOG_TimeZero	Categorical	0, 1, 2, Unknown
Diabetes at Time Zero	DM_TimeZero	Binary	Whether patients have any diabetes diagnosis within half year prior to Time Zero or on the date of Time Zero 0: No 1: Yes

Patient's combined comorbidity score at Time Zero	Comorbid_Score_TimeZero	Numerical	Calculated using the Harvard Group's algorithm of Combined comorbidity score for claims data[578,579,589] with adapted R codes specified in the Appendix 8.6. Any relevant diagnoses within half a year prior to Time Zero or on the date of Time Zero were used to classify each specific category of comorbidity as defined in the algorithm.
5. Time-varying patient characteristics			
PSA (ng/mL)	PSA_tvary	Numerical	Latest PSA level at the start of a given interval k (i.e., day k); carry over the value of a patient's last interval if no new test; carry over unknown status if no baseline PSA until any updated test
PSA (ng/mL, Categorical)	PSA_tvary_cat	Categorical	Categories based on the PSA_tvary value: < 4, $4 \leq$ PSA < 10, $10 \leq$ PSA < 20, $20 \leq$ PSA $<$ 100, > 100, Unknown
Haemoglobin (g/L)	Hb_tvary	Numerical	Latest haemoglobin level at the start of a given interval k (i.e., day k); carry over the value of a patient's last interval if no new test; carry over unknown status if no baseline haemoglobin until any updated test
Haemoglobin (g/L, Categorical)	Hb_tvary_cat	Categorical	Categories based on the Hb_tvary value: Hb < 120, $120 \leq$ Hb < 140, Hb ≥ 140, Unknown
ECOG performance status	ECOG_tvary	Categorical	Latest ECOG performance status record at the start of a given interval k (i.e., day k); carry over the value of a patient's last interval if no new test; carry over unknown status if no baseline ECOG performance status until any updated record: 0, 1, 2, 3, 4, Unknown
Diabetes status	DM_tvary	Binary	For each interval k (starting at day k), if a patient has a recorded diabetes diagnosis at any point prior or on day k, they are labelled as having diabetes (DM_tvary = 1) for that interval and all subsequent intervals. Before any diabetes diagnosis, they are labelled as not having diabetes (DM_tvary = 0).
Patient's combined comorbidity score	Comorbid_Score_tvary	Numerical	The same algorithm for calculating the patient's combined comorbidity score at Time Zero were applied, with each comorbidity subgroup separately assessed for each interval before calculating the combined comorbidity score per interval. For example, in each interval k (starting on day k), a patient with a documented hypertension diagnosis at any time up to day k is marked as having hypertension (Comorbid_HTN_tvary= 1) for that interval and all subsequent ones. Prior to a hypertension diagnosis, the patient is considered to have no hypertension (Comorbid_HTN_tvary = 0). This procedure is applied to all 20 comorbidity subgroups, with the combined comorbidity score for each interval calculated subsequently based on the validated algorithm from the Harvard group.[589]
6. Outcome			
Death status at the end of an interval	y		1: patient died at the end of the interval k (i.e. on day k +1) 0: patient remaining alive at the interval k (i.e. on day k +1) NA: patient censored for any reasons at the beginning of the interval

Status of administrative censoring at the end of an interval for PP analysis	c_admin		1: patient administrative censored at the beginning of the interval k (i.e. on day k) 0: patient remaining alive and non-administrative censored at the beginning of the interval k (i.e. on day k) NA: patient censored due to treatment strategy deviation at the beginning of the interval (i.e. on day k)
Status of censoring due to treatment strategy deviation at the end of an interval for PP analysis	c_treatdevL2		1: patient censored due to treatment strategy deviation at the beginning of the interval k (i.e. on day k) 0: patient remaining alive and non-censored due to treatment strategy deviation at the beginning of the interval k (i.e. on day k) NA: Patients administrative censored at the beginning of the interval k (i.e. on day k)

The transformation of the dataset into a long format here was specifically tailored for the application of the IPW (censor-weight) method, as detailed in Section 8.3.4.5. This approach necessitated the exclusion of any patient data beyond the point of censoring due to treatment deviation. That is, the number of rows for each patient in the long dataset was determined by the outcome types and dates outlined for the PP analysis in Table 8.5.

More precisely, Table 8.7's simplified example showcases various possible scenarios of the transformation: Patient 1 & 2, who deviated from their assigned treatment strategy on day 3, have their data retained up to the interval beginning on day 3—the date when they initiated a non-protocol second-line treatment. Both patients had an administrative censoring date on day 5 (i.e., after day 3). The key difference is that Patient 1 has a death date recorded on day 5, unlike Patient 2 who has no death record in the database. Nonetheless, both are treated similarly in the PP analysis —whether adjusted for time-varying confounding not— by discarding their data post-deviation. That is, the final interval for these patients is flagged as treatment deviation censoring = 1, with death records marked as unavailable. Retaining the interval of treatment deviation is crucial for constructing the IPW models in analyses adjusted for advanced time-varying confounding, as detailed in Section 8.3.4.3.2. However, this specific interval will receive weight of 0 in the survival analysis adjusted for time-varying confounding, as noted in Section 8.3.4.5.

Patient 3, who died on day 2 without starting any second-line treatment, has records retained up to the interval starting on day 1 (ending on day 2), and this interval is marked with death = 1. Patient 3's administrative censoring date appears (artificially) after the death date because, in my analysis, the administrative censoring date includes a 182-day grace period following the patient's last structural activity date (see Sections 8.3.3.2-8.3.3.3 for details). While this may occur to some patients, it is not universally the case. Patient 4, who died on day 3 without deviating from his assigned treatment strategy, has records retained up to the interval starting on day 2, and this interval is marked with death = 1. The distinction between Patients 3 and 4, is that Patient 4 initiated second-line

treatment before their death (on day 1), yet both complied with their assigned treatment strategies throughout the study.

Patients 5 and 6 were administratively censored before experiencing any death outcome and have their records retained up to the interval that started on the date of administrative censoring (day 8 & day 4, respectively) and flagged as administrative censoring = 1. In the final survival analysis, however, records flagged with administrative censoring = 1 were excluded (detail in Section 8.3.4.5). This is because their survival status at the end of the administrative censoring interval is unknown; it is only known that they were alive at the end of the preceding interval. In this simplified example, patient 5, with a death record on day 12, was administratively censored earlier (i.e., on day 4). In the actual data, this scenario occurs when a patient exceeds the default maximum follow-up time (i.e.,1440 days) or reaches the end of a 182-day grace period following their last structural activity date before their death record, as detailed in Sections 8.3.3.2 and 8.3.4.3. However, a shorter period is used in the simplified example for demonstration purposes. Both Patients 5 and 6 started a second-line treatment, in accordance with their assigned treatment strategies, before being administratively censored. However, it is also possible for patients to be administratively censored before initiating any second-line treatment.

In Table 8.7's simplified example, each patient has only a small number of rows (representing daily observations), while in the actual long dataset I created for the PC1 case study, each patient may have at most 1441 rows, covering intervals up to the one starting on day 1440 and ending on day 1441, provided the patient was alive on day 1440 and remained non-deviating from their assigned treatment strategy. The default maximum follow-up period of 1440 days was set to align with the maximum possible follow-time reported in Khalaf et al.'s Kaplan-Meier (KM) survival curves.[84]

For defining time-varying patient characteristics in Table 8.6, the same covariates recorded in Table 8.5's baseline dataset were used, where applicable (i.e., where characteristics could change over time and where data were available). These include PSA, ALP, LDH, Hb, ECOG score, diabetes status, and combined comorbidity score. Longitudinal data post Time Zero was linked to provide daily updates of these characteristics. The time since first-line treatment initiation and patients' age at second-line treatment start (i.e., calculated as the age at first-line treatment initiation plus the elapsed time) were inferred from the number of time intervals noted in each corresponding observation row, and thus were not separately recorded.

Table 8.6 offers detailed definitions for each time-varying variable. In the absence of new lab results, values were carried forward from the last observed interval. If a patient's records contained only unknown values up to a certain point, those intervals were marked with unknown/missing values (i.e., retaining the unknown status). Certain variables, such as race, SES, and prostate cancer status at initial diagnosis, were only available at baseline (Time Zero), with no subsequent variations over time.

Table 8.7 Simplified examples of outcome and censoring flags in the long dataset

Scenarios	ID	Tstart: start date of the interval	Tstop: end date of the interval	Initiation of second-line treatment	Administrative censor date	Tx strategy deviation date	Death date	Administrative censor flag	Tx strategy deviation flag	Death flag
Patient 1 deviated from the assigned treatment strategy before death and administrative censoring	1	0	1	0	5	3	5	0	0	0
	1	1	2	0	5	3	5	0	0	0
	1	2	3	0	5	3	5	0	0	0
	1	3	4	1	5	3	5	NA	1	NA
Patient 2 deviated from the assigned treatment strategy before death and administrative censoring	2	0	1	0	5	3	NA	0	0	0
	2	1	2	0	5	3	NA	0	0	0
	2	2	3	0	5	3	NA	0	0	0
	2	3	4	1	5	3	NA	NA	1	NA
Patient 3 experienced death without treatment deviation	3	0	1	0	6	NA	2	0	0	0
	3	1	2	0	6	NA	2	NA	NA	1
Patient 4 experienced death without treatment deviation	4	0	1	0	6	NA	3	0	0	0
	4	1	2	1	6	NA	3	0	0	0
	4	2	3	0	6	NA	3	NA	NA	1
Patient 5 was administratively censored without treatment deviation	5	0	1	0	8	NA	NA	0	0	0
	5	1	2	0	8	NA	NA	0	0	0
	5	2	3	0	8	NA	NA	0	0	0
	5	3	4	0	8	NA	NA	0	0	0
	5	4	5	1	8	NA	NA	0	0	0
	5	5	6	0	8	NA	NA	0	0	0
	5	6	7	0	8	NA	NA	0	0	0
	5	7	8	0	8	NA	NA	0	0	0
	5	8	9	0	8	NA	NA	1	NA	NA
Patient 6 was administratively censored before their death record	6	0	1	0	4	NA	12	0	0	0
	6	1	2	0	4	NA	12	0	0	0
	6	2	3	0	4	NA	12	0	0	0
	6	3	4	1	4	NA	12	0	0	0
	6	4	5	0	4	NA	12	1	NA	NA

NA: not available; Tx: treatment

8.3.4. Implementation of statistical analysis

8.3.4.1. Baseline characteristics, treatment patterns, overview of survival analyses and RCT-RWE agreement assessment

To facilitate comparisons with the Khalaf et al. study[84], I analysed patient characteristics between groups at two key points: the initiation of first-line treatment and the crossover to second-line therapy. This involved extracting patients' baseline characteristics from the wide dataset (Table 8.5) and the intervals marked as the start of second-line treatment from the long dataset (TimeCrossover =1, Table 8.6). Utilising the table1 package in R, chi-square tests and t-tests were performed to compare characteristics between the study groups for categorical and continuous variables, respectively (see the end of Appendix 8.6 for details). Additionally, a Sankey diagram was utilised to visually examine treatment patterns within the PC1 study population. Outcome labels from the ITT analysis in the wide dataset (Table 8.5) was integrated to the LOT dataset to enrich the diagram's detail. This helped distinguish patient who passed away before receiving any second- or third-line treatments from those remained alive at the study's 1440-day follow-up cut-off, providing more comprehensive insights than the preliminary checks in Section 8.2.2.

In Section 8.3.4.2, I detail survival outcome analyses without confounding adjustment using simple methods, including ITT, PP, AT. The AT analysis is divided into two parts: one including patients who did not receive any second-line treatment, and another limited to those who received second-line treatment. In Sections 8.3.4.4 to 8.3.4.5, I describe survival analysis with confounding adjustments. Section 8.3.4.3 precedes this by introducing covariate selections and models used to derive the weights for these adjustments. It details collaboration with clinical experts to identify key confounding variables for adjustment and the use of directed acyclic graph (DAG) to guide these discussions.

In Section 8.3.4.4, I describe adjusting Section 8.3.4.2's analyses for baseline confounding using the IPW method. Section 8.3.4.5 outlines detailed steps for implementing the IPW method to adjust for both baseline and time-varying confounding in the context of comparing treatment sequences, effectively refining the PP analysis in Section 8.3.4.4 by addressing bias from informative censoring. These procedures correspond to Steps II.1 to II.3 in Figure 8.4.

I assessed the findings of the emulated trial against those of Khalaf et al.[84], using the predefined RCT-RWE agreement criteria from the study protocol (Chapter 7, Section 7.5.5). The assessment aimed at evaluating the performance of the adjusted analysis in replicating the findings of the benchmark trial it aimed to emulate. The specified formulas in the protocol were followed. For criterion 4, the survival curve digitisation method by Guyot et al.[574] and survHE R package[593] were used to extract and overall survival (OS) KM curve reported in the Khalaf et al. trial for comparison.

The R code for reproducing Khalaf et al.'s survival curves and findings is detailed in Appendix 8.9.

8.3.4.2. Survival analysis with simple methods without confounding adjustments

To compare findings with those reported by Khalaf et al.[84], OS KM curves were derived for visual inspection, and Cox proportional hazards models were employed to calculate hazard ratios (HR) between two treatment groups using outcome information in the wide dataset (Table 8.5). The following four analyses comprise the set of unadjusted analyses:

- 1a. ITT analysis: This analysis included all patients starting with either abiraterone or enzalutamide, following them regardless of adherence to the assigned sequence for second-line treatment (abiraterone → enzalutamide versus enzalutamide → abiraterone, respectively). Patients were not censored for deviations due to treatment sequence strategy deviation and were followed until death or administrative censoring
- 1b. PP analysis: Using the same patient group as the 1a. ITT analysis, this approach censored patients who deviated from their assigned treatment sequence, specifically at the time when they received a non-protocol second-line treatment (non-enzalutamide or non-abiraterone, respectively).
- 1c. AT analysis: This analysis began with a subset of patients who were observed to adhere their assigned treatment sequence (abiraterone → enzalutamide versus enzalutamide → abiraterone). Any patient observed receiving an unapproved second-line treatment was fully removed from the dataset (non-enzalutamide or non-abiraterone, respectively). Those who received only one line of treatment were considered adherent.
- 1d. AT analysis, limited to patients receiving at least two lines of treatment): This analysis, a more selective subset of the 1c AT analysis, is restricted to patients who received at least two lines of treatment and strictly adhered to their assigned treatment sequence. Only those who survived long enough to start the assigned second-line treatment were included.

Analyses 1a, 1c, and 1d used death status and OS time defined for ITT analysis, while Analysis 1b used the same metrics for PP analysis (Table 8.5). These comparisons provide descriptive statistics that show differences in outcomes between two groups under possibly different conditions, thus not establishing counterfactual causality. Specifically, the 1a ITT analysis compares outcomes between first-line abiraterone versus enzalutamide groups, irrespective of subsequent treatments received. This could also be seen as outcome differences between treatment sequences (abiraterone → enzalutamide, enzalutamide → abiraterone), but these are "contaminated" by significant protocol deviations. The 1b PP analysis contrasts outcomes between groups based on periods when each patient adhered to their assigned treatment sequences, excluding data periods affected by deviations from the treatment protocol. Although it seems that effects from non-protocol treatments were

eliminated, the non-random censoring is informative and thus introduces bias, as discussed in Section 4.6.1.2.1 of Chapter 4.

Analyses 1c and 1d aimed to completely exclude patients who deviated from their assigned treatment sequences. While this approach seemingly eliminates contamination from non-protocol treatments, it introduces selection bias as the elimination was based on information observed after the initiation of first-line treatment (Time Zero). In AT Analysis 1c, excluding patients who deviated but survived long enough to receive a second-line treatment from the analysis could over represent those who did not survive to receive second-line treatment, potentially underestimating the effect of a treatment sequence strategy. In contrast, AT analysis 1d, which only includes patients surviving long enough to receive second-line treatments, likely introduces immortal time bias and could overestimate the survival benefits of a treatment sequence strategy. These revisited the simple methods discussed in Chapter 4 and were observed being used in quite a few RWD studies, primarily that resembled the 1d Analysis.[297,298,319,320,325]

These methods were set up to contrast with findings from the remaining analyses with confounding adjustments. They reflect simple methods discussed in Section 4.6.1.1 in Chapter 4, particularly those resembling Analysis 1d, which were frequently observed.[315,318,322,323,326,327,329,330] The R code for these analyses is detailed in corresponding numbered sections in Appendix 8.10. The R code for each analysis is detailed in the corresponding sections of Appendix 8.10.

8.3.4.3. Deriving IPW weights for analyses with confounding adjustments

Two types of IPW weights were derived for confounding adjustments. Baseline inverse probability of treatment weights (IPTW) were derived to emulate randomisation at Time Zero, making patients across study groups comparable at the start of first-line treatments. Additionally, inverse probability of censoring weights (IPCW) were derived to address time-varying confounding (post-Time Zero) resulting from patients deviating from their assigned treatment sequences in the study protocol ($IPCW_{txdev}$). In real-world, patients may receive a variety of second-line treatments, complicating the estimation of the effectiveness of a specific treatment sequence.

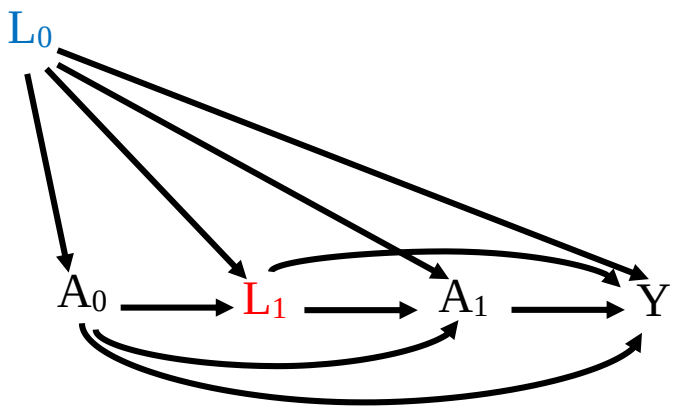
Baseline IPTW was used to refine the survival analysis of simple methods described in Section 8.3.4.2, with adjustments for baseline confounding (Section 8.3.4.4). Furthermore, baseline IPTW in conjunction with time-varying $IPCW_{txdev}$ were applied to refine the PP analysis in Section 8.3.4.2, adjusting for both baseline end time-varying confounding (Section 8.3.4.5). This is an adaptation of the advanced IPW method (i.e., $IPTW*IPCW_{txdev}$) for adjusting time-varying confounding in comparing treatment sequences in Sections 4.6.1.2.1-4.6.1.2.2 of Chapter 4. Immortality time biases and selection biases in the AT analyses (Section 8.3.4.2) cannot be mitigated using advanced statistical methods. Meanwhile, implementing the advanced IPW method requires starting with the PP analysis,

instead of the ITT analysis due to its reliance on artificial censoring (Sections 4.6.1.2.1-4.6.1.2.2, Chapter 4). Therefore, survival analysis with advanced statistical methods focuses solely on refining the PP analysis in Section 8.3.4.2.

Details of weight derivations for these IPW sets, including covariate selections, are detailed in Sections 9.1.1.1 to 9.1.1.3. The R code for weight derivation are included in Appendix 8.10. This code has adapted and extended insights from the CAUSALab courses at Harvard University as well as the causal inference course at the Private University for Health Sciences and Health Technology in Hall in Tirol (UMiT).

8.3.4.3.1. Covariates selection

A simple DAG, shown in Figure 8.6, facilitated discussions with clinicians to identify important baseline (L_0) and time-varying confounders (L_1) that may influence clinicians' choices of first-line and second-line mCRPC treatments, as well as patients' survival outcomes. An initial list of these covariates was created during the protocol design (Section 7.6.3 of Chapter 7). Further discussions with clinicians upon data checking (Section 8.2 and Section 8.3.3) refined this list by incorporating additional relevant variables available in the data and noting those that may be relevant but were unavailable. Table 8.8 outlines the importance of each variable being considered and whether they have been incorporated in the final IPW derivation model in Section 8.3.4.3.2 and Section 8.3.4.3.3. In the selection process, the methodological aspects in Sections 8.3.4.3.2 and 8.3.4.3.3 were also explained to clinicians at a conceptual level to enhance the accuracy of the selection.



A_0: First-line mCRPC treatment (time 0)

L_0: Confounders at baseline (time 0)
(e.g. performance status)

A_1: Second-line mCRPC treatment at the time of progression/treatment intolerability

L_1: Confounders at the time of treatment-switching due to progression/treatment intolerability
(e.g. performance status)

Y: Outcome
(e.g. Death)

Figure 8.6 A simple DAG of time-varying confounders in the GUTG-001 Analogue Target Trial

mCRPC: metastatic castration-resistant prostate cancer

Table 8.8 Categorisation of baseline and time-varying covariates for case study PC 1

Level of importance	Baseline covariates (L_0)	Time-varying covariates (L_1)	Included in the final model for IPW derivation
Most important factors	Age Previous treatment of docetaxel Previous treatment of other systematic agents Days since mCRPC diagnosis Days since mPC diagnosis TNM Group Stage (initial diagnosis) ECOG performance status Diabetes	Age (i.e., time since the initiation of first-line treatment) ECOG performance status Diabetes	Yes
Likely important but less critical than the most important factors	Gleason Score (initial diagnosis) M Stage (initial diagnosis) PSA Haemoglobin	PSA Haemoglobin	
Likely important but with uncertain data quality	Socialeconomic status Comorbidity score	Comorbidity score	
Ambiguous factors	Race Calendar year of first-line treatment initiation	-	
Likely important but embedded in other factors	N Stage (initial diagnosis) M Stage (initial diagnosis) Line of previous treatments	-	No
Information unavailable but potentially relevant	-	Progression status	
Information unavailable and not as relevant	Histology Morphology	-	
Unimportant factors	Days since CRPC diagnosis Alkaline phosphatase lactate dehydrogenase	Alkaline phosphatase lactate dehydrogenase	

ECOG: Eastern Cooperative Oncology Group; IPW: inverse probability weight; mCRPC: metastatic castration-resistant prostate cancer; mPC: metastatic prostate cancer; PSA: Prostate-Specific Antigen

According to Dr. Pezaro, the most important key factors influencing clinicians' decisions for both first and second-line mCRPC treatments include patients' age, performance status, diabetes status, TNM stage at initial prostate cancer diagnosis, and previous treatments. If patients have severe diabetes, abiraterone may be avoided since it requires concurrent use with prednisolone—a steroid—that can disrupt diabetes management. Consequently, diabetes was included as a significant, separate factor, despite potentially being accounted for in the comorbidity score. The days since mCRPC and mPC diagnoses are crucial as they indicate whether patients first experienced metastasis or castration resistance before receiving their first-line mCRPC treatment.

The TNM Group stage is crucial as it reflects patients' status at their initial prostate cancer diagnosis. Since the subcategories of the TNM stage at the initial diagnosis offer similar information, Dr. Pezaro suggested excluding them to avoid redundancy. The Gleason Score, while carrying similar information, provides nuanced risk categorisation and was therefore included. The M Stage was also retained because its specificity is vital for determining whether patients had metastasis at initial

diagnosis or if it was a case of disease progression. Lines of treatment received by patients were considered relevant but are generally captured by the types of previous treatments, as in most cases, treatment lines do not exceed one or two. PSA and Haemoglobin levels are also important but considered less critical than the ECOG score. Information on patient's socioeconomic status and derived comorbidity scores (Section 8.3.3.2-8.3.3.3) are relevant, yet their reliability is questioned due to potential inaccuracies in community-level socioeconomic status data and pre-diagnosis data availability in the Flatiron database, respectively.

Dr. Pezaro considered race and the calendar year of first-line treatment initiation as ambiguous factors. Specifically, she noted that the management of prostate cancer might have evolved over time, and while these elements could serve as proxy indicators of potential impacts on treatment outcomes, their actual influence or the degree of impact remains uncertain. However, she acknowledged that variations in available treatments across different periods could influence treatment choices. This is particularly true regarding the impact of the COVID-19 pandemic, which led to reduced use of docetaxel. The period of the pandemic itself may have led to varied outcomes for patients receiving cancer care, though the extent of this impact remains uncertain due to limited evidence. Several unimportant factors were identified and not included, such as days since CRPC diagnosis, alkaline phosphatase, and lactate dehydrogenase, despite the latter two characteristics being reported in the GUTG-001 trial.

The lack of a clear progression date in Flatiron data complicates determining whether patients switched treatments due to intolerance or disease progression. Dr. Pezaro highlighted that even if this information was available, recording practices could introduce significant ambiguity. Clinicians generally rely on ECOG scores and imaging/lab findings to make treatment decisions, and progression dates are may be reported inconsistently across practices, due to its retrospective nature. Dr. Pezaro noted that the time elapsed since first-line treatment initiation, by the point of treatment switching, typically indicates whether the switch was due to intolerance (early, within 1-2 months) or disease progression (later), with few exceptions. Histology and morphology at the patient's initial prostate cancer diagnosis were originally considered relevant in the study protocol but were later deemed less significant due to other available information that captured most decisions made closer to their mCRPC status.

8.3.4.3.2. Baseline inverse probability of treatment weights (IPTW)

Conceptually, baseline IPTW is used to emulate patient randomisation at the initiation of first-line treatment, refining the survival analyses with simple methods in Section 8.3.4.2. This adapted methods and formula described in Section 4.6.1.2.1, Chapter 4.

$$W^T = \begin{Bmatrix} \frac{1}{\Pr[G_0 = 1 \mid L_0]} & if\ G_0 = 1 \\ \frac{1}{1 - \Pr[G_0 = 1 \mid L_0]} & if\ G_0 = 0 \end{Bmatrix} \quad (8.1)$$

$$SW^T = \begin{Bmatrix} \frac{\Pr[G_0 = 1]}{\Pr[G_0 = 1 \mid L_0]} & if\ G_0 = 1 \\ \frac{1 - \Pr[G_0 = 1]}{1 - \Pr[G_0 = 1 \mid L_0]} & if\ G_0 = 0 \end{Bmatrix} \quad (8.2)$$

Specifically, Formula 8.1 specifies that patients who initially receive abiraterone (i.e., with the treatment strategy G = abiraterone → enzalutamide, where $G_0 = 1$ and $G_1 = 0$) are weighted by the inverse probability of receiving that specific strategy at treatment initiation (i.e., $G_0 = 1$), conditioned on all baseline covariates L_0. Similarly, patients following the treatment strategy G = enzalutamide → abiraterone ($G_0 = 0$ and $G_1 = 1$) at their initiation of first-line treatment ($G_0 = 0$) are assigned weights based on the inverse probability of receiving that specific strategy at treatment initiation ($G_0 = 0$), also conditioned on L_0. Informally, these weights create a pseudopopulation that emulates randomisation, effectively removing the influence of baseline covariates L0 on treatment assignment A_0 (i.e., G_0 in Formula 8.1), as illustrated in Figure 8.6. Formula 8.2 represents the stabilized weights, with the numerator reflecting the overall probability of receiving treatment $G_0 = 1$ versus $G_0 = 0$.

These weights are derived using logistic models, which model the log-odds of receiving treatment as a linear function of the predictors. Specifically, two models are required for weight derivation: one for the numerator and another for the denominator. Since the probability of receiving initial treatment $G_0 = 0$ is simply 1 minus $G_0 = 1$, no separate models for the two treatment groups are necessary. Covariates L_0 included in the denominator model are those labelled "yes" for inclusion as baseline covariates in Table 8.8. From these models, each patient's probability of receiving the first-line treatment, with or without conditioning on L_0, can be estimated. Then, weights for each patient can be assigned using the formulas in Formula 8.1 and 8.2 (See Appendix 8.10 for R code).

8.3.4.3.3. Time-varying inverse probability of censoring weights for treatment strategy deviations ($IPCW_{txdev}$)

Conceptually, time-varying $IPCW_{txdev}$ is used to emulate a scenario where all patients had adhered to their specific treatment sequence strategy (G = abiraterone → enzalutamide versus enzalutamide → abiraterone) throughout the follow-up period. Informally, it does this by upweighting those who did not deviate from their assigned treatment strategy whenever someone else in the same treatment sequence group did, based on the probability of the non-deviated individuals remaining non-deviated from their assigned treatment sequence strategy (i.e., those non-artificially-censored in the PP analysis). The time-varying $IPCW_{txdev}$ weights at each time interval are calculated based on the cumulative probability of remaining non-deviated up until each time interval k (default k = 0 to

1440). Here is the formula for $IPCW_{txdev}$ calculation:

$$W_t^D = \begin{cases} \prod_{k=0}^{t} \frac{1}{\Pr[D_k=0 \mid G_{k-1}, L_0, L_k, C_k=0, D_{k-1}=0, Y_{k-1}=0]} = \prod_{k=0}^{t} \frac{1}{1-\Pr[D_k=1 \mid G_{k-1}, L_0, L_k, C_k=0, D_{k-1}=0, Y_{k-1}=0]} & if\ D_k = 0 \\ 0 & if\ D_k = 1 \end{cases} \quad (8.3)$$

$$SW_t^D = \begin{cases} \prod_{k=0}^{t} \frac{\Pr[D_k=0 \mid G_{k-1}]}{\Pr[D_k=0 \mid G_{k-1}, L_0, L_k, C_k=0, D_{k-1}=0, Y_{k-1}=0]} = \prod_{k=0}^{t} \frac{1-\Pr[D_k=1 \mid G_{k-1}]}{1-\Pr[D_k=1 \mid G_{k-1}, L_0, L_k, C_k=0, D_{k-1}=0, Y_{k-1}=0]} & if\ D_k = 0 \\ 0 & if\ D_k = 1 \end{cases} \quad (8.4)$$

Specifically, Formula 8.3 represents the cumulative weight for patients remaining non-deviated from their treatment sequence strategy (W_t^D) up to each time interval t = k. It is the inverse probability of remaining cumulatively non-deviated ($D_k = 0$), and can therefore also be viewed as the cumulative inverse probability of one minus the deviation probability. These probabilities are conditioned on the treatment strategy group the patient was in, along with their baseline and time-varying covariates. Specifically, since reasons for patients to deviate from a treatment sequence strategy might differ across two groups, it is conceptually logical to have separate statistical models that model the probability of deviating from the treatment strategy for different treatment sequence groups. A simple example is that patients with diabetes may be more likely to deviate from the treatment sequence of enzalutamide → abiraterone when receiving second-line treatment (see Section 8.3.4.3.1 for details) to avoid the use of abiraterone. Conversely, for patients in the abiraterone → enzalutamide group, diabetes may not be a significant factor, or it may influence treatment deviation in a different direction compared to the enzalutamide → abiraterone group.

Formula 8.4 represents the stabilised version of the weights from Formula 8.3 by having a numerator that reflects the cumulative probability of remaining non-deviated, dependent solely on the treatment group the patient belongs to. Similar to the baseline IPTW weights in Section 8.3.4.3.2, these weights are derived using logistic models. These models calculate the log-odds of treatment deviation as a function of the predictors, where time is modelled using a cubic spline function. Data rows where TimeCrossover = 1 in the long dataset in Table 8.6 were utilised for model derivation. Specifically, four models are necessary for deriving these weights: two for the denominator and two for the numerator. That is, each treatment sequence group (i.e., G = G_{k-1} enzalutamide → abiraterone versus abiraterone → enzalutamide) requires separate models for the numerator and denominator, for reasons described above. Baseline and time-varying covariates (L_0 and L_1) included in the denominator model are those labelled "yes" for inclusion in Table 8.8. From these models, each patient's probability of treatment deviation—often referred to as the probability of treatment-switching in literature addressing unwanted treatment-switching—with or without conditioning on L_0 and L1, can be estimated. Subsequently, weights and cumulative weights (both unestablished and stabilised) for each patient at each time interval k can be assigned to append the long dataset in Table 8.6 using Formulas 8.3 and 8.4. Although the interval where a patient deviated from their treatment

sequence strategy is included for weight derivation, these rows will receive a weight of 0 in survival analyses in Section 8.3.4.5, as shown in Formulas 8.3 and 8.4. Informally, this is because other patients that do not deviate from their treatment strategy are weighted to represent them from this time point onwards. Appendix 8.10 provides detailed R code procedures for these models and weight estimations.

8.3.4.4. Survival analysis with simple methods with baseline adjustments using IPTW

All four survival analyses in Section 8.3.4.2 were refined using baseline adjustments with IPTW weights derived in Section 8.3.4.3.2. These weights, based solely on baseline characteristics, were appended to the wide dataset (Table 8.5). The same survival outcome information in the wide dataset from was used for the four baseline confounding adjusted analyses (2a-2d). The only practical difference compared to analyses without adjustments lies in the specification of a weight function in the KM and Cox model analyses in R. This adjustment creates the baseline-adjusted ITT (Analysis 2a), PP (Analysis 2b), AT (Analysis 2c), and AT limited only to patients receiving second-line treatments (Analysis 2d). The same statistics as those described in Section 8.3.4.2 were derived for inspection. These analyses primarily reflect common approaches used in the literature for adjusting for confounding. However, they do not account for time-varying confounding in the comparison of treatment sequences. This theoretical limitation likely introduces bias, prompting the exploration of advanced adjustments in the following section (Section 8.3.4.5)

For practicality, I opted to present the 95% confidence curve of the weighted KM curve and median survival estimates directly using those reported by the survfit function in R (i.e., naïve (unadjusted) 95% CI), while acknowledging that the weighted procedure might increase uncertainty. Bootstrap methods were employed with parametric pooled logistic modelling as an alternative approach for modelling survival curves. This technique produced narrower 95% confidence intervals (CIs) and smoother survival curves, and it can accommodate non-proportional hazards. However, these were not reported in the thesis as the results were closely aligned with those from the KM analysis, albeit narrower likely due to sample sizes and additional assumptions applied. Similarly, for simplicity, the conservative robust sandwich estimator was used to calculate the 95% CIs of the HRs to account for uncertainties related to the weighting procedures. The R packages sandwich and lmtest were used to derive these variances. Despite its conservative nature, the robust sandwich estimator does not directly address uncertainties arising from the weight derivation models. Bootstrap methods with 500-1000 replications were attempted, showing marginal differences of HRs limited to ≤ 0.03 (narrower 95% CI compared to those from the robust sandwich estimator); thus, they were not separately reported in the thesis. Other methods for deriving CIs associated with IPCW have recently been developed in the literature.[594] However, I did not apply them in the thesis due to the marginal

differences observed in my case study, which obscured the practicality of for the purposes of methods exploration of this thesis. Appendix 8.10 provides detailed R code procedures for these analyses.

*8.3.4.5. Survival analysis with advanced method using IPTW*IPCW$_{txdev}$*

The PP analysis in Section 8.3.4.4 was further advanced by additional adjustment for time-varying confounding with IPCW$_{txdev}$ weights derived in Section 8.3.4.3.3 (i.e., Analysis 3). Conceptually, this aims to derive a hypothetical PP estimand as if all patients had received either abiraterone → enzalutamide versus enzalutamide → abiraterone, and all patients in each group had adhered to their assigned treatment sequence strategy (see study protocol Table 7.1 in Chapter 7). The procedures in Section 8.3.4.4 emulated the randomisation at the initiation of first-line treatment, while the next step involves adjusting for confounding due to artificially censoring patients who deviate from their treatment strategy. Practically, this involves inserting the baseline IPTW weights for each patient from Section 8.3.4.3 into the long dataset in Table 8.6. These baseline weights are then multiplied by the IPCW$_{txdev}$ derived in Section 8.3.4.4, which already exists in the long dataset after weight derivations in Section 8.3.4.3.3. The formulas 8.5 and 8.6 below illustrate how the overall weights ($W_t^{T,D}$) and the stabilised weights ($SW_t^{T,D}$) are computed, respectively.

$$W_t^{T,D} = W^T \times W_t^D \quad (8.5)$$

$$SW_t^{T,D} = SW^T \times SW_t^D \quad (8.6)$$

While each patient's baseline IPTW weight ($W_t^{T,D}$) remains consistent, their IPCW$_{txdev}$ (W_t^D) is updated at each interval, suggesting that each patient's overall weights also vary across intervals. The same approach in Section 8.3.4.4 was employed to report the HRs and KM results from the advanced confounding adjustment using the IPTW*IPCW$_{txdev}$ method. Appendix 8.10 provides detailed R code procedures for these analyses.

8.4. Results

8.4.1. Patient selection, characteristics, and treatment patterns

In Section 8.4.1, the Consolidated Standards of Reporting Trials (CONSORT) diagram shows the patient selection procedures for the PC1 study (i.e., GUTG-001 Analogue Target Trial), corresponding to the cohort selection processes detailed in Section 8.3.3.1. This resulted in the inclusion of a total of 4,144 patients, with 2,172 starting abiraterone as their first-line mCRPC treatment and 1,972 receiving enzalutamide. Figure 8.7 outlines the inclusion and exclusion procedures, showing the number of patients at each step.

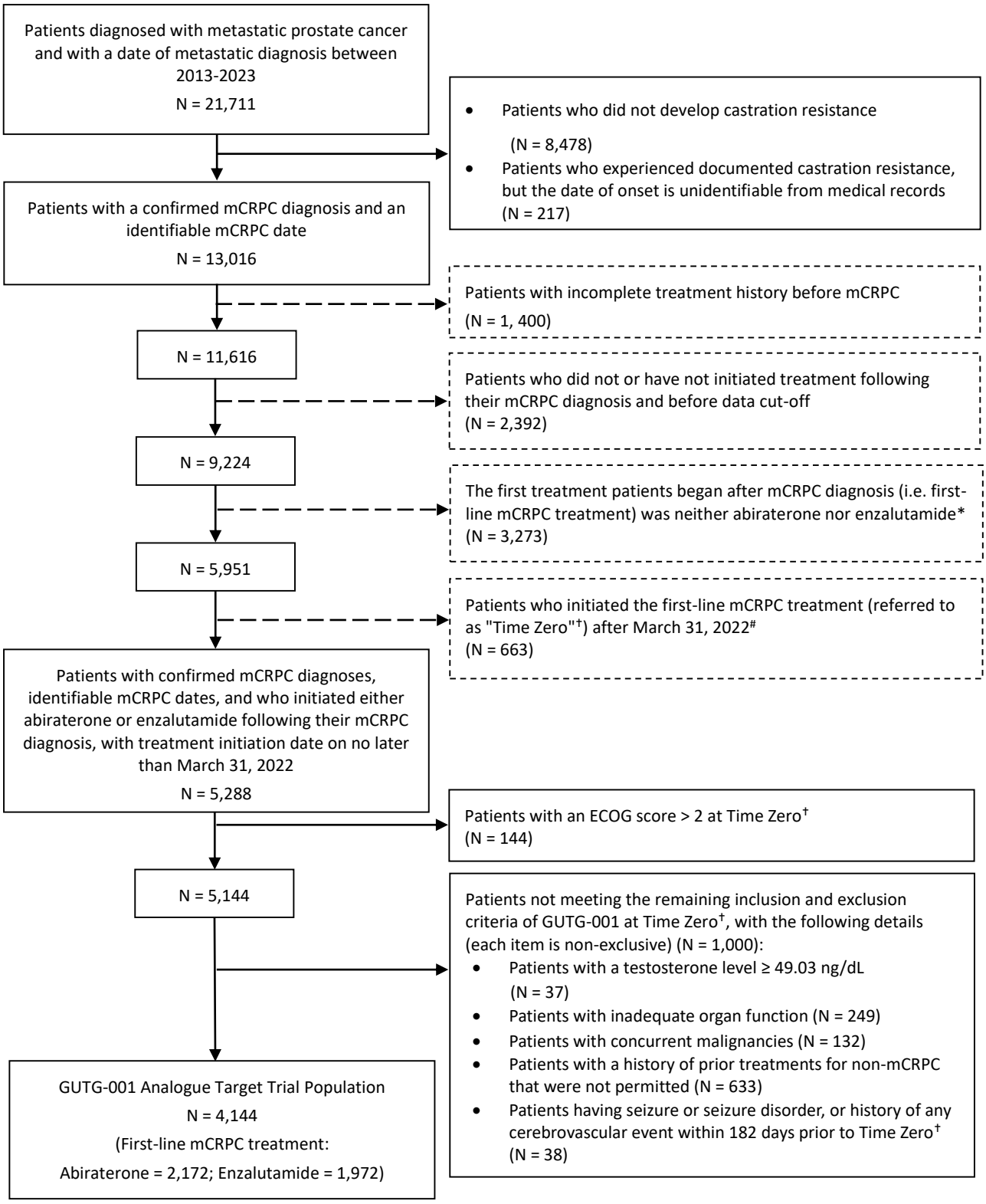


Figure 8.7 Consolidated Standards of Reporting Trials (CONSORT) diagram for the GUTG-001 Analogue Target Trial using Flatiron data

Solid Lines represent cohort selection procedures that align with the inclusion and exclusion criteria of the GUTG-001 trial. Dashed Lines represent additional cohort selection procedures not stated in the GUTG-001 trial's inclusion and exclusion criteria. However, these procedures were considered to ensure the Target Trial Population aligns more closely with the analogue of the GUTG-001 trial.
* Treatments initiated before mCRPC diagnosis and that continued through and concluded after the mCRPC diagnosis are considered as treatments prior to mCRPC.
† Time Zero: This refers to the initiation of the first line-of treatment (LOT) post-mCRPC diagnosis. Treatments initiated before mCRPC diagnosis and that continued through and concluded after the mCRPC diagnosis are considered as LOT prior to the mCRPC diagnosis (i.e. non-mCRPC LOTs) (For a detailed justification, please refer Section 8.3.3.1)
# This procedure may potentially result in including patients diagnosed with mCRPC in 2022 who had a more severe prognosis (or fit enough) and thus commenced treatment immediately upon mCRPC diagnosis or shortly after. This procedure also precludes patients diagnosed with mCRPC after August 31st, 2022

Table 8.9 outlines the characteristics of the included patients at baseline and at the time of receiving second-line treatment, comparing them with those observed in the GUTG-001 findings.[84] I reproduced the statistics reported in Khalaf et al. and displayed them at the top of the table, while the results from my emulated Target Trial are shown at the lower part of the table.

The p-values in Table 8.9 evaluate whether the characteristics of Group A (first-line abiraterone) and Group B (first-line enzalutamide) are comparable at each time point: baseline (p-values in the middle column) and the initiation of second-line treatment (p-values in the rightmost column). Chi-square tests were used for categorical variables, and t-tests for continuous variables, as outlined in the study protocol (Table 7.1, Chapter 7). While p-values for the GUTG-001 trial were not provided in Khalaf et al.'s publication[84], my PC1 case study (GUTG-001 Analogue) includes p-values for both time points. A p-value below 0.05 indicates a statistically significant difference in a specific patient characteristic between the groups (A versus B) at either baseline or the initiation of second-line treatment.

The baseline characteristics in my study are generally similar to those observed in the GUTG-001 trial, both at the time when patients received first-line treatment and when they started second-line treatment. However, there is a slightly lower percentage of patients proceed to receive second-line treatment at the 4-year cut-off (i.e., 1440 days) in my study—61% in the abiraterone first group and 56% in the enzalutamide group—compared to those in the GUTG-001 trial (72% and 74%, respectively). Dr. Pezaro suggested that this discrepancy is reasonable, as patients in trials are more likely to be offered additional treatments than real-world practice.

Although patients were not randomised to receive first-line abiraterone versus enzalutamide in the real-world setting, the characteristics of my two study groups were mostly similar. Dr. Carmel noted that this similarity was expected, as clinically, abiraterone and enzalutamide are often considered interchangeable under most conditions. However, notable differences were observed in the time from metastasis to first-line mCRPC treatment (436 versus 490 days). Dr. Pezaro highlighted that abiraterone, frequently combined with steroids, might be administered slightly earlier to being helpful in mitigating symptoms from metastasis. This indicates that patients in the abiraterone first-line group are likely to develop castration resistance before experiencing metastasis—meaning they are closer to their initial metastases when they begin mCRPC treatment. Conversely, the sequence may be reversed in the enzalutamide first-line group. For the same reasons, this likely results in the observed shorter intervals from metastasis to treatment in the abiraterone first-line group (343 versus 398 days).

Table 8.9 GUTG-001 Analogue Cohort Characteristics in Flatiron data

	Baseline (first-line treatment start date)			Crossover (second-line treatment start date)		
	Group A: First-line Abiraterone	**Group B: First-line Enzalutamide**	**P-value**	**Group A: First-line Abiraterone**	**Group B: First-line Enzalutamide**	**P-value**
GUTG-001 Trial (Figures extracted from Table of Khalaf et al.[84] study for comparison with the emulated Target Trial GUTG-001 Analogue)						
Sample Size	**N = 101**	**N = 101**		**N = 73**	**N = 75**	
Age (years)						
Median [Min, Max]	72.9 [51.3, 93.3]	77.6 [49.3, 94.1]		73.8 [51.5,92.7]	78.0 [49.8, 93.2]	
PSA (ng/mL)						
Median [Min, Max]	35.0 [2.2, 2817.0]	37.0 [1.7, 1060.0]		16.0 [0.8, 991.0]	12.0 [0.20, 1604.0]	
Alkaline phosphatase, relative to ULN						
Median [Min, Max]	0.82 [0.29, 12.50]	0.75 [0.30, 47.80]		0.88 [0.31, 6.87]	0.75 [0.31, 4.67]	
Lactate dehydrogenase, relative to ULN						
Median [Min, Max]	0.79 [0.37, 4.00]	0.80 [0.31, 12.90]		0.85 [0.22, 4.69]	0.74 [0.38, 2.46]	
Haemoglobin (g/L)						
Median [Min, Max]	130 [89, 155]	130 [89, 165]		132 [87, 152]	129 [79, 157]	
ECOG performance status 0-1						
N (%)	89 (88%)	79 (78%)		62 (85%)	57 (76%)	
Previous treatment of docetaxel for castration-sensitive disease						
N (%)	5 (5%)	6 (6%)		-	-	
Bone metastasis						
N (%)	85 (84%)	82 (81%)		61 (84%)	65 (87%)	
Lung metastasis						
N (%)	8 (8%)	9 (9%)		6 (8%)	7 (9%)	
Lung metastasis						
N (%)	5 (5%)	7 (7%)		4 (5%)	7 (9%)	
GUTG-001 Analogue (PC1 case study)						
Sample Size	**N = 2,172**	**N = 1,972**		**N = 1,317**	**N = 1,101**	
Age (years)						
Mean (SD)	74.5 (8.35)	75.1 (8.22)	0.0124	74.0 (8.37)	74.7 (8.41)	0.042
Median [Min, Max]	76.0 [43.0, 85.0]	77.0 [46.0, 88.0]		75.0 [44.5, 85.8]	75.8 [47.2, 87.5]	

PSA (ng/mL)						
Median [Min, Max]	24.3 [0, 3230000]	21.4 [0.0370, 1630000]	0.8	20.9 [0.0100, 10700000]	16.9 [0.00476, 1480000]	0.323
Missing	1207 (55.6%)	1236 (62.7%)		443 (33.6%)	494 (44.9%)	
Alkaline phosphatase, relative to ULN						
Median [Min, Max]	0.731 [0.192, 27.1]	0.708 [0.123, 29.4]	0.185	0.738 [0.131, 63.5]	0.692 [0.200, 14.7]	<0.001
Missing	701 (32.3%)	698 (35.4%)		129 (9.8%)	125 (11.4%)	
Lactate dehydrogenase, relative to ULN						
Median [Min, Max]	0.862 [0.396, 6.68]	0.853 [0.431, 15.3]	0.783	0.938 [0.396, 19.9]	0.831 [0.253, 6.73]	0.005
Missing	1943 (89.5%)	1745 (88.5%)		1020 (77.4%)	880 (79.9%)	
Haemoglobin (g/L)						
Median [Min, Max]	125 [80.0, 176]	123 [80.0, 184]	0.0541	124 [61.0, 179]	125 [57.0, 168]	0.361
Missing	677 (31.2%)	658 (33.4%)		111 (8.4%)	116 (10.5%)	
ECOG						
0	487 (22.4%)	459 (23.3%)	0.93	375 (28.5%)	312 (28.3%)	0.866
1	573 (26.4%)	523 (26.5%)		511 (38.8%)	426 (38.7%)	
2	197 (9.1%)	185 (9.4%)		132 (10.0%)	125 (11.4%)	
3	-	-		25 (1.9%)	25 (2.3%)	
4	-	-		2 (0.2%)	2 (0.2%)	
Missing	915 (42.1%)	805 (40.8%)		272 (20.7%)	211 (19.2%)	
Previous treatment lines for mHSPC, nmCRPC and/or nmHSPC						
0	1913 (88.1%)	1725 (87.5%)	0.0421			
1	246 (11.3%)	244 (12.4%)				
> 1	13 (0.6%)	3 (0.2%)				
Previous treatment of docetaxel						
Yes	234 (10.8%)	232 (11.8%)	0.337			
No	1938 (89.2%)	1740 (88.2%)				
Previous other systematic treatments						
Yes	30 (1.4%)	16 (0.8%)	0.11			
No	2142 (98.6%)	1956 (99.2%)				

Gleason Score (initial PC diagnosis)			
Low/very low risk: ≤6	153 (7.0%)	124 (6.3%)	0.468
Intermediate risk: 7	391 (18.0%)	377 (19.1%)	
High/very high risk: 8-10	1050 (48.3%)	924 (46.9%)	
Unknown / Not documented	578 (26.6%)	547 (27.7%)	
Group Stage (initial PC diagnosis)			
I	16 (0.7%)	17 (0.9%)	0.993
II	132 (6.1%)	120 (6.1%)	
III	84 (3.9%)	76 (3.9%)	
IV	969 (44.6%)	884 (44.8%)	
Unknown / Not documented	971 (44.7%)	875 (44.4%)	
T Stage (initial PC diagnosis)			
T0/T1	178 (8.2%)	181 (9.2%)	0.693
T2	362 (16.7%)	306 (15.5%)	
T3	271 (12.5%)	257 (13.0%)	
T4	65 (3.0%)	51 (2.6%)	
TX	89 (4.1%)	86 (4.4%)	
Unknown / Not documented	1207 (55.6%)	1091 (55.3%)	
N Stage (initial PC diagnosis)			
N0	501 (23.1%)	432 (21.9%)	0.482
N1	266 (12.2%)	239 (12.1%)	
NX	140 (6.4%)	149 (7.6%)	
Unknown / Not documented	1265 (58.2%)	1152 (58.4%)	
M Stage (initial PC diagnosis)			
M0	807 (37.2%)	718 (36.4%)	0.951
M1	640 (29.5%)	588 (29.8%)	
M1a	15 (0.7%)	14 (0.7%)	
M1b	156 (7.2%)	157 (8.0%)	
M1c	59 (2.7%)	51 (2.6%)	
Unknown / Not documented	495 (22.8%)	444 (22.5%)	

Combined comorbidity score						
Mean (SD)	6.14 (0.708)	6.11 (0.623)	0.149	6.35 (1.02)	6.34 (0.993)	0.946
Median [Min, Max]	6.00 [5.00, 13.0]	6.00 [5.00, 11.0]		6.00 [5.00, 16.0]	6.00 [5.00, 14.0]	
Diabetes mellitus						
Yes	88 (4.1%)	103 (5.2%)	0.085	92 (7.0%)	87 (7.9%)	0.436
No	2084 (95.9%)	1869 (94.8%)		1225 (93.0%)	1014 (92.1%)	
Race						
Asian	29 (1.3%)	26 (1.3%)	0.43			
Black or African American	231 (10.6%)	201 (10.2%)				
White	1391 (64.0%)	1220 (61.9%)				
Other Race	330 (15.2%)	334 (16.9%)				
Unknown / Not documented	191 (8.8%)	191 (9.7%)				
SES quantiles						
1 - Lowest SES	295 (13.6%)	266 (13.5%)	0.707			
2	379 (17.4%)	343 (17.4%)				
3	417 (19.2%)	378 (19.2%)				
4	443 (20.4%)	423 (21.5%)				
5 - Highest SES	412 (19.0%)	340 (17.2%)				
Unknown / Not documented	226 (10.4%)	222 (11.3%)				
Duration since mCRPC diagnosis (days)						
Mean (SD)	94.8 (180)	106 (217)	0.0659			
Median [Min, Max]	31.0 [0, 1780]	31.0 [0, 2380]				
Duration since confirmation of metastasis (days)						
Mean (SD)	436 (451)	490 (496)	<0.001			
Median [Min, Max]	322 [0, 3320]	351 [0, 3000]				
Duration since initial prostate cancer diagnosis (days)						
Mean (SD)	2360 (2400)	2430 (2440)	0.328			
Median [Min, Max]	1350 [10.0, 13900]	1390 [36.0, 15400]				
Missing	32 (1.5%)	26 (1.3%)				

Duration since first-line treatment start (days) (i.e., time-to-next-line-treatment)						
Mean (SD)				343 (271)	398 (305)	<0.001
Median [Min, Max]				265 [27.0, 1430]	310 [27.0, 1430]	
Calendar year of first-line mCRPC treatment initiation						
2013	105 (4.8%)	18 (0.9%)	<0.001			
2014	240 (11.0%)	83 (4.2%)				
2015	233 (10.7%)	184 (9.3%)				
2016	236 (10.9%)	234 (11.9%)				
2017	341 (15.7%)	258 (13.1%)				
2018	302 (13.9%)	294 (14.9%)				
2019	254 (11.7%)	348 (17.6%)				
2020	204 (9.4%)	282 (14.3%)				
2021	195 (9.0%)	222 (11.3%)				
2022	62 (2.9%)	49 (2.5%)				

ECOG, Eastern Cooperative Oncology Group (performance status score); mCRPC, metastatic castration-resistant prostate cancer; mHSPC, metastatic hormone-sensitive prostate cancer; nmCRPC, non-metastatic castration-resistant prostate cancer; nmHSPC, non-metastatic hormone-sensitive prostate cancer; PC, prostate cancer; PSA, prostate- specific antigen; SD, standardised deviation; SES, socioeconomic status; ULN, upper limit of normal.

The calendar years in which patients started each treatment varied significantly, as abiraterone was available earlier as a first-line treatment for mCRPC, capturing a higher market share in the earlier years. Over time, the numbers of patients beginning either treatment have levelled out. The level of unknown variable information was generally similar across both groups. For the most critical factor, ECOG performance, initial missingness was around 40% for both groups, decreasing to 20% at treatment crossover. For all variables included in the weighting model, missingness generally remained below 45%, with PSA being the exception. However, the documentation of PSA, also improved over the period of follow-up—from 55-63% missing at Time Zero to 33-45% missing at the time of treatment crossover

8.4.2. Sankey diagram and censoring due to treatment strategy deviation

Figure 8.8 presents the Sankey diagram for the PC1 study population, where all patients initially received either abiraterone or enzalutamide. The observational period in the diagram extends beyond the 1440-day follow-up used in the survival analyses for the purpose of matching the survival curve period of the GUTG-001 trial (see Section 8.4.3), including all available data up to the data cut-off.

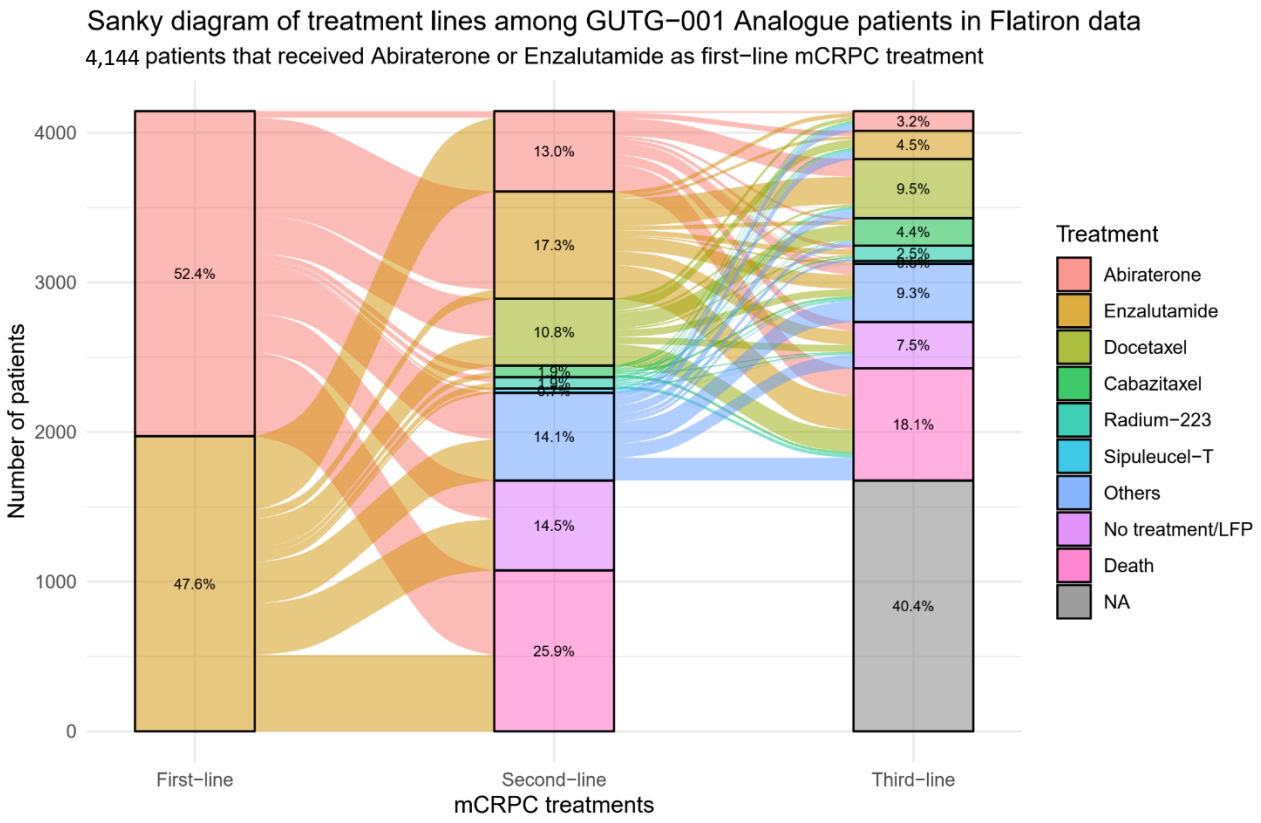


Figure 8.8 Sankey diagram for the GUTG-001 Analogue Target Trial using Flatiron data

LFP: lost follow-up

The diagram (Figure 8.8) shows that 40.4% of patients did not proceed to second-line treatments; a majority of these patients died without further treatments, while roughly a third of them (14.5% of the entire population) was lost to follow-up or remained on first-line treatment at data cut-off. These patients were considered to have adhered to (non-deviated from) their assigned treatment sequence throughout the study. Apart from patients who did not receive subsequent treatments, the predominant treatment sequences for first- and second-line treatments were abiraterone → enzalutamide and enzalutamide →abiraterone, with or without further treatments.

Other prevalent sequences involve transitions from first-line abiraterone or enzalutamide to docetaxel or various less common treatment regimens categorised as "Others" in the graph. Dr. Pezaro highlighted the unexpectedly high percentage of patients receiving "Others" as subsequent treatments, but noted that this makes sense when considering that it represents any potential combination of treatments or atypical treatments that varied across practices. For example, this group may include combinations of treatments commonly used as subsequent treatments, such as abiraterone, enzalutamide, or docetaxel paired with Radium-223. In the context of the PC1 case study, this labelling of treatment sequences was deemed sufficient for applying the intended statistical methods, as any deviation from predefined second-line treatments was considered a deviation. This resulted in the following censoring and outcome event labels in PP analyses, with or without confounding adjustments, over a four-year follow-up period:

- Censoring due to treatment deviation:
  - Abiraterone first-line group: 672 patients (31%)
  - Enzalutamide first-line group: 614 patients (31%)
- Administrative censoring:
  - Abiraterone first-line group: 615 patients (28%)
  - Enzalutamide first-line group: 612 patients (31%)
- Deaths
  - Abiraterone first-line group: 885 (41%)
  - Enzalutamide first-line group: 746 (38%)

Potentially, due to the similarities between abiraterone and enzalutamide, and thus the theoretically almost exchangeable sequences, the censoring and death outcome distribution were fairly similar in the two groups. However, comparisons, such as between abiraterone → enzalutamide versus abiraterone → docetaxel, may exhibit different censoring and outcome distributions in the analyses.

8.4.3. IPW derivation

Figure 8.9 illustrates the distribution of the baseline inverse IPTW weights derived from the procedures discussed in Section 8.3.4.3.2. Figure 8.9A displays the distribution of the propensity for being treated within either group's first-line treatment (abiraterone versus enzalutamide), contributing to the denominator in Formula 8.1 and 8.2. The x-axis displays probabilities ranging from 0 to 1, while the y-axis shows the count of patients with each probability value in a treatment group. There is a good overlap between two treatment groups, with a peak at approximately 0.5 for both groups.

This aligns with the expectation that these treatments are similar in mechanisms and are largely interchangeable to some extent. Sensitivity analyses that excluded the variable calendar year from the weight derivation model, showed an even higher degree of overlap (i.e., a larger shaded purple area, with less area shaded solely in red or blue). Conceptually, this is expected, as one significant difference between the two treatment groups was the earlier prevalence of abiraterone use as first-line mCRPC treatment. Figures 8.9B and 8.9C show the non-stabilised and stabilised baseline IPTW weights, respectively. The x-axis displays weight ranges, while the y-axis indicates the number of patients with each weight value in a treatment group. The non-stabilised weights generally remain below 2.5, with a few exceptions reaching up to approximately 11, while the stabilised weights have a peak at 1 exhibit a narrower range of weights. The stabilised weights in Figure 8.9 were used in the baseline adjusted analyses specified in Section 8.3.4.4.

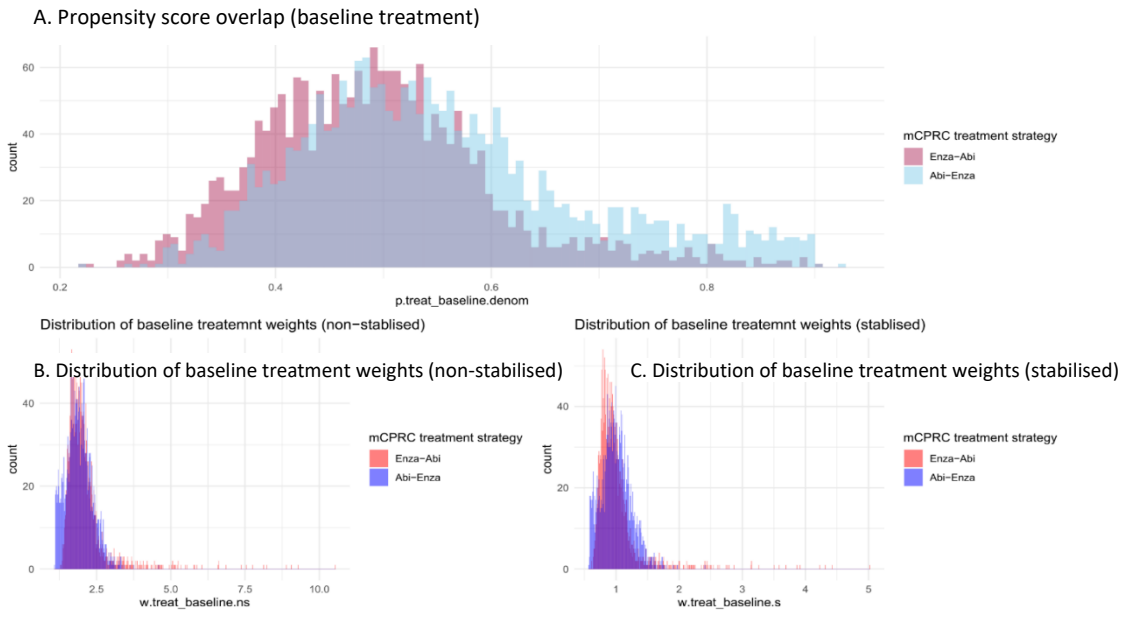


Figure 8.9 Baseline inverse probability of treatment weights (IPTW)

Abi, abiraterone; Enza, enzalutamide; mCRPC, metastatic castration-resistant prostate cancer

The denominator of the stabilised weight for IPTW at baseline (Section 8.3.4.3.2) represents the unconditional probability of receiving a treatment and was not plotted, as it reflects the proportion of

patients receiving each treatment. Unstabilised and stabilised weights are shown at a single time point for IPTW (Figure 8.9B &8.9C), while both the numerator and denominator for time-varying $IPCW_{txdev}$ weights change over time, requiring more figures to be reported (Figure 8.10–8.12).

Figure 8.10 presents hexbin graphs of the estimated time-varying probabilities of treatment strategy deviation across all groups, as detailed in Section 8.3.4.3.3. Figure 8.10A presents the probabilities of treatment strategy deviation given both baseline and time-varying covariates for each treatment group, contributing to the denominators in Formulas 8.3 and 8.4. Each hexagon's colour in this figure represents a cluster of patients, with darker blue indicating a denser cluster. The x-axis represents time points spanning four years, while the y-axis shows the distribution of treatment deviation probabilities for all patients that remained alive at each respective time point.

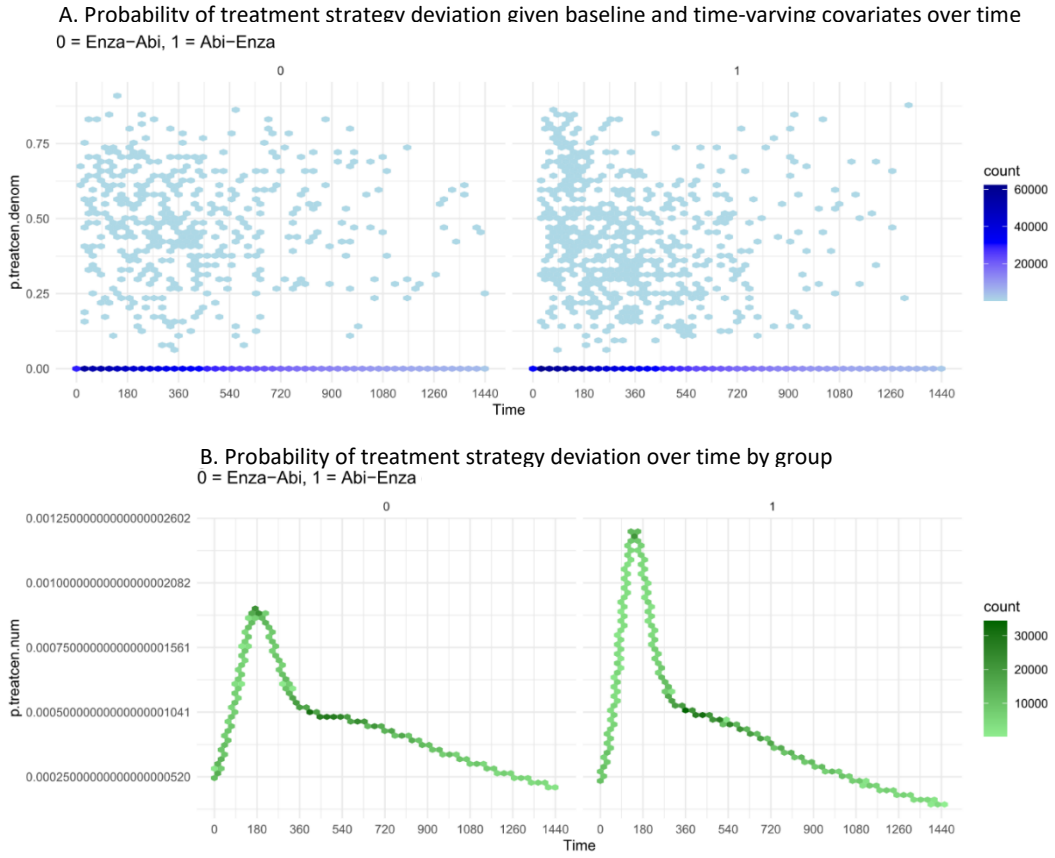


Figure 8.10 Hexbin graphs of time-varying probabilities of treatment strategy deviation

Abi, abiraterone; Enza, enzalutamideDue to the computational intensity of plotting individual points for all patients (around 2,000 patients per group, each with data at 1,440 time points), hexbin graphs are used to cluster points. Each hexbin represents the x, y location of multiple points that share similar positions. In these graphs, x represents time points, while y represents probabilities. In Figure 8.10A, the y-axis shows the probability of following the assigned treatment strategy given baseline and time-varying characteristics. In Figure 8.10B, the y-axis shows the marginal probability of treatment strategy deviation over time. Darker hexagons indicate a higher density of points within the same hexbin.

In Figure 8.10A, the dark line of hexbins from time 0 to 1,440 at y = 0 for both groups (abiraterone → enzalutamide and enzalutamide → abiraterone) indicates that most patients adhered to their assigned treatment strategy throughout, with a low probability of deviation (i.e., close to 0). Since each patient could only deviate from the treatment strategy at a single point (i.e., at the initiation of second-line treatment), higher probabilities of treatment deviation (represented by lighter hexbins above y = 0) are spread across the 1,440 days rather than occurring at the same time. These hexbins scattered above y > 0 over time represent patient-time points with an increased probability of deviating from the treatment strategy.

In Figure 8.10B, a peak of hexbins in the probability of treatment deviation occurs around six months, indicating that patients starting second-line treatment at this time are more likely to receive a non-assigned second-line treatment (i.e., deviate from the assigned strategy). A cluster of darker hexbins appears around the one-year mark, corresponding to the median progression-free survival on first-line treatment for both groups, where patients seem to be less likely to deviate from the assigned strategy (i.e., smaller y) compared to those starting second-line treatment at six months.

Throughout the follow-up period, most patients are likely to remain adherent to their treatment strategy (shown by darker clusters near zero at all-time points), as compared to those with higher probabilities of treatment strategy deviation (lighter blue clusters above zero). This was expected, as per the study design, patients were considered to have deviated from the treatment protocol only when they received non-protocol second-line treatment. Hence, each patient likely experienced only a brief period of increased likelihood of treatment strategy deviation given their time-varying characteristics, while most of the time, they were likely to remain adherent to their assigned treatment strategy. Most of the non-zero probabilities of treatment strategy deviation appeared within the first 2 years of follow-up for both groups. Additionally, in both groups, the count of patients decreased over time, as depicted by the overall fading darkness of blue on the graph, mainly due to censoring or death.

Figure 8.10B shows the unconditioned probability of treatment strategy deviations over time, contributing to the numerator in Formula 8.4. As this probability was expected to be the same for all patient at a given time point, the graph appears to look like a single curve. The peak of this curve, conceptually, indicates the period when patients in each group were most likely to deviate from their assigned treatment sequence, with the highest probability occurring around 6 months for both groups.

Figure 8.11A displays hexbin graphs of the non-stabilised time-varying inverse probability of censoring weights ($IPCW_{txdev}$) for treatment strategy deviations, informed by the probabilities shown in Figure 8.10A. The figure is similar to Figure 10A, but with the y-axis now indicating the weights' values. It was expected that at each time point, most patients' weights would remain close to zero, indicating patients at most time points tend to stay adherent to their treatment sequence. However, if any patients deviated from their treatment strategy, conceptually, patients with similar characteristics at that time point (e.g., likely just starting second-line treatment but adhering to the protocol) would be weighted to represent the records of those who had been censored. Higher-value weights were more likely to appear in the early follow-up before 6 months, or in the later periods after 2 years. This was also expected because fewer patients started receiving second-line treatment during these times, whether they deviated from the treatment sequence strategy or not. Consequently, higher weights were assigned to the smaller number of patients who began second-line treatment around the same time but remained adherent to the protocol, resulting in these patients receiving higher weights due to their limited numbers.

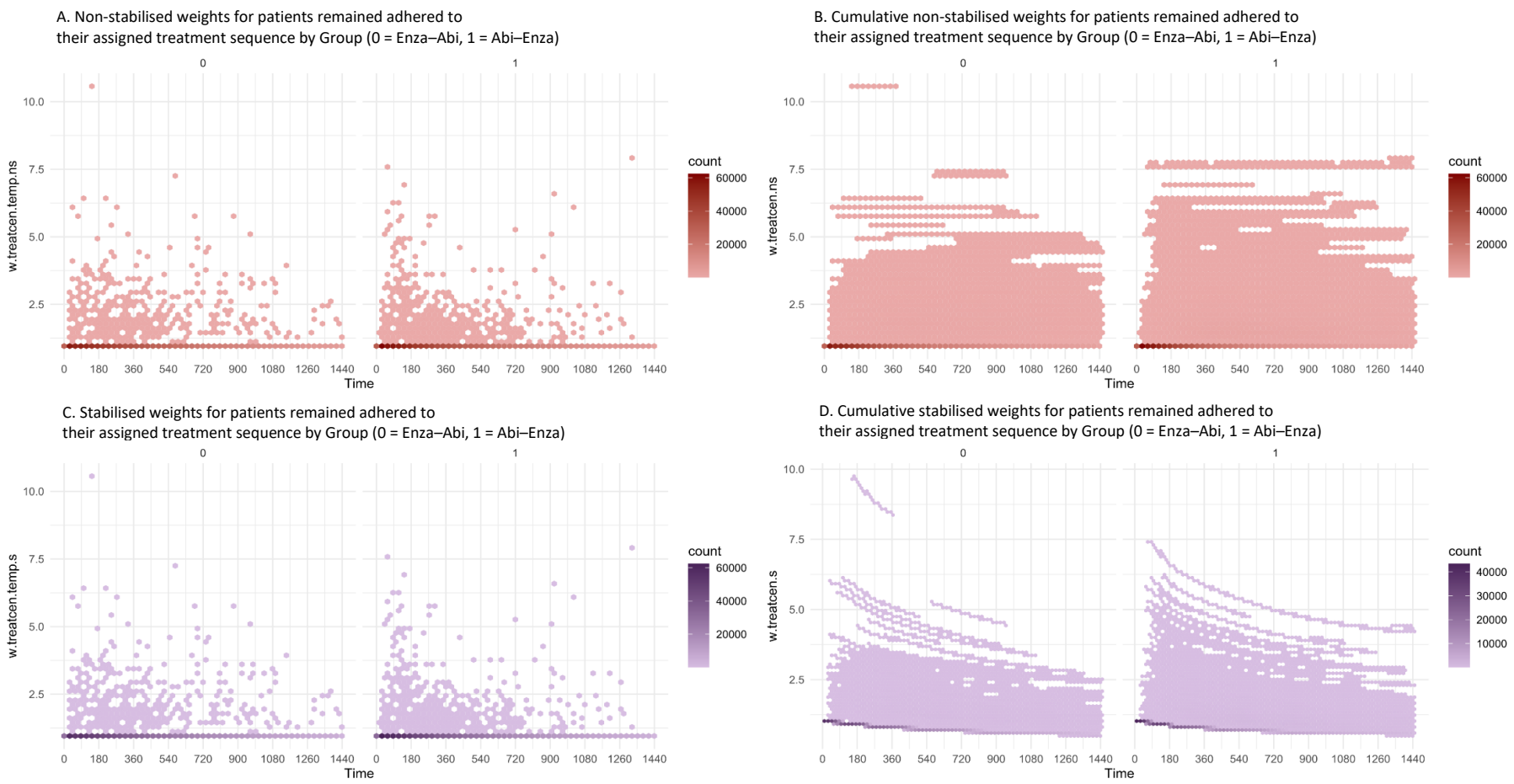


Figure 8.11 Hexbin graphs of time-varying inverse probability of censoring weights for treatment strategy deviations ($IPCW_{txdev}$)

Abi, abiraterone; Enza, enzalutamide

Due to the computational intensity of plotting individual points for all patients (around 2,000 patients per group, each with data at 1,440 time points), hexbin graphs are used to cluster points. Each hexbin represents the x, y location of multiple points that share similar positions. In these graphs, x represents time points, while y represents weights. Figure 8.11A shows the non-stabilised time-varying IPCWtxdev weights for patients who adhered to their assigned treatment sequence, essentially the inverse of one minus the values shown in Figure 8.10A (i.e., one minus the probability of deviating from the assigned treatment sequence given baseline and time-varying characteristics). Figure 8.11B displays the cumulative weights from Figure 8.10A over time. Figure 8.11C presents the stabilised weights for patients who remained on their assigned treatment sequence. These weights are calculated by multiplying the values in Figure 8.11A by a numerator, which is derived by subtracting the values in Figure 8.10B (i.e., the probability of deviating from the assigned treatment sequence given time) from one. Figure 8.11D shows the cumulative weights from Figure 8.11C over time. Cumulative weights (Figures 8.11B & 8.11D) are necessary for IPCWtxdev, as the probability of adhering to a treatment strategy depends on having adhered in the previous time intervals, requiring multiplication over time.

This weighting process is further explained in Figure 8.11B, which shows the cumulative non-stabilised $IPCW_{txdev}$ weights. Each horizontal line roughly represents one patient or a cluster of patients. That is, those that had been upweighted were likely to maintain similar weights in subsequent follow-up times, representing the forgone subsequent records of others with similar characteristics who had been censored. The rationale for these weights not likely increasing further per patient was that those who had been weighted were typically those who had already started their second-line treatment. Conceptually, their subsequent records should not have been weighted to represent those about to be censored or newly starting second-line treatment. Each line in Figure 11.B stops at the point where the corresponding patient likely either died or was censored. These conceptual explanations of the graphs helped examine the theoretical correctness of the weight derivation models.

Figures 8.11C & 11D display the stabilised version of $IPCW_{txdev}$ weights (Formula 8.4.) of those in Figures 8.11A & 8.11B (Formula 8.3). The major visual difference is that each patient who has been upweighted shows a decrease in cumulative stabilised $IPCW_{txdev}$ weights over time (Figure 11D). This was expected because the numerator in the stabilised weights accounted for the gradually decreasing unconditional probability of treatment deviation over time (Figure 8.10B). Overall, the stabilised weights exhibit a less extreme weight distribution.

Figure 8.12 shows the hexbin graphs of the overall time-varying IPW for advanced confounding adjustment, as required for the analyses in Section 8.3.4.5. Figure 8.12A shows the non-stabilised weights and Figure 8.12B the stabilised weights, calculated using Formulas 8.5 and 8.6, respectively. Specifically, Figure 8.12A is the result of multiplying the weights from Figure 8.9B with those from Figure 8.11C, while Figure 8.12B was created by multiplying the weights in Figure 8.9C with those from Figure 8.11D. The baseline IPTW stays consistent over time, and consequently, the overall pattern of the weights in Figure 8.12 reflects the trends seen in the cumulative $IPCW_{txdev}$ weights from Figures 8.11B and 8.11D. In Figure 8.12, I configured each hexagon to represent a smaller group of patients, resulting in finer and more distinct separation per smaller groups of patients compared to previous graphs (Figure 10-11), whereas earlier settings were designed to highlight clusters of weights with distinct colour shading. The previous settings were necessary because each timepoint should contain, at most, the same number of sample size dots as the study population (n = 4,144). With up to 1440 time points on the graph, presenting individual dots would be computationally exhaustive and provide less information, such as not being able to distinctly identify where the extreme weights lie due to small clusters.

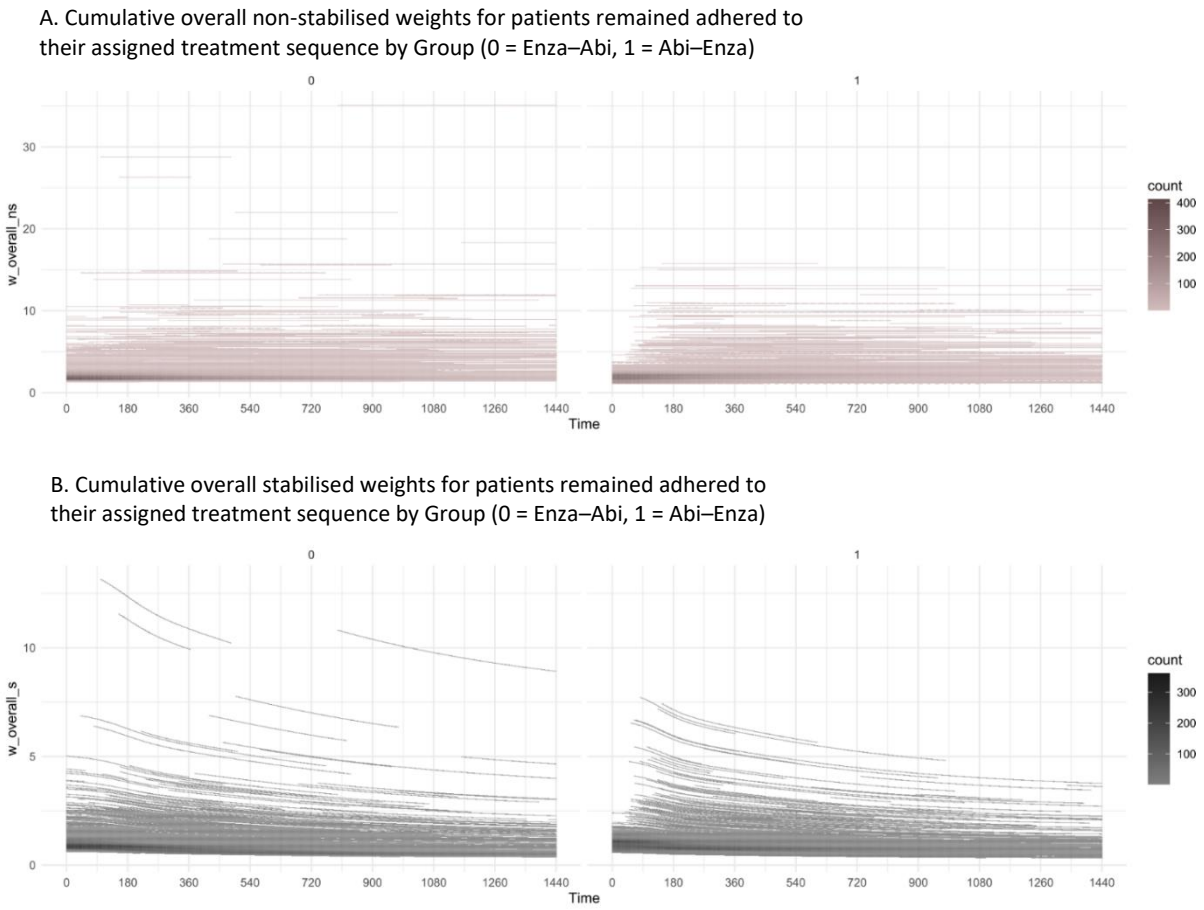


Figure 8.12 Hexbin graphs of time-varying overall inverse probabilities weights for advanced confounding adjustment

Abi, abiraterone; Enza, enzalutamide

Figure 8.12A represents the non-stabilised weight of IPTW*IPCWtxdev, which is the product of the non-stabilised IPTW weight (Figure 8.9B) and the cumulative non-stabilised time-varying IPCWtxdev weight at each time point in Figure 8.11B. Figure 8.12B represents the stabilised weight of IPTW*IPCWtxdev, which is the product of the stabilised IPTW weight (Figure 8.9C) and the cumulative stabilised time-varying IPCWtxdev weight at each time point in Figure 8.11D.

8.4.4. Survival analyses

Table 8.10 presents the results of all survival analyses, including:

- Four analyses using simple methods without adjustment for confounding (Analysis 1a-1d)
- Four analyses using simple methods whilst adjusted for baseline confounding with IPTW (Analysis 2a-2d)
- The final analysis (Analysis 3), aiming to closely emulate the GUTG-001 trial results by adjusting for baseline and time-varying confounding with IPTW*IPCW$_{txdev}$.

Table 8.10 GUTG-001 Analogue Target Trial Emulation results comparison table

Models (population)	Sources of bias	Median follow-up time (months, IQR)	Median OS (months, 95 CI%), using Kaplan-Meier estimator†		Hazard ratio (95 CI%), using Cox models¶	Cox model performance		
			ABI-ENZA (n = 2,172)	ENZA-ABI (n = 1,972)		AIC	BIC	Log likelihood
Cox model summary statistics from Khalaf 2019 Study (ITT, Khalaf 2019[84])	-	30.7 (25.1-36.2)	28.8 (25.4, not reached)	24.7 (18.8, 34.0)	0.79 (0.54, 1.16)	-	-	-
0. Cox model with digitally reconstructed survival data from Khakaf 2019 (ITT, Khalaf 2019[84])	-	22.2 (14.5-31.1)	29.0 (25.5, not reached)	24.8 (19.9, 32.5)	0.78 (0.54, 1.15)	1034.59	1037.26	-516.29
1a. Unadjusted ITT	Confounded by baseline and time-varying covariates (treatment regimen protocol violation)	21.4 (11.6-35.4)	25.4 (24.0, 26.9)	28.8 (27.4, 30.7)	1.14 (1.05, 1.23)	37887.01	37892.81	-19824.54 (df=1)
1b. Unadjusted PP	Confounded by baseline and time-varying covariates (informative censoring)	15.4 (7.1-28.9)	31.0 (29.5, 33.1)	35.3 (32.5, 38.6)	1.17 (1.06, 1.29)	24672.21	24677.61	-12335.11 (df=1)
1c. Unadjusted AT	Confounded by baseline and time-varying covariates (due to conditioning on post-treatment covariates)	19.8 (9.5-34.2)	25.2 (23.7, 28.0) (n = 1, 500)	28.5 (26.9, 31.2) (n = 1, 358)	1.14 (1.03, 1.25)	24011.29	24016.68	-12004.61 (df=1)
1d. Unadjusted AT, limited to second-line treatment recipients	Confounded by baseline and time-varying covariates (due to conditioning on post-treatment covariates), and immortal time bias	27.1 (17.6-39.1)	31.9 (30.7, 34.7) (n = 645)	29.9 (27.7, 32.9) (n = 487)	0.92 (0.80, 1.07)	9062.31	9066.87	-4530.156 (df=1)
2a. Baseline IPTW weighted ITT	Confounded by time-varying covariates (treatment regimen protocol violation)	20.6 (10.8-33.8)	25.9 (24.3, 27.8)	28.5 (27.0, 30.6)	1.11 (1.02, 1.20)	37929.09	37934.89	-18963.54 (df=1)
2b. Baseline IPTW weighted PP	Confounded by time-varying covariates (informative censoring)	14.8 (6.8-28.0)	31.7 (29.9, 33.9)	35.0 (32.5, 38.6)	1.16 (1.04, 1.28)	24505.77	24511.17	-12251.89 (df=1)
2c. Baseline IPTW weighted AT	Confounded by time-varying covariates (due to conditioning on post-treatment covariates)	19.0 (9.0-32.4)	25.4 (23.7, 28.4) (n = 1, 500)	28.9 (27.4, 32.3) (n = 1,358)	1.14 (1.03, 1.27)	23907.89	23913.29	-11952.94 (df=1)
2d. Baseline IPTW weighted AT, limited to second-line treatment recipients	Confounded by time-varying covariates (due to conditioning on post-treatment covariates), and immortal time bias	25.4 (15.6-36.6)	31.9 (30.4, 34.8) (n = 645)	30.6 (28.0, 34.5) (n = 487)	0.97 (0.82, 1.14)	8951.784	8956.341	-4474.892 (df=1)
3. Baseline IPTW and time-varying $IPCW_{txdev}$ weighted PP (hypothetical PP effect with perfect compliance to the treatment sequence assignment)	Possible residual confounding due to unmeasured confounders	-	28.7 (25.6, 31.2)	28.9 (27.5, 32.4)	1.07 (0.96, 1.20)	26697.83	26703.23	-13347.92 (df=1)

ABI: abiraterone; AIC: Akaike Information Criterion; CI: confidence interval; ENZA: enzalutamide; AT: as-treated, BIC: Bayesian Information Criterion; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; IQR: interquartile Range; ITT: intention-to-treat; KM: Kaplan-Meier; OS: overall survival; PP: per-protocol, 4-year is defining as 1440 post Time Zero

* cubic spline for time

† All 95% CIs for median overall survival presented were directly obtained from the survfit function in R (i.e., naïve (unadjusted) 95% CI), including those for weighted analyses.

¶ 95% CIs for HRs from unadjusted analyses (Analysis 0 to 1d) were directly obtained using the coxph function in R. For weighted analyses (Analysis 2a to 3), 95% CIs were derived from the robust sandwich estimator

The table details HRs and median OS for each group (i.e., starting with first-line abiraterone versus enzalutamide) after a four-year follow-up period. It also summarises sources of bias for each analysis as explained in Sections 8.3.4.2 to 8.3.4.5. Additionally, the table reproduced the GUTG-001 trial findings from the Khalaf et al. study at the top.[84] While unnumbered entries were directly taken from their reported figures, Analysis 0 presents results from reanalysing the digitised, reconstructed patient-level data from the reported OS graphs. Analysis 0 demonstrates high concordance with the original findings reported in the Khalaf et al. paper, despite minimal variations potentially due to digitisation variance where a high number of censoring events occured in later follow-up periods. The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the log-likelihood of each Cox model were also reported in Table 8.10.

Figure 8.13 illustrates the HR comparisons of all methods (in black) against the benchmark GUTG-001 estimates (in red) in a forest plot[84], while Figure 8.14 visualises the comparison of absolute median OS for the abiraterone-first (Figure 8.14A) and enzalutamide-first groups (Figure 8.14B) against the benchmark. These plots not only visualise the point estimate comparisons, but also the 95% CIs. Additionally, the size of each point estimate indicates the sample size of each analysis. AT analyses (Analyses 1c-1d and 2c-2d) feature smaller sample sizes due to the exclusion of patients who deviated from their assigned treatment sequences. In contrast, all ITT (Analyses 1a and 2a) and PP analyses (Analyses 1b, 2b, and 3) maintained the same population as at enrolment, with no patient exclusions, although record censoring may have been involved.

For the HRs (Figure 8.13), only the AT analyses—specifically Analyses 1d and 2d, which were restricted to patients who survived to receive second-line treatment—, and the final adjusted analyses using the advanced IPTW*IPCW$_{txdev}$ method (Analysis 3), showed a non-significant difference between the two treatment sequence groups as seen in the benchmark GUTG-001 trial.[84] All other analyses, regardless of whether they adjusted for baseline confounding, significantly favoured the sequence of enzalutamide → abiraterone. Although the HR from the IPTW*IPCW$_{txdev}$ adjusted analysis (1.07, 95% CI 0.96-1.20) was closer to those observed in the GUTG-001 trial (compared with ITT and PP analyses), the HR point estimate still did favour the sequence of abiraterone → enzalutamide as seen in the GUTG-001 trial (0.79, 95% CI 0.54-1.16). Meanwhile, AT Analyses 1d and 2d not only reflected the non-significant HR observed in the GUTG-001 trial, but their HR point estimates also favoured the sequence of abiraterone followed by enzalutamide. However, these AT analyses are inherently flawed due to immortal time biases (Table 8.10), and this is evident from the comparison of absolute outcomes, specifically median OS comparisons in Figure 8.14.

For the median OS of the abiraterone → enzalutamide group, the IPTW*IPCW$_{txdev}$ adjusted analysis (Figure 8.14A: 28.7 months, 95% CI 25.6-31.2) yielded a point estimate closest to those from the GUTG-001 trial, compared to all other methods. Although all methods showed 95% CIs that

partially overlap with those from the Khalaf trial, their point estimates were consistently more distant. Only Analysis 3's 95% CI completely fell within the 95% CI of the Khalaf trial (28.8 months, 95%CI 25.4-not reached). While the 95% CI for the OS median of the abiraterone → enzalutamide group was wide due to immature OS data, the CIs from all the other analyses did not completely fall within with this range. For the median OS of the enzalutamide → abiraterone group, the IPTW*IPCW$_{txdev}$ adjusted analyses also yielded 95% CIs (Figure 8.14A: 28.9 months, 95% CI 27.5-32.4) that completely fell within those in the GUTG-001 trial (24.7 months, 95%CI 18.8-34.0). However, its point estimate was roughly 4 months higher than that of the GUTG-001 trial. Several other analyses demonstrated overlapping 95% CIs for the median OS of the enzalutamide → abiraterone group. However, none of them fully fell within the 95% CI of the GUTG001 study.

Since abiraterone and enzalutamide are very similar, all paired analyses—whether unadjusted or baseline-only adjusted (e.g., Analyses 1a & 2a, 1b & 2b, 1c & 2c, 1d & 2d)—consistently showed very similar results (Table 8.10). PP analyses, which artificially excluded death information for some patients, tend to inflate median OS estimates (Figure 8.14). AT analyses that only excluded patients who received non-protocol subsequent treatments (Analyses 1c-2c) tend to over-represent those who died early without further treatment, consequently resulting in a shorter OS in all groups (Figure 8.14). This results from selection bias based on post-enrolment information, as detailed in Section 8.3.4.2. On the other hand, AT analyses limited to patients who survived long enough to receive second-line treatment (Analyses 1d-2d) likely underrepresent those unable to receive further treatment, resulting in an overestimation of OS (Figure 8.13) due to immortal time bias, as noted in in Section 8.3.4.2. The advanced IPTW*IPCW$_{txdev}$ method theoretically corrects the baseline confounding and informative censoring bias in the PP analysis (Analysis 1b). It showed improved alignment of both the comparative effectiveness (i.e., HRs) and absolute outcome estimates (i.e., median OS) for comparing treatment sequences when benchmarked against the GUTG-001 trial, compared to all other analyses using simpler methods.

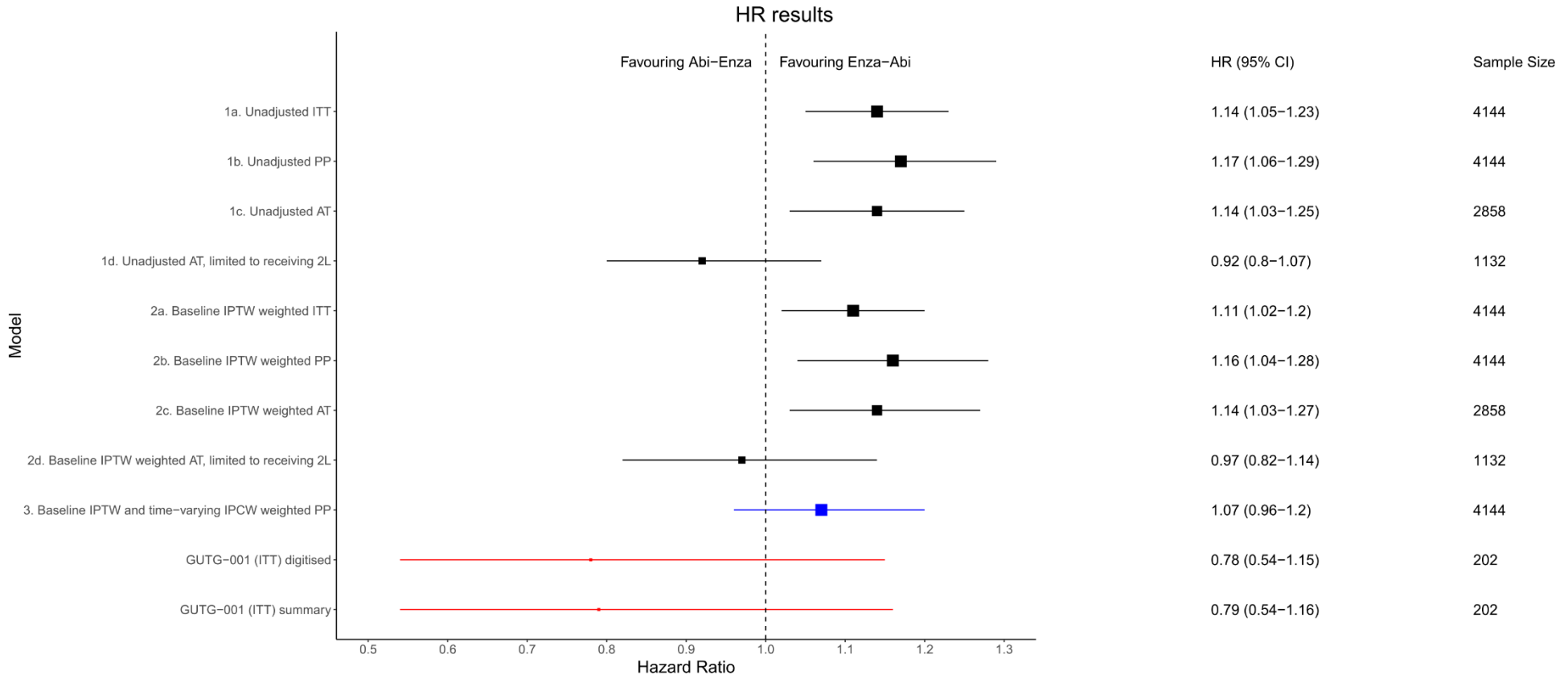
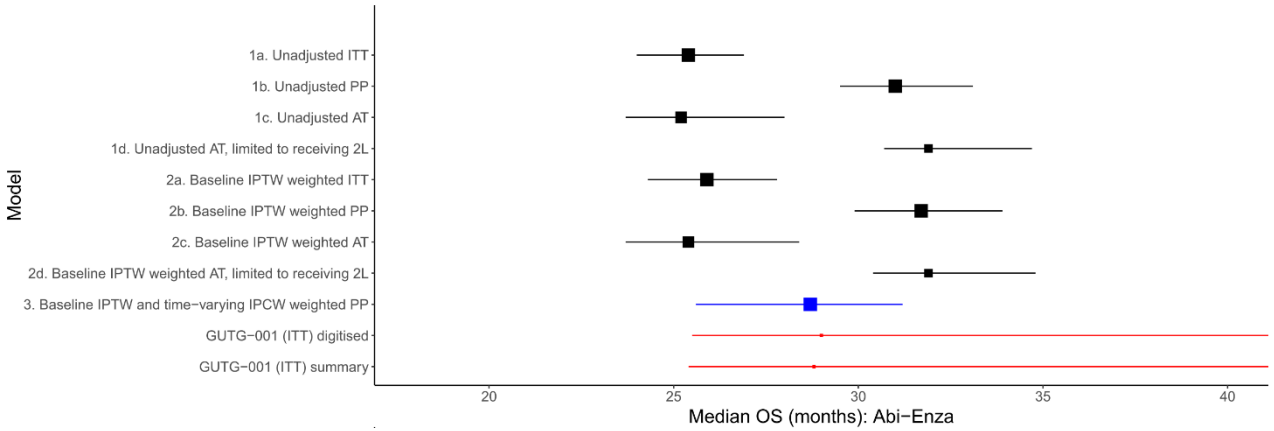


Figure 8.13 Comparison of hazard ratios against the GUTG-001 trial

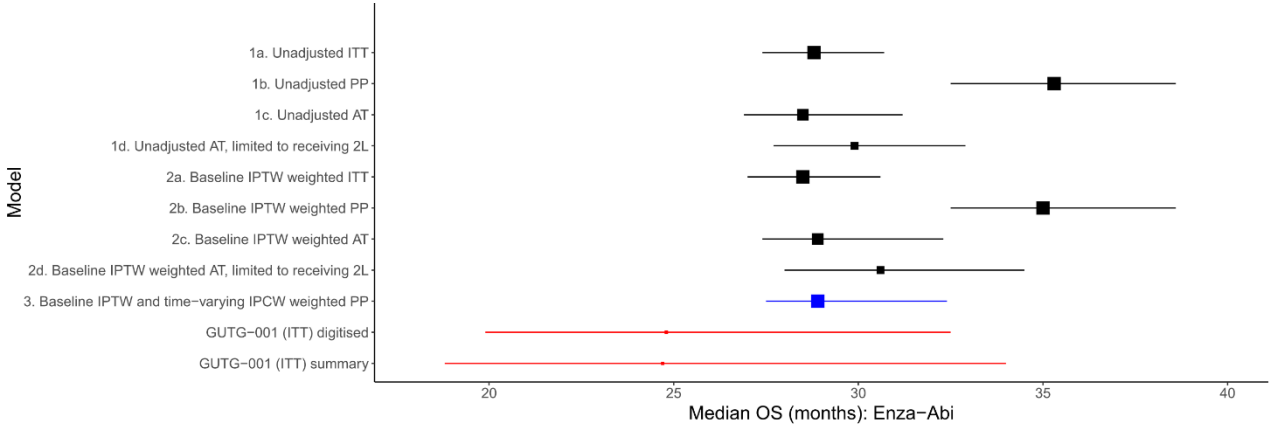
2L, second-line; Abi, abiraterone; AT, as-treated; Enza, enzalutamide; HR, hazard ratio; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; ITT: intention to treat; mCRPC, metastatic castration-resistant; PP, per-protocol

A. Absolute median overall survival for the abiraterone-enzalutamide group (Figure 8.14A)



Median OS (Abi−Enza)	Sample Size
25.4 (24.0, 26.9)	2172
31.0 (29.5, 33.1)	2172
25.2 (23.7, 28.0)	1500
31.9 (30.7, 34.7)	645
25.9 (24.3, 27.8)	2172
31.7 (29.9, 33.9)	2172
25.4 (23.7, 28.4)	1500
31.9 (30.4, 34.8)	645
28.7 (25.6, 31.2)	2172
29.0 (25.5, not reached)	101
28.8 (25.4, not reached)	101

B. Absolute median overall survival for the enzalutamide-abiraterone group (Figure 8.14B)



Median OS (Enza−Abi)	Sample Size
28.8 (27.4, 30.7)	1972
35.3 (32.5, 38.6)	1972
28.5 (26.9, 31.2)	1358
29.9 (27.7, 32.9)	487
28.5 (27.0, 30.6)	1972
35.0 (32.5, 38.6)	1972
28.9 (27.4, 32.3)	1358
30.6 (28.0, 34.5)	487
28.9 (27.5, 32.4)	1972
24.8 (19.9, 32.5)	101
24.7 (18.8, 34.0)	101

Figure 8.14 Comparison of median overall survival against the GUTG-001 trial

2L, second-line; Abi, abiraterone; AT, as-treated; Enza, enzalutamide; HR, hazard ratio; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; ITT: intention to treat; mCRPC, metastatic castration-resistant prostate cancer; PP, per-protocol

Figures 8.15 to 8.18 demonstrate the "evolution" of KM curves from an unadjusted RWD analysis to the final adjusted analysis incorporating advanced confounding adjustment using IPTW*IPCW$_{txdev}$. This series includes analyses for ITT (Analysis 1a), PP (Analysis 2a), PP adjusted for baseline (Analysis 2b), and PP adjusted for both baseline and time-varying confounding (Analysis 3). Initially, the ITT analysis shown in Figure 8.15 illustrates a scenario in RWD studies comparing first-line treatments, where two groups of patients may not be comparable, and the impact of different subsequent treatments on OS is not factored in. The median OS for both groups in unadjusted ITT ranges between 720-900 days. Then, the unadjusted PP analysis aimed to eliminate the contaminated effects of non-protocol treatments for assessing the impact of treatment sequences (Figure 8.16). This results in a higher median OS seen compared to the ITT analysis (Figure 8.15). Compared to the unadjusted ITT analysis, the unadjusted PP analysis resulted in KM curves becoming closer together before 900 days but beginning to diverge more after 1080 days. In the unadjusted PP analysis, the OS for both groups is extended to over 900 days. The baseline-adjusted PP analysis led to KM curves of both groups becoming closer (Figure 8.17) than those without baseline adjustments (Figure 8.16), showing greater overlap between the 95% CIs of KM curves of both groups. The median OS of both groups remained over 900 days in this analysis.

Finally, the PP analysis adjusted with the advanced IPTW*IPCW$_{txdev}$ method, as shown in Figure 8.18, reduced the median OS of both groups to just below 900 days. Importantly, the gap in median OS between the groups narrowed significantly, and the 95% CI curves show much greater overlap. Notably, the point estimate where the abiraterone → enzalutamide sequence surpasses that of the enzalutamide → abiraterone occurred around 900 days. While earlier analyses using simpler methods also displayed tangled survival curves between the two groups before 180 days (Figures 8.15-8.17), it was only with the adjustments using advanced methods that further potential tangled survival curve point estimates emerged after 800 days (Figure 8.18).

Figure 8.19 displays how the KM curves from my survival analysis using the IPTW*IPCW$_{txdev}$ method compared with those from the benchmark GUTG-001 trial. Figures 8.20 and 8.21 separate the graphs into two treatment groups: abiraterone → enzalutamide and enzalutamide → abiraterone, respectively. The survival point estimates from my analysis mostly fell within the 95% CIs of the GUTG-001 trial, with an exception of those within the first 3 months of follow-up in the abiraterone → enzalutamide group. The 95% CI curves in my analysis are narrower, likely due to larger sample sizes, and mostly fell within those of the GUTG-001 trial. For both groups, the point estimates tangled with those from the GUTG-001 trial before 540 days. After 540 days, the survival point estimate for the abiraterone → enzalutamide group is slightly lower but still close to that of the GUTG-001 trial. In contrast, the enzalutamide → abiraterone group shows higher point estimates after 540 days compared to the GUTG-001 trial, resulting in the median OS about 4 months longer (Table 8.10).

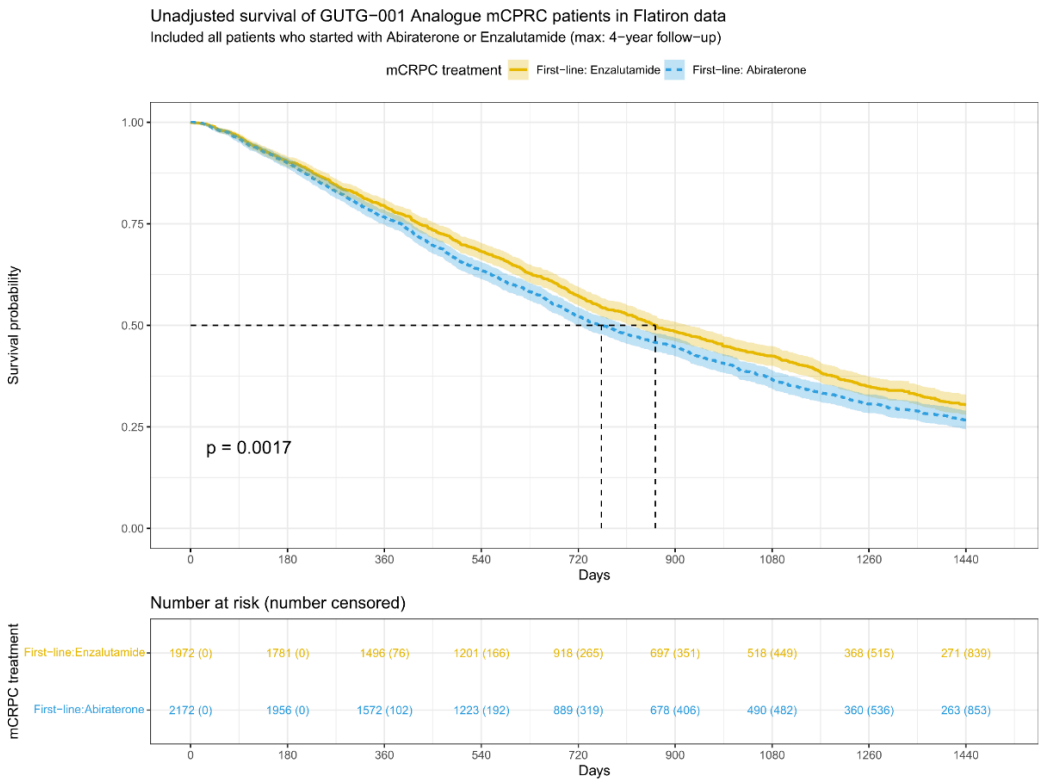


Figure 8.15 Kaplan-Meier curve of intention-to-treat analysis

mCRPC, metastatic castration-resistant prostate cancer

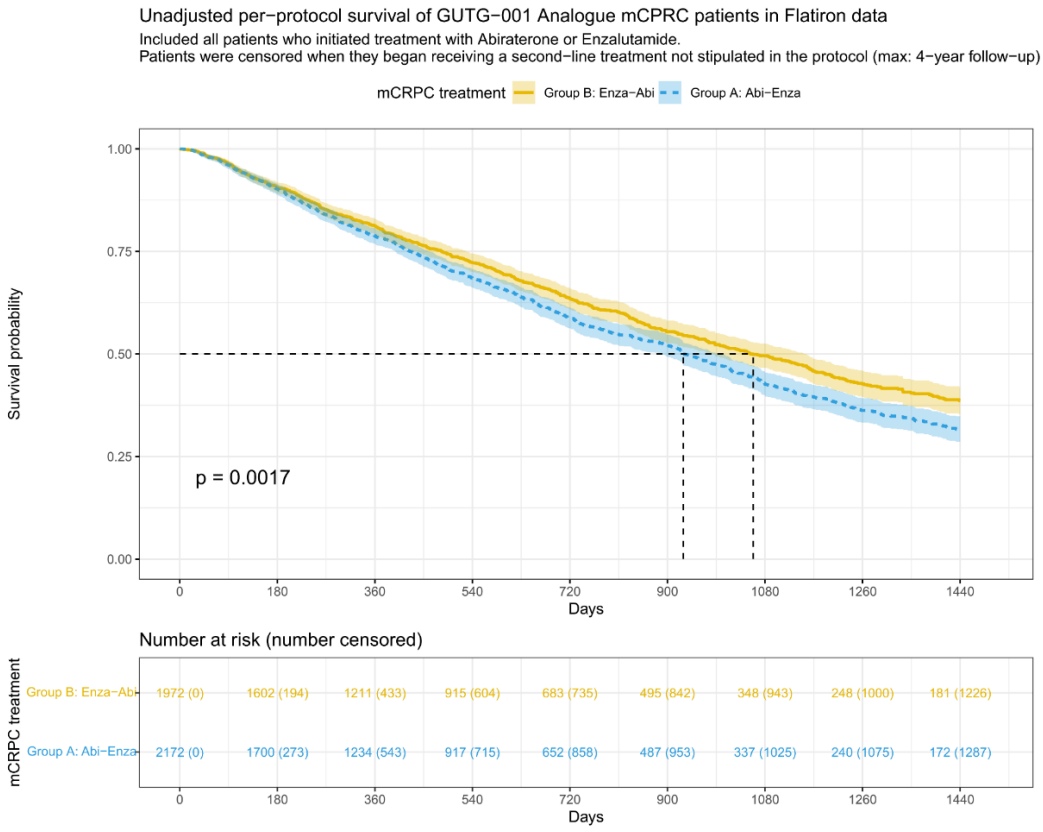


Figure 8.16 Kaplan-Meier curve of per-protocol analysis

mCRPC, metastatic castration-resistant prostate cancer

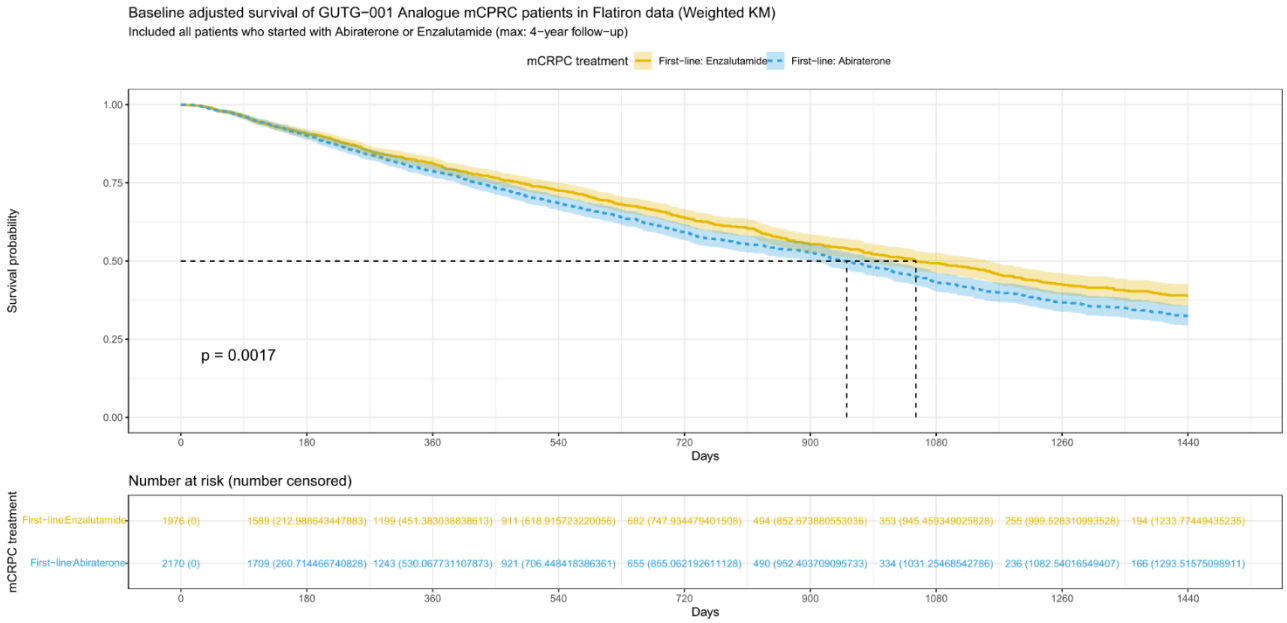


Figure 8.17 Kaplan-Meier curve of per-protocol analysis, adjusted for baseline confounding with inverse probability of treatment weights

KM, Kaplan-Meier; mCRPC, metastatic castration-resistant prostate cancer

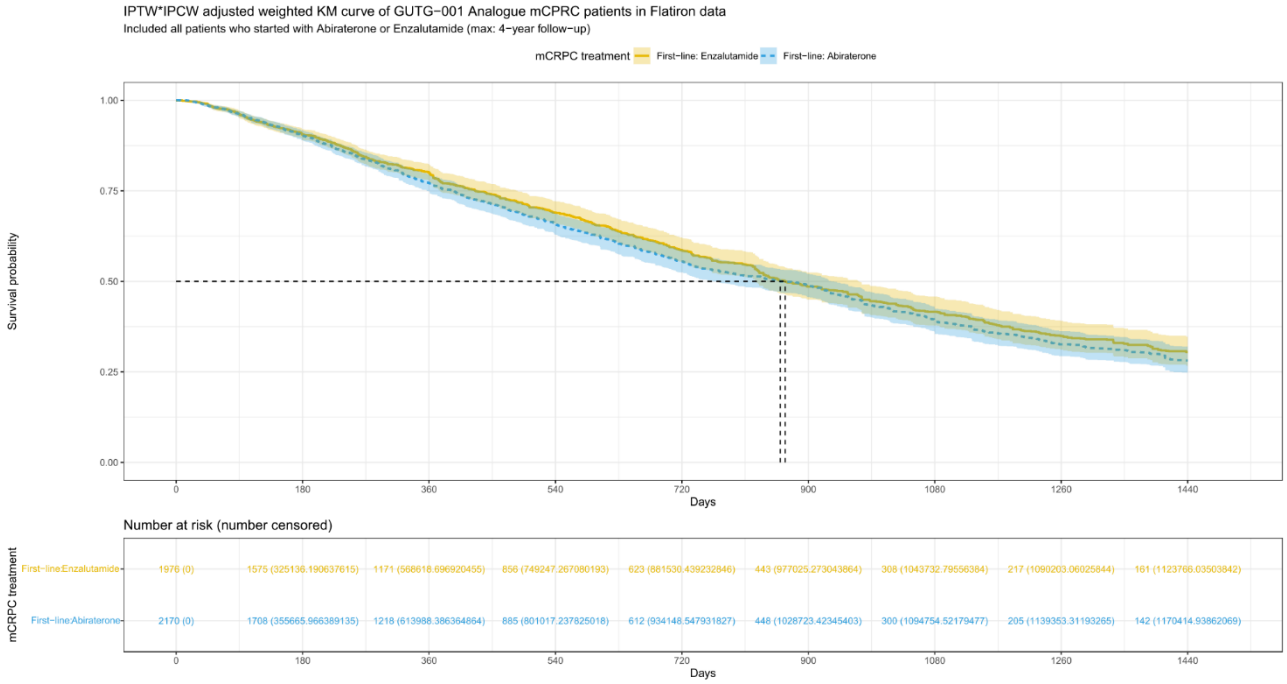


Figure 8.18 Kaplan-Meier curve of per-protocol analysis, adjusted for baseline and time-varying confounding with inverse probability weights

IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; mCRPC, metastatic castration-resistant prostate cancer

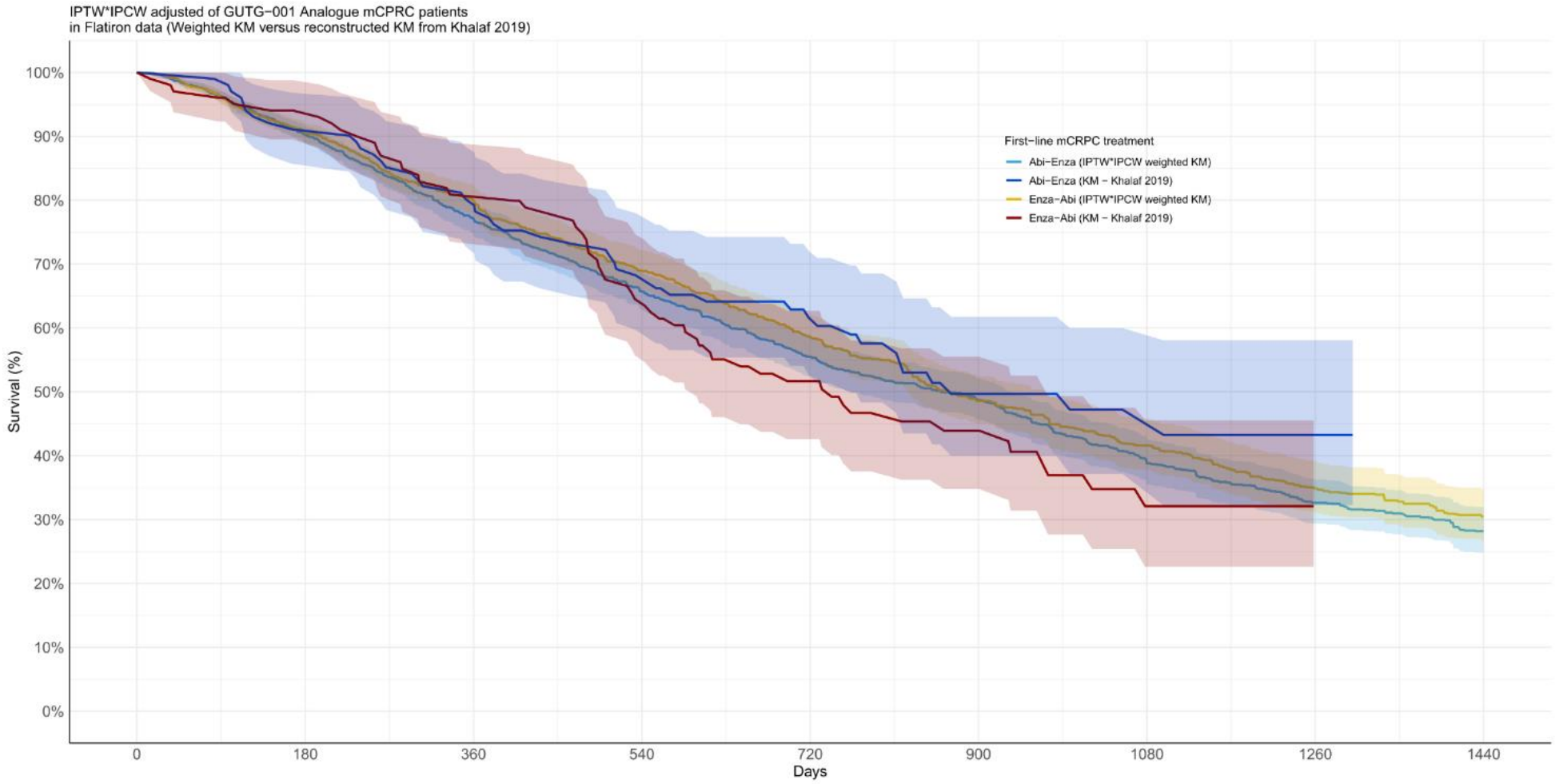


Figure 8.19 Comparison of Kaplan-Meier curves: per-protocol analysis with IPTW*IPCW$_{txdev}$ versus benchmark GUTG-001 trial

Abi, abiraterone; Enza, enzalutamide; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; KM, Kaplan-Meier; mCRPC, metastatic castration-resistant prostate cancer

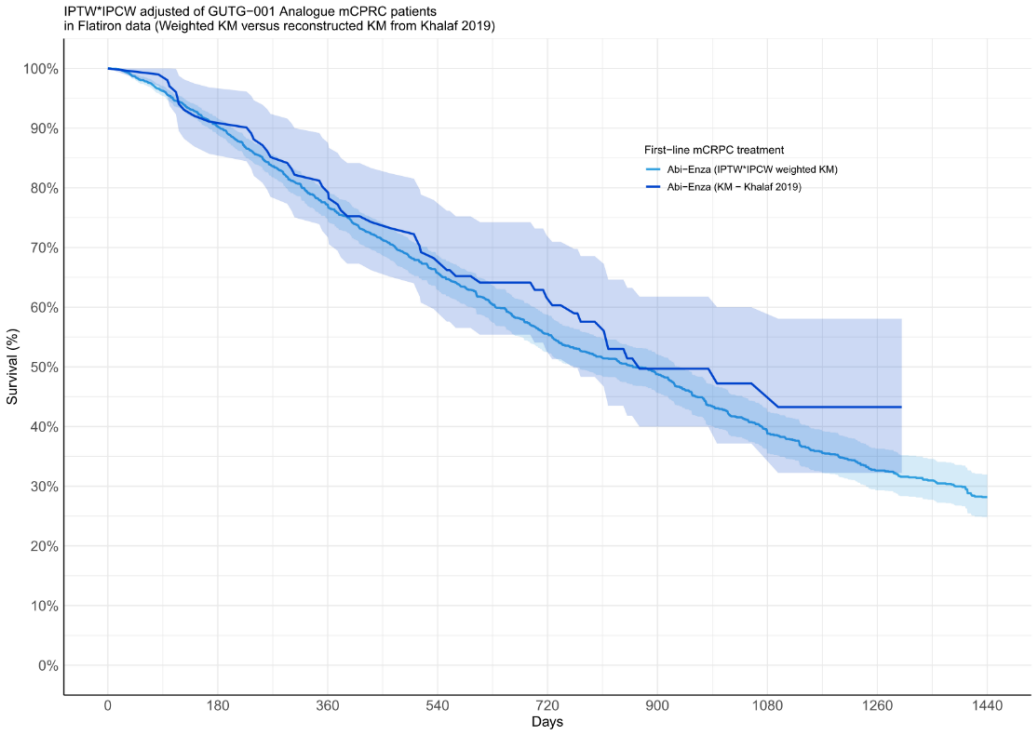


Figure 8.20 Comparison of Kaplan-Meier curves: per-protocol analysis with IPTW*IPCW$_{txdev}$ versus benchmark GUTG-001 trial: abiraterone → enzalutamide group

Abi, abiraterone; Enza, enzalutamide; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; KM, Kaplan-Meier; mCRPC, metastatic castration-resistant prostate cancer

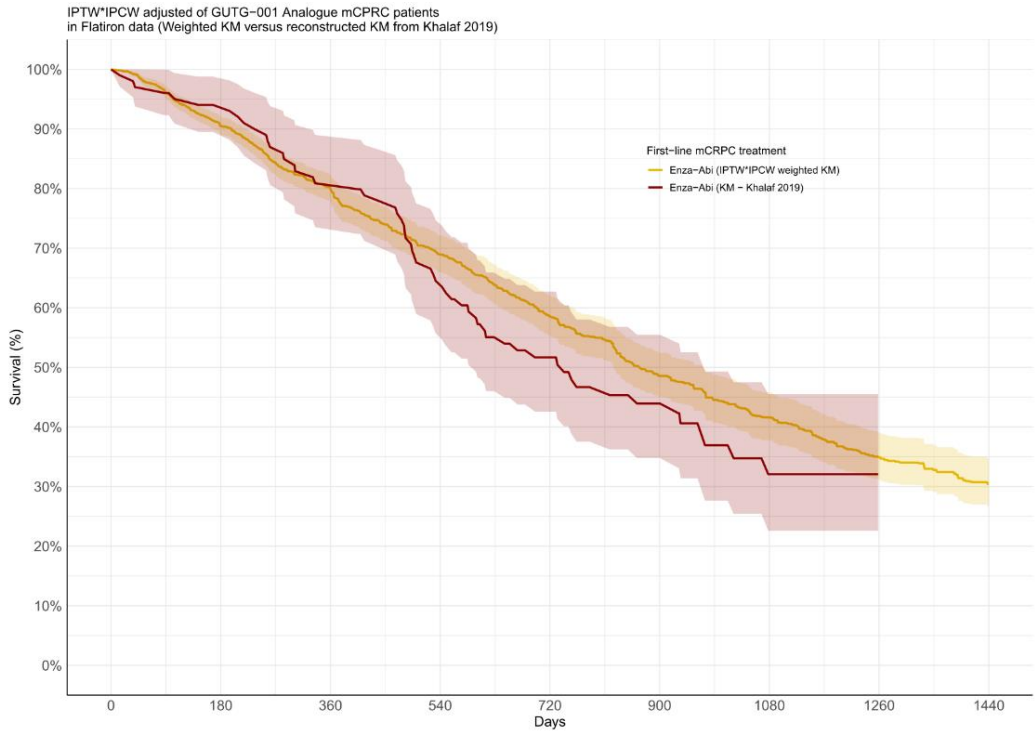


Figure 8.21 Comparison of Kaplan-Meier curves: per-protocol analysis with IPTW*IPCW$_{txdev}$ versus benchmark GUTG-001 trial: enzalutamide → abiraterone group

Abi, abiraterone; Enza, enzalutamide; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; KM, Kaplan-Meier; mCRPC, metastatic castration-resistant prostate cancer

Table 8.11 shows the ad-hoc analyses conducted to compare the restricted mean survival time (RMST) between my analyses and those of the GUTG-001 trial at 12, 24, and 36 months of follow-up. This includes deriving the RMST for both the abiraterone → enzalutamide and enzalutamide → abiraterone groups, as well as point estimates for RMST differences in each analysis (Analysis 1a-1d, 2a-2d, and 3 in Table 8.11). The RMST values for both groups and the difference between them were then compared with the results from the GUTG-001 trial across all analyses.[84] Since Khalaf et al. did not report the RMST for OS in the GUTG-001 trial[84], the RMST results from my analyses were compared to estimates derived from the digitised survival curve (i.e., Analysis 0 in Table 8.11).

The RMST for the unweighted KM analyses (Analysis 0 and 1a-1d) was derived using the rmst2 package in R. However, no R packages support the calculation of RMST for weighted KM. Therefore, I wrote a function "perform_rmst_trapezoids_wKM" in R to calculate the RMST for each group by determining the area under the weighted KM curve for the remaining five weighted analyses: baseline IPTW adjusted (Analysis 2a-2d) and IPTW*$IPCW_{texdev}$ adjusted (Analysis 3). Detailed steps of the function and the R code for deriving RMST for each group and RMST differences between groups for each analysis (1a-1d, 2a-2d, and 3) are provided in Appendix 8.11.

RMST is used in health economic evaluations to estimate the area under the survival curve and indicate the time patients spend in each health state. For example, in a 3-state partitioned survival model (progression-free, post-progression, and death), the RMST for progression-free survival (PFS) reflects time spent in the PFS state, while the difference between RMST for OS and PFS represents time spent in the post-progression state. The RMST ad-hoc analyses in Table 8.11 provide additional insights into assessing the agreement between RCT and RWE in this context.

For the RMST differences between the two treatment sequence groups (shown in the right column for RMST intervals in Table 8.11), Analysis 1d (AT, limited to second-line treatment recipients), 2d (baseline IPTW weighted AT, limited to second-line treatment recipients) and Analysis 3 (advanced IPTW*$IPCW_{txdev}$) produced some of the closest numerical results compared to the GUTG-001 trial (Analysis 0). Howver, most RMST point estimates for the two treatment groups in Analysis 1d and Analysis 2d at 12-month, 24-month, and 36-month did not fall within the 95% CIs of the GUTG-001 trial. For example, the 12-month RMST for abiraterone → enzalutamide and enzalutamide → abiraterone in GUTG-001 was 10.99 (10.52, 11.46) and 11.01 (10.53, 11.50) months, respectively, while the point estimates for both groups in Analysis 2d were 11.61 and 11.59 months. In contrast, the RMST point estimates for all groups in Analysis 3 fell within the 95% CI of GUTG-001, with 12-month RMST estimates of 10.77 and 10.85 months, respectively. This pattern also holds true for RMST estimates from Analysis 3 at other time intervals.

Table 8.11: GUTG-001 Analogue Target Trial Emulation results comparison table - restricted mean survi-val time (RMST) ad-hoc analyses

Models (population)	Restricted mean survival time (months, 95% CI)†								
	12-month			24-month			36-month		
	ABI-ENZA (n = 2,172)	ENZA-ABI (n = 1,972)	Differences	ABI-ENZA (n = 2,172)	ENZA-ABI (n = 1,972)	Differences	ABI-ENZA (n = 2,172)	ENZA-ABI (n = 1,972)	Differences
Summary statistics from Khalaf 2019 Study (ITT, Khalaf 2019)	NA	NA	NA	NA	NA	NA	NA	NA	NA
0. Summary statistics using digitally reconstructed survival data from Khakaf 2019 (ITT, Khalaf 2019)	10.99 (10.52, 11.46)	11.01 (10.53, 11.50)	-0.03	19.30 (17.93, 20.67)	18.93 (17.60, 20.25)	0.38	25.55 (23.20, 27.90)	24.02 (21.76, 26.28)	1.53
1a. Unadjusted ITT	10.73 (10.61, 10.84)	10.85 (10.73, 10.96)	-0.12	18.41 (18.09,18.72)	19.04 (18.72, 19.36)	-0.63	23.73 (23.20, 24.25)	24.91 (24.37, 25.45)	-1.18
1b. Unadjusted PP	10.80 (10.68, 10.91)	10.90 (10.78, 11.02)	-0.10	19.04 (18.71, 19.38)	19.57 (19.23, 19.91)	-0.53	25.20 (24.61, 25.78)	26.30 (25.70, 26.89)	-1.10
1c. Unadjusted AT	10.41 (10.25, 10.56) (n = 1, 500)	10.51 (10.34, 10.67) (n = 1, 358)	-0.10	17.88 (17.48, 18.28) (n = 1, 500)	18.35 (17.93, 18.77) (n = 1, 358)	-0.47	23.24 (22.58, 23.91) (n = 1, 500)	24.19 (23.49, 24.88) (n = 1, 358)	-0.94
1d. Unadjusted AT, limited to second-line treatment recipients	11.66 (11.57, 11.76) (n = 645)	11.59 (11.47, 11.72) (n = 487)	0.07	20.91 (20.51, 21.31) (n = 645)	20.80 (20.32, 21.29) (n = 487)	0.11	27.35 (26.57, 28.12) (n = 645)	26.89 (25.99, 27.79) (n = 487)	0.46
2a. Baseline IPTW weighted ITT◈	10.71 (10.57, 10.85)	10.81 (10.65, 10.97)	-0.10	18.43 (18.04, 18.83)	18.97 (18.55, 19.40)	-0.54	23.84 (23.18, 24.53)	24.78 (24.06, 25.52)	-0.93
2b. Baseline IPTW weighted PP◈	10.77 (10.62, 10.91)	10.89 (10.74, 11.05)	-0.13	19.02 (18.61, 19.45)	19.59 (19.16, 20.02)	-0.56	25.26 (24.52, 26.03)	26.29 (25.52, 27.08)	-1.03
2c. Baseline IPTW weighted AT◈	10.36 (10.17, 10.56) (n = 1, 500)	10.5 (10.29, 10.71) (n = 1, 358)	-0.13	17.81 (17.31, 18.32) (n = 1, 500)	18.39 (17.85, 18.94) (n = 1, 358)	-0.58	23.23 (22.4, 24.09) (n = 1, 500)	24.26 (23.37, 25.20) (n = 1, 358)	-1.03
2d. Baseline IPTW weighted AT, limited to second-line treatment recipients◈	11.61 (11.47, 11.73) (n = 645)	11.59 (11.41, 11.76) (n = 487)	0.02	20.77 (20.24, 21.32) (n = 645)	20.87 (20.24, 21.52) (n = 487)	-0.10	27.20 (26.19, 28.27) (n = 645)	27.03 (25.84, 28.30) (n = 487)	0.17
3. Baseline IPTW and time-varying $IPCWt_{xdev}$* weighted PP (hypothetical per-protocol effect with perfect compliance to the treatment sequence assignment)◈	10.77 (10.61, 10.93)	10.85 (10.68, 11.02)	-0.08	18.68 (18.19, 19.19)	19.11 (18.59, 19.66)	-0.43	24.42 (23.55, 25.34)	25.03 (24.08, 26.04)	-0.61

ABI: abiraterone; AT: as-treated, CI: confidence interval; ENZA: enzalutamide; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; ITT: intention-to-treat; KM: Kaplan-Meier; PP: per-protocol; RMST: restricted mean survival times

* cubic spline for time

† 12-month, 24-month, and 36-month are defined as 360 days, 720 days, and 1080 days days post Time Zero

◈Due to the R function rmst2 not supporting the calculation of RMST for weighted survival objects, RMST of the weighted Kaplan-Meier curves are calculated using the trapezoid rule with all available data points for each survival curves. The 95% CI for RMST are reported from the naïve (i.e., unadjusted) 95% CI curve. The difference in RMST between the two arms was calculated using the point estimates of RMST from each arm.

In summary, the ad-hoc RMST results (Table 8.11) aligned with the median OS and HR findings (Table 8.10), showing that the advanced IPTW*IPCW$_{txdev}$ method (Analysis 3) provided a better assessment of treatment sequence effectiveness compared to simpler methods (Analysis 1a-1d and 2a-2d). While the RMST results from Analysis 1d seemed to offer the best estimates of RMST differences at 24 and 36 months (i.e., closest to those in GUTG-001), the RMST point estimates for both treatment groups (abiraterone → enzalutamide and enzalutamide → abiraterone) mostly fell outside the 95% CIs of the GUTG-001 trial estimates, likely due to the inherent immortal time bias of the AT methods.

The direction of the RMST differences at 24 and 36 months in the GUTG-001 trial (positive at 0.38 and 1.53, respectively) contrasted with those in Analysis 3 (negative at -0.43 and -0.61), even though they were aligned at 12 months. This discrepancy may be due to the median time from treatment initiation to second-line progression in GUTG-001 being around 16-20 months (see Figure 2 in Khalat et al.'s GUTG-001 trial publication[84]), with second-line treatment discontinuation and third-line treatment initiation likely occurring around this time. That is, smaller sample sizes in the GUTG-001 trial and unadjusted third-line treatments in my GUTG-001 Analogue may have contributed to the discrepancy, as third-line treatments might differ between the GUTG-001 trial and those in RWD. Since no third-line treatment data were provided in the GUTG-001 trial, further adjustments were not possible. Additionally, without an existing method to calculate RMST 95% CIs for weighted KM (other than bootstrapping), it is unclear if the differences are statistically significant. Further research is needed to develop methods for assessing the agreement of RMST differences between treatment groups in RCTs and RWE.

8.4.5. RCT-RWE agreement assessment

Table 8.12 summarises the RCT-RWE agreement assessment metrics, which are pre-defined in my study protocol (Section 7.5.5 of Chapter 7). For the first three criteria, adapted from the RCT DUPLICATE studies[172], my emulation of the GUTG-001 Analogue Target Trial using the IPTW*IPCW$_{txdev}$ method showed full agreement. Specifically, the HR from my emulation indicated a no significant difference between the two treatment sequence strategies (1.07, 95%CI 0.96-1.20), which is consistent with findings from the GUTG-001 trial (0.79, 95%CI 0.54-1.16).[84] Although the point estimate in my emulation seemed to favour the enzalutamide → abiraterone sequence instead, this still represents a full agreement. Additionally, the second criterion also demonstrates full agreement, as the point estimates of my HR fell within the 95% CI of the GUTG-001 trial. For the third criterion, the standardizsed difference between the HRs in my emulation and the benchmark shows no significant difference, indicating full agreement.

Table 8.12 RCT-RWE agreement assessment: **GUTG-001 trial versus emulated GUTG-001 Analogue Target Trial with the IPTW*IPCW$_{txdev}$ method**

Criteria	Findings
1. Regulator agreement: assessing whether the direction and statistical significance are in alignment with the benchmark trial	Agreed: The GUTG-001 trial indicates no statistical difference in HR between the two treatment sequences (0.79, 0.54-1.16), and the HR from my final emulation with the advanced confounding adjustment method (1.07, 0.96-1.20) concurs with this finding.
2. Estimate agreement: assessing whether the point estimate falls within the 95% CIs of the benchmark trial	Agreed: The point estimate of HR from my final emulation with the advanced confounding adjustment method (1.07) falls within the 95% CI of the GUTG-001 trial (0.54-1.16). Further, the point estimates of the median OS for both treatment groups— abiraterone → enzalutamide: 28.7, and enzalutamide group → abiraterone: 28.9—also fall within the 95% respective CIs in the GUTG-001 trial (28.8-not reached, and 18.8-34.0, respectively). In the ad-hoc RMST analysis, the 12-, 24-, and 36-month point estimates for both treatment groups—abiraterone → enzalutamide: 10.77, 18.68, and 24.42, and enzalutamide → abiraterone: 10.85, 19.11, and 25.03—also fall within the respective 95% CIs from the GUTG-001 trial: abiraterone → enzalutamide (10.52-11.46), (17.93-20.67), and (23.20-27.90); enzalutamide → abiraterone (10.53-11.50), (17.60-20.25), and (21.76-26.28), respectively.
3. Exploratory – standardised difference: $Z = \frac{\hat{\theta}_{RWE} - \hat{\theta}_{RCT}}{\sqrt{\sigma^2_{RWE} + \sigma^2_{RCT}}}$, where the absolute value of Z-value less than 1.96 indicates no significant difference between the estimates from RWE and RCT	Agreed: Z = -1.48 (for HR from my final emulation versus GUTG-001), indicating no significant difference between the estimates from RWE and RCT
4. Exploratory - survival curve comparison: assessing whether the point estimate of the RWE survival curve for each treatment-sequence group fall within the 95% CI of the benchmark trial.	Largely aligned: The survival point estimate of my final emulation with the advanced confounding adjustment method mostly fell within the 95% CIs of the GUTG-001 trial, with an exception of those within the first 3 months of follow-up in the abiraterone → enzalutamide group.

CI: confidence interval; HR: hazard ratio; OS: overall survival; RCT: randomised controlled trial; RMST: restricted mean survival time; RWE: real-world evidence

Additionally, leveraging the same logic as criterion two, I assessed the concordance of absolute outcomes in addition to the relative outcomes between groups, which were not specified in the RCT DUPLICATE's criteria (i.e., exploratory assessment in criterion 2, Table 8.12).[172] The point estimate of the median OS for both groups in my emulated Target Trial—abiraterone → enzalutamide: 28.7, and enzalutamide → abiraterone: 28.9—fell within the respective 95% CIs in the GUTG-001 trial

(28.8-not reached, and 18.8-34.0, respectively). Further, in the ad-hoc RMST analysis, the 12-, 24-, and 36-month point estimates for both treatment groups—abiraterone → enzalutamide: 10.77, 18.68, and 24.42, and enzalutamide → abiraterone: 10.85, 19.11, and 25.03—also fell within the respective 95% CIs from the GUTG-001 trial: abiraterone → enzalutamide (10.52-11.46), (17.93-20.67), and (23.20-27.90); enzalutamide → abiraterone (10.53-11.50), (17.60-20.25), and (21.76-26.28), respectively.

For the added fourth exploratory criterion, I visually examined the concordance of the emulated survival curves with those of the GUTG-001 trial. The survival point estimates from my emulation mostly fell within the 95% CIs of the GUTG-001 trial, with the exception of the initial three-month follow-up period in the abiraterone → enzalutamide group, where they were marginally lower.

8.4.6. Sensitivity analysis

Table 8.13 summarises the results of several sensitivity analyses, all of which modified operational procedures in implementing of the final emulation using the advanced IPTW*IPCW$_{txdev}$ method to evaluate its robustness. Sensitivity analyses (1) and (2) investigated the impact of excluding ambiguous factors, as listed in Table 8.8, on the emulated outcomes. The results show that even without these factors, the HR between the two treatment groups remained non-significant, aligning with the findings of the original model. However, the HR point estimates increased after excluding the "calendar year" covariate, or when it was excluded alongside the "M Stage" factor. This suggests that the original model may have more effectively accounted for unmeasured confounding represented by these covariates.

Sensitivity analyses (3) and (4) addressed challenges related to defining the administrative censoring date in analysing the Flatiron data, an electronic health record (EHR) database. Specifically, beyond the default follow-up period (1,440 days) and data cutoff date (August 31st, 2023), an additional date was defined for administrative censoring, the last recorded structural activity of patients in the database plus a 182-day period, whichever comes first (Section 8.3.3.2). This accounts for the fact that patient records in this EHR database were often affirmed from diverse sources. Unless patients died during treatment or were hospitalised, their deaths might only be captured during periods following their last structural activities at the clinics/hospitals. There are exceptions, such as cases where treatment was prescribed with a pre-existing refill order that lasts beyond the patient's death, or instances where patients were discovered deceased during a tele-visit to their home, resulting in a structural activity occurring after death. Though these exceptions are likely to be rare. Hence, administratively censoring patients at their last structural activity date likely misses many death events, especially those occurring shortly after. Sensitivity analyses (4) demonstrated this by showing the median OS could not be calculated for both treatment groups due to immature survival endpoints.

Table 8.13 Sensitivity analyses

Models (population)	Sensitivity analysis	Median OS (months, 95 CI%), using Kaplan-Meier estimator†		Hazard ratio (95 CI%), using Cox models¶
		ABI-ENZA (n = 2,172)	ENZA-ABI (n = 1,972)	
Cox model summary statistics from Khalaf 2019 Study (ITT, Khalaf 2019[84])	-	28.8 (25.4, not reached)	24.7 (18.8, 34.0)	0.79 (0.54, 1.16)
0. Cox model with digitally reconstructed survival data from Khakaf 2019 (ITT, Khalaf 2019[84])	-	29.0 (25.5, not reached)	24.8 (19.9, 32.5)	0.78 (0.54, 1.15)
3. Baseline IPTW and time-varying $IPCW_{txdev}$ weighted PP (hypothetical PP effect with perfect compliance to the treatment sequence assignment)	Original model	28.7 (25.6, 31.2)	28.9 (27.5, 32.4)	1.07 (0.96, 1.20)
	(1) Variations in weight derivation models: Excluding the covariate "calendar year" from the weight derivation models"	27.3 (24.6-30.4)	28.9 (27.6-32.4)	1.11 (0.99-1.24)
	(2) Variations in weight derivation models: Excluding the covariate "calendar year" and "M Stage" from the weight derivation models	27.7 (24.7-30.4)	28.9 (27.6-32.4)	1.10 (0.99-1.23)
	(3) Alternative administrative censoring criteria: Patients are administratively censored on their last structural activity date in the database without any grace period, at the maximum follow-up period in the survival analysis, or up to the data-cutoff date (August 31st, 2023), whichever comes first.	cannot be calculated	cannot be calculated	0.93 (0.72-1.21)
	(4) Alternative administrative censoring criteria: Patients are administratively censored at the maximum follow-up period in the survival analysis or up to the data-cutoff date (August 31st, 2023), whichever comes first.	29.7 (26.2-31.7)	30.6 (27.9-33.5)	1.07 (0.96-1.19)

ABI: abiraterone; AIC: Akaike Information Criterion; CI: confidence interval; ENZA: enzalutamide; AT: as-treated, BIC: Bayesian Information Criterion; IPCW: inverse probability of censoring weighting; IPTW: inverse probability of treatment weighting; IQR: interquartile Range; ITT: intention-to-treat; KM: Kaplan-Meier; OS: overall survival; PP: per-protocol, 4-year is defining as 1440 post Time Zero

* cubic spline for time

† All 95% confidence intervals for median overall survival presented were directly obtained from the survfit function in R, including those for weighted analyses.

¶ 95% CI for HR from the unadjusted analysis (Analysis 0) was directly obtained using the coxph function in R. For weighted analyses, 95% CIs were derived from the robust sandwich estimator

On the other hand, counting all confirmed deaths in the database without administratively censoring patients their last structural activity date would provide comprehensive death records. However, it might overlook critical information on subsequent treatments for patients lost to follow-up for extended periods (e.g., deaths inferred from an obituary without recent clinical follow-up). This could obscure relevant data on deviations from treatment strategies during the periods lost to follow-up, potentially biasing comparisons of treatment sequences.

Ideally, using each patient's prescription refill dates, date of pre-planned treatment cycles, or scheduled follow-up dates as customised individual grace periods would better track follow-up loss. However, since such information was unavailable, I opted for a six-month grace period following the last known activity date. This approach aimed to balance the challenges of missing information on subsequent treatments during periods of lost follow-up and the risk of failing to capture deaths if

administratively censored at the last hospital/clinic visit date. This approach is based on the assumption that patients are likely to visit their physicians every 3 to 12 months for treating prostate cancer, depending on their condition's stability.

Sensitivity analysis (4) tested a different extreme scenario where patients were not administratively censored based on their last structural activity date. The results showed HR almost identical to those in the original model, while the median OS lengths increased by 1 to 1.5 months for both groups. This suggests that although incorporating the last structural period with a grace period in the original model was theoretically rational, the sensitivity analyses further confirmed that not considering the last structural activity date does not significantly alter the relative outcomes.

8.5. Discussion

The PC1 case study findings underscore the feasibility of using advanced statistical methods to obtain reliable effectiveness estimates for treatment sequences from RWD, specifically Flatiron database. These methods should be adopted as the preferred approach when aiming to assess the effectiveness of a specific treatment sequence strategy or the comparative effectiveness of different treatment sequences. Conceptually, the IPTW*IPCW$_{txdev}$ approach conceptually eliminated the effects contaminated by non-protocol subsequent treatments without introducing additional biases. This effectively resolves major issues present in simpler methods, including ITT, PP, and AT (with or without baseline baseline confounding adjustments), which either do not address or incorrectly account for time-varying confounding. The complexities of time-varying confounding in treatment sequences were effectively managed using causal inference-guided methods, including TTE, DAGs, and advanced statistical methods. These tools were instrumental in planning the RWD analyses and facilitating discussions with clinicians to identify relevant covariates. The IPTW*IPCW$_{txdev}$ approach stands out as a pragmatic initial step due to its straightforward explanations, which significantly facilitate communication with clinicians.

The TTE using the advanced IPTW*IPCW$_{txdev}$ method demonstrated good agreement in the RCT-RWE concordance assessment, validating this approach as a robust proof of concept. Importantly, I expanded the criteria for RCT-RWE agreement assessment beyond those initially developed by RCT DUPLICATE[172], pioneers in TTE benchmarking studies. Specifically, Criterion 2 was expanded to evaluate both absolute and relative outcomes. This expansion proved crucial, as the HR from AT analyses (Analysis 1d and 2d in Table 8.10) would have met all RCT-RWE agreement Criteria (1) to (3) in Table 8.12. However, these analyses were significantly flawed due to evident selection or immortal time bias. By examining absolute outcomes, such as the accordance of median OS, the additional criteria helped identify these underlying issues. Additionally, an extra Criterion 4 was developed to evaluate the agreement of survival curves between the benchmark and the emulated

Target Trial. The reliability of the emulated survival curves is particularly relevant in HTA, where they are often used to perform parametric survival modelling for extrapolating longer-term outcomes, making their credibility essential. Overall, this expansion extends insights in both the comparison of treatment sequences and the broader scope of RWE generation with benchmarking. This is particularly vital in the use of RWE in HTA, where absolute outcomes like treatment duration or progression survival time are crucial for calculating treatment costs and quality-adjusted life years, highlighting the importance of benchmarking more than just relative effect estimates.

The value of employing a TTE approach lies in the necessity to critically assess whether any aspect of the Target Trial design and its emulation—including protocol implementation and necessary adaptations—introduces biases in analysing RWD. This is crucial for addressing biases that cannot be mitigated by advanced statistical methods, such as selection or immortal time bias. Properly addressing these biases through carefully defining "Time Zero" is key in mitigating these biases by preventing selecting patients based on post-enrolment factors (i.e., AT analyses in Table 8.10). Section 8.3 actualised these concepts, offering detailed procedures and coding examples, with a focus on comparing treatment sequences. Further, it also extends the general insights of NICE's RWE framework by demonstrating the practical considerations necessary for preparing fit-for-purpose, analysis-ready datasets to accommodate advanced IPW methods. It also highlights considerations specific to the use of the Flatiron databases and ensuring transparent documentation.

The study findings do not only offer insights into comparing the effectiveness of treatment sequences, but also into comparisons that may not explicitly assess sequences while still involving them, potentially leading to biased decisions. Dr. Pezaro noted that the baseline-adjusted ITT analysis (Analysis 2a in Table 8.10) was conceptually intended to compare the effectiveness of abiraterone versus enzalutamide as first-line treatments for mCRPC. Although the analysis significantly favoured enzalutamide as first-line, Dr. Pezaro cautioned against this interpretation in clinical decision making. Particularly, the Sankey diagram in Figure 8.8 revealed that patients receiving either first-line treatment had very different choices of subsequent treatments, making them incomparable unless the effects of these subsequent treatments were factored in. For instance, the final emulated results of abiraterone → enzalutamide versus enzalutamide → abiraterone showed no significant difference with advanced adjusted analyses. Hence, Dr. Pezaro highlighted the value of Sankey diagram in presenting how imbalances in subsequent treatments can affect OS outcomes. For instance, one might want to investigate the effects of abiraterone versus enzalutamide as first-line treatments, with docetaxel as the primary subsequent treatment in both arms. However, obtaining this information from the data requires advanced adjustments, as it is not readily available.

In fact, as identified in Chapter 2, TA 377 included a comparison of two treatment sequences: abiraterone → docetaxel versus enzalutamide → docetaxel.[126] Here, the OS of first-line treatments

from trials was used to represent the OS for the entire treatment sequences. This approach implicitly assumes subsequent treatments of these first-line treatments were reflective of those in the NHS and were comparable between study groups. This assumption was likely valid at the time of TA 377 when subsequent treatment choices in mCRPC were relatively limited. However, for example, sufficient comparative evidence does not exist and an external control arm is needed, which is increasingly common in oncology TA. In this case, the Sankey diagram proves indispensable in illustrating the variations in subsequent treatments across different arms in HTA, whether using RWD or trial data. Specifically, it visually emphasises the potential pitfalls of excluding selected patients based on post-enrolment criteria (i.e., subsequent treatments), as noted in the AT analyses in Table 8.10 and commonly seen in RWD studies comparing treatment sequences with simple methods, as noted in Chapter 4.

Several challenges and limitations should be acknowledged. First, the benchmarking process is constrained by the relatively small sample size of the benchmark GUTG-001 trial, increasing the likelihood that point estimates from the emulate Target Trial will fall within the 95% CIs of the benchmark trial. Despite this, advanced analyses have significantly improved the findings by directing point estimates toward more accurate directions, both in terms of both absolute and relative effect estimates. Specifically, the 95% CIs of median OS for both treatment groups fully fell within the 95% CI of the benchmark, which was not the case in analyses with simple methods. This represents a crucial improvement over simpler analyses in comparing treatment sequences, which were conceptually and practically flawed and have not previously been examined in HTA. The divergence of survival curves after a 1.5-year follow-up between the emulated Target Trial and the benchmark, particularly in the enzalutamide-first group (Figure 8.20-8.21), may result from the small sample size of the GUTG-001 trial and more prevalent use of enzalutamide in later calendar years. These factors suggest that a potentially proportionally higher use of novel treatments from the third line onwards in the enzalutamide-first group could contribute to seemingly better survival outcomes in real-world compared to the benchmark (Figure 8.21). However, without details on subsequent treatments in the GUTG-001 trial, these observations cannot be fully verified.

Furthermore, the lack of precise progression date information and incomplete patient data, notably missing values in key prognostic factors like ECOG scores, may lead to residual confounding. Inconsistencies in historical records could potentially result in a biased definition of comorbidities across patients. These limitations could likely be improved by using customised add-on variables from Flatiron's service, which provide comprehensive manually abstracted information on comorbidities and progression dates.[438] However, these add-on variables are not included in the standard mPC dataset. Additionally, the absence of comprehensive progression information in the standard Flatiron mPC dataset and variably recorded PSA levels among patients hinder the ability to

conduct alternative outcome analyses specified in the protocol in Chapter 7, such as progression-free survival defined by PSA progression. Given these data constraints, employing sensitivity analyses such as time-to-next-line treatment initiation and time-to-treatment discontinuation could be practical approaches in this research context. Additionally, E-values may be used to demonstrate the impact of unmeasured confounding.[595] While multiple imputation could be considered to mitigate the impact of missing data, caution is advised in its use for causal inference questions, with ongoing investigation in the field.[596,597]

Despite its limitations, this study has several notable strengths. It is the first to design and benchmark TTE analyses against an existing RCT that compares treatment sequences. The study extends the status quo in RCT-RWE agreement assessment by introducing additional criteria essential for evaluating absolute outcome estimates and visualising survival curve agreements, critical for HTA. It offers a detailed, step-by-step tutorial on implementing TTE implementation in comparing treatment sequences, hereby complementing and extending the insights of the NICE RWE framework. Moreover, the research highlights the indispensable role of the Sankey diagram and preliminary data checks in the design phase, upon data receipt, and during the TTE implementation. These checks guide necessary and appropriate adaptations to the designed Target Trial protocol. Overall, the study demonstrates a logical step forward in using RWD as an alternative means to inform the effectiveness of treatment sequences, offering novel insights into tackling the challenges of evidence scarcity in HTA involving the evaluation of treatment sequences.

Although several sensitivity analyses have been prioritised for investigation within the constrained analysis timeframe of this thesis (Section 8.4.6), several additional sensitivity analyses, as mentioned in the implementation and protocol, could further be planned in the same line of future research to test the robustness of the results. These include restricting the study to pre-COVID patient cohorts, varying the period for identifying baseline characteristics based on diagnosis (i.e., default = 6 months), and merging treatment sequences, such as abiraterone → abiraterone into one LOT (see Section 8.2.2).

Additionally, a critical exploration would be conducting a sequential emulation of a set of Target Trials with varying Time Zeros instead of using the initiation of first-line treatment as Time Zero. Specifically, the variability in the time from metastatic diagnosis to first-line treatment in mCRPC across groups (Table 8.9) could lead to potential immortal time bias in patient enrolment. However, it was considered acceptable in the current emulation because the time from mCRPC to first-line treatment initiation was not significantly different between groups. Specifically, the shorter interval from mCRPC to first-line treatment initiation, compared to the time from metastasis to treatment initiation in both groups, mitigates these concerns. Nonetheless, conducting sensitivity analyses with different Time Zeros could provide valuable insights, although this would deviate from the GUTG-

001 trial's design, where follow-up and randomisation align with treatment initiation, and making RCT-RWE benchmarking impossible. Essentially, this variation could be deemed a generalisability study, falling within the scope of the second prostate cancer case study (PC2) as specified in the original study protocol (Chapter 7). Implementing sequential emulation would require a significant reconfiguration of the analysis and expansion of the dataset, potentially increasing the computational load. In the same line of research regarding generalisability, the cloning technique outlined in Chapter 4 could be explored to compare alternative treatment sequences that start with the same first-line treatment. Additionally, alternative advanced methods like, such as g-formula (Chapter 4) may be investigated to assess their performance in emulating the same trial. Then, the IPTW*IPCW$_{txdev}$ method, along with the same set of covariates and implementation procedures, could be explored to evaluate its feasibility in emulating the same Target Trial in alternative databases.

8.6. Chapter summary

This study represents a pioneering effort to benchmark TTE analyses against an existing RCT that compare treatment sequences. It contributes to evaluating treatment sequences in HTA by demonstrating the utility of RWD as an alternative method for providing unbiased estimates that could inform the effectiveness of treatment sequences. This chapter builds on insights from previous chapters, including reviews of advanced statistical methods guided by causal inference (Chapter 4), evaluation of appropriate oncology databases suitable for implementing these methodologies and pertinent for English HTA (Chapter 5), and identification of RCT benchmarks for establishing proof-of-concept studies (Chapter 6). Importantly, it followed the predefined protocol from Chapter 7 with necessary adaptions being transparently documented. Hence, this chapter is the realisation of a combination of efforts from all previous chapters. In addition to offering fresh insights into treatment sequence comparisons in HTA, this chapter also provided a step-by-step tutorial on implementing TTE analyses and introduced additional criteria to the existing RCT-RWE agreement assessment metrics, extending the practical aspects of the NICE RWE framework.

Chapter 9 Research summary and recommendations

9.1. Overview

This chapter provides a comprehensive summary and conclusion of the thesis. It begins by revisiting the key findings from each chapter, highlighting the new insights they have contributed to the topic of evaluating treatment sequences in HTA (Section 9.2). Section 9.3 then spotlights the thesis's key contributions to the field, emphasising the main advances. Section 9.4 discusses the strengths and limitations of my research. Section 9.5 outlines potential areas for future research. The chapter concludes with recommendations and final remarks in Section 9.6 and 9.7, emphasising the thesis's implications to the field.

9.2. Summary of chapter findings

- ***Chapter 2***

Chapter 2's scoping review fulfilled my Research Aim 1 (Section 1.5, Chapter 1), contributing to the field by laying out a roadmap of key literature on evaluating treatment sequences in HTA. The review was recently published in a peer-reviewed journal.[87] It traces the evolution of this field's literature, highlighting the multifaceted challenges and their interplay. Methods for structuring treatment-sequencing economic models have been established[1,2], yet the scarcity of clinical evidence for populating these models introduces structural and parametric considerable uncertainties. Recent publications have increasingly focused on addressing this issue.[34-36,125] Specifically, clinical trials typically do not compare treatment sequences but treatments at specific lines, necessitating merging clinical evidence from different sources to populate models across different treatment lines. Lewis et al. and Huang et al. proposed applying adjustment methods to refine the derivation of line-of-treatment (LOT) effect (e.g. meta-regression, subgroup meta-analysis for adjusting patient's treatment history) and to account for indirect treatment comparisons (ITC) when merging evidence from different sources, respectively.[34,35] However, NICE HTA guidelines have yet to adopt these approaches for evaluations involving treatment sequences.[142-145] Further, while pragmatic, these methods may fall short in certain contexts as they often rely on the assumption that patient characteristics align across evidence sources for different LOTs, an assumption that is unlikely to hold where the primary outcome is survival.

These findings solidified the scope of my doctoral research, which aimed to address the scarcity of clinical evidence on the unbiased comparative effectiveness of treatment sequences by leveraging real-world data (RWD). RWD is particularly adept at capturing entire treatment sequences, an area that previously unexplored due to its susceptibility to confounding biases. My review offers new perspectives by discussing how Target Trial Emulation (TTE)[76]—a methodological framework

recently endorsed by the National Institute for Health and Care Excellence (NICE) real-world evidence (RWE) framework[37]— may shed light on the evaluations involving treatment sequences. It spotlights the niche of innovative efforts to employ TTE[3] and advanced causal inference guided statistical methods to mitigate biases in analysing non-randomised patient-level data. Subsequently, my review identified several research directions, including applying causal inference methods in analysing RWD and/or trial data harbouring sequences to derive unbiased estimates of effectiveness, whether as a standalone evidence source or combined with others (e.g. synthesising an external control arm involving sequences). My thesis then focused on exploring the first scenario (i.e., comparing treatment sequences using RWD as a standalone source) in depth, as the second scenario would require accessing two distinct data sources, and comparing sequences within a single dataset represents a logical first step in this area of research.

- ***Chapter 3***

Chapter 3 fulfilled my Research Aim 2 (Section 1.5, Chapter 1). It systematically reviewed NICE technology appraisals (TAs) to understand the current practices and challenges in comparing treatment sequences within health economic evaluations. The review specifically sought to (1) explore how frequently treatment sequence discussions occurred in English HTA, and (2) examine how effectiveness estimates for these sequences were selected and derived. This aims to shed light on the significance of clinical evidence scarcity in treatment sequencing HTA and explore how it has been tackled with distinct types of data sources, particularly the utilisation of RWD.

Previous reviews on this topic has focused on TAs that explicitly employed a treatment sequencing economic model.[1,2] My review contributes to the field by revealing that nearly all TAs address treatment sequences to some degree, which arises as a natural consequence of the lifetime horizon required in many decision scenarios. However, explicit estimation of effects from different treatment sequences was not commonly performed. Effectiveness simplifying assumptions were commonly made, assuming treatment effects stay consistent across different LOTs and calculating costs of subsequent treatments without adjusting for potential changes in treatment effectiveness.

The in-depth review of 35 TAs that explicitly compared treatment sequences highlighted autoimmune diseases (n = 22) and oncology (n = 7) as the most prevalent disease areas, with notable differences in how clinical evidence was utilised in these fields. Most TAs leveraged a mix of clinical evidence sources, including clinical trials, (network) meta-analyses (NMA/MA), RWE, and simplifying assumptions to either directly or indirectly inform the effectiveness of treatment sequences. Specifically, autoimmune TAs predominantly relied on NMA/MA often assumed that a given treatment's effectiveness remains constant or decreases with a fixed modifier across different LOT. In contrast, oncology TAs commonly used progression-free survival (PFS) from LOT-specific

trials to calculate the cumulative treatment duration across LOTs for costs calculation, rather than applying a one-size-fits-all effectiveness assumption for all positions. However, estimates of overall survival (OS) for the entire treatment sequences were typically based solely on first-line trial results.

RWE has been used in many TAs to address gaps left by trial data, particularly in estimating the duration of later-line treatments for cost calculations and informing structures. However, there were no cases where RWD was used solely to evaluate the effectiveness of an entire treatment sequence or to conduct comparisons between two sequences. Similarly, trials that directly compared two entire treatment sequences was absent. Arguably, the studies using first-line treatment OS from trials as a proxy for the OS of an entire sequence are an exception, however, this method relies on assumptions that may be difficult to verify (i.e. subsequent treatments received in the trial are relevant to those used in clinical practices), especially when details on subsequent treatments might not be consistently documented in trials.[126,203,249,265] These instances were primarily seen in oncology TAs.

Key challenges identified involve integrating LOT-specific evidence from various sources, particularly biases from indirect comparisons and inconsistent patient populations across LOTs. Solutions such as matching-adjusted indirect comparisons or subgroup analysis of trials, including within NMA/MA (to account for "position" effect), emerged as common strategies to tackle these issues. However, the implementation of these methods has been inconsistent, aligning with findings from parallel reviews by Lewis et al. and Huang et al.[34,35] Moreover, these fall short in addressing challenges like crossing cumulative treatment duration and OS curves, particularly in oncology—a challenge which has not yet been explicitly addressed in the existing literature.[126] This challenge arises when cumulative treatment duration appears to be longer than OS, and stems largely from collating evidence from different treatment lines to proxy the survival outcome of an entire treatment sequence. The current solution of arbitrarily truncating one of the crossing curves is suboptimal. Further, there has been little effort to adjust for time-varying confounding in analysing trial or RWD involving sequences; only one oncology TA attempted to adjust for subsequent treatments in trials that did not reflect the sequence intended for comparison[203], while another declared that it was not possible to perform adjustments despite recognising the issue.[126]

In summary, the scarcity of RCTs directly comparing treatment sequences and the limitations of existing methods to combine survival evidence across LOTs highlighted a gap in evidence. This suggests that when treatment sequences have been compared in HTA, they might have been done sub-optimally. My review adds new insights to the field by examining treatment sequence evidence derivation methods separately for different diseases and evidence types, areas that were previously overlooked. It underscores the challenges of current statistical approaches and data sources in providing unbiased sequence estimates, emphasising the potential of RWD, capable of capturing the longitudinal detail of entire treatment sequences. However, harnessing RWD's promise requires

investigating strategies to address its inherent confounding bias effectively, especially time-varying confounding in comparing treatment sequences, as this has been largely under explored.

- ***Chapter 4***

Chapter 4 builds on the previous two chapters, fulfilling Research Aim 3 (Section 1.5, Chapter 1) by systematically exploring causal inference-guided statistical methods that may be used for estimating treatment sequence effectiveness using RWD and their applicability in HTA. This exploration particularly drew on disciplines rich in relevant methodologies, including statistics and epidemiology. This review not only identified key statistical methods for designing my proof-of-concept studies using RWD to inform treatment sequence effectiveness in HTA (Chapter 7), but also summarised taxonomies that integrate knowledge from other disciplines into the context of treatment sequences in HTA.

Although the concept of dynamic treatment regimens (DTR) from epidemiology seemed relevant to HTA treatment sequences, the specifics, nuances, and their applicability to HTA comparisons were not well-defined. For example, the comparison of treatment sequences in HTA typically involves examining the effect of a uniform treatment sequence for the entire population (Chapter 3). This approach represents a specific type of DTR, which focuses on comparing the effects of all patients in the population receiving the same treatment sequence against another. The dynamic aspect of this is that each patient's timing of receiving a subsequent treatment may vary, depending on specific triggering events, such as disease progression.

The review revealed that statistical methods for deriving the most common type of estimands for treatment sequences in HTA (i.e., uniform treatment sequence for the entire population) exist, but methods for other types of treatment sequences are also available. These include methods for estimating the effect of an optimal uniform treatment sequence for the entire population or optimal individualised treatment sequences. However, these are not typically the focus of HTA, which generally compared a given treatment sequence with the appraised technology against the most prevalent sequence in clinical practice, rather than comparing two optimal, or personalised, treatment sequences (Chapter 3). My review, therefore, summarised a taxonomy of different treatment sequence strategies (time-related static versus dynamic) and estimand types of treatment sequences (uniform, optimal uniform, or optimal individualised), and connected these to a discussion of the statistical methods identified, highlighting the most readily applicable ones in the context of HTA.

Additionally, the review contributes to knowledge by summarising parallels between various data sources, especially RCTs that randomise patients to receive treatment sequences, sequential multiple assignment randomized trial (SMARTs), and RWD. It provided conceptual links that demonstrate how causal inference methods used in different studies are relevant to analysing RWD.

This supports the perspective from a recent NICE DSU review identified in Chapter 2, which highlights the relevance of causal inference methods developed for correcting unwanted treatment-switching in HTA to the context of treatment sequencing.[43,44] In fact, all methods identified in the review can trace their relevance back to the contemporary causal inference methods established by Robins (i.e., g-methods).[54]

Most importantly, the review also contributed a taxonomy of the identified methods and provided comprehensive explanations, including the origin of each method, their theoretical properties, and application studies. These encompass simple methods, including intention-to-treat (ITT), per-protocol (PP), and as-treated (AT) and their caveats, along with advanced methods, primarily including inverse probability weighting (IPW), structural nested failure time models (SNFTM), g-formula, and Q-learning. IPW was selected for my proof-of-concept studies due to its straightforward nature and ease of communication with clinicians, as well as fitting well into the TTE framework[3,37], which, as introduced in Chapters 1 and 2, plays a crucial role in mitigating biases in RWD study designs. Most importantly, only IPW methods have been utilised in the context of comparing uniform treatment sequences, making them more readily applicable compared to other methods that focus on identifying optimal individualised treatment sequences. Notably, its variation, inverse probability of censoring weighting (IPCW), has been extensively applied in tackling unwanted treatment switching in HTA and is well understood. Contrastingly, g-formula and Q-learning have not been as widely adopted in HTA as IPW and have been primarily employed to identify optimal personalised treatment sequences in the reviewed studies. Meanwhile, SNFTM methods have been used exclusively to determine the optimal uniform treatment sequences for the population. While all of these other approaches require further adaptation to be suitable for comparing two predefined uniform treatment sequences, PW stands out as more readily applicable due to its matching estimand in the reviewed studies and theoretically lower computational expense." Hence, these make the use of IPW as a logical first step for investigation, while the research line may be continued with the support of PhD transition funding.

- ***Chapter 5 & 6***

Chapters 5 and 6 collaboratively investigated suitable RWD sources and benchmarks for conducting proof-of-concept studies aimed at generating unbiased comparative effectiveness estimates of treatment sequences for English HTA. This effort fulfilled my Research Aim 4 and partly met Aim 5 (Section 1.5, Chapter 1).

Chapter 5 highlighted the English National Cancer Registration and Analysis Service (NCRAS) and US Flatiron databases as key oncology RWD sources relevant for informing English HTA.[82,83,391] It detailed the procedures to identify these databases and compared their similarities, strengths, and

limitations, particularly focusing on their ability to accommodate the statistical methods identified in Chapter 4 for comparing treatment sequences. These insights shaped the design of the case study (Chapter 7) and facilitated data application. The exploration of data specifications also informed the necessary adaptation of criteria for implementing Target Trial case studies in Chapter 8, ensuring the studies accurately reflect the planned objectives and take into account data limitations.

Chapter 6 systematically reviewed RCTs comparing treatment sequences and identified the GUTG-001 prostate cancer trial and RECORD-3 trial as relevant benchmarks for designing proof-of-concept case studies.[84,86] The limitations of these trials inspired the innovative design of my interconnected case studies in Chapter 7. While no direct benchmark RCT exists for NCRAS data, the treatment sequences in GUTG-001 can serve as an indirect benchmark for validation against the English NCRAS database, supported by proof-of-concept studies using Flatiron data.

These efforts collectively contributed to the knowledge base by providing a detailed and transparent account of how pertinent oncology databases and relevant benchmark trials were identified, contrasting with TTE studies that do not have a benchmarking purpose to assess the feasibly and reliability of methods in a given database before applying them to other research questions. Specifically, this extended the NICE RWE framework by adding insights into assessing the suitability of databases[37], particularly for comparing treatment sequences. It highlighted the importance of considering treatment patterns within database population, which can be achieved by examining published statistics, as well as local treatment guidelines, and engaging with clinicians. These steps are essential for estimating sample sizes and assessing the feasibility of designing a study using RWD.

The approach of how I selected the benchmark trials also differs from the United States (US) Food and Drug Administration (FDA)-funded pioneering RCT DUPLICATE initiative, which assess the feasibility of using RWD for drug repurposing with regulatory implications at the outset, and therefore have focused on benchmarking trials in pertinent diseases, such as hypertension, diabetes, and cardiovascular diseases.[172] It is important to acknowledge that Chapters 2 and 3 identified oncology as a field with significant treatment sequencing considerations in HTA and challenges in using LOT evidence for survival outcomes, areas that may benefit from RWE support. Further, methods identified in Chapter 4 predominantly support oncology applications. The pilot review of benchmarks in Chapter 6 further underscored the relevance of focusing on oncology by revealing a disproportionately high number of RCTs in oncology that compare treatment sequences. These together led to a focus of searching oncology databases for my case studies.

At the end of Chapter 5, I highlighted the critical challenge of database access, which is, needless to say, the first and most crucial step in assessing the feasibility of using RWD to derive unbiased estimates of treatment sequences. It details the nuances and procedural hurdles encountered,

exacerbated by COVID-related delays and complex administrative processes. To improve this, the chapter proposed the critical use of more transparent, publicly accessible timelines, concise flowchart guides from both data custodians and academic institutions, and clear communication to eliminate any hidden expectations to streamline the approval process and minimise unnecessary revisions. Further, proactively engaging senior legal experts early in the process in reviewing complex agreements, such as international agreements or those that an intuition is unfamiliar with can navigate potential legal complexities more effectively and efficiently.

- ***Chapter 7 & 8***

Building on the insights established in the previous chapters, Chapters 7 and 8 addressed Research Aims 5 and 6, respectively. Chapter 7 detailed a proof-of-concept case study protocol for comparing treatment sequences using NCRAS and Flatiron data, and has been published online.[170] Chapter 8 offered a detailed, step-by-step tutorial on executing a TTE using advanced statistical methods for deriving unbiased estimates of treatment sequence comparisons from RWD, with Prostate Cancer Study 1 (PC1) serving as an example.

Chapter 7 contributes to the field by detailing the rationale and design of a set of interconnected proof-of-concept TTE case studies, specifically designed for examining the feasibility of deriving unbiased estimates for comparing treatment sequences using the NCRAS and Flatiron data. My case study design is innovative because, while the TTE framework and advanced statistical methods have been previously used to compare various types of DTR, they have not been specifically applied to comparing treatment sequences with implications for HTA. Furthermore, the aforementioned oncology databases have not been calibrated with benchmark studies for English HTA, especially for comparing treatment sequences. As mentioned in Chapter 6, this protocol introduced a novel approach—indirect benchmarking—to assess the feasibility of using NCRAS for comparing treatment sequences, with support from direct benchmarking with Flatiron data (Case Study PC2-3). Given the recent partnership between NICE and Flatiron, Flatiron data has gained relevance for English HTA.[83] Hence, direct benchmarking with Flatiron data itself also presents its own distinct value (Case Study PC1). These efforts contribute to the field by offering practical approaches to explore alternative means of addressing the challenge of evidence scarcity in evaluating treatment sequences, as highlighted in Chapters 2 and 3.

The study protocol also contributed to the NICE RWE framework by highlighting the need for benchmarking/calibration, which has been mentioned in a recent publication by Gomes et al. as an area that should be prioritised but was not explicitly instructed in the framework.[78] The significance of benchmarking has been highlighted—for example, it has been adopted by the US FDA-funded research to assess the feasibility of using real-world data (RWD) to inform evidence in drug

repurposing from a regulatory perspective.[471]

Chapter 8 presents the outcomes of Chapter 7's PC1 case study (abiraterone → enzalutamide versus enzalutamide → abiraterone) with a tutorial-style implementation guide. This work complements and extends the NICE RWE framework in several ways: First, it underscored the necessity of preliminary data checks, highlighting essential protocol adaptations when data limitations are encountered, and ensuring transparent documentation throughout. These steps are crucial to ensure that the analysis accurately aligns with the original objectives but were, however, not explicitly emphasised in the framework. Preliminary treatment pattern assessments with Sankey diagrams were identified as particularly crucial for treatment sequence analysis. These checks are essential to assess the viability of the study, even with previous sample size estimations made at the design stage. Further, Sankey diagrams also proved to be valuable in reporting the outcomes of treatment sequence analyses and in facilitating communication with clinicians. Moreover, I detailed the preparation of analysis-ready datasets and the implementation of advanced statistical methods, especially the development of statistical models for weighting and outcome assessment. This highlights a crucial, yet commonly underestimated aspect of framework, especially relevant in the analysis with time-varying confounders. Notably, methodological papers typically lack accompanying implementation codes, making this a noteworthy addition.

My case studies mark a valuable addition to the field of TTE and RWE for HTA by being the first attempt to benchmark TTE analyses comparing static treatment sequences. It demonstrates the versatility of the TTE framework in designing studies and applying advanced statistical methods in the context of comparing treatment sequences. Most notably, the PC1 study has demonstrated the ability to derive unbiased estimates of treatment sequence effectiveness using Flatiron data. This was benchmarked against RCT findings that PC1 aimed to replicate. The methodologies developed in this study offer a solid foundation for exploring similar questions across alternative databases, such as NCRAS, and those used internationally.

The findings from my PC1 case study demonstrated that simple methods (ITT, PP, AT) were insufficient for accurately estimating the causal effects of different treatment sequences with RWD, even with adjustments for baseline confounders. In contrast, the advanced IPW method not only improved the estimates of hazard ratios (HR) but also median OS estimates and survival curves, aligning these estimates more closely with those observed in the benchmark GUTG-001 trial.[84] These results were evaluated against the predefined RCT-RWE agreement criteria outlined in Chapter 7, where analyses using the IPW method met all quantitative criteria and showed good agreement with the survival curves of the GUTG-001. In contrast, ITT analysis yielded biased estimates due to not accounting for effects influenced by subsequent treatments. PP analysis, aimed at eliminating these contaminated effects through censoring, introduced bias due to informative censoring. The AT

analyses attempted to eliminate bias by excluding patients who deviated from the assigned treatment sequence, but this introduced immortal time and selection bias by selecting patients based on post-enrolment observations. This led to an incorrect representation of the percentage of patients who did not survive to receive second-line treatments. Despite HRs from the AT analyses aligning with benchmarks, the absolute outcomes are biased.

Equally importantly, my study contributed to the field of TTE benchmarking by introducing additional RTW-RWE agreement criteria, beyond those adopted from the RCT DUPLICATE studies.[172] Specifically, my study went beyond the assessment of relative estimates between emulation and the benchmark. It also investigated whether the emulated absolute outcomes (e.g. median OS) in both treatment groups aligned with those of the benchmark trial and evaluated the visual concordance of survival curves. This comprehensive approach aimed to address potential biases in estimating absolute outcomes, which might not have been identified using standard criteria, such as the consistent immortal bias across groups in AT analyses. This is especially crucial for HTA, given the implications of survival curve extrapolation.[598]

Challenges and limitations in the PC1 study include the lack of exact progression date information, complicating differentiation between treatment switches due to intolerance or disease progression. Nevertheless, Dr. Carmel suggests that early switches typically indicate intolerance, while later switches often suggest disease progression, with few exceptions. Thus, accounting for the timing of treatment switches should theoretically partially address unmeasured confounding due to this lack of information, especially when Eastern Cooperative Oncology Group (ECOG) scores and other lab results highly correlated with disease progression were included in the IPW model. Nevertheless, the analysis was potentially compromised by incomplete and inconsistent covariate information for each patient. Missing variable information was categorised “unknown”, based on the assumption that the reasons for being unknown were similar between the two comparison groups due to their similarities. Nevertheless, future analyses using other methods for dealing with missing data should be sought to test the robustness of the results.

9.3. Key contributions to the field

The key contributions of the thesis are outlined in bullet points, highlighted in bold and italic, with a few explanations provided. These cover various areas, including informing the assessment of treatment sequences in HTA, generating RWE for HTA, and advancing broader insights in TTE study design and benchmarking.

- ***Curated a roadmap of treatment sequence research in HTA***

A significant contribution of my work to the field was organising scattered references about methodological advancements in evaluating treatment sequences in HTA into a structured format

(Chapter 2). This summarised key research themes and highlighted their relationships. This enables researchers with specific interests to easily locate relevant references. The roadmap also visualises the prominent and growing research demands to address the scarcity of effectiveness data and summarises actionable approaches to move forward, with my thesis being one of them.

- ***Summarised the landscape of English HTA involving treatment sequences***

Another contribution of my work was updating the status quo of English HTA involving treatment sequences by offering additional insights beyond previous reviews on treatment-sequencing economic models. Chapter 3 summarised the current landscape, showing that treatment sequences were implicitly mentioned in almost all TAs, even those without an explicit treatment-sequencing model. Importantly, the chapter highlighted substantial disparities in modelling practices and data sources between the two primary disease areas: oncology and autoimmune diseases. Notably, treatment sequence considerations are prevalent in both areas; however, the approaches vary significantly. Pervious research primarily recommended using NMA/MA to derive LOT-specific effectiveness for treatment-sequencing economic models, mainly based on insights from autoimmune disease models[34], and thus may not be fully applicable to the complexities in oncology. Further research suggested using indirect treatment comparison adjustments when merging LOT-specific evidence, but this still did not properly account for the assumption of exchangeability of populations from different LOTs and can be problematic when merging survival evidence.[35]

The chapter therefore reinforced the potential of exploring RWD as an alternative evidences source as RWD are adept at capturing treatment sequences. My review identified a lack of previous TAs utilising advanced causal inference statistical techniques to address the challenges of using RWD in the context of comparing treatment sequences. This highlighted the need for further exploration of this topic, which formed an integral part of my thesis.

- ***Proposed benchmarked methodologies to inform unbiased estimates of comparative treatment effectiveness with RWD for HTA, complementing treatment-sequencing modelling frameworks***

The PC1 Case Study 1 has demonstrated good RCT-RWE agreement with the benchmark it aimed to emulate using Flatiron data, showing consistency in both relative and absolute effectiveness estimates of treatment sequences.[84] The study closely followed the principles of the TTE framework to mitigate biases from RWD study design and employed the advanced IPTW*IPTW$_{txdev}$ method to handle time-varying confounding. Extensive clinical inputs were incorporated to mitigate biases from unmeasured confounding by including pertinent covariates in the statistical model. Additionally, several sensitivity analyses were conducted to examine the robustness of the study findings.

My approach represents a significant step forward in using an alternative approach (i.e., RWD) to inform the effectiveness of treatment sequences in HTA, as population exchangeability

assumptions for existing methods that rely on merging LOT-specific evidence may not always hold. This field was previously underexplored. Hence, my research serves as a valuable starting point for further research in the same line, not only for comparative treatment sequence effects using RWD as a standalone source but also for exploring the utility of deriving an external synthesis controlled arm with relevant treatment sequences. Specifically, insights regarding mitigating biases in real-world study design and advanced statistical methods in comparing treatment sequences are transferable.

- ***Extended the NICE RWE framework for treatment sequence comparisons with a tutorial-style guide for implementation, especially on analytical dataset derivation, advanced IPW implementation, and data check adaptions***

This thesis not only provided insights into using RWD to inform treatment sequence effectiveness in HTA but also extends the NICE RWE framework. Specifically, it offers a tutorial-style guide for implementing TTE studies to compare treatment sequences, including analytical dataset derivation and advanced IPW implementation, along with relevant R code for each step and transparent documentation. The practical aspects, such as the analytical dataset and statistical model for advanced methods, were not detailed in the RWE framework but are crucial in actualising the TTE study planned to match the objectives set. To accommodate specific advanced statistical methods, accuracy in labelling outcome and censoring rows is paramount, as mislabelling could lead to significant errors. This aspect is not often emphasised but is indispensable, especially considering that not all methodological papers provide code and example datasets for verification.

Furthermore, while the framework mentioned that the TTE approach and advanced statistical methods can be used in understanding DTR, the recommendation was very generic. My research sheds light on the nuances and provides a practical example, especially regarding adaptations necessary for comparing treatment sequences. For instance, when assessing the suitability of a RWD source, considerations should be made for treatment patterns in the database population and changes over the years. In preliminary data checks, the use of Sankey diagrams and counts of treatment sequences, as well as the relative proportions of the prevalence of the interested sequences, are crucial for estimating the viable sample size before initiating the treatment sequence analysis. Sankey diagrams were also found to be a valuable visual tool for communicating with clinicians and in finding reporting.

- ***Improved the status quo of RCT-RWE agreement assessment metrics in TTE benchmarking***

My study introduced additional criteria for assessing absolute outcomes in TTE benchmarking, beyond the standard criteria from RCT DUPLICATE studies.[172] Specifically, it evaluated whether the emulated absolute outcomes (e.g., median OS) in both treatment groups aligned with those of the benchmark trial and assessed the visual concordance of survival curves. This approach aimed to

address potential biases in estimating absolute outcomes, especially crucial for HTA due to survival curve implications for extrapolation. This was proven indispensable, especially when biases such as immortal time bias are consistent across study groups, as evident in my AT analyses. Standard criteria on relative outcomes could overlook biased absolute outcome findings, yet still consider the emulation adequate due to acceptance of agreement in relative outcomes. This not only contributes to insights in HTA but also more broadly in the field of benchmarking TTE studies.

- ***Innovative proof-of-concept TTE study design leveraging a mix of direct and indirect benchmarking in the face of scarcity of benchmark trials***

The design of my interconnected case studies uniquely contributed to the field of TTE and benchmarking by leveraging the strengths of two databases to complement their respective weaknesses. Specifically, I introduced a novel approach—indirect benchmarking—to assess the feasibility of using NCRAS for comparing treatment sequences, supported by direct benchmarking with Flatiron data (Case Study PC2-3). Additionally, given the recent partnership between NICE and Flatiron, Flatiron data has gained relevance for English HTA. Hence, direct benchmarking with Flatiron data itself also presents its own distinct value (Case Study PC1). Although this thesis did not delve into the PC2-PC3 case studies due to time constrains, it provides a foundational step forward, as the methodologies have been validated in the PC1 study with Flatiron and can be carried forward for examining the use of NCRAS data for the same purpose. Additionally, another direct planned benchmarking study on renal cell carcinoma (RCC) single-arm with NCRAS data can further complement insights learned from indirect benchmarking. Importantly, these proof-of-concept studies were designed to compare treatment sequences. To my knowledge, no TTE study has attempted to assess uniform treatment sequences for HTA, let alone benchmarking. Therefore, it marks the first benchmark TTE for this purpose, adding valuable insights to the TTE study sphere.

- ***Uncovered further pertinent elements deserving of inclusion in the NICE RWE framework: benchmarking, data checks and necessary protocol adaptions***

One crucial step, not necessarily related to treatment sequences, is benchmarking (or calibration) a given database with relevant benchmark studies before using it to answer further questions. While advocated by a recent study by Gomes et al., this step was not explicitly mentioned in the NICE RWE framework.[78] My Chapters 5 and 6 collectively presented an example of systematic search approach to identifying pertinent benchmark trials. My Chapters 5 and 6 collectively presented a systematic approach to identifying pertinent benchmark trials, providing an example for such searches.

Furthermore, understanding the data prior to TTE design are important, as necessary adaptations may be required to better align with study objectives encountering unexpected data idiosyncrasies. These adjustments may become evident during data checks upon data reception. Transparent

documentation of protocol adaptations is essential. These were highlighted as crucial steps that would affect the emulation in the implementation of my study but were not thoroughly elaborated upon within the NICE RWE framework. This point was underscored in a recent publication by researchers from the RCT DUPLICATE group, stressing the importance of thorough pre-study data checks, carefully considering necessary adaptations at study initiation to avoid biases, and transparent documentation.[599] While some research groups may possess extensive collective experience with similar databases, ensuring better study design, this is not universally the case. Access to databases typically requires a pre-defined protocol approved by ethics committees before data application, posing a dilemma in thorough pre-study data checks. Thus, in my view, it is prudent to plan as much as possible using available published studies and data dictionaries, while transparent documentation of any adaptations along with their justifications, is crucial for upholding research integrity and navigating a pragmatic path forward. My case study provides an example of this, although not as extensively as those in the RCT DUPLICATE group, where each step was registered on clinical.gov. However, it is as transparent as possible.

- ***Systematically identified advanced statistical methods for comparing treatment sequences***

My review systematically identified key advanced statistical methods for comparing treatment sequences. It revealed various advanced statistical methods and provided insights into DTR terminology, relevance to HTA treatment sequences, and different types of treatment sequence estimands. Common methods include the advanced IPW method used in my case study, along with g-formula, SNFTM, and Q-learning. Furthermore, the review reveals the potential adaptation of these methods for analysing trial data involving both randomised and non-randomised subsequent treatments, thereby opening avenues for further exploration in the same line of research.

9.4. Strength and limitations

Several limitations of the study should be acknowledged. Firstly, the restricted sample size in the benchmark GUTG-001 trial may have reduced the challenge of attaining RCT-RWE agreement. However, it was evident that advanced methods produced results closer to theoretical correctness compared to those of simple methods, which were flawed. Thus, it highlights the appropriate use of IPW methods. Secondly, time constraints hindered a systematic review of covariates from the literature and restricted the exploration of a full spectrum of sensitivity analyses to test the robustness of the statistical methods applied. However, considerable effort was invested in reviewing literature to identify pertinent covariates, as well as iterative discussions with clinicians to facilitate the elicitation of pertinent covariates for statistical models. Additionally, several critical sensitivity analyses were conducted, demonstrating robust results. Thirdly, the study faced challenges with inconsistent missing data across patients, underscoring the necessity for further investigation to tackle

this. However, due to time constraints and the complexity of employing techniques such as multiple imputation in causal research (an evolving area in the literature itself), addressing this issue was deemed beyond the scope of the thesis. Fourthly, only one advanced statistical method was examined. Nevertheless, its prioritisation was supported by a strong justification as a logical first step in this line of research due to its relevance in HTA. However, efforts should certainly be made to compare the IPW with other available methods, notably the g-formula and Q-learning.

There are several strengths of my research. Its greatest strength lies in being the first attempt to conduct a TTE benchmark study for comparing uniform treatment sequences for HTA. Specifically, the design of the case study was supported by a series of systematic reviews, guiding the selection of suitable databases, advanced analytical methods, and benchmarks. Secondly, the innovative design of interconnected case studies enabled benchmarking with both Flatiron and NCRAS data. Specifically, the utilisation of Flatiron was benchmarked for its ability to inform unbiased estimates of treatment sequences, assessed with pre-defined RCT-RWE agreement assessment metrics. The emulated PC1 study demonstrates good agreement with the benchmark trial GUTG-001. Significantly, I adopted improved assessment metrics, which went beyond the status quo by also examining absolute outcome agreement, including survival curves concordance, which has implications in HTA for survival extrapolation. Furthermore, these investigations were guided by a set of systematic reviews, with oncology being identified as the most pertinent disease area in treatment sequencing that research should prioritise due to caveats in merging evidence for survival outcomes. The literature roadmap of treatment sequence methodology studies in HTA revealed a growing body of literature highlighting the scarcity of evidence. It was also identified that research on unbiased RWE to inform the effectiveness of treatment sequences was severely underexplored but holds significant value. Finally, my research provides a tutorial-style guide for TTE implementation, complementing and extending the practical aspects of NICE RWE and related research in this area.

9.5. Areas for future research

(1) Examine alternative advanced statistical methods: Additional advanced statistical methods, particularly g-formula and Q-learning, may be further explored in comparison with the IPTW*IPCW$_{texdev}$ approach.

(2) Survival curve extrapolation with parametric models for HTA: Investigating the use of parametric models, such as Weibull, for extrapolation of the survival curves in the current emulation is a crucial step forward to demonstrate the practicality of estimates derived from RWD to be used in HTA. Specifically, an initial step forward could be using validated methodologies to emulate survival curves for treatment sequences in TA377 (prostate cancer TA: abiraterone → docetaxel versus enzalutamide → docetaxel) using Flatiron data with parametric model extrapolation and

comparing them with those originally used in company submission models.[126] Pooled logistic models were tested for their capability to account for non-proportional hazards but not reported in the thesis. The survival curves with bootstrap generally showed narrower 95% CI, potentially due to the large sample size and additional assumptions of the parametric model. Hence, further investigation into this area with parametric model extrapolation with RWE for treatment sequences is warranted.

(3) Sensitivity analysis: Further exploration of sensitivity analyses is warranted to examine the robustness of the methods. Key sensitivity analysis include examining the impact of unmeasured confounders using E-values[595], implementing varying Time Zero with sequential Target Trial emulation to assess potential immortal time bias, and multiple imputation for missing data. Each of these avenues requires substantial planning and merits separate research efforts.

(4) Simulation study: Given the limited sample sizes and specific settings of the benchmark trials, simulation studies offer a complementary alternative to RCT-RWE agreement assessments for examining the performance of statistical methods for comparing treatment sequences. The advantage of simulation studies lies in their ability to provide an understanding of the behaviour of statistical methods. This is achieved by analysing the simulated data with the statistical methods of interest and comparing the results they provide to the empirical "true answers", which are known due to the pre-specified data generation mechanism. Additionally, simulation studies can be conducted across diverse scenarios, allowing researchers to test method performance by varying one or more study conditions to assess their sensitivity and limitations.

While Huang et al. have conducted a simulation study on comparing treatment sequences with advanced IPW methods and demonstrated good agreement between results they obtained using advanced IPW methods and the empirical outcomes of the simulated data[289] (Section 4.6.1.2.1, Chapter 4), further simulation studies with varying assumptions could examine the robustness of the results. For instance, exploring varying proportions of patients censored due to treatment deviation could result in different weights and potentially influence the performance of the IPW method.

Here, I describe several potential applications of simulation studies in evaluating statistical methods for comparing treatment sequences. To assess the performance of statistical methods for comparing treatment sequences in RWD, pseudo-RWD datasets that mimic real-world treatment sequences first need to be simulated. These datasets would be generated with known relationships between baseline confounders, time-varying confounders, treatments, and outcomes (i.e., no unmeasured confounders). Specific treatment changes would be determined by functions of baseline and time-dependent characteristics, which are themselves influenced by treatment. These characteristics, along with the treatments, would affect survival times, thereby making the time-

dependent characteristics serve as time-dependent confounders. One can then analyse these datasets using the statistical methods of interest and evaluate them with performance metrics, such as percentage bias in estimating specific estimands (e.g. restricted mean survival times if all patients had received the same treatment sequence). This helps us determine how close the results from the tested methods are to the true answers. This approach can be used to test the relative performance of a range of statistical methods across a range of scenarios

A key assumption of causal inference methods used to analyse observational data is that there is no unmeasured confounding. Simulation studies could be used to examine how sensitive these methods are to the impact of unmeasured confounders. For example, one could test how a method performs when all known confounders are included in the analysis versus omitting one or two. One can also generate multiple simulated datasets with varying confounder relationships and strengths to test the method's behaviour under different conditions of unmeasured confounders.

Simulation studies may also be used to explore how a method performs under different levels of censoring due to treatment deviation. For example, IPCW-related methods are known to be sensitive to high censoring levels in other scenarios.[43] This evaluation is particularly relevant for assessing method performance when only small sample sizes have followed a complete treatment sequence of interest (e.g., sunitinib → everolimus in the RCC study mentioned in Section 8.2.3), compared to the total number of patients who received the same first-line treatment. That is, estimating the marginal effect of that treatment sequence will result in high censoring rates due to deviations at the second-line. Testing the impact on method performance of varying censoring levels is unlikely to be possible when using RCT-RWE benchmark agreement assessments, as this would require the existence of multiple benchmark RCTs with varying censoring levels. The same is true for other dataset characteristics— in a benchmark RCT setting, only a limited number of RCTs may be available, but in a simulation study, an analyst can investigate any scenario of interest.

Furthermore, simulation studies may also be used to evaluate how different methods (e.g., g-formula, Q-learning) perform under the same and varying conditions (e.g., different censoring rates) and determine which methods may produce less biases in specific scenarios.

To operationalise a simulation study for evaluating statistical methods, here are the key steps to consider in designing and reporting, adhering to the ADEMP framework (aims; data-generation; estimands; methods; performance measures), as summarised by Morris et al.[600]

- Aim: Establish a specific objective, such as evaluating the performance of methods for estimating the relative effectiveness of different treatment sequences in a real-world setting.
- Data-generating mechanisms: Determine the relationships to be simulated between confounders, treatments, and outcomes. Outline scenarios for investigation and report the

parameters and assumptions used in the simulation process. For example, one option for creating pseudo-RWD with treatment sequences is using the SNFTM, as used by Huang et al. in their simulation study.[289]

- Methods: Select the statistical method(s) to be evaluated.
- Estimands: Clearly define measures to be estimated in the study.
- Performance measures: Determine how to assess the performance of each method by comparing simulated results with known empirical outcomes.

Although simulation studies have advantages in assessing the performance of statistical methods, there are also caveats that need to be acknowledged. One major limitation of simulation studies is that the data are constructed in a way that the relationships between confounders, treatments, and outcomes without loss-to-follow-up (or mechanisms of loss-to-follow-up) are known. In RWD, the performance of methods may vary due to additional complexities and variations. Furthermore, simulating pseudo-RWD with treatment sequences presents particular challenges. Defining all the relationships between prognostic factors, patients' likelihood of receiving each line of treatment, and timing of treatment switching can be complex. Starting with simulation studies that have limited treatment options and specific time points for patient switching may provide a foundation, though these scenarios are not entirely realistic.

Despite these challenges, simulation studies are valuable for understanding the nuanced properties of methods. They are particularly useful for testing method performance that would be difficult to assess with a single source of RWD (e.g., varying censoring rates). Thus, simulation studies, when coupled with validation studies using RWD, provide a comprehensive approach to statistical method evaluation.

(5) Single-arm treatment sequencing evidence: Exploring the same methodology in emulating single-arm treatment sequences holds significance. For example, my direct benchmarking RCC case study with NCRAS can be taken forward. Furthermore, investigating the feasibility of an external control arm to compare with those in TA377 (prostate cancer TA: abiraterone → docetaxel versus enzalutamide → docetaxel) could be a step even further (using steps planned in generalisbility PC2). However, additional methods such as indirect treatment comparisons or transportability methods merit separate study.

(6) Methodology application in other diseases and public health: The developed methodologies have the potential to be extrapolated for investigating treatment sequencing effects in other contexts. Given the complexities of interventions and outcome measurement in alternative areas, these investigations merit separate research, including autoimmune diseases (e.g., comparing sequences with numerous LOTs), rare diseases (e.g., dealing with small sample sizes even with RWD), or

public health interventions (e.g., navigating intricate smoking cessation strategies), where evidence on sequential interventions is also likely scarce but relevant.

9.6. Recommendations

Insights gleaned from the thesis give rise to following recommendations:

- Research aiming to use RWD in assessing the comparative effectiveness of treatment sequences for HTA would benefit from adopting the implementation tutorial developed within the thesis (Chapter 8). This tutorial leveraged the TEE framework and the IPTW*IPTWtxdev method to mitigate biases arising from RWD study designs and time-varying confounding.
- Benchmarking is considered a valuable approach before utilising a given RWD database to inform HTA decisions, particularly in cases involving advanced statistical methods that have not been previously examined.
- When benchmarking RCT-RWE comparisons, it is essential to consider additional criteria for assessing absolute outcome results against the RCT benchmark, in addition to evaluating the standard criteria for relative effect estimates. Survival curves should also be examined where applicable
- Sankey diagrams should be integrated into data checks, study result reports, and communications with clinicians for evaluating treatment sequences.
- Merely designing TTE and registering a pre-defined protocol does not guarantee unbiased real-world study results. It is crucial to conduct thorough fit-for-purpose data assessment during pre-study design, perform data checks upon receipt of data, implement necessary adaptations, and ensure transparent documentation.

9.7. Conclusions

In summary, this thesis has consolidated methodological literature on treatment sequences in HTA and highlights the significant issue of data scarcity. While consideration of treatment sequences is nearly ubiquitous in HTA and often pertinent in the development of clinical guidelines, explicit comparisons of treatment sequences remain relatively uncommon despite the prevalence of treatment-sequencing models. The standard approach of merging effectiveness evidence from different LOTs often relies on exchangeability assumptions across populations, which can be particularly problematic in oncology when merging survival evidence. RWE emerges as a valuable alternative to address these challenges, but often suffers from confounding bias.

This thesis demonstrates how RWE can enrich the existing health economic treatment-sequencing modelling framework. Specifically, through carefully designed and implemented TTE case studies, it illustrates how RWD, combined with causal inference methods, can be leveraged to provide unbiased effectiveness estimates for comparing treatment sequences. It extends the NICE's

RWE framework by providing specific insights into comparing treatment sequences, a detail not extensively outlined in the framework. Notably, my benchmark study comparing the sequence of abiraterone → enzalutamide versus enzalutamide → abiraterone in prostate cancer has shown good agreement with the benchmark RCT it aimed to replicate (i.e., the GUTG-001 trial).[84] In addition to the standard criteria for assessing RCT-RWE agreement in TTE benchmarking studies, this study went beyond the status quo by not only evaluating the agreement of relative effect estimates between the emulation and the benchmark, but also scrutinised the concordance of absolute estimates and survival curves. These bear significant implications in HTA, ensuring accurate survival extrapolations and treatment duration estimates for cost calculations.

Crucially, my thesis offers a tutorial-style guide for designing and implementing the aforementioned benchmarked methodology. Importantly, success in TTE does no solely hinge on study designs and its execution fully adhered to the pre-defined protocol. It requires careful, necessary, and reasonable adaption of study protocols based on data checks, and specific data features pertinent to a given database. My case study implementations provide examples of such adaptations using Flatiron data. In light of NICE's piloting of the Pathway project[27], the importance of evidence regarding treatment sequences is expected to increase. My studies offer a comprehensive exploration of how RWE can be utilised to address evidence scarcity in evaluating treatment sequences and provide a practical way forward.

References

1. Zheng Y, Pan F, Sorensen S. Modeling Treatment Sequences in Pharmacoeconomic Models. *Pharmacoeconomics.* 2017;35(1):15-24.
2. Viola MG, Diamantopoulos A. Determining the Baseline Strategy in a Cost-Effectiveness Analysis with Treatment Sequences. *Appl Health Econ Health Policy.* 2020;18(1):17-29.
3. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol.* 2016;183(8):758-764.
4. Chakraborty B, Murphy SA. Dynamic treatment regimes. *Annual review of statistics and its application.* 2014;1:447.
5. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: FL. Chapman & Hall/CRC; 2020.
6. Kaner EF, Beyer FR, Garnett C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. *Cochrane Database of Systematic Reviews.* 2017(9).
7. Bush T, Lovejoy J, Javitz H, et al. Comparative effectiveness of adding weight control simultaneously or sequentially to smoking cessation quitlines: study protocol of a randomized controlled trial. *BMC Public Health.* 2016;16(1):1-12.
8. Tosh J, Stevenson M, Akehurst R. Health Economic Modelling of Treatment Sequences for Rheumatoid Arthritis: A Systematic Review. *Curr Rheumatol Rep.* 2014;16(10):447.
9. Al M, Jin H, Robins J, Sculpher M. Virtual ISPOR Europe 2020: Novel Methods for Modeling Treatment Sequencing (Spotlight session 3). 2020; https://www.ispor.org/conferences-education/conferences/past-conferences/ispor-europe-2020/program/spotlight-sessions. Accessed Nov 19, 2020.
10. Benedict A, Stevenson M, Sorensen S. ISPOR Europe 2017: Modeling Separate Lines of Treatment Versus Treatment Sequences in Cancer (Workshop 8). 2017; https://www.ispor.org/docs/default-source/presentations/1221.pdf?sfvrsn=6f1f29a1_1. Accessed June 20, 2020.
11. Mathes T, Jacobs E, Morfeld J-C, Pieper D. Methods of international health technology assessment agencies for economic evaluations- a comparative analysis. *BMC Health Serv Res.* 2013;13(1):371.
12. Vreman RA, Mantel-Teeuwisse AK, Hövels AM, Leufkens HGM, Goettsch WG. Differences in Health Technology Assessment Recommendations Among European Jurisdictions: The Role of Practice Variations. *Value Health.* 2020;23(1):10-16.
13. Allen N, Walker SR, Liberti L, Salek S. Health Technology Assessment (HTA) Case Studies: Factors Influencing Divergent HTA Reimbursement Recommendations in Australia, Canada, England, and Scotland. *Value Health.* 2017;20(3):320-328.
14. Kamae I, Thwaites R, Hamada A, Fernandez JL. Health technology assessment in Japan: a work in progress. *J Med Econ.* 2020;23(4):317-322.
15. Oortwijn W, Mathijssen J, Banta D. The role of health technology assessment on pharmaceutical reimbursement in selected middle-income countries. *Health Policy.* 2010;95(2):174-184.

16. Allen N, Liberti L, Walker SR, Salek S. A comparison of reimbursement recommendations by European HTA agencies: is there opportunity for further alignment? *Front Pharmacol.* 2017;8:384.
17. O'Donnell JC, Pham SV, Pashos CL, Miller DW, Smith MD. Health technology assessment: lessons learned from around the world—an overview. *Value Health.* 2009;12:S1-S5.
18. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes.* Oxford university press; 2015.
19. Clement FM, Harris A, Li JJ, Yong K, Lee KM, Manns BJ. Using Effectiveness and Cost-effectiveness to Make Drug Coverage Decisions: A Comparison of Britain, Australia, and Canada. *JAMA.* 2009;302(13):1437-1443.
20. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes (4th ed.).* Oxford: Oxford University Press; 2015.
21. NICE. Briefing paper for methods review working party on choosing comparators. 2012; https://www.sheffield.ac.uk/media/34080/download?attachment. Accessed Jan 26, 2023.
22. NICE. Briefing paper for methods review working party on treatment sequences and downstream costs 2012; https://www.sheffield.ac.uk/media/34080/download?attachment. Accessed Jan 26, 2023.
23. NICE. NICE Process and methods: Guide to the methods of technology appraisal 2013 (PMG 9). 2013; https://www.nice.org.uk/process/pmg9/chapter/foreword. Accessed June 2, 2020.
24. NICE. NICE health technology evaluations: the manual (PMG 36). 2022; https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation. Accessed Aug 17, 2023.
25. Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the Economic Evaluation of Health Technologies: Canada — 4th Edition. 2017; https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf. Accessed March 11, 2023.
26. NICE. History of NICE. 2020; https://www.nice.org.uk/about/who-we-are/history-of-nice. Accessed June 2, 2020.
27. NICE. Taking a proportionate approach to technology appraisals. 2023; https://www.nice.org.uk/about/what-we-do/proportionate-approach-to-technology-appraisals. Accessed Jan 26, 2024.
28. NICE. Renal cell carcinoma Pathways Pilot [ID6186]. 2024; https://www.nice.org.uk/guidance/indevelopment/gid-ta11186. Accessed April 5, 2024.
29. NICE. Treatments for non-small-cell lung cancer [ID6234]. 2024; https://www.nice.org.uk/guidance/indevelopment/gid-ta11289. Accessed April 5, 2024.
30. Tosh J. *Simulation optimisation to inform economic evaluations of sequential therapies for chronic conditions: a case study in rheumatoid arthritis [Doctoral Thesis]* [thesis], University of Sheffield; 2015.
31. Kim E. *Sequential drug decision problems in long-term medical conditions: a case study of primary hypertension [Doctoral Thesis]* [thesis], University of Sheffield; 2015.

32. Navarro F, Martinez-Sesmero J, Balsa A, et al. Cost-effectiveness analysis of treatment sequences containing tofacitinib for the treatment of rheumatoid arthritis in Spain. *Clin Rheumatol.* 2020;39(10):2919-2930.
33. Incerti D, Jansen JP. The Innovation and Value Initiative's (IVI's) non-small cell lung cancer (NSCLC) simulation model (the IVI-NSCLC model). 2019; https://innovationvalueinitiative.github.io/IVI-NSCLC/model-doc/model-doc.pdf. Accessed May 18, 2020.
34. Lewis RA, Hughes D, Sutton AJ, Wilkinson C. Quantitative Evidence Synthesis Methods for the Assessment of the Effectiveness of Treatment Sequences for Clinical and Economic Decision Making: A Review and Taxonomy of Simplifying Assumptions. *Pharmacoeconomics.* 2021;39:25-61.
35. Huang M, Ramsey S, Xue W, Xie J, Pellissier J, Briggs A. Conceptual Framework and Methodological Challenges for Modeling Effectiveness in Oncology Treatment Sequence Models. *Pharmacoeconomics.* 2022;40:257-268.
36. Lewis R. *Quantitative evidence synthesis methods for the assessment of the effectiveness of treatment sequences for clinical and economic decision-making [Doctoral Thesis]* [thesis], Bangor University; 2019.
37. National Institute for Health and Care Excellence. NICE real-world evidence framework [ECD 9]. 2021; https://www.nice.org.uk/corporate/ecd9/chapter/overview. Accessed June 30, 2023.
38. Pazzagli L, Linder M, Zhang M, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. *Pharmacoepidemiol Drug Saf.* 2018;27(2):148-160.
39. Bell H, Wailoo A, Hernandez M, et al. The use of real world data for the estimation of treatment effects in NICE decision making (report by the NICE Decision Support Unit). 2016; http://nicedsu.org.uk/wp-content/uploads/2018/05/RWD-DSU-REPORT-Updated-DECEMBER-2016.pdf. Accessed June 2, 2020.
40. Faria R, Hernadez Alava M, Manca A, Wailoo A. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. 2015; http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD17-DSU-Observational-data-FINAL.pdf. Accessed 2020, June 2.
41. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. Criteria for distinguishing effectiveness from efficacy trials in systematic reviews. 2010.
42. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med.* 2009;151(3):203-205.
43. Latimer NR, Abrams KR. NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching. 2014; http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD16_Treatment_Switching.pdf. Accessed June 2, 2020.
44. Welton N, Phillippo D, Owen R, et al. CHTE2020 sources and synthesis of evidence; update to evidence synthesis methods. *NICE Decision Support Unit, Sheffield, UK* 2020; https://rees-

france.com/wp-content/uploads/2020/12/CHTE-2020_synthesis-of-evidence.pdf. Accessed Aug 29, 2022.

45. Gangl M. Causal inference in sociological research. *Annu Rev Sociol.* 2010;36:21-47.
46. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. 2010.
47. Pearl J. Causal Inference: history, perspectives, adventures, and unification (an interview with Judea Pearl). *Observational Studies.* 2022;8(2):1-14.
48. Keele L. The statistics of causal inference: A view from political methodology. *Political Analysis.* 2015;23(3):313-335.
49. Hoover KD. Economic theory and causal inference. *Philosophy of economics.* 2012;13:89-113.
50. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association.* 1986;81(396):945-960.
51. Pearl J, Glymour M, Jewell NP. *Causal inference in statistics: A primer.* John Wiley & Sons; 2016.
52. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health.* 2005;95(S1):S144-S150.
53. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* 2006;17(4):360-372.
54. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling.* 1986;7(9):1393-1512.
55. Shinozaki E, Makiyama A, Kagawa Y, et al. Treatment sequences of patients with advanced colorectal cancer and use of second-line FOLFIRI with antiangiogenic drugs in Japan: A retrospective observational study using an administrative database. *PLoS One.* 2021;16(2):e0246160.
56. Hochfelder CG, McGinn AP, Mehta V, Castellucci E, Kabarriti R, Ow TJ. Treatment sequence and survival in locoregionally advanced hypopharyngeal cancer: A surveillance, epidemiology, and end results–based study. *The Laryngoscope.* 2020;130(11):2611-2621.
57. Girard N, Moro-Sibilot D, Bouée S, et al. Treatment sequence of first and second generation tyrosine kinase inhibitor followed by osimertinib in EGFR-mutated non-small-cell lung cancer: a real life study. *Future Oncol.* 2020;16(16):1115-1124.
58. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550-560.
59. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health.* 2018;108(5):616-619.
60. Robins J. Interview with Jamie Robins. *Observational Studies.* 2022;8(2):37-76.
61. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. *Ann Agricultural Sciences.* 1923:1-51.
62. Imbens GW, Rubin DB. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press; 2015.

63. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol.* 1974;66(5):688.
64. Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics.* 1978:34-58.
65. Rubin DB. Matching to remove bias in observational studies. *Biometrics.* 1973:159-183.
66. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS.* 1989:113-159.
67. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology.* 1992;3(4):319-336.
68. Robins JM. Association, causation, and marginal structural models. *Synthese.* 1999;121(1/2):151-179.
69. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association.* 2001;96(454):440-448.
70. Hernán MA, Heeg B, Thorlund K. Virtual ISPOR 2020: Head-to-head comparisons using real world data (RWD) - is the era of network meta analysis (NMA) over? (Issue Panel 5). 2020; https://www.ispor.org/heor-resources/presentations-database/presentation/intl2020-3196/11698. Accessed June 5, 2020.
71. Pearce N, Vandenbroucke JP. Are Target Trial Emulations the Gold Standard for Observational Studies? *Epidemiology.* 2023;34(5):614-618.
72. De Stavola BL, Gomes M, Katsoulis M. Transparency and Rigor: Target Trial Emulation Aims to Achieve Both. *Epidemiology.* 2023;34(5):624-626.
73. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Emulating a target trial in case-control designs: an application to statins and colorectal cancer. *Int J Epidemiol.* 2020;49(5):1637-1646.
74. Stürmer T, Wang T, Golightly YM, Keil A, Lund JL, Jonsson Funk M. Methodological considerations when analysing and interpreting real-world data. *Rheumatology (Oxford).* 2020;59(1):14-25.
75. Maringe C, Benitez Majano S, Exarchakou A, et al. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *Int J Epidemiol.* 2020;49(5):1719-1729.
76. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol.* 2016;79:70-75.
77. Kutcher SA, Brophy JM, Banack HR, Kaufman JS, Samuel M. Emulating a randomised controlled trial with observational data: an introduction to the target trial framework. *Can J Cardiol.* 2021;37(9):1365-1377.

78. Gomes M, Latimer N, Soares M, et al. Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *Pharmacoeconomics.* 2022;40(6):577-586.
79. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2019; https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Accessed April 6, 2024.
80. Kahan BC, Hindley J, Edwards M, Cro S, Morris TP. The estimands framework: a primer on the ICH E9 (R1) addendum. *BMJ.* 2024;384.
81. Luijken K, van Eekelen R, Gardarsdottir H, Groenwold RH, van Geloven N. Tell me what you want, what you really really want: estimands in observational pharmacoepidemiologic comparative effectiveness and safety studies. *Pharmacoepidemiol Drug Saf.* 2023;32(8):863-872.
82. England PH. A guide to NCRAS data and its availability. 2020; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1073359/NCRAS_data_and_its_availability-October_2020.pdf. Accessed Apr 30, 2022.
83. Mpofu P, Kent S, Jónsson P, et al. Evaluation of US oncology electronic health record real-world data to reduce uncertainty in health technology appraisals: a retrospective cohort study. *BMJ open.* 2023;13(10):e074559.
84. Khalaf DJ, Annala M, Taavitsainen S, et al. Optimal sequencing of enzalutamide and abiraterone acetate plus prednisone in metastatic castration-resistant prostate cancer: a multicentre, randomised, open-label, phase 2, crossover trial. *The Lancet Oncology.* 2019;20(12):1730-1739.
85. Flatiron Health. Flatiron Health: Publications. 2020; https://flatiron.com/publications/. Accessed June 2, 2020.
86. Knox JJ, Barrios CH, Kim TM, et al. Final overall survival analysis for the phase II RECORD-3 study of first-line everolimus followed by sunitinib versus first-line sunitinib followed by everolimus in metastatic RCC. *Ann Oncol.* 2017;28(6):1339-1345.
87. Chang J-YA, Chilcott JB, Latimer NR. Challenges and Opportunities in Interdisciplinary Research and Real-World Data for Treatment Sequences in Health Technology Assessments. *Pharmacoeconomics.* 2024.
88. NICE. Position statement: consideration of products recommended for use in the Cancer Drugs Fund as comparators, or in a treatment sequence, in the appraisal of a new cancer product. 2019; https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/cancer-drugs-fund/CDF-comparator-position-statement.pdf. Accessed Mar 23, 2020.
89. Claxton K. Exploring uncertainty in cost-effectiveness analysis. *Pharmacoeconomics.* 2008;26:781-798.
90. Kaufman HL, Atkins MB, Subedi P, et al. The promise of Immuno-oncology: implications for defining the value of cancer treatment. *Journal for immunotherapy of cancer.* 2019;7:1-11.

91. Chehayeb RJ, Hood A, Wang X, et al. Treatment Sequencing Patterns and Associated Direct Medical Costs of Metastatic Breast Cancer Care in the United States, 2011 to 2021. *JAMA Network Open.* 2022;5(11):e2244204-e2244204.
92. Roux J, Grimaud O, Leray E. Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis. *Stat Methods Med Res.* 2019;28(6):1651-1663.
93. Kim AH, Jang JE, Han J. Current status on the therapeutic strategies for heart failure and diabetic cardiomyopathy. *Biomed Pharmacother.* 2022;145:112463.
94. Kern DM, Cepeda MS, Defalco F, Etropolski M. Treatment patterns and sequences of pharmacotherapy for patients diagnosed with depression in the United States: 2014 through 2019. *BMC Psychiatry.* 2020;20(1):1-10.
95. Faria R. Problem Structuring in Economic Evaluation. *In: Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press2021.
96. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ.* 2005;331(7524):1064-1065.
97. Schlosser RW, Wendt O, Bhavnani S, Nail-Chiwetalu B. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. *Int J Lang Commun Disord.* 2006;41(5):567-582.
98. Hernandez L, O'Donnell M, Postma M. Modeling Approaches in Cost-Effectiveness Analysis of Disease-Modifying Therapies for Relapsing-Remitting Multiple Sclerosis: An Updated Systematic Review and Recommendations for Future Economic Evaluations. *Pharmacoeconomics.* 2018;36(10):1223-1252.
99. Jansen JP, Incerti D, Mutebi A, et al. Cost-effectiveness of sequenced treatment of rheumatoid arthritis with targeted immune modulators. *J Med Econ.* 2017;20(7):703-714.
100. Badampudi D, Wohlin C, Petersen K. Experiences from using snowballing and database searches in systematic literature studies. Paper presented at: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE)2015; Nanjing, China.
101. Zwakman M, Verberne LM, Kars MC, Hooft L, van Delden JJM, Spijker R. Introducing PALETTE: an iterative method for conducting a literature search for a review in palliative care. *BMC Palliat Care.* 2018;17(1):82.
102. Haywood P. *Increasing complexity of cancer care: How displaced treatments impact efficiency, cost-effectiveness and equity [Doctoral Thesis]* [thesis], University of Technology Sydney; 2018.
103. Hawkins N, Sculpher M, Epstein D. Cost-effectiveness analysis of treatments for chronic disease: using R to incorporate time dependency of treatment response. *Med Decis Making.* 2005;25(5):511-519.
104. Saramago P, Manca A, Sutton AJ. Deriving input parameters for cost-effectiveness modeling: taxonomy of data types and approaches to their statistical synthesis. *Value Health.* 2012;15(5):639-649.

105. Tordrup D, Chouaid C, Cuijpers P, et al. Priorities for Health Economic Methodological Research: Results of an Expert Consultation. *Int J Technol Assess Health Care.* 2017;33(6):609-619.
106. Heeg BMS. Developing health economic models of chronic diseases for reimbursement purposes [Doctoral Thesis]. *University of Groningen: Groningen, The Netherlands.* 2015.
107. Brixner DI, Watkins JB. Can CER be an effective tool for change in the development and assessment of new drugs and technologies? *J Manag Care Pharm.* 2012;18(5 Supp A):S06-S11.
108. Tappenden P, Chilcott J, Brennan A, Squires H, Stevenson M. Whole disease modeling to inform resource allocation decisions in cancer: a methodological framework. *Value Health.* 2012;15(8):1127-1136.
109. Lord J, Willis S, Eatock J, et al. Economic modelling of diagnostic and treatment pathways in National Institute for Health and Care Excellence clinical guidelines: the Modelling Algorithm Pathways in Guidelines (MAPGuide) project. *Health Technol Assess.* 2013;17(58):1-192.
110. Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. *Health Econ.* 2006;15(12):1295-1310.
111. Rauner MS, Gutjahr WJ, Heidenberger K, Wagner J, Pasia J. Dynamic policy modeling for chronic diseases: metaheuristic-based identification of pareto-optimal screening strategies. *Oper Res.* 2010;58(5):1269-1286.
112. Cressman S, Browman GP, Hoch JS, Kovacic L, Peacock SJ. A Time-Trend Economic Analysis of Cancer Drug Trials. *Oncologist.* 2015;20(7):729-736.
113. de Mello-Sampayo F. The timing and probability of treatment switch under cost uncertainty: an application to patients with gastrointestinal stromal tumor. *Value Health.* 2014;17(2):215-222.
114. Vernon JA, Hughen WK. A Primer on Dynamic Optimization and Optimal Control in Pharmacoeconomics. *Value Health.* 2006;9(2):106-113.
115. Chick SE, Inoue K. New two-stage and sequential procedures for selecting the best simulated system. *Oper Res.* 2001;49(5):732-743.
116. Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology.* 2007;32(2):257-262.
117. Cain LE, Saag MS, Petersen M, et al. Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. *Int J Epidemiol.* 2016;45(6):2038-2049.
118. Messner DA. Evaluating the comparative effectiveness of treatment sequences in oncology: a new approach. *Journal of Comparative Effectiveness Research.* 2015;4(6):537-540.
119. Dekker J, Van HL, Hendriksen M, et al. What is the best sequential treatment strategy in the treatment of depression? Adding pharmacotherapy to psychotherapy or vice versa? *Psychother Psychosom.* 2013;82(2):89-98.
120. Weintraub M, Waldman E, Koplewitz B, et al. A sequential treatment algorithm for infants with stage 4s neuroblastoma and massive hepatomegaly. *Pediatr Blood Cancer.* 2012;59(1):182-184.

121. Trappe R, Oertel S, Leblond V, et al. Sequential treatment with rituximab followed by CHOP chemotherapy in adult B-cell post-transplant lymphoproliferative disorder (PTLD): the prospective international multicentre phase 2 PTLD-1 trial. *The lancet oncology.* 2012;13(2):196-206.
122. Grigor C, Capell H, Stirling A, et al. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial. *The Lancet.* 2004;364(9430):263-269.
123. Wang L, Rotnitzky A, Lin X, Millikan RE, Thall PF. Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer. *Journal of the American Statistical Association.* 2012;107(498):493-508.
124. Briggs A, Sidhu M, Baker T. A framework for analysing treatment sequences: Incorporating time dependent transitions through partitioned survival analysis. *Value Health.* 2014;17(3):A204-A205.
125. Simpson A, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to HTA: part 5. *Journal of Comparative Effectiveness Research.* 2022;11(4):213-215.
126. NICE. Enzalutamide for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated (TA 377). 2016; https://www.nice.org.uk/guidance/ta377. Accessed Apr 30, 2021.
127. NICE. Abiraterone for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated (TA 387). 2016; https://www.nice.org.uk/guidance/ta387. Accessed Apr 30, 2021.
128. Pan F, Reifsnider O, Zheng Y, et al. Modeling clinical outcomes in prostate cancer: application and validation of the discrete event simulation approach. *Value Health.* 2018;21(4):416-422.
129. Parikh RC, Du XL, Robert MO, Lairson DR. Cost-effectiveness of treatment sequences of chemotherapies and targeted biologics for elderly metastatic colorectal cancer patients. *Journal of Managed Care & Specialty Pharmacy.* 2017;23(1):64-73.
130. Rognoni C, Ciani O, Sommariva S, Tarricone R. Cost–effectiveness analysis of treatments involving radioembolization in intermediate-stage hepatocellular carcinoma. *Journal of Comparative Effectiveness Research.* 2018;7(3):209-221.
131. Woods BS, Sideris E, Palmer S, Latimer N, Soares M. Partitioned survival and state transition models for healthcare decision making in oncology: where are we now? *Value Health.* 2020;23(12):1613-1621.
132. Spelman T, Magyari M, Piehl F, et al. Treatment escalation vs immediate initiation of highly effective treatment for patients with relapsing-remitting multiple sclerosis: data from 2 different national strategies. *JAMA neurology.* 2021;78(10):1197-1204.
133. Simonsen CS, Flemmen HØ, Broch L, et al. Early high efficacy treatment in multiple sclerosis is the best predictor of future disease activity over 1 and 2 years in a Norwegian population-based registry. *Front Neurol.* 2021;12:693017.
134. Brown JWL, Coles A, Horakova D, et al. Association of initial disease-modifying therapy with later conversion to secondary progressive multiple sclerosis. *JAMA.* 2019;321(2):175-187.

135. He A, Merkel B, Brown JWL, et al. Timing of high-efficacy therapy for multiple sclerosis: a retrospective observational cohort study. *The Lancet Neurology.* 2020;19(4):307-316.
136. Vertosick EA, Assel M, Vickers AJ. A systematic review of instrumental variable analyses using geographic region as an instrument. *Cancer Epidemiol.* 2017;51:49-55.
137. National Institute for Clinical Excellence. Capecitabine and oxaliplatin in the adjuvant treatment of stage III (Dukes' C) colon cancer (Technology Appraisal 100). 2006; https://www.nice.org.uk/guidance/ta100. Accessed Jun 30, 2020.
138. Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. *Health Technol Assess.* 2004;8(11):iii, 1-91.
139. Tran-Duy A, Boonen A, van de Laar MA, Franke AC, Severens JL. A discrete event modelling framework for simulation of long-term outcomes of sequential treatment strategies for ankylosing spondylitis. *Ann Rheum Dis.* 2011;70(12):2111-2118.
140. Tappenden P. Conceptual modelling for health economic model development. *HEDS Discussion Paper* 2012; https://eprints.whiterose.ac.uk/74464/1/HEDSDP1205.pdf. Accessed March 30, 2023.
141. Kaltenthaler E, Tappenden P, Paisley S, Squires H. NICE DSU Technical Support Document 13: Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models. 2011; https://www.ncbi.nlm.nih.gov/books/NBK425832/. Accessed March 30, 2023.
142. Dias S, Welton NJ, Sutton AJ, Ades A. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. *University of Sheffield, Decision Support Unit* 2011; https://www.ncbi.nlm.nih.gov/books/NBK395870/pdf/Bookshelf_NBK395870.pdf. Accessed March 30, 2023.
143. Dias S, Sutton AJ, Welton NJ, Ades A. NICE DSU Technical Support Document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011; https://www.ncbi.nlm.nih.gov/books/NBK395886/pdf/Bookshelf_NBK395886.pdf. Accessed March 30, 2023.
144. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades A. NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials. 2011; https://www.ncbi.nlm.nih.gov/books/NBK310372/pdf/Bookshelf_NBK310372.pdf. Accessed March 30, 2023.
145. Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. 2016; https://www.sheffield.ac.uk/nice-dsu/tsds/population-adjusted. Accessed 15 Oct, 2023.
146. Woods B, Sideris E, Palmer S, Latimer N, Soares M. NICE DSU Technical Support Document 19: partitioned survival analysis for decision modelling in health care: a critical review. 2017; https://www.sheffield.ac.uk/media/34205/download?attachment. Accessed July 4, 2023.

147. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol.* 2017;45(6):1866-1886.
148. Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. *Psychol Med.* 2021;51(4):563-578.
149. Vanni T, Karnon J, Madan J, et al. Calibrating Models in Economic Evaluation. *Pharmacoeconomics.* 2011;29(1):35-49.
150. Mandrik O, Thomas C, Whyte S, Chilcott J. Calibrating Natural History of Cancer Models in the Presence of Data Incompatibility: Problems and Solutions. *Pharmacoeconomics.* 2022;40(4):359-366.
151. Ryan CJ, Smith MR, Fizazi K, et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. *The Lancet Oncology.* 2015;16(2):152-160.
152. Latimer N, White I, Tilling K, Siebert U. Improved two-stage estimation to adjust for treatment switching in randomised trials: g-estimation to address time-dependent confounding. *Stat Methods Med Res.* 2020;29(10):2900-2918.
153. Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. *BMC Med Res Methodol.* 2012;12:1-16.
154. Jansen JP, Incerti D, Trikalinos TA. Multi-state network meta-analysis of progression and survival data. *Stat Med.* 2023;42(19):3371-3391.
155. NICE. Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor (part review of NICE technology appraisal guidance 36, review of NICE technology appraisal guidance 126 and 141) (TA 195). 2010; https://www.nice.org.uk/guidance/ta195. Accessed Mar 23, 2020.
156. National Institute for Health and Care Excellence. Rituximab in combination with glucocorticoids for treating anti-neutrophil cytoplasmic antibody-associated vasculitis (Technology Appraisal 308). 2014; https://www.nice.org.uk/guidance/ta308. Accessed Jun 30, 2020.
157. National Institute for Health and Care Excellence. Apremilast for treating moderate to severe plaque psoriasis (Technology Appraisal 419). 2016; https://www.nice.org.uk/guidance/ta419. Accessed Jun 30, 2020.
158. National Institute for Health and Care Excellence. Ixekizumab for treating moderate to severe plaque psoriasis (Technology Appraisal 442). 2017; https://www.nice.org.uk/guidance/ta442. Accessed Jun 30, 2020.
159. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol.* 2017;46(2):756-762.
160. Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. *Med Decis Making.* 2011;31(4):625-641.

161. Bell Gorrod H, Kearns B, Stevens J, et al. A Review of Survival Analysis Methods Used in NICE Technology Appraisals of Cancer Treatments: Consistency, Limitations, and Areas for Improvement. *Med Decis Making.* 2019;39(8):899-909.
162. Latimer NR, Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ medicine.* 2022;1(1):e000094.
163. Bullement A, Stevenson MD, Baio G, Shields GE, Latimer NR. A systematic review of methods to incorporate external evidence into trial-based survival extrapolations for health technology assessment. *Med Decis Making.* 2023;43(5):610-220.
164. Jaksa A, Louder A, Maksymiuk C, et al. A Comparison of Seven Oncology External Control Arm Case Studies: Critiques From Regulatory and Health Technology Assessment Agencies. *Value Health.* 2022;25(12):1967-1976.
165. Burger HU, Gerlinger C, Harbron C, et al. The use of external controls: To what extent can it currently be recommended? *Pharm Stat.* 2021;20(6):1002-1016.
166. Colnet B, Mayer I, Chen G, et al. Causal inference methods for combining randomized trials and observational studies: a review. arXiv preprint. 2020; https://doi.org/10.48550/arXiv.2011.08047. Accessed June 30, 2023.
167. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. *Eur J Epidemiol.* 2019;34:719-722.
168. NHS. Cancer Drugs Fund. 2016; https://www.england.nhs.uk/cancer/cdf/. Accessed May 22, 2020.
169. Kang J, Cairns J. "Don't think twice, It's all right": using additional data to reduce uncertainty regarding oncologic drugs provided through managed access agreements in England. *PharmacoEconomics-Open.* 2022;7:77-91.
170. Chang J-YA, Chilcott JB, Latimer NR. Leveraging real-world data to assess treatment sequences in health economic evaluations: a study protocol for emulating target trials using the English Cancer Registry and US Electronic Health Records-Derived Database. *HEDS Discussion Paper.* 2024;24.01:1-61.
171. Franklin JM, Patorno E, Desai RJ, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation.* 2021;143(10):1002-1013.
172. RCT Duplicate. RCT duplicate: randomized controlled trials duplicated using prospective longitudinal insurance claims: applying techniques of epidemiology 2020; www.rctduplicate.org. Accessed Jan 15, 2021.
173. Regan MM, Neven P, Giobbie-Hurder A, et al. Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8·1 years median follow-up. *The Lancet Oncology.* 2011;12(12):1101-1108.

174. Eichelberg C, Vervenne WL, De Santis M, et al. SWITCH: A Randomised, Sequential, Open-label Study to Evaluate the Efficacy and Safety of Sorafenib-sunitinib Versus Sunitinib-sorafenib in the Treatment of Metastatic Renal Cell Cancer. *Eur Urol.* 2015;68(5):837-847.
175. Retz M, Bedke J, Bogemann M, et al. SWITCH II: Phase III randomized, sequential, open-label study to evaluate the efficacy and safety of sorafenib-pazopanib versus pazopanib-sorafenib in the treatment of advanced or metastatic renal cell carcinoma (AUO AN 33/11). *Eur J Cancer.* 2019;107:37-45.
176. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A" SMART" design for building individualized treatment sequences. *Annu Rev Clin Psychol.* 2012;8:21-48.
177. Bujkiewicz S, Singh J, Wheaton L, et al. Bridging disconnected networks of first and second lines of biologic therapies in rheumatoid arthritis with registry data: bayesian evidence synthesis with target trial emulation. *J Clin Epidemiol.* 2022;150:171-178.
178. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics.* 2002;58(1):48-57.
179. Zang Y, Lee JJ. Adaptive clinical trial designs in oncology. *Chinese clinical oncology.* 2014;3(4):49.
180. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med.* 2018;16(1):29.
181. Mahajan R, Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect. *Indian J Pharmacol.* 2010;42(4):201-207.
182. Thorlund K, Haggstrom J, Park JJ, Mills EJ. Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ.* 2018;360:k698.
183. Crown W, Buyukkaramikli N, Thokala P, et al. Constrained optimization methods in health services research—an introduction: report 1 of the ISPOR optimization methods emerging good practices task force. *Value Health.* 2017;20(3):310-319.
184. Chang J, Lin F, Wang C. Changes in Treatment Patterns of Tyrosine Kinase Inhibitors among Non-Small Cell Lung Cancer Patients under the National Health Insurance in Taiwan. International Society for Pharmacoeconomics and Outcomes Research 20th Annual European Conference; Nov 4-8, 2017; Glasgow, UK.
185. Chang J, Lin F, Wang C. Chemotherapy Utilization among NSCLC Patients after First-/Second-Line Tyrosine Kinase Inhibitor Treatment Failure: Findings from Nationwide Data in Taiwan. ISPE 10th Asian Conference on Pharmacoepidemiology; Oct 29-31, 2017; Brisbane, Australia.
186. NICE. Guidance, NICE advice and quality standards: Technology appraisal guidance. 2019; https://www.nice.org.uk/guidance/published?to=2019-11-30&from=2000-01-01&ngt=Technology%20appraisal%20guidance&ndt=Guidance. Accessed Dec 1, 2019.
187. Chang J-YA, Latimer NR, Gillespie D, Chilcott J. Prevalence, characteristics and key issues of modelling treatment sequences in health economic evaluations – a systematic review of NICE technology appraisals. CRD42020171834. 2020;

https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020171834. Accessed May 5, 2020.

188. National Institute for Health and Care Excellence. Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for rheumatoid arthritis not previously treated with DMARDs or after conventional DMARDs only have failed (Technology Appraisal 375). 2016; https://www.nice.org.uk/guidance/ta375. Accessed Jun 30, 2020.
189. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med.* 2009;151(4):264-269.
190. NICE. Xeomin (botulinum neurotoxin type A) for treating chronic sialorrhoea (Technology Appraisal 605). 2019; https://www.nice.org.uk/guidance/ta605. Accessed Jun 30, 2020.
191. NICE. Rivaroxaban for preventing atherothrombotic events in people with coronary or peripheral artery disease (Technology Appraisal 607). 2019; https://www.nice.org.uk/guidance/ta607. Accessed Jun 30, 2020.
192. NICE. Pentosan polysulfate sodium for treating bladder pain syndrome (Technology Appraisal 610). 2019; https://www.nice.org.uk/guidance/ta610. Accessed Jun 30, 2020.
193. NICE. Rucaparib for maintenance treatment of relapsed platinum-sensitive ovarian, fallopian tube or peritoneal cancer (Technology Appraisal 611). 2019; https://www.nice.org.uk/guidance/ta611. Accessed Jun 30, 2020.
194. National Institute for Health and Care Excellence. Neratinib for extended adjuvant treatment of hormone receptor-positive, HER2-positive early stage breast cancer after adjuvant trastuzumab (Technology Appraisal 612). 2019; https://www.nice.org.uk/guidance/ta612. Accessed Jun 30, 2020.
195. National Institute for Clinical Excellence. Dronedarone for the treatment of non-permanent atrial fibrillation (Technology Appraisal 197). 2012; https://www.nice.org.uk/guidance/ta197. Accessed Jun 30, 2020.
196. National Institute for Health and Care Excellence. Aflibercept for treating visual impairment caused by macular oedema after branch retinal vein occlusion (Technology Appraisal 409). 2016; https://www.nice.org.uk/guidance/ta409. Accessed Jun 30, 2020.
197. National Institute for Health and Care Excellence. Dasatinib, nilotinib and imatinib for untreated chronic myeloid leukaemia (Technology Appraisal 426). 2016; https://www.nice.org.uk/guidance/ta448. Accessed Jun 30, 2020.
198. National Institute for Clinical Excellence. Telbivudine for the treatment of chronic hepatitis B (Technology Appraisal 154). 2008; https://www.nice.org.uk/guidance/ta154. Accessed Jun 30, 2020.
199. NICE. Lenalidomide for the treatment of multiple myeloma in people who have received at least 2 prior therapies (TA 171). 2009; https://www.nice.org.uk/guidance/ta171. Accessed Apr 30, 2021.
200. National Institute for Clinical Excellence. Tocilizumab for the treatment of systemic juvenile idiopathic arthritis (Technology Appraisal 238). 2011; https://www.nice.org.uk/guidance/ta238. Accessed Jun 30, 2020.

201. National Institute for Clinical Excellence. Aripiprazole for treating moderate to severe manic episodes in adolescents with bipolar I disorder (Technology Appraisal 292). 2013; https://www.nice.org.uk/guidance/ta292. Accessed Jun 30, 2020.
202. National Institute for Clinical Excellence. Eltrombopag for treating chronic immune (idiopathic) thrombocytopenic purpura (Technology Appraisal 293). 2013; https://www.nice.org.uk/guidance/ta293. Accessed Jun 30, 2020.
203. National Institute for Health and Care Excellence. Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma (Technology Appraisal 319). 2014; https://www.nice.org.uk/guidance/ta319. Accessed Jun 30, 2020.
204. National Institute for Clinical Excellence. Bevacizumab in combination with oxaliplatin and either fluorouracil plus folinic acid or capecitabine for the treatment of metastatic colorectal cancer (Technology Appraisal 212). 2010; https://www.nice.org.uk/guidance/ta212. Accessed Jun 30, 2020.
205. National Institute for Clinical Excellence. Bevacizumab (first-line), sorafenib (first- and second-line), sunitinib (second-line) and temsirolimus (first-line) for the treatment of advanced and/or metastatic renal cell carcinoma (Technology Appraisal 178). 2009; https://www.nice.org.uk/guidance/ta178. Accessed Jan 01, 2021.
206. National Institute for Health and Care Excellence. Golimumab for treating non-radiographic axial spondyloarthritis (Technology Appraisal 497). 2018; https://www.nice.org.uk/guidance/ta497. Accessed Jun 30, 2020.
207. National Institute for Health and Care Excellence. Ocrelizumab for treating relapsing–remitting multiple sclerosis (Technology Appraisal 533). 2018; https://www.nice.org.uk/guidance/ta533. Accessed Jun 30, 2020.
208. National Institute for Health and Care Excellence. Dupilumab for treating moderate to severe atopic dermatitis (Technology Appraisal 534). 2018; https://www.nice.org.uk/guidance/ta534. Accessed Jun 30, 2020.
209. National Institute for Health and Care Excellence. Etelcalcetide for treating secondary hyperparathyroidism (Technology Appraisal 448). 2017; https://www.nice.org.uk/guidance/ta448. Accessed Jun 30, 2020.
210. NICE. Pirfenidone for treating idiopathic pulmonary fibrosis (TA 504). 2018; https://www.nice.org.uk/guidance/ta504. Accessed Apr 30, 2021.
211. National Institute for Health and Care Excellence. Ribociclib with an aromatase inhibitor for previously untreated, hormone receptor-positive, HER2-negative, locally advanced or metastatic breast cancer (Technology Appraisal 496). 2017; https://www.nice.org.uk/guidance/ta496. Accessed Jun 30, 2020.
212. National Institute for Clinical Excellence. Febuxostat for the management of hyperuricaemia in people with gout (Technology Appraisal 164). 2008; https://www.nice.org.uk/guidance/ta164. Accessed Jun 30, 2020.

213. National Institute for Health and Care Excellence. Adalimumab, etanercept and ustekinumab for treating plaque psoriasis in children and young people (Technology Appraisal 455). 2017; https://www.nice.org.uk/guidance/ta455. Accessed Jun 30, 2020.
214. National Institute for Clinical Excellence. Ustekinumab for the treatment of adults with moderate to severe psoriasis (Technology Appraisal 180). 2009; https://www.nice.org.uk/guidance/ta180. Accessed Jun 30, 2020.
215. National Institute for Clinical Excellence. Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma (Technology Appraisal 137). 2008; https://www.nice.org.uk/guidance/ta137. Accessed Jun 30, 2020.
216. National Institute for Clinical Excellence. Infliximab and adalimumab for the treatment of Crohn's disease (Technology Appraisal 187). 2010; https://www.nice.org.uk/guidance/ta187. Accessed Jun 30, 2020.
217. National Institute for Clinical Excellence. Adalimumab and dexamethasone for treating non-infectious uveitis (Technology Appraisal 460). 2010; https://www.nice.org.uk/guidance/ta460. Accessed Jun 30, 2020.
218. National Institute for Health and Care Excellence. Nivolumab for treating relapsed or refractory classical Hodgkin lymphoma (Technology Appraisal 462). 2017; https://www.nice.org.uk/guidance/ta462. Accessed Jun 30, 2020.
219. National Institute for Health and Care Excellence. Trastuzumab emtansine for treating HER2-positive advanced breast cancer after trastuzumab and a taxane (Technology Appraisal 458). 2017; https://www.nice.org.uk/guidance/ta458. Accessed Jun 30, 2020.
220. National Institute for Clinical Excellence. Eluxadoline for treating irritable bowel syndrome with diarrhoea (Technology Appraisal 471). 2017; https://www.nice.org.uk/guidance/ta471. Accessed Jun 30, 2020.
221. National Institute for Clinical Excellence. Lenvatinib with everolimus for previously treated advanced renal cell carcinoma (Technology Appraisal 498). 2018; https://www.nice.org.uk/guidance/ta498. Accessed Jun 30, 2020.
222. National Institute for Clinical Excellence. Ertugliflozin as monotherapy or with metformin for treating type 2 diabetes (Technology Appraisal 572). 2019; https://www.nice.org.uk/guidance/ta572. Accessed Jun 30, 2020.
223. National Institute for Clinical Excellence. Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis (Technology Appraisal 199). 2010; https://www.nice.org.uk/guidance/ta199. Accessed Jun 30, 2020.
224. National Institute for Clinical Excellence. Topotecan for the treatment of recurrent and stage IVB cervical cancer (Technology Appraisal 183). 2009; https://www.nice.org.uk/guidance/ta183. Accessed Jun 30, 2020.

225. National Institute for Clinical Excellence. Raloxifene and teriparatide for the secondary prevention of osteoporotic fragility fractures in postmenopausal women (Technology Appraisal 161). 2008; https://www.nice.org.uk/guidance/ta161. Accessed Jun 30, 2020.
226. NICE. Mannitol dry powder for inhalation for treating cystic fibrosis (Technology Appraisal 266). 2012; https://www.nice.org.uk/guidance/ta266. Accessed Jun 30, 2020.
227. NICE. Ranibizumab for treating choroidal neovascularisation associated with pathological myopia (Technology Appraisal 298). 2013; https://www.nice.org.uk/guidance/ta298. Accessed Jun 30, 2020.
228. National Institute for Health and Care Excellence. Naloxegol for treating opioid-induced constipation (Technology Appraisal 345). 2015; https://www.nice.org.uk/guidance/ta345. Accessed Jun 30, 2020.
229. NICE. Immunosuppressive therapy for kidney transplant in children and young people (Technology Appraisal 482). 2017; https://www.nice.org.uk/guidance/ta482. Accessed Jun 30, 2020.
230. NICE. Lenalidomide for treating myelodysplastic syndromes associated with an isolated deletion 5q cytogenetic abnormality (Technology Appraisal 322). 2014; https://www.nice.org.uk/guidance/ta322. Accessed Jun 30, 2020.
231. National Institute for Health and Care Excellence. Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab (Technology Appraisal 357). 2015; https://www.nice.org.uk/guidance/ta357. Accessed Jun 30, 2020.
232. NICE. Azacitidine for treating acute myeloid leukaemia with more than 30% bone marrow blasts (Technology Appraisal 399). 2016; https://www.nice.org.uk/guidance/ta399. Accessed Jun 30, 2020.
233. NICE. Everolimus for advanced renal cell carcinoma after previous treatment (TA 432). 2017; https://www.nice.org.uk/Guidance/TA432. Accessed Jan 1, 2021.
234. NICE. Vandetanib for treating medullary thyroid cancer (Technology Appraisal 550). 2018; https://www.nice.org.uk/guidance/ta550. Accessed Jun 30, 2020.
235. National Institute for Health and Care Excellence. Dimethyl fumarate for treating relapsing-remitting multiple sclerosis (Technology Appraisal 320). 2014; https://www.nice.org.uk/guidance/ta320. Accessed Jun 30, 2020.
236. National Institute for Health and Care Excellence. Ustekinumab for moderately to severely active Crohn's disease after previous treatment (Technology Appraisal 456). 2017; https://www.nice.org.uk/guidance/ta456. Accessed Jun 30, 2020.
237. NICE. Cabazitaxel for hormone-relapsed metastatic prostate cancer treated with docetaxel (TA 391). 2016; https://www.nice.org.uk/guidance/ta391. Accessed Apr 30, 2021.
238. National Institute for Health and Care Excellence. Trametinib in combination with dabrafenib for treating unresectable or metastatic melanoma (Technology Appraisal 396). 2016; https://www.nice.org.uk/guidance/ta396. Accessed Jun 30, 2020.
239. National Institute for Health and Care Excellence. Idelalisib for treating chronic lymphocytic leukaemia (Technology Appraisal 359). 2015; https://www.nice.org.uk/guidance/ta359. Accessed Jun 30, 2020.

240. National Institute for Health and Care Excellence. Ombitasvir–paritaprevir–ritonavir with or without dasabuvir for treating chronic hepatitis C (Technology Appraisal 365). 2015; https://www.nice.org.uk/guidance/ta365. Accessed Jun 30, 2020.
241. NICE. Elbasvir–grazoprevir for treating chronic hepatitis C (Technology Appraisal 413). 2016; https://www.nice.org.uk/guidance/ta413. Accessed Jun 30, 2020.
242. NICE. Sofosbuvir–velpatasvir for treating chronic hepatitis C (Technology Appraisal 430). 2017; https://www.nice.org.uk/guidance/ta430. Accessed Jun 30, 2020.
243. NICE. Glecaprevir–pibrentasvir for treating chronic hepatitis C (Technology Appraisal 499). 2018; https://www.nice.org.uk/guidance/ta499. Accessed Jun 30, 2020.
244. National Institute for Health and Care Excellence. Apremilast for treating active psoriatic arthritis (Technology Appraisal 433). 2017; https://www.nice.org.uk/guidance/ta433. Accessed Jun 30, 2020.
245. National Institute for Clinical Excellence. Etanercept and efalizumab for the treatment of adults with psoriasis (Technology Appraisal 103). 2006; https://www.nice.org.uk/guidance/ta103. Accessed Jun 30, 2020.
246. National Institute for Health and Care Excellence. Certolizumab pegol and secukinumab for treating active psoriatic arthritis after inadequate response to DMARDs (Technology Appraisal 445). 2017; https://www.nice.org.uk/guidance/ta445. Accessed Jun 30, 2020.
247. National Institute for Health and Care Excellence. Guselkumab for treating moderate to severe plaque psoriasis (Technology Appraisal 521). 2018; https://www.nice.org.uk/guidance/ta521. Accessed Jun 30, 2020.
248. National Institute for Health and Care Excellence. Pegaspargase for treating acute lymphoblastic leukaemia (Technology Appraisal 408). 2016; https://www.nice.org.uk/guidance/ta408. Accessed Jun 30, 2020.
249. NICE. Enzalutamide for hormone-relapsed non-metastatic prostate cancer (TA 580). 2019; https://www.nice.org.uk/guidance/ta580. Accessed Apr 30, 2021.
250. National Institute for Health and Care Excellence. Dimethyl fumarate for treating moderate to severe plaque psoriasis (Technology Appraisal 475). 2017; https://www.nice.org.uk/guidance/ta475. Accessed Jun 30, 2020.
251. National Institute for Health and Care Excellence. Brodalumab for treating moderate to severe plaque psoriasis (Technology Appraisal 511). 2018; https://www.nice.org.uk/guidance/ta511. Accessed Jun 30, 2020.
252. National Institute for Health and Care Excellence. Ixekizumab for treating active psoriatic arthritis after inadequate response to DMARDs (Technology Appraisal 537). 2018; https://www.nice.org.uk/guidance/ta543. Accessed Jun 30, 2020.
253. National Institute for Health and Care Excellence. Tofacitinib for treating active psoriatic arthritis after inadequate response to DMARDs (Technology Appraisal 543). 2018; https://www.nice.org.uk/guidance/ta543. Accessed Jun 30, 2020.

254. National Institute for Health and Care Excellence. Certolizumab pegol for treating moderate to severe plaque psoriasis (Technology Appraisal 574). 2019; https://www.nice.org.uk/guidance/ta574. Accessed Jun 30, 2020.
255. National Institute for Health and Care Excellence. Tildrakizumab for treating moderate to severe plaque psoriasis (Technology Appraisal 575). 2019; https://www.nice.org.uk/guidance/ta575. Accessed Jun 30, 2020.
256. National Institute for Clinical Excellence. Golimumab for the treatment of rheumatoid arthritis after the failure of previous disease-modifying anti-rheumatic drugs (Technology Appraisal 225). 2011; https://www.nice.org.uk/guidance/ta225. Accessed Jun 30, 2020.
257. National Institute for Health and Care Excellence. Certolizumab pegol for treating rheumatoid arthritis after inadequate response to a TNF-alpha inhibitor (Technology Appraisal 415). 2016; https://www.nice.org.uk/guidance/ta415. Accessed Jun 30, 2020.
258. National Institute for Health and Care Excellence. Baricitinib for moderate to severe rheumatoid arthritis (Technology Appraisal 466). 2017; https://www.nice.org.uk/guidance/ta466. Accessed Jun 30, 2020.
259. National Institute for Health and Care Excellence. Sarilumab for moderate to severe rheumatoid arthritis (Technology Appraisal 485). 2017; https://www.nice.org.uk/guidance/ta485. Accessed Jun 30, 2020.
260. National Institute for Health and Care Excellence. Tofacitinib for moderately to severely active ulcerative colitis (Technology Appraisal 547). 2018; https://www.nice.org.uk/guidance/ta547. Accessed Jun 30, 2020.
261. <standard_treatment_programme.pdf>.
262. National Institute for Clinical Excellence. Dabigatran etexilate for the prevention of stroke and systemic embolism in atrial fibrillation (Technology Appraisal 249). 2012; https://www.nice.org.uk/guidance/ta249. Accessed Jun 30, 2020.
263. National Institute for Clinical Excellence. Tenofovir disoproxil for the treatment of chronic hepatitis B (Technology Appraisal 173). 2009; https://www.nice.org.uk/guidance/ta173. Accessed Jun 30, 2020.
264. National Institute for Health and Care Excellence. Dapagliflozin in triple therapy for treating type 2 diabetes (Technology Appraisal 418). 2016; https://www.nice.org.uk/guidance/ta418. Accessed Jun 30, 2020.
265. National Institute for Clinical Excellence. Dasatinib, nilotinib and standard-dose imatinib for the first-line treatment of chronic myeloid leukaemia (Technology Appraisal 251). 2012; https://www.nice.org.uk/guidance/ta251. Accessed Jun 30, 2020.
266. Cheah C, Chihara D, Horowitz S, et al. Patients with classical Hodgkin lymphoma experiencing disease progression after treatment with brentuximab vedotin have poor outcomes. *Ann Oncol.* 2016;27(7):1317-1323.

267. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA.* 2022;328(24):2446-2447.
268. Miyake H, Hara T, Tamura K, et al. Comparative Assessment of Efficacies Between 2 Alternative Therapeutic Sequences With Novel Androgen Receptor-Axis-Targeted Agents in Patients With Chemotherapy-Naïve Metastatic Castration-Resistant Prostate Cancer. *Clin Genitourin Cancer.* 2017;15(4):e591-e597.
269. Matsubara N, Yamada Y, Tabata KI, et al. Abiraterone Followed by Enzalutamide Versus Enzalutamide Followed by Abiraterone in Chemotherapy-naive Patients With Metastatic Castration-resistant Prostate Cancer. *Clin Genitourin Cancer.* 2018;16(2):142-148.
270. Terada N, Maughan BL, Akamatsu S, et al. Exploring the optimal sequence of abiraterone and enzalutamide in patients with chemotherapy-naïve castration-resistant prostate cancer: The Kyoto-Baltimore collaboration. *Int J Urol.* 2017;24(6):441-448.
271. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health.* 2004;58(4):265-271.
272. Höfler M. Causal inference based on counterfactuals. *BMC Med Res Methodol.* 2005;5(1):1-12.
273. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association.* 2005;100(469):322-331.
274. Mahar RK, McGuinness MB, Chakraborty B, Carlin JB, Ijzerman MJ, Simpson JA. A scoping review of studies using observational data to optimise dynamic treatment regimens. *BMC Med Res Methodol.* 2021;21(1):39.
275. Simoneau G, Moodie EEM, Azoulay L, Platt RW. Adaptive Treatment Strategies With Survival Outcomes: An Application to the Treatment of Type 2 Diabetes Using a Large Observational Database. *Am J Epidemiol.* 2020;189(5):461-469.
276. Kosorok MR, Moodie EE. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine.* SIAM; 2015.
277. Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: statistical and decision-making challenges and approaches. *Int J Technol Assess Health Care.* 2016;32(3):160-166.
278. Alshreef A, Latimer N, Tappenden P, et al. Statistical Methods for Adjusting Estimates of Treatment Effectiveness for Patient Nonadherence in the Context of Time-to-Event Outcomes and Health Technology Assessment: A Systematic Review of Methodological Papers. *Med Decis Making.* 2019;39(8):910-925.
279. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol.* 2006;98(3):237-242.
280. Cole SR, Hernán MA, Margolick JB, Cohen MH, Robins JM. Marginal Structural Models for Estimating the Effect of Highly Active Antiretroviral Therapy Initiation on CD4 Cell Count. *Am J Epidemiol.* 2005;162(5):471-478.
281. Latimer NR. *The role of treatment crossover adjustment methods in the context of economic evaluation*, University of Sheffield; 2012.

282. Alshreef A. *A methodological framework to account for the impact of non-adherence on the cost-effectiveness of chronic medications*, University of Sheffield; 2021.
283. Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and prognostic research.* 2021;5(1):1-16.
284. Farmer RE, Kounali D, Walker AS, et al. Application of causal inference methods in the analyses of randomised controlled trials: a systematic review. *Trials.* 2018;19(1):23.
285. Thall PF, Mueller P, Xu Y, Guindani M. Bayesian nonparametric statistics: A new toolkit for discovery in cancer research. *Pharm Stat.* 2017;16(6):414-423.
286. London WB, Frantz CN, Campbell LA, et al. Phase II randomized comparison of topotecan plus cyclophosphamide versus topotecan alone in children with recurrent or refractory neuroblastoma: a Children's Oncology Group study. *J Clin Oncol.* 2010;28(24):3808.
287. Wahed AS, Thall PF. Evaluating joint effects of induction–salvage treatment regimes on overall survival in acute leukaemia. *Journal of the Royal Statistical Society Series C: Applied Statistics.* 2013;62(1):67-83.
288. Huang X, Ning J, Wahed AS. Optimization of individualized dynamic treatment regimes for recurrent diseases. *Stat Med.* 2014;33(14):2363-2378.
289. Huang X, Cormier JN, Pisters PW. Estimation of the causal effects on survival of two-stage nonrandomized treatment sequences for recurrent diseases. *Biometrics.* 2006;62(3):901-909.
290. Huang X, Ning J. Analysis of multi-stage treatments for recurrent diseases. *Stat Med.* 2012;31(24):2805-2821.
291. Krakow EF, Hemmer M, Wang T, et al. Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations From Cohort and Registry Data Using Q-Learning. *Am J Epidemiol.* 2017;186(2):160-172.
292. Miyahara S, Wahed AS. Weighted Kaplan–Meier estimators for two-stage treatment regimes. *Stat Med.* 2010;29(25):2581-2591.
293. Chao Y-C, Tran Q, Tsodikov A, Kidwell KM. Joint modeling and multiple comparisons with the best of data from a SMART with survival outcomes. *Biostatistics.* 2022;23(1):294-313.
294. Van Leeuwen R, Bonsel GJ, Reiss P, Danner SA, Lange JM. Additive or sequential nucleoside analogue therapy compared with continued zidovudine monotherapy in human immunodeficiency virus-infected patients with advanced disease does not prolong survival: an observational study. *The Journal of infectious diseases.* 1997;175(6):1344-1351.
295. AGIS investigators. The Advanced Glaucoma Intervention Study (AGIS): 12. Baseline risk factors for sustained loss of visual field and visual acuity in patients with advanced glaucoma. *Am J Ophthalmol.* 2002;134(4):499-512.
296. Benchalal M, Le Prisé E, de Lafontan B, et al. Influence of the time between surgery and radiotherapy on local recurrence in patients with lymph node-positive, early-stage, invasive breast carcinoma undergoing breast-conserving surgery: Results of the French Adjuvant Study Group.

Cancer: Interdisciplinary International Journal of the American Cancer Society. 2005;104(2):240-250.

297. Busch J, Seidel C, Kempkensteffen C, et al. Sequence therapy in patients with metastatic renal cell carcinoma: comparison of common targeted treatment options following failure of receptor tyrosine kinase inhibitors. *Eur Urol.* 2011;60(6):1163-1170.
298. Busch J, Seidel C, Erber B, et al. Retrospective comparison of triple-sequence therapies in metastatic renal cell carcinoma. *Eur Urol.* 2013;64(1):62-70.
299. Kumada T, Toyoda H, Yasuda S, Ito T, Tanaka J. Impact of switching to tenofovir alafenamide fumarate in patients with entecavir-treated chronic hepatitis B. *Eur J Gastroenterol Hepatol.* 2021;33(1S):e898-e904.
300. Feng W, Wahed AS. Supremum weighted log-rank test and sample size for comparing two-stage adaptive treatment strategies. *Biometrika.* 2008;95(3):695-707.
301. Guo X, Tsiatis A. A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics.* 2005;1(1).
302. Chen L-W, Yavuz I, Cheng Y, Wahed AS. Cumulative incidence regression for dynamic treatment regimens. *Biostatistics.* 2020;21(2):e113-e130.
303. Yavuz I, Cheng Y, Wahed AS. Estimating the cumulative incidence function of dynamic treatment regimes. *Journal of the Royal Statistical Society Series A: Statistics in Society.* 2018;181(1):85-106.
304. Tang X, Wahed AS. Cumulative hazard ratio estimation for treatment regimes in sequentially randomized clinical trials. *Statistics in biosciences.* 2015;7:1-18.
305. Vilakati S, Cortese G, Dlamini T. Estimating survival distributions for two-stage adaptive treatment strategies: A simulation study. *Communications for Statistical Applications and Methods.* 2021;28(5):411-424.
306. Goldberg Y, Kosorok MR. Q-learning with censored data. *Annals of statistics.* 2012;40(1):529.
307. Tang X, Wahed AS. Comparison of treatment regimes with adjustment for auxiliary variables. *Journal of Applied Statistics.* 2011;38(12):2925-2938.
308. Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics.* 2011;67(4):1422-1433.
309. Ruppert A, Yin J, Davidian M, et al. Application of a sequential multiple assignment randomized trial (SMART) design in older patients with chronic lymphocytic leukemia. *Ann Oncol.* 2019;30(4):542-550.
310. Wahed AS. Inference for two-stage adaptive treatment strategies using mixture distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics.* 2010;59(1):1-18.
311. Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics.* 2004;60(1):124-133.
312. Kidwell KM, Wahed AS. Weighted log-rank statistic to compare shared-path adaptive treatment strategies. *Biostatistics.* 2013;14(2):299-312.

313. Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat Med.* 2007;26(26):4687-4702.
314. Early Breast Cancer Trialists' Collaborative Group. Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. *The Lancet.* 2015;386(10001):1341-1352.
315. Lu X, Qian C. Efficacy, Safety, and Prognosis of Sequential Therapy with Tamoxifen and Letrozole versus Letrozole Monotherapy for Breast Carcinoma. *Comput Math Methods Med.* 2022;2022.
316. Recht A, Come SE, Henderson IC, et al. The sequencing of chemotherapy and radiation therapy after conservative surgery for early-stage breast cancer. *N Engl J Med.* 1996;334(21):1356-1361.
317. Rabaglio M, Sun Z, Maibach R, et al. Cumulative incidence of cardiovascular events under tamoxifen and letrozole alone and in sequence: a report from the BIG 1-98 trial. *Breast Cancer Res Treat.* 2021;185:697-707.
318. Jonasch E, Signorovitch JE, Lin PL, et al. Treatment patterns in metastatic renal cell carcinoma: a retrospective review of medical records from US community oncology practices. *Curr Med Res Opin.* 2014;30(10):2041-2050.
319. Iacovelli R, Cartenì G, Sternberg CN, et al. Clinical outcomes in patients receiving three lines of targeted therapy for metastatic renal cell carcinoma: results from a large patient cohort. *Eur J Cancer.* 2013;49(9):2134-2142.
320. Giuliani J, Drudi F. Immunotherapy and targeted therapies in metastatic renal cell carcinoma: is there a preferred sequence? *Cancer Biotherapy and Radiopharmaceuticals.* 2012;27(8):513-518.
321. BIG 1-98 Collaborative Group. Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. *N Engl J Med.* 2009;361(8):766-776.
322. Stenner F, Chastonay R, Liewen H, et al. A pooled analysis of sequential therapies with sorafenib and sunitinib in metastatic renal cell carcinoma. *Oncology.* 2012;82(6):333-340.
323. Herrmann E, Marschner N, Grimm M, et al. Sequential therapies with sorafenib and sunitinib in advanced or metastatic renal cell carcinoma. *World J Urol.* 2011;29:361-366.
324. Derks MGM, Blok EJ, Seynaeve C, et al. Adjuvant tamoxifen and exemestane in women with postmenopausal early breast cancer (TEAM): 10-year follow-up of a multicentre, open-label, randomised, phase 3 trial. *The Lancet Oncology.* 2017;18(9):1211-1220.
325. Buchler T, Klapka R, Melichar B, et al. Sunitinib followed by sorafenib or vice versa for metastatic renal cell carcinoma—data from the Czech registry. *Ann Oncol.* 2012;23(2):395-401.
326. Dudek AZ, Zolnierek J, Dham A, Lindgren BR, Szczylik C. Sequential therapy with sorafenib and sunitinib in renal cell carcinoma. *Cancer.* 2009;115(1):61-67.
327. Ishihara H, Takagi T, Kondo T, et al. Efficacy and safety of third-line molecular-targeted therapy in metastatic renal cell carcinoma resistant to first-line vascular endothelial growth factor receptor tyrosine kinase inhibitor and second-line therapy. *Int J Clin Oncol.* 2018;23:559-567.

328. Chirgwin JH, Giobbie-Hurder A, Coates AS, et al. Treatment Adherence and Its Impact on Disease-Free Survival in the Breast International Group 1-98 Trial of Tamoxifen and Letrozole, Alone and in Sequence. *J Clin Oncol.* 2016;34(21):2452-2459.
329. Li J-R, Yang C-K, Wang S-S, et al. First-line treatment result influence second-line regimen selection in targeted therapy for metastatic renal cell carcinoma. *Anticancer Res.* 2014;34(10):5643-5647.
330. Paglino C, Procopio G, Sabbatini R, et al. A retrospective analysis of two different sequences of therapy lines for advanced kidney cancer. *Anticancer Res.* 2013;33(11):4999-5004.
331. Xu Y, Müller P, Wahed AS, Thall PF. Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association.* 2016;111(515):921-950.
332. Nishino K, Imamura F, Morita S, et al. A retrospective analysis of 335 Japanese lung cancer patients who responded to initial gefitinib treatment. *Lung Cancer.* 2013;82(2):299-304.
333. Hernán MA. How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ.* 2018;360:k182.
334. Zhao SS, Lyu H, Yoshida K. Versatility of the clone-censor-weight approach: response to "trial emulation in the presence of immortal-time bias". *Int J Epidemiol.* 2021;50(2):694-695.
335. Gaber CE, Hanson KA, Kim S, Lund JL, Lee TA, Murray EJ. The Clone-Censor-Weight Method in Pharmacoepidemiologic Research: Foundations and Methodological Implementation. *Current Epidemiology Reports.* 2024.
336. Nahum-Shani I, Qian M, Almirall D, et al. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychol Methods.* 2012;17(4):457-477.
337. Kidwell KM. SMART designs in cancer research: past, present, and future. *Clinical trials.* 2014;11(4):445-456.
338. Cormier JN, Huang X, Xing Y, et al. Cohort analysis of patients with localized, high-risk, extremity soft tissue sarcoma treated at two cancer centers: chemotherapy-associated outcomes. *J Clin Oncol.* 2004;22(22):4567-4574.
339. Estey EH, Thall PF, Pierce S, et al. Randomized phase II study of fludarabine+ cytosine arabinoside+ idarubicin±all-trans retinoic acid±granulocyte colony-stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome. *Blood, The Journal of the American Society of Hematology.* 1999;93(8):2478-2484.
340. Robins JM. Optimal structural nested models for optimal sequential decisions. Paper presented at: Proceedings of the second seattle Symposium in Biostatistics2004.
341. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2003;65(2):331-355.
342. Brumback BA, London WB. Causal inference in cancer clinical trials. *Modern Clinical Trial Analysis.* 2013:203-225.
343. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association.* 1994;89(427):846-866.

344. Stone RM, Berg DT, George SL, et al. Granulocyte–macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. *N Engl J Med.* 1995;332(25):1671-1677.
345. Kolitz JE, George SL, Marcucci G, et al. P-glycoprotein inhibition using valspodar (PSC-833) does not improve outcomes for patients younger than age 60 years with newly diagnosed acute myeloid leukemia: Cancer and Leukemia Group B study 19808. *Blood, The Journal of the American Society of Hematology.* 2010;116(9):1413-1421.
346. Matthay KK, Reynolds CP, Seeger RC, et al. Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: a children's oncology group study. *J Clin Oncol.* 2009;27(7):1007-1013.
347. Fisher LD, Dixon DO, Herson J, Frankowski RF, Hearron MS, Peace KE. Intention to treat in clinical trials. *Statistical issues in drug research and development*: Routledge; 2017:331-350.
348. Lasagna L. The controlled clinical trial: theory and practice. *J Chronic Dis.* 1955;1(4):353-367.
349. Pétavy F, Guizzaro L, Antunes dos Reis I, Teerenstra S, Roes KC. Beyond "Intent-to-treat" and "Per protocol": Improving assessment of treatment effects in clinical trials through the specification of an estimand. *Br J Clin Pharmacol.* 2020;86(7):1235-1239.
350. Latimer NR, Abrams K, Lambert P, et al. Adjusting for treatment switching in randomised controlled trials–a simulation study and a simplified two-stage method. *Stat Methods Med Res.* 2017;26(2):724-751.
351. Ellenberg JH. Intent-to-treat analysis versus as-treated analysis. *Drug Inf J.* 1996;30(2):535-544.
352. Ellenberg JH. Selection bias in observational and experimental studies. *Stat Med.* 1994;13(5-7):557-567.
353. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med.* 1991;10(10):1595-1605.
354. Horwitz RI, Horwitz SM. Adherence to treatment and health outcomes. *Arch Intern Med.* 1993;153(16):1863-1868.
355. White IR, Walker S, Babiker AG, Darbyshire JH. Impact of treatment changes on the interpretation of the Concorde trial. *AIDS.* 1997;11(8):999-1006.
356. Platt RW, Hutcheon JA, Suissa S. Immortal Time Bias in Epidemiology. *Current Epidemiology Reports.* 2019;6(1):23-27.
357. Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. *BMC Med.* 2021;19(1):279.
358. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS epidemiology: methodological issues*: Springer; 1992:297-331.
359. Robins JM. Analytic methods for estimating HIV-treatment and cofactor effects. *Methodological issues in AIDS behavioral research.* 1993:213-288.
360. Richardson TS, Rotnitzky A. Causal etiology of the research of James M. Robins. 2014.

361. Robins JM. Marginal structural models. 1997.
362. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Stat Med.* 2008;27(23):4678-4721.
363. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *Int J Biostat.* 2010;6(2):Article 18.
364. Keogh RH, Gran JM, Seaman SR, Davies G, Vansteelandt S. Causal inference in survival analysis using longitudinal observational data: Sequential trials and marginal structural models. *Stat Med.* 2023;42(13):2191-2225.
365. Scola G, Chis Ster A, Bean D, Pareek N, Emsley R, Landau S. Implementation of the trial emulation approach in medical research: a scoping review. *BMC Med Res Methodol.* 2023;23(1):186.
366. Robins JM. Structural nested failure time models. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons; 1998.
367. Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association.* 1994;89(427):737-749.
368. Bellman R. Dynamic programming, princeton univ. *Press Princeton, New Jersey.* 1957.
369. Lin D, Sun W, Ying Z. Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika.* 1999;86(1):59-70.
370. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. Paper presented at: Proceedings of the American Statistical Association2000.
371. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics.* 1998:948-963.
372. Robins JM. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications.* 1987;14(9-12):923-945.
373. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol.* 2009;38(6):1599-1611.
374. Robins J. The control of confounding by intermediate variables. *Stat Med.* 1989;8(6):679-701.
375. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology.* 2014;25(6):889-897.
376. Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med.* 2012;31(18):2000-2009.
377. Wang L, Zhang W, He X, Zha H. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. Paper presented at: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining2018.

378. Baniya A, Herrmann S, Qiao Q, Lu H. Adaptive interventions treatment modelling and regimen optimization using sequential multiple assignment randomized trials (SMART) and Q-learning. Paper presented at: IIE Annual Conference. Proceedings2017.
379. Schomaker M, Luque-Fernandez MA, Leroy V, Davies MA. Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Stat Med.* 2019;38(24):4888-4911.
380. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. *Journal of causal inference.* 2014;2(2):147-185.
381. Dahabreh IJ, Steingrimsson JA, Robins JM, Hernán MA. Randomized trials and their observational emulations: a framework for benchmarking and joint analysis. *arXiv preprint arXiv:220314857.* 2022.
382. Forbes SP, Dahabreh IJ. Benchmarking Observational Analyses Against Randomized Trials: a Review of Studies Assessing Propensity Score Methods. *J Gen Intern Med.* 2020:1-9.
383. Nolan J, Bolton T. British Heart Foundation DSC Monthly Webinar: Constructing Research Ready Datasets. 2022; https://www.youtube.com/watch?v=8dDYGeBD_fA. Accessed Apr 30, 2022.
384. Mc Grath-Lone L, Jay MA, Blackburn R, et al. What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *International Journal of Population Data Science.* 2022;7(1).
385. Gatto NM, Campbell UB, Rubinstein E, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. *Clin Pharmacol Ther.* 2022;111(1):122-134.
386. Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, LoCasale R. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf.* 2012;21(1):1-10.
387. NCRAS. About the National Cancer Registration and Analysis Service. http://www.ncin.org.uk/about_ncin/. Accessed June 2, 2020.
388. Henson KE, Elliss-Brookes L, Coupland VH, et al. Data Resource Profile: National Cancer Registration Dataset in England. *Int J Epidemiol.* 2020;49(1):16-16h.
389. Bright CJ, Lawton S, Benson S, et al. Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset. *Int J Epidemiol.* 2019;49(1):15-15l.
390. The SACT Team. The SACT Systemic Anti-Cancer Therapy Dataset website. 2020; http://www.chemodataset.nhs.uk/home. Accessed June 2, 2020.
391. NDRS. Cancer data and the Cancer Drugs Fund. 2022; https://digital.nhs.uk/ndrs/our-work/ncras-partnerships/cancer-data-and-the-cancer-drugs-fund. Accessed Aug 31, 2022.
392. NDRS. Systemic Anti-Cancer Therapy (SACT) Data Newsletter: June 2022. 2022; https://myemail.constantcontact.com/SACT-Newsletter---June-2022.html?soid=1108552320760&aid=BBGAqhXGzsc. Accessed July 31, 2022.

393. NICE. Pembrolizumab for adjuvant treatment of completely resected stage 3 melanoma (TA 766). 2022; https://www.nice.org.uk/Guidance/TA766. Accessed Apr 30, 2022.
394. NICE. Nivolumab with ipilimumab for untreated advanced renal cell carcinoma (TA 780). 2022; https://www.nice.org.uk/Guidance/TA780. Accessed Apr 30, 2022.
395. NICE. Niraparib for maintenance treatment of relapsed, platinum-sensitive ovarian, fallopian tube and peritoneal cancer (TA 784). 2022; https://www.nice.org.uk/Guidance/TA784. Accessed Apr 30, 2022.
396. NICE. Daratumumab monotherapy for treating relapsed and refractory multiple myeloma (TA 783). 2022; https://www.nice.org.uk/Guidance/TA783. Accessed Apr 30, 2022.
397. NICE. Ibrutinib for treating Waldenstrom's macroglobulinaemia (TA 795). 2022; https://www.nice.org.uk/Guidance/TA795. Accessed June 30, 2022.
398. NICE. Venetoclax for treating chronic lymphocytic leukaemia (TA 796). 2022; https://www.nice.org.uk/Guidance/TA796. Accessed June 30, 2022.
399. Morrell L, Wordsworth S, Schuh A, Middleton MR, Rees S, Barker RW. Will the reformed Cancer Drugs Fund address the most common types of uncertainty? An analysis of NICE cancer drug appraisals. *BMC Health Serv Res.* 2018;18(1):1-9.
400. PHE. Guidance: Accessing PHE data through the Office for Data Release. 2020; https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data. Accessed May 23, 2020.
401. CPRD. CPRD linked data. 2022; https://cprd.com/cprd-linked-data. Accessed June 28, 2022.
402. CPRD, Dedman D. Clinical Practice Research Datalink (CPRD): data access and application process. 2019; https://statistics-group.nihr.ac.uk/wp-content/uploads/2019/02/Dedman_CPRD_RoutineDataWorkshop_Jan2019.pdf. Accessed Apr 30, 2022.
403. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol.* 2019;48(6):1740-1740g.
404. CPRD. CPRD: Pricing. 2022; https://cprd.com/pricing. Accessed Jun 22, 2022.
405. ClinicalTrials.gov. Efficacy and Safety Comparison of RAD001 Versus Sunitinib in the First-line and Second-line Treatment of Patients With Metastatic Renal Cell Carcinoma (RECORD-3) (NCT00903175). 2011; https://clinicaltrials.gov/study/NCT00903175. Accessed March 31, 2021.
406. Motzer RJ, Barrios CH, Kim TM, et al. Phase II randomized trial comparing sequential first-line everolimus and second-line sunitinib versus first-line sunitinib and second-line everolimus in patients with metastatic renal cell carcinoma. *J Clin Oncol.* 2014;32(25):2765-2772.
407. Leeds Institute for Data Analytics. DATA-CAN - harnessing, enhancing and using cancer datasets for research and improved patient outcomes. 2022; https://lida.leeds.ac.uk/data-can/. Accessed Dec 18, 2023.
408. DATA-CAN. DATA-CAN: Health Data Research Hub for Cancer. 2022; https://www.data-can.org.uk/. Accessed Dec 18, 2023.

409. NPCA. National Prostate Cancer Audit: Annual Report 2017. 2018; https://www.npca.org.uk/reports/npca-annual-report-2017/. Accessed Apr 30, 2021.
410. NPCA. National Prostate Cancer Audit: Annual Report 2019. 2020; https://www.npca.org.uk/reports/npca-annual-report-2019/. Accessed Apr 30, 2021.
411. NPCA. National Prostate Cancer Audit: Annual Report 2020. 2021; https://www.npca.org.uk/content/uploads/2021/02/NPCA-Short-Report-2021_Using-the-CPG-in-the-NPCA_Final-11.02.21.pdf. Accessed Apr 30, 2021.
412. NPCA. National Prostate Cancer Audit: First Year Annual Report – Organisation of Services and Analysis of Existing Clinical Data. 2014; https://www.npca.org.uk/content/uploads/2018/02/NPCA-Annual-Report-FINAL-10_11_14-1.pdf. Accessed Apr 30, 2021.
413. NPCA. National Prostate Cancer Audit (NPCA): Annual Report 2020 - Methodology Supplement. 2021; https://www.npca.org.uk/content/uploads/2021/01/NPCA-Annual-Report-2020_Methodology-Supplement-140121.pdf. Accessed Apr 30, 2021.
414. England PH. Cancer Outcome and Services Data (COSD) set - User guide v9.0.3, 2019. 2020; http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd. Accessed Jan, 2021.
415. Rich AL, Tata LJ, Stanley RA, et al. Lung cancer in England: information from the national lung cancer audit (LUCADA). *Lung Cancer.* 2011;72(1):16-22.
416. Khakwani A, Jack RH, Vernon S, et al. Apples and pears? A comparison of two sources of national lung cancer audit data in England. *ERJ open research.* 2017;3(3).
417. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate survival in patients with colorectal cancer: cohort study. *BMJ.* 2017;357.
418. Clift AK, Coupland CA, Hippisley-Cox J. Prostate-specific antigen testing and opportunistic prostate cancer screening: a cohort study in England, 1998–2017. *Br J Gen Pract.* 2021;71(703):e157-e165.
419. Hippisley-Cox J, Coupland C. Predicting the risk of prostate cancer in asymptomatic men: a cohort study to develop and validate a novel algorithm. *Br J Gen Pract.* 2021;71(706):e364-e371.
420. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf.* 2007;16(4):393-401.
421. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. *Pharmacoepidemiol Drug Saf.* 2009;18(8):730-736.
422. Cea Soriano L, Zong J, García Rodríguez LA. Feasibility and validity of The Health Improvement Network database of primary care electronic health records to identify and characterise patients with small cell lung cancer in the United Kingdom. *BMC Cancer.* 2019;19(1):91.
423. Leahy TP, Ramagopalan S, Sammon C. The use of UK primary care databases in health technology assessments carried out by the National Institute for health and care excellence (NICE). *BMC Health Serv Res.* 2020;20(1):675.
424. Young GJ, Harrison S, Turner EL, et al. Prostate-specific antigen (PSA) testing of men in UK general practice: a 10-year longitudinal cohort study. *BMJ open.* 2017;7(10):e017729.

425. Vinogradova Y, Dening T, Hippisley-Cox J, Taylor L, Moore M, Coupland C. Hormone replacement therapy and dementia risk: nested case-control studies using CPRD and QResearch. *East Midlands Research into Ageing Network (EMRAN) Discussion Paper Series.* 2021;38.
426. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015;44(3):827-836.
427. Nuffield department of primary care health. QResearch: Now Available to Researchers at All UK Universities. 2024; https://www.phc.ox.ac.uk/research/qresearch-open-to-all-researchers. Accessed Apr 7, 2024.
428. Akehurst R, Murphy LA, Solà-Morales O, Cunningham D, Mestre-Ferrandiz J, de Pouvourville G. Using real-world data in the health technology assessment of pharmaceuticals: strengths, difficulties and a pragmatic way forward. Elsevier; 2023.
429. Sandhu S, Sharpe M, Findlay Ú, et al. Cohort profile: radiotherapy dataset (RTDS) in England. *BMJ open.* 2023;13(6):e070699.
430. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). *Int J Epidemiol.* 2017;46(4):1093-1093i.
431. Boyd A, Cornish R, Johnson L, Simmonds S, Syddall H, Westbury L. Understanding Hospital episode statistics (HES). *London, UK: CLOSER* 2017; https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf. Accessed Dec 29, 2023.
432. Di Girolamo C, Walters S, Gildea C, et al. Which patients are not included in the English Cancer Waiting Times monitoring dataset, 2009–2013? Implications for use of the data in research. *Br J Cancer.* 2018;118(5):733-737.
433. Di Girolamo C, Walters S, Gildea C, Benitez Majano S, Rachet B, Morris M. Can we assess Cancer Waiting Time targets with cancer survival? A population-based study of individually linked data from the National Cancer Waiting Times monitoring dataset in England, 2009-2013. *PLoS One.* 2018;13(8):e0201288.
434. NCRAS. NCRAS Cancer Registration Data Dictionary v4.4. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1054575/NCRAS_ODR_data_dictionary_v4.4.xlsx. Accessed June 2, 2021.
435. Henson KE, Brock R, Shand B, et al. Cohort profile: prescriptions dispensed in the community linked to the national cancer registry in England. *BMJ open.* 2018;8(7):e020980.
436. Simeone JC, Nordstrom BL, Patel K, Klein AB. Treatment patterns and overall survival in metastatic non-small-cell lung cancer in a real-world, US setting. *Future Oncol.* 2019;15(30):3491-3502.
437. George DJ, Sartor O, Miller K, et al. Treatment patterns and outcomes in patients with metastatic castration-resistant prostate cancer in a real-world clinical practice setting in the united states. *Clin Genitourin Cancer.* 2020.

438. Parrinello CM, Seidl-Rathkopf KN, Bourla AB, Nussbaum NC, Carson KR, Abernethy AP. Comparison of structured versus abstracted comorbidities using oncology EHR data from cancer patients in the flatiron health network. *Value Health.* 2018;21:S14.
439. England PH. A guide to completing the NCRAS data dictinoary. 2020; https://assets.publishing.service.gov.uk/media/61df0d6f8fa8f5058a4b2dd0/Aguide_to_completing_the_NCRAS_data_dictionary.pdf. Accessed Apr 30, 2022.
440. NDRS. CAS-SOP #4.8 Linking treatment tables: chemotherapy, tumour resections and radiotherapy 2023; https://digital.nhs.uk/ndrs/our-work/ncras-work-programme/treatment-data/cas-sop-4.8. Accessed Aug 30, 2023.
441. NDRS. CAS-SOP #4.6 Linking treatment tables: chemotherapy, tumour resections and radiotherapy 2017; https://webarchive.nationalarchives.gov.uk/ukgwa/20230422111656mp_/http://www.ncin.org.uk/view?rid=4299. Accessed Aug 30, 2023.
442. Thorn JC, Turner E, Hounsome L, et al. Validation of the Hospital Episode Statistics Outpatient Dataset in England. *Pharmacoeconomics.* 2016;34(2):161-168.
443. Benedum CM, Sondhi A, Fidyk E, et al. Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning. *Cancers (Basel).* 2023;15(6):1853.
444. Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *Journal of Healthcare Informatics Research.* 2023.
445. PHE. Calculating Treatment Duration for Oral Drugs - Cancer Drugs Fund (CDF) methodology document 2020; http://www.chemodataset.nhs.uk/view?rid=280#:~:text=Treatment%20duration%20is%20calculated%20by,that%20the%20prescription%20is%20filled. Accessed May 23, 2019.
446. PHE. Methodologies applied when using the Systemic-Anti-Cancer Therapy (SACT) dataset to evaluate treatments in the Cancer Drugs Fund (CDF). 2021; http://www.chemodataset.nhs.uk/view?rid=274. Accessed July 15, 2021.
447. FDA. FDA approves abiraterone acetate in combination with prednisone for high-risk metastatic castration-sensitive prostate cancer. 2018; https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-abiraterone-acetate-combination-prednisone-high-risk-metastatic-castration-sensitive. Accessed June 21.
448. NICE. Abiraterone for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen (TA 259). 2012; https://www.nice.org.uk/guidance/ta259. Accessed Apr 30, 2021.
449. Arora S, Charlesworth A, Kelly E, Stoye G. *Public payment and private provision: the changing landscape of health care in the 2000s.* Nuffield Trust London; 2013.

450. Davies EA, Coupland VH, Dixon S, Mokbel K, Jack RH. Comparing the case mix and survival of women receiving breast cancer care from one private provider with other London women with breast cancer: pilot data exchange and analyses. *BMC Cancer.* 2016;16(1):421.
451. Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ.* 2021;373.
452. Macnair A, Love SB, Murray ML, et al. Accessing routinely collected health data to improve clinical trials: recent experience of access. *Trials.* 2021;22(1):340.
453. NHS Digital. Infrastructure operations and support. 2022; https://digital.nhs.uk/services/cloud-centre-of-excellence/infrastructure-operations-and-support. Accessed June 30, 2023.
454. NHS Digital. Independent Group Advising on the Release of Data (IGARD). 2023; https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data. Accessed June 30, 2023.
455. Zhang J, Morley J, Gallifant J, et al. Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation. *The Lancet Digital Health.* 2023;5(10):e737-e748.
456. J. Schuemie M, Soledad Cepede M, A. Suchard M, et al. How Confident Are We About Observational Findings in Health Care: A Benchmark Study. *Harvard Data Science Review.* 2020.
457. Matthews AA, Dahabreh IJ, Fröbert O, et al. Benchmarking observational analyses before using them to address questions trials do not answer: an application to coronary thrombus aspiration. *Am J Epidemiol.* 2022.
458. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet.* 2007;370(9596):1453-1457.
459. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342(25):1887-1892.
460. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000;342(25):1878-1886.
461. MacLehose R, Reeves B, Harvey I, Sheldon T, Russell I, Black A. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health technology assessment (Winchester, England).* 2000;4(34):1-154.
462. Ioannidis JP, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA.* 2001;286(7):821-830.
463. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (RECORD): study design and protocol. *Diabetologia.* 2005;48(9):1726-1735.
464. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone evaluated for cardiovascular outcomes—an interim analysis. *N Engl J Med.* 2007;357(1):28-38.

465. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. *The Lancet.* 2009;373(9681):2125-2135.
466. Mahaffey KW, Hafley G, Dickerson S, et al. Results of a reevaluation of cardiovascular outcomes in the RECORD trial. *Am Heart J.* 2013;166(2):240-249. e241.
467. Rosen CJ. Revisiting the rosiglitazone story—lessons learned. *N Engl J Med.* 2010;363(9):803-806.
468. Rosen CJ. The rosiglitazone story—lessons from an FDA Advisory Committee meeting. *N Engl J Med.* 2007;357(9):844-846.
469. Pouwels KB, van Grootheest K. The rosiglitazone decision process at FDA and EMA. What should we learn? *Int J Risk Saf Med.* 2012;24(2):73-80.
470. Flory JH, Mushlin AI. Observational and clinical trial findings on the comparative effectiveness of diabetes drugs showed agreement. *J Clin Epidemiol.* 2015;68(2):200-210.
471. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation.* 2020.
472. Wang SV, Schneeweiss S, Franklin JM, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. *JAMA.* 2023;329(16):1376-1385.
473. Thompson D. Replication of Randomized, Controlled Trials Using Real-World Data: What Could Go Wrong? : Elsevier; 2020.
474. Crown WH, Bierer BE. Real-World Evidence: Understanding Sources of Variability Through Empirical Analysis. *Value Health.* 2020.
475. Kelly C. FDA's Real-World Evidence Project Finds That Pragmatic Clinical Trials Are Easier To DUPLICATE. 2022; https://pink.pharmaintelligence.informa.com/PS146169/FDAs-Real-World-Evidence-Project-Finds-That-Pragmatic-Clinical-Trials-Are-Easier-To-DUPLICATE. Accessed June 15, 2023.
476. OptumLabs. Using RWD in regulatory decision-making. 2020; https://www.optumlabs.com/work/data-regulatory-decision.html. Accessed Jan 15, 2021.
477. Morris TP, van Smeden M. Causal analyses of existing databases: the importance of understanding what can be achieved with your data before analysis (commentary on Hernán). *J Clin Epidemiol.* 2022;142:261-263.
478. Antonarakis ES, Tagawa ST, Galletti G, et al. Randomized, Noncomparative, Phase II Trial of Early Switch From Docetaxel to Cabazitaxel or Vice Versa, With Integrated Biomarker Analysis, in Men With Chemotherapy-Naïve, Metastatic, Castration-Resistant Prostate Cancer. *J Clin Oncol.* 2017;35(28):3181-3188.
479. Shitara K, Yamanaka T, Denda T, et al. REVERCE: a randomized phase II study of regorafenib followed by cetuximab versus the reverse sequence for previously treated metastatic colorectal cancer patients. *Ann Oncol.* 2019;30(2):259-265.

480. Cascinu S, Rosati G, Nasti G, et al. Treatment sequence with either irinotecan/cetuximab followed by FOLFOX-4 or the reverse strategy in metastatic colorectal cancer patients progressing after first-line FOLFIRI/bevacizumab: An Italian Group for the Study of Gastrointestinal Cancer phase III, randomised trial comparing two sequences of therapy in colorectal metastatic patients. *Eur J Cancer.* 2017;83:106-115.
481. Modest DP, Stintzing S, von Weikersthal LF, et al. Impact of Subsequent Therapies on Outcome of the FIRE-3/AIO KRK0306 Trial: First-Line Therapy With FOLFIRI Plus Cetuximab or Bevacizumab in Patients With KRAS Wild-Type Tumors in Metastatic Colorectal Cancer. *J Clin Oncol.* 2015;33(32):3718-3726.
482. Chibaudel B, Bonnetain F, Tournigand C, et al. STRATEGIC-1: A multiple-lines, randomized, open-label GERCOR phase III study in patients with unresectable wild-type RAS metastatic colorectal cancer. *BMC Cancer.* 2015;15:496.
483. Wu YL, Zhou C, Lu S, et al. Erlotinib versus gemcitabine/cisplatin in Chinese patients with EGFR mutation-positive advanced non-small-cell lung cancer: Crossover extension and post-hoc analysis of the ENSURE study. *Lung Cancer.* 2019;130:18-24.
484. Genovese MC, van Vollenhoven RF, Wilkinson B, et al. Switching from adalimumab to tofacitinib in the treatment of patients with rheumatoid arthritis. *Arthritis Res Ther.* 2016;18:145.
485. Yim HJ, Kim W, Ahn SH, et al. Besifovir Dipivoxil Maleate 144-Week Treatment of Chronic Hepatitis B: An Open-Label Extensional Study of a Phase 3 Trial. *Am J Gastroenterol.* 2020;115(8):1217-1225.
486. Tanaka S, Mori S, Hagino H, Sugimoto T. Design of a randomized trial of teriparatide followed by alendronate: Japanese Osteoporosis Intervention Trial-05 (JOINT-05). *J Bone Miner Metab.* 2020;38(3):412-417.
487. Kendler DL, Bone HG, Massari F, et al. Bone mineral density gains with a second 12-month course of romosozumab therapy following placebo or denosumab. *Osteoporos Int.* 2019;30(12):2437-2448.
488. Naar S, Ellis D, Idalski Carcone A, et al. Outcomes From a Sequential Multiple Assignment Randomized Trial of Weight Loss Strategies for African American Adolescents With Obesity. *Ann Behav Med.* 2019;53(10):928-938.
489. Schoenfelder EN, Chronis-Tuscano A, Strickland J, Almirall D, Stein MA. Piloting a Sequential, Multiple Assignment, Randomized Trial for Mothers with Attention-Deficit/Hyperactivity Disorder and Their At-Risk Young Children. *J Child Adolesc Psychopharmacol.* 2019;29(4):256-267.
490. Nelson B, Amminger GP, Yuen HP, et al. Staged Treatment in Early Psychosis: A sequential multiple assignment randomised trial of interventions for ultra high risk of psychosis patients. *Early intervention in psychiatry.* 2018;12(3):292-306.
491. Yoon H, Lee DH, Jang ES, et al. Effects of N-acetylcysteine on First-Line Sequential Therapy for Helicobacter pylori Infection: A Randomized Controlled Pilot Trial. *Gut and liver.* 2016;10(4):520-525.
492. Sibbald B, Roberts C. Understanding controlled trials. Crossover trials. *BMJ.* 1998;316(7146):1719.

493. Alliance NG. *Early and locally advanced breast cancer: diagnosis and management - NICE guideline [NG101]*. London: National Institute for Health and Care Excellence,; 2018.
494. NICE. NICE guideline: Lung cancer: diagnosis and management (NG 122). 2019; https://www.nice.org.uk/guidance/ng122. Accessed May 16, 2020.
495. NICE. Prostate cancer: diagnosis and management (NG 131). 2019; https://www.nice.org.uk/guidance/ng131. Accessed Apr 30, 2021.
496. Gradishar WJ, Anderson BO, Abraham J, et al. Breast cancer, version 3.2020, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw.* 2020;18(4):452-478.
497. Armstrong DK, Alvarez RD, Bakkum-Gamez JN, et al. NCCN Guidelines Insights: Ovarian Cancer, Version 1.2019: Featured Updates to the NCCN Guidelines. *J Natl Compr Canc Netw.* 2019;17(8):896-909.
498. Jonasch E. NCCN guidelines updates: management of metastatic kidney cancer. *J Natl Compr Canc Netw.* 2019;17(5.5):587-589.
499. Tempero MA. NCCN guidelines updates: pancreatic cancer. *J Natl Compr Canc Netw.* 2019;17(5.5):603-605.
500. van de Velde CJH, Rea D, Seynaeve C, et al. Adjuvant tamoxifen and exemestane in early breast cancer (TEAM): a randomised phase 3 trial. *The Lancet.* 2011;377(9762):321-331.
501. LeCaer H, Barlesi F, Corre R, et al. A multicentre phase II randomised trial of weekly docetaxel/gemcitabine followed by erlotinib on progression, vs the reverse sequence, in elderly patients with advanced non small-cell lung cancer selected with a comprehensive geriatric assessment (the GFPC 0504 study). *Br J Cancer.* 2011;105(8):1123-1130.
502. Pignata S, Scambia G, Bologna A, et al. Randomized Controlled Trial Testing the Efficacy of Platinum-Free Interval Prolongation in Advanced Ovarian Cancer: The MITO-8, MaNGO, BGOG-Ov1, AGO-Ovar2.16, ENGOT-Ov1, GCIG Study. *J Clin Oncol.* 2017;35(29):3347-3353.
503. Dahan L, Bonnetain F, Ychou M, et al. Combination 5-fluorouracil, folinic acid and cisplatin (LV5FU2-CDDP) followed by gemcitabine or the reverse sequence in metastatic pancreatic cancer: final results of a randomised strategic phase III trial (FFCD 0301). *Gut.* 2010;59(11):1527-1534.
504. Heinemann V, Vehling-Kaiser U, Waldschmidt D, et al. Gemcitabine plus erlotinib followed by capecitabine versus capecitabine plus erlotinib followed by gemcitabine in advanced pancreatic cancer: final results of a randomised phase 3 trial of the 'Arbeitsgemeinschaft Internistische Onkologie' (AIO-PK0104). *Gut.* 2013;62(5):751-759.
505. NICE. NICE Pathways: Early and locally advanced breast cancer. 2021; https://pathways.nice.org.uk/pathways/early-and-locally-advanced-breast-cance. Accessed Jan 5, 2021.
506. Matthews A, Stanway S, Farmer RE, et al. Long term adjuvant endocrine therapy and risk of cardiovascular disease in female breast cancer survivors: systematic review. *BMJ.* 2018;363:k3845.

507. Hess LM, Cui ZL, Mytelka DS, Han Y, Goodloe R, Schelman W. Treatment patterns and survival outcomes for patients receiving second-line treatment for metastatic colorectal cancer in the USA. *Int J Colorectal Dis.* 2019;34(4):581-588.
508. Lester J, Escriu C, Khan S, et al. Retrospective analysis of real-world treatment patterns and clinical outcomes in patients with advanced non-small cell lung cancer starting first-line systemic therapy in the United Kingdom. *BMC Cancer.* 2021;21(1):515.
509. Tan C-S, Gilligan D, Pacey S. Treatment approaches for EGFR-inhibitor-resistant patients with non-small-cell lung cancer. *The lancet oncology.* 2015;16(9):e447-e459.
510. Li Y, Appius A, Pattipaka T, Feyereislova A, Cassidy A, Ganti AK. Real-world management of patients with epidermal growth factor receptor (EGFR) mutation-positive non–small-cell lung cancer in the USA. *PLoS One.* 2019;14(1):e0209709.
511. NICE. NICE Pathways: Managing pancreatic cancer. 2021; https://pathways.nice.org.uk/pathways/pancreatic-cancer. Accessed Jan 5.
512. O'Reilly EM, Surinach A, Wu Z, Cockrum P. Real-world patterns of care among patients with metastatic pancreatic cancer (mPC). American Society of Clinical Oncology; 2020.
513. NICE. Docetaxel for the treatment of hormone-refractory metastatic prostate cancer (TA 101). 2006; https://www.nice.org.uk/guidance/ta101. Accessed Apr 30, 2021.
514. NICE. Enzalutamide for metastatic hormone-relapsed prostate cancer previously treated with a docetaxel-containing regimen (TA 316). 2014; https://www.nice.org.uk/guidance/ta316. Accessed Apr 30, 2021.
515. McGrane J, Rowe M, Hidayat A, et al. The Impact of Treatment Crossover and Prostate-specific Antigen Flare with Enzalutamide and Abiraterone in Metastatic Castrate Resistant Prostate Cancer. *Clin Oncol (R Coll Radiol).* 2021;33(1):e92.
516. Mori K, Miura N, Mostafaei H, et al. Sequential therapy of abiraterone and enzalutamide in castration-resistant prostate cancer: a systematic review and meta-analysis. *Prostate Cancer Prostatic Dis.* 2020;23(4):539-548.
517. Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data. *BMJ open.* 2020;10(8):e037719.
518. National Institute for Clinical Excellence. Everolimus for the second-line treatment of advanced renal cell carcinoma (Technology Appraisal 219). 2011

https://www.nice.org.uk/guidance/ta219. Accessed Jan 1, 2021.
519. National Institute for Clinical Excellence. Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma (Technology Appraisal 169). 2009; https://www.nice.org.uk/guidance/ta169. Accessed Jan 01, 2021.
520. NICE. NICE Pathways: Renal cancer overview. 2021; https://pathways.nice.org.uk/pathways/renal-cancer. Accessed Jan 5.

521. Hawkins R, Fife K, Hurst M, et al. Treatment patterns and health outcomes in metastatic renal cell carcinoma patients treated with targeted systemic therapies in the UK. *BMC Cancer.* 2020;20(1):670.
522. Motzer RJ, Jonasch E, Boyle S, et al. NCCN Guidelines Insights: Kidney Cancer, Version 1.2021: Featured Updates to the NCCN Guidelines. *J Natl Compr Canc Netw.* 2020;18(9):1160-1170.
523. Motzer RJ, Jonasch E, Agarwal N, et al. Kidney cancer, version 3.2015. *J Natl Compr Canc Netw.* 2015;13(2):151-159.
524. Webster-Clark M, Ross RK, Lund JL. Initiator types and the causal question of the prevalent new-user design: A simulation study. *Am J Epidemiol.* 2021;190(7):1341-1348.
525. NCRAS. CancerData. https://www.cancerdata.nhs.uk/. Accessed Jan 4, 2021.
526. Hirst C, Cabrera C, Kirby M. Epidemiology of castration resistant prostate cancer: a longitudinal analysis using a UK primary care database. *Cancer Epidemiol.* 2012;36(6):e349-e353.
527. NICE. NICE Pathways: Prostate cancer overview. 2021; https://pathways.nice.org.uk/pathways/prostate-cancer. Accessed Feb 5, 2021.
528. NICE. Radium-223 dichloride for treating hormone-relapsed prostate cancer with bone metastases (TA 412). 2016; https://www.nice.org.uk/guidance/ta412. Accessed Apr 30, 2021.
529. NICE. Darolutamide with androgen deprivation therapy for treating hormone-relapsed non-metastatic prostate cancer (TA 660). 2020; https://www.nice.org.uk/guidance/ta660. Accessed Apr 30, 2021.
530. NICE. Apalutamide with androgen deprivation therapy for treating high-risk hormone-relapsed non-metastatic prostate cancer (TA 740). 2021; https://www.nice.org.uk/guidance/ta740. Accessed Dec 2, 2023.
531. NHS England. National Cancer Drugs Fund (CDF) list. 2023; https://www.england.nhs.uk/publication/national-cancer-drugs-fund-list/. Accessed Dec 3, 2023.
532. Vale CL, Fisher DJ, Godolphin PJ, et al. Which patients with metastatic hormone-sensitive prostate cancer benefit from docetaxel: a systematic review and meta-analysis of individual participant data from randomised trials. *The Lancet Oncology.* 2023;24(7):783-797.
533. NICE. Enzalutamide for treating hormone-sensitive metastatic prostate cancer (TA 712). 2021; https://www.nice.org.uk/guidance/ta712. Accessed Dec 2, 2023.
534. NICE. Apalutamide with androgen deprivation therapy for treating hormone-sensitive metastatic prostate cancer (TA 741). 2021; https://www.nice.org.uk/guidance/ta741. Accessed Dec 2, 2023.
535. NICE. Darolutamide with androgen deprivation therapy and docetaxel for treating hormone-sensitive metastatic prostate cancer (TA 903). 2023; https://www.nice.org.uk/guidance/ta903. Accessed Apr 30, 2021.
536. Gravis G, Boher J-M, Joly F, et al. Androgen deprivation therapy (ADT) plus docetaxel versus ADT alone in metastatic non castrate prostate cancer: impact of metastatic burden and long-term survival analysis of the randomized phase 3 GETUG-AFU15 trial. *Eur Urol.* 2016;70(2):256-262.

537. Gravis G, Fizazi K, Joly F, et al. Androgen-deprivation therapy alone or with docetaxel in non-castrate metastatic prostate cancer (GETUG-AFU 15): a randomised, open-label, phase 3 trial. *The lancet oncology.* 2013;14(2):149-158.
538. Kyriakopoulos CE, Chen Y-H, Carducci MA, et al. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer: long-term survival analysis of the randomized phase III E3805 CHAARTED trial. *J Clin Oncol.* 2018;36(11):1080.
539. Clarke NW, Ali A, Ingleby F, et al. Addition of docetaxel to hormonal therapy in low-and high-burden metastatic hormone sensitive prostate cancer: long-term survival results from the STAMPEDE trial. *Ann Oncol.* 2019;30(12):1992-2003.
540. James ND, Sydes MR, Clarke NW, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet.* 2016;387(10024):1163-1177.
541. Group PCTC. Maximum androgen blockade in advanced prostate cancer: an overview of the randomised trials. *The Lancet.* 2000;355(9214):1491-1498.
542. FDA. FDA approves enzalutamide for metastatic castration-sensitive prostate cancer. 2019; https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-enzalutamide-metastatic-castration-sensitive-prostate-cancer. Accessed June 21.
543. Cancer Research UK. Kidney cancer statistics. 2021; https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/kidney-cancer. Accessed Apr 30, 2021.
544. NICE. Cabozantinib for untreated advanced renal cell carcinoma (TA 542). 2018; https://www.nice.org.uk/Guidance/TA417. Accessed Apr 30, 2021.
545. NICE. Nivolumab for previously treated advanced renal cell carcinoma (TA 417). 2016; https://www.nice.org.uk/Guidance/TA417. Accessed Apr 30, 2021.
546. Martinez Chanza N, Tripathi A, Harshman LC. Adjuvant Therapy Options in Renal Cell Carcinoma: Where Do We Stand? *Curr Treat Options Oncol.* 2019;20(5):44.
547. Labriola MK, George DJ. Setting a new standard for long-term survival in metastatic kidney cancer. Vol 128: Wiley Online Library; 2022:2058-2060.
548. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Value Health.* 2017;20(8):1003-1008.
549. Flatiron Health. NICE Partners with Flatiron Health to Develop Real-World Evidence Research Methodologies. 2020; https://flatiron.com/resources/press-release/nice-partnership-2020. Accessed June 30, 2023.
550. Flatiron Health. Flatiron Health Presents Research at ISPOR Europe. 2021; https://flatiron.com/resources/press-release/flatiron-health-presents-research-at-ispor-europe. Accessed June 30, 2023.

551. NICE. Dabrafenib plus trametinib for treating BRAF V600 mutation-positive advanced non-small-cell lung cancer (TA 898). 2023; https://www.nice.org.uk/guidance/ta898. Accessed Oct 20, 2023.
552. Crown W, Dahabreh IJ, Li X, Toh S, Bierer B. Can Observational Analyses of Routinely Collected Data Emulate Randomized Trials? Design and Feasibility of the Observational Patient Evidence for Regulatory Approval Science and Understanding Disease Project. *Value Health.* 2023;26(2):176-184.
553. Franklin JM, Pawar A, Martin D, et al. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin Pharmacol Ther.* 2020;107(4):817-826.
554. Lorente D, Mateo J, Perez-Lopez R, de Bono JS, Attard G. Sequencing of agents in castration-resistant prostate cancer. *The Lancet Oncology.* 2015;16(6):e279-e292.
555. Tran K, McGill S. Treatment Sequences of Androgen Receptor–Targeted Agents for Prostate Cancer. *Canadian Journal of Health Technologies.* 2021;1(3).
556. Cassinello J, Dominguez-Lubillo T, Gomez-Barrera M, et al. Optimal treatment sequencing of abiraterone acetate plus prednisone and enzalutamide in patients with castration-resistant metastatic prostate cancer: A systematic review and meta-analysis. *Cancer Treat Rev.* 2021;93:102152.
557. Cattrini C, España R, Mennitto A, et al. Optimal sequencing and predictive biomarkers in patients with advanced prostate cancer. *Cancers (Basel).* 2021;13(18):4522.
558. NICE. Padeliporfin for untreated localised prostate cancer (TA 546). 2018; https://www.nice.org.uk/guidance/ta546. Accessed Apr 30, 2021.
559. NICE. Olaparib for preOlaparib for previously treated BRCA mutation-positive hormone-relapsed metastatic prostate cancer (TA 887). 2023; https://www.nice.org.uk/guidance/ta887. Accessed Dec 3, 2023.
560. NICE. Lutetium-177 vipivotide tetraxetan for treating PSMA-positive hormone-relapsed metastatic prostate cancer after 2 or more treatments (TA 930). 2023; https://www.nice.org.uk/guidance/ta930. Accessed Dec 3, 2023.
561. Schaeffer EM, Srinivas S, Adra N, et al. NCCN Guidelines® insights: prostate cancer, version 1.2023: featured updates to the NCCN guidelines. *J Natl Compr Canc Netw.* 2022;20(12):1288-1298.
562. Chi KN, Eigl B, Seto L, et al. A Randomized Phase II Study of Sequencing Abiraterone Acetate and Enzalutamide in Metastatic Castration-Resistant Prostate Cancer: GUTG-001 (Veresion 6.0, 16Jul2018). 2018; https://classic.clinicaltrials.gov/ProvidedDocs/57/NCT02125357/Prot_SAP_000.pdf. Accessed June 30, 2023.
563. Annala M, Vandekerkhove G, Khalaf D, et al. Circulating tumor DNA genomics correlate with resistance to abiraterone and enzalutamide in prostate cancer. *Cancer Discov.* 2018;8(4):444-457.

564. Khalaf DJ, Sunderland K, Eigl BJ, et al. Health-related quality of life for abiraterone plus prednisone versus enzalutamide in patients with metastatic castration-resistant prostate cancer: results from a phase II randomized trial. *Eur Urol.* 2019;75(6):940-947.
565. Pereira-Salgado A, Kwan EM, Tran B, Gibbs P, De Bono J, IJzerman M. Systematic review of efficacy and health economic implications of real-world treatment sequencing in prostate cancer: Where do the newer agents enzalutamide and abiraterone fit in? *European urology focus.* 2021;7(4):752-763.
566. NICE. Avelumab with axitinib for untreated advanced renal cell carcinoma (TA 645). 2020; https://www.nice.org.uk/guidance/ta377. Accessed Apr 30, 2021.
567. Iacovelli R, Ciccarese C, Bria E, et al. Immunotherapy versus standard of care in metastatic renal cell carcinoma. A systematic review and meta-analysis. *Cancer Treat Rev.* 2018;70:112-117.
568. NICE. Axitinib for treating advanced renal cell carcinoma after failure of prior systemic treatment (TA 333). 2015; https://www.nice.org.uk/guidance/ta333. Accessed Mar 23, 2020.
569. Géczi L, Bodoky G, Rokszin G, Fábián I, Torday L. Survival benefits of second-line axitinib versus everolimus after first line sunitinib treatment in metastatic renal cell carcinoma. *Pathology & Oncology Research.* 2020;26(4):2201-2207.
570. Iacovelli R, Cartenì G, Milella M, et al. Clinical outcomes in patients with metastatic renal cell carcinoma receiving everolimus or temsirolimus after sunitinib. *Canadian Urological Association Journal.* 2014;8(3-4):E121.
571. Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. *Clin Epidemiol.* 2020;12:457-467.
572. ESMO. Performance scales: Karnofsky & ECOG scores. https://oncologypro.esmo.org/oncology-in-practice/practice-tools/performance-scales. Accessed Mar 15, 2021.
573. Heyard R, Held L, Schneeweiss S, Wang SV. Design differences explain variation in results between randomized trials and their non-randomized emulations. *medRxiv.* 2023.
574. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol.* 2012;12:9.
575. Minassian C, Williams R, Meeraus WH, Smeeth L, Campbell OM, Thomas SL. Methods to generate and validate a Pregnancy Register in the UK Clinical Practice Research Datalink primary care database. *Pharmacoepidemiol Drug Saf.* 2019;28(7):923-933.
576. Nuttall M, van der Meulen J, Emberton M. Charlson scores based on ICD-10 administrative data were valid in assessing comorbidity in patients undergoing urological cancer surgery. *J Clin Epidemiol.* 2006;59(3):265-273.
577. Austin SR, Wong Y-N, Uzzo RG, Beck JR, Egleston BL. Why summary comorbidity measures such as the Charlson comorbidity index and Elixhauser score work. *Med Care.* 2015;53(9):e65.
578. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol.* 2011;64(7):749-759.

579. Sun JW, Rogers JR, Her Q, et al. Adaptation and Validation of the Combined Comorbidity Score for ICD-10-CM. *Med Care.* 2017;55(12):1046-1051.
580. Kwee SA, Wong LL, Ludema C, et al. Target Trial Emulation: A Design Tool for Cancer Clinical Trials. *JCO Clinical Cancer Informatics.* 2023(7):e2200140.
581. Zuo H, Yu L, Campbell SM, Yamamoto SS, Yuan Y. The implementation of target trial emulation for causal inference: a scoping review. *J Clin Epidemiol.* 2023;162:29-37.
582. Bakker LJ, Goossens LMA, O'Kane MJ, Uyl-de Groot CA, Redekop WK. Analysing electronic health records: The benefits of target trial emulation. *Health Policy and Technology.* 2021;10(3):100545.
583. NHS. Cancer Data: Detailed Statistics: prostate cancer. 2022; https://www.cancerdata.nhs.uk/getdataout/prostate. Accessed March 22, 2022.
584. NHS. Cancer Data: Detailed Statistics: kidney cancer. 2022; https://www.cancerdata.nhs.uk/getdataout/kidney. Accessed June 30, 2022.
585. Ljungberg B, Campbell SC, Cho HY, et al. The Epidemiology of Renal Cell Carcinoma. *Eur Urol.* 2011;60(4):615-621.
586. Choueiri TK, Tomczak P, Park SH, et al. Adjuvant pembrolizumab after nephrectomy in renal-cell carcinoma. *N Engl J Med.* 2021;385(8):683-694.
587. Zhao SS, Lyu H, Solomon DH, Yoshida K. Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework. *Ann Rheum Dis.* 2020;79(7):883-890.
588. ClinicalTrials.gov. Replication of the TRITON-TIMI Antiplatelet Trial in Healthcare Claims Data (NCT04237922). 2019; https://classic.clinicaltrials.gov/ct2/show/NCT04237922. Accessed March 29, 2024.
589. Division of Pharmacoepidemiology and Pharmacoeconomics at the Brigham and Women's Hospital. Combined Comorbidity Score for Claims Data. https://assets.website-files.com/5ddc21a55d412bcfa838a006/5f4520e0df1b213317f0604e_combined_comorbidity_score_ICD9_10_code.txt. Accessed Sep 1, 2023.
590. North Bristol NHS Trust. Test Information - Alkaline phosphatase. 2023; https://www.nbt.nhs.uk/severn-pathology/requesting/test-information/alp. Accessed Sep 30, 2023.
591. York and Scarborough Teaching Hospitals NHS Foundation Trust. Laboratory Medicine - Test Directory - Lactate Dehydrogenase. 2023; https://www.yorkhospitals.nhs.uk/our-services/a-z-of-services/lab-med/test-directory/clinical-biochemistry/lactate-dehydrogenase/. Accessed Sep 30, 2023.
592. Alberto IRI, Alberto NRI, Ghosh AK, et al. The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet Digital health.* 2023;5(5):e288-e294.
593. Baio G. survHE: survival analysis for health economic evaluation and cost-effectiveness modeling. *Journal of Statistical Software.* 2020;95:1-47.

594. Shu D, Young JG, Toh S, Wang R. Variance estimation in inverse probability weighted Cox models. *Biometrics.* 2021;77(3):1101-1117.
595. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med.* 2017;167(4):268-274.
596. Mainzer R, Moreno-Betancur M, Nguyen C, Simpson J, Carlin J, Lee K. Handling of missing data with multiple imputation in observational studies that address causal questions: protocol for a scoping review. *BMJ open.* 2023;13(2):e065576.
597. Tompsett D, Zylbersztejn A, Hardelid P, De Stavola B. Target Trial Emulation and Bias Through Missing Eligibility Data: An Application to a Study of Palivizumab for the Prevention of Hospitalization Due to Infant Respiratory Illness. *Am J Epidemiol.* 2023;192(4):600-611.
598. Latimer NR. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Med Decis Making.* 2013;33(6):743-754.
599. Wang SV, Schneeweiss S. Data Checks Before Registering Study Protocols for Health Care Database Analyses. *JAMA.* 2024.
600. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med.* 2019;38(11):2074-2102.

Appendix 3.1 R code for systematic full-text retrieval of NICE TA documents

```
################################################################
# Project: Review NICE TA regarding sequencing problem
#         automation of downloading TA document
#         manually downloading 60 files can take 2 hours and there
#         might be mistake due to manual incorrect entry
# Create: JY Amy Chang
# Date: 02Mar2020
################################################################

library(bitops)
library(RCurl)
library(XML)
library(httr)
library(xml2)
library(rvest)
library(stringr)
library(pagedown)
library(truncnorm)

# import active TA list for review
TA_list <- read.csv(file = "raw/TA_list_20191201.csv", header = FALSE)
TA_vector <- as.vector(TA_list[,"V1"])

# transform TA list into vectors for creating ulr and file name for bulk download
# sort vector to download from lastest to the oldest
TA_vector <- sort(TA_vector, decreasing = T)

# create a function for downloading files, return data frame error_vec that indicates which files are not downloaded
file_download <- function(TA_number, url, file_name) [
error_vec <- rep("NA", times = length(TA_number)) # 460
download_status <- data.frame (TA = TA_number,
                        Status = error_vec)
download_status$Status <- as.character(download_status$Status)
for (n in 1:length(TA_number)) [
   tryCatch([
   download.file(url[n], destfile = file_name[n], mode="wb")
   download_status$Status[n] <- "downloaded"
   ]
, error = function(e)[
       cat("ERROR :",conditionMessage(e), "\n")
       ])
]
download_status[download_status$Status != "downloaded", "Status"] <- "NA"
return(download_status)
]

##############################
# FAD
##############################

# download FAD #not all file use the same logic of url
url <- paste("https://www.nice.org.uk/guidance/ta",TA_vector,"/documents/final-appraisal-determination-document", sep = "")
file_name <- paste("TA", TA_vector, "_FAD.pdf", sep = "")
Fulltext_FAD <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
length(Fulltext_FAD[Fulltext_FAD$Status == "NA","TA"]) # 259 undownloaded
write.csv(Fulltext_FAD, "Fulltext_FAD.csv")

# download FAD #not all file use the same logic of url
# url <- paste("https://www.nice.org.uk/guidance/ta",
#           Fulltext_FAD[Fulltext_FAD$Status == "NA", "TA"],
#           "/documents/final-appraisal-determination-document-2", sep = "")
# file_name <- paste("TA", Fulltext_FAD[Fulltext_FAD$Status == "NA", "TA"], "_FAD.pdf", sep = "")
# Fulltext_FAD_temp <- file_download(TA_number = Fulltext_FAD[Fulltext_FAD$Status == "NA", "TA"],
#                         url = url, file_name = file_name)
#???Fulltext_FAD_temp <- Fulltext_FADtemp[Fulltext_FAD_temp$Status != "NA", ] #8 [1] 606 557 527 515 510 421 405 374
# update original Fulltext_FAD
# Fulltext_FAD[Fulltext_FAD$TA %in% Fulltext_FAD_temp$TA, "Status"] <- "downloaded"
# length(Fulltext_FAD[Fulltext_FAD$Status == "NA","TA"]) #259 downloaded (251 missing)
# write.csv(Fulltext_FAD, "Fulltext_FAD.csv")

# url with document 2
url <- paste("https://www.nice.org.uk/guidance/ta",TA_vector,"/documents/final-appraisal-determination-document-2", sep = "")
file_name <- paste("TA", TA_vector, "_FAD2.pdf", sep = "")
```

```
Fulltext_FAD2 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_FAD2[Fulltext_FAD2$Status == "downloaded", "TA"] # 42 downloaded
Fulltext_FAD$Status2 <- Fulltext_FAD2$Status

# test if there is document 1
# url <- paste("https://www.nice.org.uk/guidance/ta",TA_vector,"/documents/final-appraisal-determination-document-1", sep = "")
# file_name <- paste("TA", TA_vector, "_FAD1.pdf", sep = "")
# Fulltext_FAD1 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
# Fulltext_FAD1[Fulltext_FAD1$Status == "downloaded", "TA"] # 0 downloaded

# url with document 3
url <- paste("https://www.nice.org.uk/guidance/ta",TA_vector,"/documents/final-appraisal-determination-document-3", sep = "")
file_name <- paste("TA", TA_vector, "_FAD3.pdf", sep = "")
Fulltext_FAD3 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_FAD3[Fulltext_FAD3$Status == "downloaded", "TA"] # 9 downloaded
Fulltext_FAD$Status3 <- Fulltext_FAD3$Status

# url with document 4
url <- paste("https://www.nice.org.uk/guidance/ta",TA_vector,"/documents/final-appraisal-determination-document-4", sep = "")
file_name <- paste("TA", TA_vector, "_FAD4.pdf", sep = "")
Fulltext_FAD4 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_FAD4[Fulltext_FAD4$Status == "downloaded", "TA"] # 2 downloaded
Fulltext_FAD$Status4 <- Fulltext_FAD4$Status #[1] 491 487 these two has more documents due to CDF and managed access


##############################
# CP
##############################

# Committee paper is more complicated than FAD as the first committee paper is usually consultation document

# CP document 1
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers", sep = "")
file_name <- paste("TA", TA_vector, "_CP.pdf", sep = "")
Fulltext_CP <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
length(Fulltext_CP[Fulltext_CP$Status == "downloaded","TA"]) # 204 undownloaded
#write.csv(Fulltext_FAD, "Fulltext_CP.csv")

# CP document 2
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-2", sep = "")
file_name <- paste("TA", TA_vector, "_CP2.pdf", sep = "")
Fulltext_CP2 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
length(Fulltext_CP2[Fulltext_CP2$Status == "downloaded","TA"]) # 136 undownloaded
Fulltext_CP$Status2 <- Fulltext_CP2$Status

# CP document 3
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-3", sep = "")
file_name <- paste("TA", TA_vector, "_CP3.pdf", sep = "")
Fulltext_CP3 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
length(Fulltext_CP3[Fulltext_CP3$Status == "downloaded","TA"]) # 63 undownloaded
Fulltext_CP$Status3 <- Fulltext_CP3$Status

# CP document 4 (sometimes it can be just slides)
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-4", sep = "")
file_name <- paste("TA", TA_vector, "_CP4.pdf", sep = "")
Fulltext_CP4 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
length(Fulltext_CP4[Fulltext_CP4$Status == "downloaded","TA"]) # 32 downloaded
Fulltext_CP$Status4 <- Fulltext_CP4$Status

# CP document 5 (can be managing access agreement)
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-5", sep = "")
file_name <- paste("TA", TA_vector, "_CP5.pdf", sep = "")
Fulltext_CP5 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP5[Fulltext_CP5$Status == "downloaded","TA"] # 17 downloaded
# [1] 588 541 510 502 495 491 484 483 479 474 473 472 445 432 423 417 402
Fulltext_CP$Status5 <- Fulltext_CP5$Status

# CP document 6 (can be CDF glossary)
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-6", sep = "")
file_name <- paste("TA", TA_vector, "_CP6.pdf", sep = "")
Fulltext_CP6 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP6[Fulltext_CP6$Status == "downloaded","TA"] # 7 downloaded
# [1] 510 484 483 479 474 473 402
Fulltext_CP$Status6 <- Fulltext_CP6$Status
```

```
# CP document 7 (can be CDF glossary) (for those who had 2 times consultation: seems like the maximum)
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-7", sep = "")
file_name <- paste("TA", TA_vector, "_CP7.pdf", sep = "")
Fulltext_CP7 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP7[Fulltext_CP7$Status == "downloaded","TA"] # 3 downloaded
# [1] 484 474 473
Fulltext_CP$Status7 <- Fulltext_CP7$Status

# CP document 8 sorabenib, email
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-8", sep = "")
file_name <- paste("TA", TA_vector, "_CP8.pdf", sep = "")
Fulltext_CP8 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP8[Fulltext_CP8$Status == "downloaded","TA"] # 2 downloaded
# [1] 474 473
Fulltext_CP$Status8 <- Fulltext_CP8$Status

# CP document 9 sorabenib, email
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-9", sep = "")
file_name <- paste("TA", TA_vector, "_CP9.pdf", sep = "")
Fulltext_CP9 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP9[Fulltext_CP9$Status == "downloaded","TA"] # 2 downloaded
# [1] 474 473
Fulltext_CP$Status9 <- Fulltext_CP9$Status

# CP document 10 sorabenib, email
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/documents/committee-papers-10", sep = "")
file_name <- paste("TA", TA_vector, "_CP10.pdf", sep = "")
Fulltext_CP10 <- file_download(TA_number = TA_vector, url = url, file_name = file_name)
Fulltext_CP10[Fulltext_CP10$Status == "downloaded","TA"] # 2 downloaded
# [1] 473
Fulltext_CP$Status10 <- Fulltext_CP10$Status

# There are further documents of TA 473 up to document 12 (but will not download it due to irrelavence)

######################################################################
# FIND TERMINATED appraisals
######################################################################

# create url
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/history", sep = "")

# create dataframe for storage
title <- data.frame(TA = TA_vector,
                    title = rep(NA, times = length(TA_vector)))
for(n in 1:length(TA_vector))[
   tryCatch([
   # read html
   webpage <- read_html(url[n])
   # Using CSS selectors to scrape the rankings section
   title_data_html <- html_nodes(webpage,'#content-start')
   title$title[n]  <- html_text(title_data_html)
   ], error = function(e)[
   cat("ERROR :",conditionMessage(e), "\n")
   ])

  # random sleeping time
  sleepy = sample(c(0.5:2.5), 1)
  cat("\n let's just wait for",sleepy,"seconds...")
  Sys.sleep(sleepy) # website will ban IP when too many queries are sent too quickly...
]

title[title$title == "NA","TA"] # none (all titles are downloaded)

# update FAD & CP document

Fulltext_CP[grep("terminated", title$title), 2:10] <- "terminated"
write.csv(Fulltext_CP, "Fulltext_CP_20200305.csv")

Fulltext_FAD[grep("terminated", title$title), 2:5 ] <- "terminated"
write.csv(Fulltext_FAD, "Fulltext_FAD_20200302.csv")
```

```
##############################################
# Create a file to store TA titles
##############################################

# create url
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/history", sep = "")

# create dataframe for storage
documents <- data.frame(TA = TA_vector,
                title = rep(NA, times = length(TA_vector)))
for(n in 1:length(TA_vector))[
  tryCatch([
    # read html
    webpage <- read_html(url[n])
    # Using CSS selectors to scrape the rankings section
    title_data_html <- html_nodes(webpage,'#content-start')
    title$title[n]  <- html_text(title_data_html)
  ], error = function(e)[
    cat("ERROR :",conditionMessage(e), "\n")
  ])

  # random sleeping time
  sleepy = sample(c(0.5:2.5), 1)
  cat("\n let's just wait for",sleepy,"seconds...")
  Sys.sleep(sleepy) # website will ban IP when too many queries are sent too quickly...
]

write.csv(documents, "Title_20200302.csv")


##############################################
# trying to download earlier FAD & CP
# less regularly named
##############################################
# the key point is to find everything is under the big trunk of <ul class="media-list">
base <- 'https://www.nice.org.uk'
url  <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/history", sep = "")
# not using the following code because it produces only the relative link
# links <- read_html(url) %>% html_nodes(., ".media-list a") %>% html_attr(., "href")

# default to download committee-papers (".media-list a" need to be changed in the function if it is not stored there)
link_download <- function(TA_number = 1:length(TA_vector), url = url, search_term = "committee-papers") [
  # create vacant table for storing links
  links_df <- data.frame(TA = TA_vector,
                         link   = rep(NA, times = length(TA_vector)),
                         link2  = rep(NA, times = length(TA_vector)),
                         link3  = rep(NA, times = length(TA_vector)),
                         link4  = rep(NA, times = length(TA_vector)),
                         link5  = rep(NA, times = length(TA_vector)),
                         link6  = rep(NA, times = length(TA_vector)),
                         link7  = rep(NA, times = length(TA_vector)),
                         link8  = rep(NA, times = length(TA_vector)),
                         link9  = rep(NA, times = length(TA_vector)),
                         link10 = rep(NA, times = length(TA_vector)))

for(n in c(TA_number))[
    tryCatch([
      # output links of
      links <- url_absolute(read_html(url[n]) %>% html_nodes(., ".media-list a") %>% html_attr(., "href"), base)
      links_df_temp <- links[grep(search_term, links)]
      links_df[n, 2:(length(links_df_temp)+1)] <- t(links_df_temp)
    ], error = function(e)[
      cat("ERROR :",conditionMessage(e), "\n")
    ])

    # random sleeping time
    sleepy = rtruncnorm(n = 1, a = 0.000001, mean = 0.8, sd = 0.3)
    cat("\n let's just wait for",sleepy,"seconds...")
    Sys.sleep(sleepy) # website will ban IP when too many queries are sent too quickly...
  ]
  return(links_df)
]

title <- read.csv(file = "output/Title_20200302.csv", header = TRUE)
```

```
title$X <- NULL
TA_vector_index <- grep("terminated", title$title)

# download FAD links (FAD links normall will link to pdf document, html FAD overview will not have "final-appraisal-determination" as link )
FAD_links <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "final-appraisal-determination")
summary(as.factor(rowSums(!is.na(FAD_links[ , 2:22])))) # 61 TAs that has no links (less than TA that terminated)
# 0  1  2  3  4  5  9 12 21
# 61 184 185  11  11  4  2  1  1
# reduced link up to 5 and manual download those more than 5 links of not corretly downloaded
FAD_links[ , 7:22] <- NULL

#download function with links matrix
file_download2 <- function(TA_number = c(1:length(TA_vector)), url = FAD_links, file_name)[
 for (n in c(TA_number)) [
 url_vector <- na.omit(unlist(url[n, 2:length(url[1,])]))
  if (length(url_vector) != 0) [
   for (k in 1:length(url_vector)) [
    tryCatch([
     download.file(url_vector[k], destfile = file_name[n, k], mode="wb")
     download_status[n, k + 1] <- "downloaded"
    ], error = function(e)[
     cat("ERROR :",conditionMessage(e), "\n")])
   ]
  ]
 # random sleeping time
 sleepy = round(rtruncnorm(n = 1, a = 0.000001, mean = 0.8, sd = 0.3), 1)
 cat("\n let's just wait for",round(sleepy, 1),"seconds...")
 Sys.sleep(sleepy) # website will ban IP when too many queries are sent too quickly...
 ]
 download_status <- data.frame(TA = download_status[ , 1],
                               Status = download_status[ , 2:length(url[1,])])
 return(download_status)
]

# find where files need to be re-downloaded (where there is no TA at all)
FAD <- read.csv(file = "output/Fulltext_FAD_20200302.csv", header = TRUE)
FAD$X <- NULL

# initial file download of FAD
# use FAD as base case and add one more column
FAD_temp <- FAD
FAD_temp$Status5 <- NA
download_status <- as.matrix(FAD_temp)

file_name      <- matrix (c(paste("TA", TA_vector, "_FAD1.pdf", sep = ""),
                            paste("TA", TA_vector, "_FAD2.pdf", sep = ""),
                            paste("TA", TA_vector, "_FAD3.pdf", sep = ""),
                            paste("TA", TA_vector, "_FAD4.pdf", sep = ""),
                            paste("TA", TA_vector, "_FAD5.pdf", sep = "")),
                            ncol = 5, byrow = F)


# Download FAD file with links # starting from TA404 there is no FAD (index 204)
TA_vector_index <- 1:length(TA_vector)
Fulltext_FAD_amend <- file_download2(TA_number = TA_vector_index[rowSums(!is.na(FAD[ , 2:5])) == 0],
                                     url = FAD_links, file_name)


    ########################################################
    # create exclusion dataframe
    # (based on fetch FAD link result)
    ########################################################

    # output TAnumber where there is no FAD and not terminated
    `%notin%` <- Negate(`%in%`)
    No_FAD_link <- FAD_links$TA[rowSums(!is.na(FAD_links[ , 2:22])) == 0]
    No_FAD_link[No_FAD_link %notin% TA_vector[TA_vector_terminated]]

    length(FAD_links$TA[rowSums(!is.na(FAD_links[ , 2:11])) == 0]) #18
    # [1] *532(withdrawn no longer on market)
    #     *493(replaced by NG616)
    #     *459(withdrawn)
    #     404(outlier "fad-document, suggest maual download)
    #     *394(review: no FAD, no ERG)
```

```
    #    *381(replaced by TA620 in Jan2020)
    #    366(wrongly named, "appraisal-consultation-document", suggest manual download)
    #    292(wrongly named, "final-appraisal-determintation-document2", manual download)
    #    266(wrongly named, "final-appraisal-determinaton")
    #    264(wrongly named, "final-appriasal-determination", manual download)
    #    55(no FAD, but assessment report)
    #    38(no FAD, but HTA report)
    #    34(no FAD, but Assessmen report)
    #    29(no FAD, but Assessmen report)
    #    23(no FAD, but Assessmen report)
    #    20(no FAD, but Assessmen report)
    #    10(no FAD, but Assessmen report)
    #    *1(review: no FAD, no ERG)


    TA_withdrawn  <- c(532, 459)
    TA_replaced   <- c(493, 381)
    TA_noCSFADERG <- c(394, 1)
    #43 (43 terminated + 6)

test <- Fulltext_FAD_amend
test[2:6] <- sapply(test[2:6], as.character)
test[TA_vector_terminated, 2:6] <- "terminated"
test[TA_vector %in% TA_withdrawn, 2:6] <- "withdrawn"
test[TA_vector %in% TA_replaced, 2:6] <- "replaced"
test[TA_vector %in% TA_noCSFADERG, 2:6] <- "no_CS_FAD_ERG"
Fulltext_FAD_amend <- test

Fulltext_FAD_amend$TA[rowSums(!is.na(Fulltext_FAD_amend[ , 2:6])) == 0] #12
#  [1] 404 366 292 266 264  55  38  34  29  23  20  10 (TAs without FAD) see explanation above

# output UPDATED FAD
write.csv(Fulltext_FAD_amend, "Fulltext_FAD_20200305.csv")
write.csv(FAD_links, "Links_FAD_20200305.csv")


### download CP links (some CP link can be linked to html website but not pdf)
CP_links  <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "committee-papers")
TA_vector[rowSums(!is.na(CP_links[ , 12:13])) != 0] # TA 473 will have some undownloaded documents but it is too long

# e.g. TA 209evaluation report (seems like CP changed names)
ER_links    <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "evaluation-report")
rowSums(!is.na(ER_links[ , 2:10]))
# e.g. TA 192
ERGR_links  <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "erg-report")
rowSums(!is.na(ERGR_links[ , 2:10]))
# e.g. TA 123
ERGR_links2 <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "evidence-review-group-report")
rowSums(!is.na(ERGR_links2[ , 2:10]))
# TA191
ERGR_links3 <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "evidence-review-groups-report")
rowSums(!is.na(ERGR_links3[ , 2:10]))
# e.g. TA38
HTA_links   <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "hta-report")
rowSums(!is.na(HTA_link[ , 2:10]))
# e.g. TA75
HTA_links2  <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "health-technology-assessment")
rowSums(!is.na(HTA_link2[ , 2:10]))
# e.g. TA278, TA61, TA 59
AR_links    <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "assessment-report")
rowSums(!is.na(AR_links[ , 2:10]))
# TA195 really long
# e.g. TA188
AR_links2   <- link_download(TA_number = 1:length(TA_vector), url = url, search_term = "assessment-group-report")
rowSums(!is.na(AR_links2[ , 2:10]))
#


# check how many links (including CP) won't exceed 8 download spacces (can use the data frame of CP_amend to update)
summary(as.factor(rowSums(!is.na(CP_links[ , 2:11]))    +
            rowSums(!is.na(ER_links[ , 2:11]))    +
            rowSums(!is.na(ERGR_links[ , 2:11]))  +
            rowSums(!is.na(ERGR_links2[ , 2:11])) +
            rowSums(!is.na(ERGR_links3[ , 2:11])) +
```

```
            rowSums(!is.na(HTA_links[ , 2:11]))   +
            rowSums(!is.na(HTA_links2[ , 2:11]))  +
            rowSums(!is.na(AR_links[ , 2:11]))    +
            rowSums(!is.na(AR_links2[ , 2:11])))))
# 0  1  2  3  4  5  6  7  8  9 10 11 13
# 50 98 139 67 44 21 15 12  5  3  4  1  1
# links can be stored together

summary(as.factor(  rowSums(!is.na(ER_links[ , 2:10]))    +
              rowSums(!is.na(ERGR_links[ , 2:10]))  +
              rowSums(!is.na(ERGR_links2[ , 2:10]))  +
              rowSums(!is.na(ERGR_links3[ , 2:10]))  +
              rowSums(!is.na(HTA_links[ , 2:10]))   +
              rowSums(!is.na(HTA_links2[ , 2:10]))  +
              rowSums(!is.na(AR_links[ , 2:10]))    +
              rowSums(!is.na(AR_links2[ , 2:10])))))
#  0  1  2  3  4  5  6  7  8  9 10
# 252 58 65 27 24  9  9  7  4  4  1

# still no files for CP at all #50
no_CP <- TA_vector_index [(as.factor(rowSums(!is.na(CP_links[ , 2:11]))    +
        rowSums(!is.na(ER_links[ , 2:11]))    +
        rowSums(!is.na(ERGR_links[ , 2:11]))  +
        rowSums(!is.na(ERGR_links2[ , 2:11]))  +
        rowSums(!is.na(ERGR_links3[ , 2:11]))  +
        rowSums(!is.na(HTA_links[ , 2:11]))   +
        rowSums(!is.na(HTA_links2[ , 2:11]))  +
        rowSums(!is.na(AR_links[ , 2:11]))    +
        rowSums(!is.na(AR_links2[ , 2:11])))) == 0]
length(no_CP)
no_CP
no_CP <- no_CP[no_CP %notin% TA_vector_terminated] # delete those terminated
no_CP <- no_CP[no_CP %notin% TA_vector_index[TA_vector %in% TA_noCSFADERG]] # NO CSFADERG
no_CP <- no_CP[no_CP %notin% TA_vector_index[TA_vector %in% TA_replaced]] # NO replaced
no_CP <- no_CP[no_CP %notin% TA_vector_index[TA_vector %in% TA_withdrawn]] # NO wthdrawn # 65
no_CP
# [1] 440 447
TA_vector[no_CP]
# [1] 77 (protocol-newer-hypnotic-drugs-for-shortterm-pharmacotherapy-for-insomnia2)
#    64 (report-by-a-consortium)


no_CP <- TA_vector_index [(as.factor(rowSums(!is.na(CP_links[ , 2:11]))    +
                           rowSums(!is.na(ER_links[ , 2:11]))    +
                           rowSums(!is.na(ERGR_links[ , 2:11]))  +
                           rowSums(!is.na(ERGR_links2[ , 2:11]))  +
                           rowSums(!is.na(ERGR_links3[ , 2:11]))  +
                           rowSums(!is.na(HTA_links[ , 2:11]))   +
                           rowSums(!is.na(HTA_links2[ , 2:11]))  +
                           rowSums(!is.na(AR_links[ , 2:11]))    +
                           rowSums(!is.na(AR_links2[ , 2:11])))) == 0]

# find where files need to be re-downloaded (where there is no TA at all)
CP <- read.csv(file = "output/Fulltext_CP_20200305.csv", header = TRUE)
CP$X <- NULL
CP[TA_vector_terminated , 2:11] <- NA

summary(as.factor(rowSums(!is.na(CP_links[ , 2:13])))) # 204 TAs that has no links in original CP download
# 0  1  2  3  4  5  6  7  9 12
# 229 60 86 41 26  9  4  3  1  1
rowSums(!is.na(CP_links[ , 2:11]))
# reduced link up to 5 and manual download those more than 5 links of not corretly downloaded
CP_links[ , 12:13] <- NULL # manual download those who had 9 and & links

# initial file download of FAD
# use FAD as base case and add one more column
CP_temp <- CP
download_status <- as.matrix(CP_temp)

file_name      <- matrix (c(paste("TA", TA_vector, "_CP1.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP2.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP3.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP4.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP5.pdf", sep = ""),
```

```
                    paste("TA", TA_vector, "_CP6.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP7.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP8.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP9.pdf", sep = ""),
                    paste("TA", TA_vector, "_CP10.pdf", sep = "")),
                    ncol = 10, byrow = F)


# Download CP file with links
TA_vector_index <- 1:length(TA_vector)
length(TA_vector_index[rowSums(!is.na(CP[ , 2:11])) == 0]) # 247 index need to be download use link (including termination)
Fulltext_CP_amend <- file_download2(TA_number = TA_vector_index[rowSums(!is.na(CP[ , 2:11])) == 0],
                        url = CP_links, file_name)
length(TA_vector_index[rowSums(!is.na(Fulltext_CP_amend[ , 2:11])) == 0]) # 225 index still need to be download use link
TA_vector_terminated <- grep("terminated", title$title)
test <- Fulltext_CP_amend
test[2:11] <- sapply(test[2:11], as.character)
test[TA_vector_terminated, 2:11] <- "terminated"
test[TA_vector %in% TA_withdrawn, 2:11] <- "withdrawn"
test[TA_vector %in% TA_replaced, 2:11] <- "replaced"
test[TA_vector %in% TA_noCSFADERG, 2:11] <- "no_CS_FAD_ERG"
Fulltext_CP_amend <- test

length(TA_vector_index[rowSums(!is.na(Fulltext_CP_amend[ , 2:11])) == 0]) # 180 index still need to be download use link

# output UPDATED CP
write.csv(Fulltext_CP_amend, "Fulltext_CP_20200305.csv")
write.csv(CP_links, "Links_CP_20200305.csv")

########################################################################
# download the rest of ERG
########################################################################
# create url matrix for download
url_ERG <- matrix (rep(NA, 11*length(TA_vector)), byrow = T, ncol = 11)
url_ERG[ , 1] <- TA_vector
for (n in c(TA_vector_index))[
 x <- c(unname(unlist(ER_links[n, 2:11])[!is.na(unlist(ER_links[n, 2:11]))]),
      unname(unlist(ERGR_links[n, 2:11])[!is.na(unlist(ERGR_links[n, 2:11]))]),
      unname(unlist(ERGR_links2[n, 2:11])[!is.na(unlist(ERGR_links2[n, 2:11]))]),
      unname(unlist(ERGR_links3[n, 2:11])[!is.na(unlist(ERGR_links3[n, 2:11]))]),
      unname(unlist(HTA_links[n, 2:11])[!is.na(unlist(HTA_links[n, 2:11]))]),
      unname(unlist(HTA_links2[n, 2:11])[!is.na(unlist(HTA_links2[n, 2:11]))]),
      unname(unlist(AR_links[n, 2:11])[!is.na(unlist(AR_links[n, 2:11]))]),
      unname(unlist(AR_links2[n, 2:11])[!is.na(unlist(AR_links2[n, 2:11]))])
 )
 x <- x[0:min(10, length(x))]

 if (length(x) != 0)[
  url_ERG[n, 2:(length(x)+1)] <- x
 ]
]

url_ERG <- as.data.frame(url_ERG)

file_name      <- matrix (c(paste("TA", TA_vector, "_ERG1.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG2.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG3.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG4.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG5.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG6.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG7.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG8.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG9.pdf", sep = ""),
                    paste("TA", TA_vector, "_ERG10.pdf", sep = "")
                    ), ncol = 10, byrow = F)
# Download ERG file with links
download_status[ , 2:11] <- NA
url_ERG[2:11] <- sapply(url_ERG[2:11], as.character)
Fulltext_ERG <- file_download2(TA_number = TA_vector_index,
                        url = url_ERG, file_name)



length(TA_vector_index[rowSums(!is.na(Fulltext_ERG[ , 2:11])) == 0]) # 252 index still need to be download use link
```

```
TA_vector_terminated <- grep("terminated", title$title)
test <- Fulltext_ERG
test[2:11] <- sapply(test[2:11], as.character)
test[TA_vector_terminated, 2:11] <- "terminated"
test[TA_vector %in% TA_withdrawn, 2:11] <- "withdrawn"
test[TA_vector %in% TA_replaced, 2:11] <- "replaced"
test[TA_vector %in% TA_noCSFADERG, 2:11] <- "no_CS_FAD_ERG"
Fulltext_ERG <- test

# output UPDATED CP
write.csv(Fulltext_ERG, "Fulltext_ERG_20200305.csv")
write.csv(url_ERG, "Links_ERG_20200305.csv")


####################################
# Download TA history file
# convert html to pdf
####################################
# ref: https://rdrr.io/cran/pagedown/man/chrome_print.html
# install.packages("pagedown")

# create url
url <- paste("https://www.nice.org.uk/guidance/ta", TA_vector, "/history", sep = "")
file_name <- paste("TA", TA_vector, "_history.pdf", sep = "")
history <- data.frame (TA = TA_vector,
                       Status = rep(NA, length(TA_vector)))
indices <- 1:length(TA_vector)

download_history <- function(url, file_name, indices)[
 for(n in c(indices))[
   tryCatch([
     # read html
     chrome_print(url[n], output = file_name[n])
     history$Status[n]  <- "downloaded"
   ], error = function(e)[
     cat("ERROR :",conditionMessage(e), "\n")
   ])
   # random sleeping time
   sleepy = sample(c(0.5:3), 1)
   cat("\n let's just wait for",sleepy,"seconds...")
   Sys.sleep(sleepy) # website will ban IP when too many queries are sent too quickly...
 ]
return(history)
]

# download all history
download_history(url = url, file_name = file_name, indices = indices)
history[is.na(history$Status), "Status"] <- "NA"
history[history$Status == "NA", "TA"]
# [1] 556 550 547 507 435 434 431 362 359 353 351 350 169 167 161  34  20  10   1

# output the indicies where files are not downloaded and try again
indices <- which(grepl("NA", history$Status))
download_history(url = url, file_name = file_name, indices = indices)
history[history$Status == "NA", "TA"]
# [1] 556 434 431 167 161
# manual download these files!
history[history$Status == "NA", "Status"] <- "downloaded"

# save file
write.csv(history, "History_20200303.csv")
```

Appendix 3.2 Excluded studies in the systematic review of NICE Technology Appraisals (Chapter 3)

This appendix table lists basic demographics for all technology appraisals (TAs) excluded from the NICE TA systematic review in Chapter 3, detailing the title, disease, disease area, and exclusion reason for each TA. An abbreviation list for table contents is provided at the end of the table

TA	TA Title	Disease Area	Disease	Reason of Exclusion
613	Fluocinolone acetonide intravitreal implant for treating chronic diabetic macular oedema in phakic eyes after an inadequate response to previous therapy	Other	DMO	implant
609	Ramucirumab for treating unresectable hepatocellular carcinoma after sorafenib (terminated appraisal)	Oncology	HCC	terminated
608	Ibrutinib with rituximab for treating Waldenstrom's macroglobulinaemia (terminated appraisal)	Oncology	WM	terminated
603	Lenalidomide with bortezomib and dexamethasone for untreated multiple myeloma (terminated appraisal)	Oncology	MM	terminated
602	Pomalidomide with bortezomib and dexamethasone for treating relapsed or refractory multiple myeloma (terminated appraisal)	Oncology	MM	terminated
601	Bezlotoxumab for preventing recurrent Clostridium difficile infection (terminated appraisal)	Infectious disease	C. diff infection	terminated
594	Brentuximab vedotin for untreated advanced Hodgkin lymphoma (terminated appraisal)	Oncology	lymphoma	terminated
582	Cabozantinib for previously treated advanced hepatocellular carcinoma (terminated appraisal)	Oncology	HCC	terminated
576	Bosutinib for untreated chronic myeloid leukaemia (terminated appraisal)	Oncology	leukaemia	terminated
570	Pembrolizumab for treating recurrent or metastatic squamous cell carcinoma of the head and neck after platinum-based chemotherapy (terminated appraisal)	Oncology	HNC	terminated
568	Abatacept for treating psoriatic arthritis after DMARDs (terminated appraisal)	Autoimmune	psoriasis	terminated
566	Cochlear implants for children and adults with severe to profound deafness	Other	deafness	medical device
565	Benralizumab for treating severe eosinophilic asthma	Other	asthma	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
564	Dabrafenib with trametinib for treating advanced metastatic BRAF V600E mutation-positive non-small-cell lung cancer (terminated appraisal)	Oncology	NSCLC	terminated
560	Bevacizumab with carboplatin, gemcitabine and paclitaxel for treating the first recurrence of platinum-sensitive advanced ovarian cancer (terminated appraisal)	Oncology	ovarian cancer	terminated
555	Regorafenib for previously treated advanced hepatocellular carcinoma	Oncology	HCC	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
549	Denosumab for preventing skeletal-related events in multiple myeloma (terminated appraisal)	Oncology	MM	terminated
548	Decitabine for untreated acute myeloid leukaemia (terminated appraisal)	Oncology	leukaemia	terminated
532	Cenegermin for treating neurotrophic keratitis	Other	neurotrophic keratitis	withdrawn
508	Autologous chondrocyte implantation using chondrosphere for treating symptomatic articular cartilage defects of the knee	Other	articular cartilage defects	procedure

TA	TA Title	Disease Area	Disease	Reason of Exclusion
501	Intrabeam radiotherapy system for adjuvant treatment of early breast cancer	Oncology	breast cancer	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
494	Naltrexone–bupropion for managing overweight and obesity	Other	obesity	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
493	Cladribine tablets for treating relapsing–remitting multiple sclerosis	Autoimmune	MS	replaced by TA616
486	Aflibercept for treating choroidal neovascularisation	Other	CNV	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
480	Tofacitinib for moderate to severe rheumatoid arthritis	Autoimmune	RA	No key terms in CS, ERG/AG or FAD
479	Reslizumab for treating severe eosinophilic asthma	Other	asthma	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
477	Autologous chondrocyte implantation for treating symptomatic articular cartilage defects of the knee	Other	articular cartilage defects	procedure
474	Sorafenib for treating advanced hepatocellular carcinoma	Oncology	HCC	No key terms in CS, ERG/AG or FAD
473	Cetuximab for treating recurrent or metastatic squamous cell cancer of the head and neck	Oncology	HNC	No key terms in CS, ERG/AG or FAD
469	Idelalisib with ofatumumab for treating chronic lymphocytic leukaemia (terminated appraisal)	Oncology	leukaemia	terminated
468	Methylnaltrexone bromide for treating opioid-induced constipation (terminated appraisal)	Other	constipation	terminated
467	Holoclar for treating limbal stem cell deficiency after eye burns	Other	LSCD	stemp cell therapy
464	Bisphosphonates for treating osteoporosis	Other	osteoporosis	No key terms in CS, ERG/AG or FAD
461	Roflumilast for treating chronic obstructive pulmonary disease	Other	COPD	No key terms in CS, ERG/AG or FAD
459	Collagenase clostridium histolyticum for treating Dupuytren's contracture	Other	Dupuytren's Contracture	withdrawn
454	Daratumumab with lenalidomide and dexamethasone for treating relapsed or refractory multiple myeloma (terminated appraisal)	Oncology	MM	terminated
453	Bortezomib for treating multiple myeloma after second or subsequent relapse (terminated appraisal)	Oncology	MM	terminated
452	Ibrutinib for untreated chronic lymphocytic leukaemia without a 17p deletion or TP53 mutation (terminated appraisal)	Oncology	leukaemia	terminated
444	Afatinib for treating advanced squamous non-small-cell lung cancer after platinum-based chemotherapy (terminated appraisal)	Oncology	NSCLC	terminated
443	Obeticholic acid for treating primary biliary cholangitis	Other	PBC	No key terms in CS, ERG/AG or FAD
438	Alectinib for previously treated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer (terminated appraisal)	Oncology	NSCLC	terminated
437	Ibrutinib with bendamustine and rituximab for treating relapsed or refractory chronic lymphocytic leukaemia after systemic therapy (terminated appraisal)	Oncology	leukaemia	terminated
436	Bevacizumab for treating EGFR mutation-positive non-small-cell lung cancer (terminated appraisal)	Oncology	NSCLC	terminated
435	Tenofovir alafenamide for treating chronic hepatitis B (terminated appraisal)	Infectious disease	hepatitis B	terminated
434	Elotuzumab for previously treated multiple myeloma (terminated appraisal)	Oncology	MM	terminated

TA	TA Title	Disease Area	Disease	Reason of Exclusion
431	Mepolizumab for treating severe refractory eosinophilic asthma	Other	asthma	No key terms in CS, ERG/AG or FAD
398	Lumacaftor–ivacaftor for treating cystic fibrosis homozygous for the F508del mutation	Other	cystic fibrosis	No key terms in CS, ERG/AG or FAD
397	Belimumab for treating active autoantibody-positive systemic lupus erythematosus	Autoimmune	lupus erythematosus	No key terms in CS, ERG/AG or FAD
394	Evolocumab for treating primary hypercholesterolaemia and mixed dyslipidaemia	Cardiovascular disease	dyslipidaemia	No reports available
392	Adalimumab for treating moderate to severe hidradenitis suppurativa	Other	hidradenitis suppurativa	No key terms in CS, ERG/AG or FAD
385	Ezetimibe for treating primary heterozygous-familial and non-familial hypercholesterolaemia	Cardiovascular disease	dyslipidaemia	No key terms in CS, ERG/AG or FAD
382	Eltrombopag for treating severe aplastic anaemia refractory to immunosuppressive therapy (terminated appraisal)	Other	anaemia	terminated
381	Olaparib for maintenance treatment of relapsed, platinum-sensitive, BRCA mutation-positive ovarian, fallopian tube and peritoneal cancer after response to second-line or subsequent platinum-based chemotherapy	Oncology	ovarian cancer	replaced by TA620
379	Nintedanib for treating idiopathic pulmonary fibrosis	Autoimmune	IPF	No key terms in CS, ERG/AG or FAD
369	Ciclosporin for treating dry eye disease that has not improved despite treatment with artificial tears	Other	dry eye	implant
362	Paclitaxel as albumin-bound nanoparticles with carboplatin for untreated non-small-cell lung cancer (terminated appraisal)	Oncology	NSCLC	terminated
358	Tolvaptan for treating autosomal dominant polycystic kidney disease	Other	ADPKD	No key terms in CS, ERG/AG or FAD
356	Ruxolitinib for treating polycythaemia vera (terminated appraisal)	Oncology	polycythaemia vera	terminated
353	Bevacizumab for treating relapsed, platinum-resistant epithelial ovarian, fallopian tube or primary peritoneal cancer (terminated appraisal)	Oncology	ovarian cancer	terminated
351	Cangrelor for reducing atherothrombotic events in people undergoing percutaneous coronary intervention or awaiting surgery requiring interruption of anti-platelet therapy (terminated appraisal)	Cardiovascular disease	TE	terminated
349	Dexamethasone intravitreal implant for treating diabetic macular oedema	Other	DMO	implant
348	Everolimus for preventing organ rejection in liver transplantation	Other	immunosuppressive therapy	No key terms in CS, ERG/AG or FAD
339	Omalizumab for previously treated chronic spontaneous urticaria	Other	urticaria	No key terms in CS, ERG/AG or FAD
337	Rifaximin for preventing episodes of overt hepatic encephalopathy	Other	hepatic encephalopathy	No key terms in CS, ERG/AG or FAD
334	Regorafenib for metastatic colorectal cancer after treatment for metastatic disease (terminated appraisal)	Oncology	colorectal cancer	terminated
330	Sofosbuvir for treating chronic hepatitis C	Infectious disease	hepatitis C	No key terms in CS, ERG/AG or FAD
324	Dual-chamber pacemakers for symptomatic bradycardia due to sick sinus syndrome without atrioventricular block	Cardiovascular disease	arrhythmia	medical device
323	Erythropoiesis-stimulating agents (epoetin and darbepoetin) for treating anaemia in people with cancer having chemotherapy	Other	anaemia	No key terms in CS, ERG/AG or FAD
317	Prasugrel with percutaneous coronary intervention for treating acute coronary syndromes	Cardiovascular disease	ACS	No key terms in CS, ERG/AG or FAD
314	Implantable cardioverter defibrillators and cardiac resynchronisation therapy for arrhythmias and heart failure	Cardiovascular disease	arrhythmia	medical device
304	Total hip replacement and resurfacing arthroplasty for end-stage arthritis of the hip	Other	arthritis	procedure
302	Canakinumab for treating systemic juvenile idiopathic arthritis (terminated appraisal)	Autoimmune	JIA	terminated

TA	TA Title	Disease Area	Disease	Reason of Exclusion
301	Fluocinolone acetonide intravitreal implant for treating chronic diabetic macular oedema after an inadequate response to prior therapy	Other	DMO	implant
287	Rivaroxaban for treating pulmonary embolism and preventing recurrent venous thromboembolism	Cardiovascular disease	TE	No key terms in CS, ERG/AG or FAD
286	Loxapine inhalation for treating acute agitation and disturbed behaviours associated with schizophrenia and bipolar disorder (terminated appraisal)	Neurology/mental health	bipolar disorder	terminated
283	Ranibizumab for treating visual impairment caused by macular oedema secondary to retinal vein occlusion	Other	visual impairment	No key terms in CS, ERG/AG or FAD
281	Canakinumab for treating gouty arthritis attacks and reducing the frequency of subsequent attacks (terminated appraisal)	Other	gout	terminated
279	Percutaneous vertebroplasty and percutaneous balloon kyphoplasty for treating osteoporotic vertebral compression fractures	Other	osteoporotic vertebral compression fractures	procedure
278	Omalizumab for treating severe persistent allergic asthma	Other	asthma	No key terms in CS, ERG/AG or FAD
277	Methylnaltrexone for treating opioid-induced bowel dysfunction in people with advanced illness receiving palliative care (terminated appraisal)	Other	opioid-induced bowel dysfunction	terminated
273	Tadalafil for the treatment of symptoms associated with benign prostatic hyperplasia (terminated appraisal)	Other	benign prostatic hyperplasia	terminated
272	Vinflunine for the treatment of advanced or metastatic transitional cell carcinoma of the urothelial tract	Oncology	urothelial cancer	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
270	Decitabine for the treatment of acute myeloid leukaemia (terminated appraisal)	Oncology	leukaemia	terminated
267	Ivabradine for treating chronic heart failure	Cardiovascular disease	HF	No key terms in CS, ERG/AG or FAD
265	Denosumab for the prevention of skeletal-related events in adults with bone metastases from solid tumours	Oncology	skeletal-related events in cancer	No key terms in CS, ERG/AG or FAD
264	Alteplase for treating acute ischaemic stroke	Cardiovascular disease	TE	No key terms in CS, ERG/AG or FAD
261	Rivaroxaban for the treatment of deep vein thrombosis and prevention of recurrent deep vein thrombosis and pulmonary embolism	Cardiovascular disease	TE	No key terms in CS, ERG/AG or FAD
258	Erlotinib for the first-line treatment of locally advanced or metastatic EGFR-TK mutation-positive non-small-cell lung cancer	Oncology	NSCLC	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
246	Pharmalgen for the treatment of bee and wasp venom allergy	Other	venom allergy	No key terms in CS, ERG/AG or FAD
245	Apixaban for the prevention of venous thromboembolism after total hip or knee replacement in adults	Cardiovascular disease	TE	No key terms in CS, ERG/AG or FAD
240	Panitumumab in combination with chemotherapy for the treatment of metastatic colorectal cancer (terminated appraisal)	Oncology	colorectal cancer	terminated
236	Ticagrelor for the treatment of acute coronary syndromes	Cardiovascular disease	ACS	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
235	Mifamurtide for the treatment of osteosarcoma	Oncology	osteosarcoma	No key terms in CS, ERG/AG or FAD
231	Agomelatine for the treatment of major depressive episodes (terminated appraisal)	Neurology/mental health	depression	terminated
230	Bivalirudin for the treatment of ST-segment-elevation myocardial infarction	Cardiovascular disease	ACS	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
229	Dexamethasone intravitreal implant for the treatment of macular oedema secondary to retinal vein occlusion	Other	DMO	implant

TA	TA Title	Disease Area	Disease	Reason of Exclusion
221	Romiplostim for the treatment of chronic immune (idiopathic) thrombocytopenic purpura	Autoimmune	ITP	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
218	Azacitidine for the treatment of myelodysplastic syndromes, chronic myelomonocytic leukaemia and acute myeloid leukaemia	Oncology	leukaemia	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
211	Prucalopride for the treatment of chronic constipation in women	Other	constipation	No key terms in CS, ERG/AG or FAD
208	Trastuzumab for the treatment of HER2-positive metastatic gastric cancer	Oncology	gastric cancer	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
207	Temsirolimus for the treatment of relapsed or refractory mantle cell lymphoma (terminated appraisal)	Oncology	lymphoma	terminated
206	Bendamustine for the treatment of indolent (low grade) non-Hodgkin's lymphoma that is refractory to rituximab (terminated appraisal)	Oncology	lymphoma	terminated
204	Denosumab for the prevention of osteoporotic fractures in postmenopausal women	Other	osteoporosis	No key terms in CS, ERG/AG or FAD
200	Peginterferon alfa and ribavirin for the treatment of chronic hepatitis C	Infectious disease	hepatitis C	No key terms in CS, ERG/AG or FAD
192	Gefitinib for the first-line treatment of locally advanced or metastatic non-small-cell lung cancer	Oncology	NSCLC	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
191	Capecitabine for the treatment of advanced gastric cancer	Oncology	gastric cancer	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
188	Human growth hormone (somatropin) for the treatment of growth failure in children	Other	growth failure	No key terms in CS, ERG/AG or FAD
185	Trabectedin for the treatment of advanced soft tissue sarcoma	Oncology	soft tissue sarcoma	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
184	Topotecan for the treatment of relapsed small-cell lung cancer	Oncology	SCLC	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
179	Sunitinib for the treatment of gastrointestinal stromal tumours	Oncology	GIST	No key terms in CS, ERG/AG or FAD
177	Alitretinoin for the treatment of severe chronic hand eczema	Other	eczema	No key terms in CS, ERG/AG or FAD
170	Rivaroxaban for the prevention of venous thromboembolism after total hip or total knee replacement in adults	Cardiovascular disease	TE	No key terms in CS, ERG/AG or FAD
168	Amantadine, oseltamivir and zanamivir for the treatment of influenza	Infectious disease	influenza	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
167	Endovascular stent–grafts for the treatment of abdominal aortic aneurysms	Cardiovascular disease	abdominal aortic aneurysms	medical device
165	Machine perfusion systems and cold static storage of kidneys from deceased donors	Other	transplant	medical device
163	Infliximab for acute exacerbations of ulcerative colitis	Autoimmune	UC	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
160	Raloxifene for the primary prevention of osteoporotic fragility fractures in postmenopausal women	Other	osteoporosis	No key terms in CS, ERG/AG or FAD

TA	TA Title	Disease Area	Disease	Reason of Exclusion
159	Spinal cord stimulation for chronic pain of neuropathic or ischaemic origin	Other	chronic pain	medical device
158	Oseltamivir, amantadine (review) and zanamivir for the prophylaxis of influenza	Infectious disease	influenza	No key terms in CS, ERG/AG or FAD
157	Dabigatran etexilate for the prevention of venous thromboembolism after hip or knee replacement surgery in adults	Cardiovascular disease	TE	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
156	Routine antenatal anti-D prophylaxis for women who are rhesus D negative	Other	anti-D prophylaxis	No key terms in CS, ERG/AG or FAD
155	Ranibizumab and pegaptanib for the treatment of age-related macular degeneration	Other	AMD	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
152	Drug-eluting stents for the treatment of coronary artery disease	Cardiovascular disease	ACS	medical device
151	Continuous subcutaneous insulin infusion for the treatment of diabetes mellitus	Diabetes mellitus	DM	No key terms in CS, ERG/AG or FAD
149	Carmustine implants for the treatment of recurrent glioblastoma multiforme (terminated appraisal)	Oncology	glioblastoma	terminated
148	Bevacizumab for the treatment of non-small-cell lung cancer (terminated appraisal)	Oncology	NSCLC	terminated
145	Cetuximab for the treatment of locally advanced squamous cell cancer of the head and neck	Oncology	HNC	No key terms in CS, ERG/AG or FAD
139	Continuous positive airway pressure for the treatment of obstructive sleep apnoea/hypopnoea syndrome	Other	apnoea	procedure
136	Structural neuroimaging in first-episode psychosis	Neurology/mental health	psychosis	procedure
135	Pemetrexed for the treatment of malignant pleural mesothelioma	Oncology	mesothelioma	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
131	Inhaled corticosteroids for the treatment of chronic asthma in children under the age of 12 years	Other	asthma	No key terms in CS, ERG/AG or FAD
129	Bortezomib monotherapy for relapsed multiple myeloma	Oncology	MM	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
128	Stapled haemorrhoidopexy for the treatment of haemorrhoids	Other	haemorrhoids	procedure
127	Natalizumab for the treatment of adults with highly active relapsing–remitting multiple sclerosis	Autoimmune	MS	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
124	Pemetrexed for the treatment of non-small-cell lung cancer	Oncology	NSCLC	No key terms in CS, ERG/AG or FAD
123	Varenicline for smoking cessation	Other	smoking cessation	No key terms in CS, ERG/AG or FAD
117	Cinacalcet for the treatment of secondary hyperparathyroidism in patients with end-stage renal disease on maintenance dialysis therapy	Other	hyperparathyroidism	No key terms in CS, ERG/AG or FAD
115	Naltrexone for the management of opioid dependence	Other	opioid dependence	No key terms in CS, ERG/AG or FAD
114	Methadone and buprenorphine for the management of opioid dependence	Other	opioid dependence	No key terms in CS, ERG/AG or FAD
106	Peginterferon alfa and ribavirin for the treatment of mild chronic hepatitis C	Infectious disease	hepatitis C	Key terms in CS, ERG/AG or FAD not referring to consideration of treatment sequences
105	Laparoscopic surgery for colorectal cancer	Oncology	colorectal cancer	procedure

TA	TA Title	Disease Area	Disease	Reason of Exclusion
92	HealOzone for the treatment of tooth decay (occlusal pit and fissure caries and root caries)	Other	tooth problem	procedure
88	Dual-chamber pacemakers for symptomatic bradycardia due to sick sinus syndrome and/or atrioventricular block	Cardiovascular disease	arrhythmia	medical device
86	Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours	Oncology	GIST	No key terms in CS, ERG/AG or FAD
83	Laparoscopic surgery for inguinal hernia repair	Other	inguinal hernia	procedure
81	Frequency of application of topical corticosteroids for atopic eczema	Other	eczema	No key terms in CS, ERG/AG or FAD
78	Fluid-filled thermal balloon and microwave endometrial ablation techniques for heavy menstrual bleeding	Other	menstrual bleeding	procedure
77	Guidance on the use of zaleplon, zolpidem and zopiclone for the short-term management of insomnia	Other	insomnia	No key terms in CS, ERG/AG or FAD
75	Interferon alfa (pegylated and non-pegylated) and ribavirin for the treatment of chronic hepatitis C	Infectious disease	hepatitis C	No key terms in CS, ERG/AG or FAD
74	Pre-hospital initiation of fluid replacement therapy in trauma	Other	trauma	No key terms in CS, ERG/AG or FAD
73	Myocardial perfusion scintigraphy for the diagnosis and management of angina and myocardial infarction	Cardiovascular disease	ACS	procedure
71	Guidance on the use of coronary artery stents	Cardiovascular disease	ACS	medical device
69	Guidance on the use of liquid-based cytology for cervical screening	Other	cervical screening	procedure
64	Human growth hormone (somatropin) in adults with growth hormone deficiency	Other	growth failure	No key terms in CS, ERG/AG or FAD
61	Guidance on the use of capecitabine and tegafur with uracil for metastatic colorectal cancer	Oncology	colorectal cancer	No key terms in CS, ERG/AG or FAD
59	Guidance on the use of electroconvulsive therapy	Other	electroconvulsive therapy	procedure
49	Guidance on the use of ultrasound locating devices for placing central venous catheters	Other	central venous catheters	medical device
47	Guidance on the use of glycoprotein IIb/IIIa inhibitors in the treatment of acute coronary syndromes	Cardiovascular disease	ACS	No key terms in CS, ERG/AG or FAD
38	Inhaler devices for routine treatment of chronic asthma in older children (aged 5–15 years)	Other	asthma	No key terms in CS, ERG/AG or FAD
29	Guidance on the use of fludarabine for B-cell chronic lymphocytic leukaemia	Oncology	leukaemia	No key terms in CS, ERG/AG or FAD
20	Guidance on the use of Riluzole (Rilutek) for the treatment of Motor Neurone Disease	Neurology/mental health	Motor Neurone Disease	No key terms in CS, ERG/AG or FAD
10	Guidance on the use of inhaler systems (devices) in children under the age of 5 years with chronic asthma	Other	asthma	No key terms in CS, ERG/AG or FAD
1	Guidance on the Extraction of Wisdom Teeth	Other	tooth problem	No reports available

ACS: acute coronary syndrome (including myocardial infarction), ADPKD: autosomal dominant polycystic kidney disease, AG: Assessment Group reports, AMD: age-related macular degeneration, C. diff : Clostridium difficile, CNV: choroidal neovascularisation, COPD: chronic obstructive pulmonary disease, CS: company submissions, DM: diabetes mellitus, DMO: diabetic macular oedema, ERG: Evidence Review Group reports, FAD: Final Appraisal Determination document, GIST: gastrointestinal stromal tumour, HCC: hepatocellular carcinoma, HF: heart failure: HNC: head and neck cancer, IPF: idiopathic pulmonary fibrosis, ITP: immune thrombocytopenia, JIA: juvenile arthritis, LSCD: limbal stem cell deficiency, MM: multiple myeloma, MS: multiple sclerosis, NETs: neuroendocrine tumours, NSCLC: non-small cell lung cancer, RA: rheumatoid arthritis, RCC: renal cell carcinoma, SCLC: small cell lung cancer, TE: thromboembolism, UC: ulcerative colitis, WM: Waldenstrom macroglobulinemia.

Appendix 3.3 Included studies in Part A of the systematic review of NICE TAs

This appendix table outlines basic demographics for all technology appraisals (TAs) included in the NICE TA systematic review from Chapter, covering the title, disease, disease area, and indicators of key documents' presence (i.e., Company Submissions (CS), Evidence Review Group (ERG)/Assessment Group (AG) reports, and Final Appraisal Determination (FAD) documents), an indicator of whether a document listed at least two treatment sequences for comparison, and an indicator of whether a de novo treatment-sequencing model was employed. Each row represents information for a single TA.

The symbols in each column are defined as follows:

† for the second-to-right column: "-" signifies the document is not applicable or unavailable for the TA. "Y" and "N" indicate if the report lists at least two treatment sequences for comparison, yes or no, respectively. For multiple TAs, each company's submission is detailed in brackets. "Details unavailable" is noted for unassessable detailed reports. Sepcial conditions, such as a TA acts as a review of previous TAs are also bracketed.

¶ for the rightmost column: "-" again means the document is not applicable or unavailable. "Y" and "N" show if a de novo treatment-sequencing model was employed, yes or no, respectively, with "na" indicating no model was built, typically in ERG reports that review rather than create models. Information is separated by brackets for multiple TAs, and "details unavailable" is used where further details of the report could not be assessed. Brackets indicate special conditions, such as supplemental models in amendments.

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
612	Neratinib for extended adjuvant treatment of hormone receptor-positive, HER2-positive early stage breast cancer after adjuvant trastuzumab	Oncology	breast cancer	N	N	-	N	na	-
611	Rucaparib for maintenance treatment of relapsed platinum-sensitive ovarian, fallopian tube or peritoneal cancer	Oncology	ovarian cancer	N	N	-	N	na	-
610	Pentosan polysulfate sodium for treating bladder pain syndrome	Other	bladder pain syndrome	N	N	-	Y	na	-
607	Rivaroxaban for preventing atherothrombotic events in people with coronary or peripheral artery disease	Cardiovascular disease	TE	N	N	-	N	na	-
606	Lanadelumab for preventing recurrent attacks of hereditary angioedema	Other	hereditary angioedema	N	N	-	N	na	-
605	Xeomin (botulinum neurotoxin type A) for treating chronic sialorrhoea	Neurology/mental health	sialorrhoea	N	N	-	N	na	-
604	Idelalisib for treating refractory follicular lymphoma	Oncology	lymphoma	N	N	-	N	na	-
600	Pembrolizumab with carboplatin and paclitaxel for untreated metastatic squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
599	Sodium zirconium cyclosilicate for treating hyperkalaemia	Other	hyperkalaemia	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
598	Olaparib for maintenance treatment of BRCA mutation-positive advanced ovarian, fallopian tube or peritoneal cancer after response to first-line platinum-based chemotherapy	Oncology	ovarian cancer	N	N	-	N	na	-
597	Dapagliflozin with insulin for treating type 1 diabetes	Diabetes mellitus	DM	N	N	-	N	na	-
596	Risankizumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	N	N	-	N	na	-
595	Dacomitinib for untreated EGFR mutation-positive non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
593	Ribociclib with fulvestrant for treating hormone receptor-positive, HER2-negative, advanced breast cancer	Oncology	breast cancer	N	N	-	Y	na	-
592	Cemiplimab for treating metastatic or locally advanced cutaneous squamous cell carcinoma	Oncology	skin cancer	N	N	-	N	na	-
591	Letermovir for preventing cytomegalovirus disease after a stem cell transplant	Infectious disease	CMV	N	N	-	N	na	-
590	Fluocinolone acetonide intravitreal implant for treating recurrent non-infectious uveitis	Other	uveitis	N	N	-	Y	na	-
589	Blinatumomab for treating acute lymphoblastic leukaemia in remission with minimal residual disease activity	Oncology	leukaemia	N	N	-	N	na	-
588	Nusinersen for treating spinal muscular atrophy	Other	spinal muscular atrophy	N	N	-	N	na	-
587	Lenalidomide plus dexamethasone for previously untreated multiple myeloma	Oncology	MM	N	N	-	N	na	-
586	Lenalidomide plus dexamethasone for multiple myeloma after 1 treatment with bortezomib	Oncology	MM	N	N	-	N	na	-
585	Ocrelizumab for treating primary progressive multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
584	Atezolizumab in combination for treating metastatic non-squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
583	Ertugliflozin with metformin and a dipeptidyl peptidase-4 inhibitor for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	na	na	-
581	Nivolumab with ipilimumab for untreated advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-
580	Enzalutamide for hormone-relapsed non-metastatic prostate cancer	Oncology	prostate cancer	Y	N	-	Y	na	-
579	Abemaciclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy	Oncology	breast cancer	N	N	-	N	na	-
578	Durvalumab for treating locally advanced unresectable non-small-cell lung cancer after platinum-based chemoradiation	Oncology	NSCLC	N	N	-	N	na	-
577	Brentuximab vedotin for treating CD30-positive cutaneous T-cell lymphoma	Oncology	lymphoma	N	N	-	N	na	-
575	Tildrakizumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-
574	Certolizumab pegol for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
573	Daratumumab with bortezomib and dexamethasone for previously treated multiple myeloma	Oncology	MM	N	N	-	Y	na	-
572	Ertugliflozin as monotherapy or with metformin for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	N	na	-
571	Brigatinib for treating ALK-positive advanced non-small-cell lung cancer after crizotinib	Oncology	NSCLC	N	N	-	N	na	-
569	Pertuzumab for adjuvant treatment of HER2-positive early stage breast cancer	Oncology	breast cancer	N	N	-	Y	na	-
567	Tisagenlecleucel for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic therapies	Oncology	lymphoma	N	N	-	N	na	-
563	Abemaciclib with an aromatase inhibitor for previously untreated, hormone receptor-positive, HER2-negative, locally advanced or metastatic breast cancer	Oncology	breast cancer	N	N	-	Y	na	-
562	Encorafenib with binimetinib for unresectable or metastatic BRAF V600 mutation-positive melanoma	Oncology	melanoma	N	N	-	Y	na	-
561	Venetoclax with rituximab for previously treated chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
559	Axicabtagene ciloleucel for treating diffuse large B-cell lymphoma and primary mediastinal large B-cell lymphoma after 2 or more systemic therapies	Oncology	lymphoma	N	N	-	N	na	-
558	Nivolumab for adjuvant treatment of completely resected melanoma with lymph node involvement or metastatic disease	Oncology	melanoma	N	N	-	N	na	-
557	Pembrolizumab with pemetrexed and platinum chemotherapy for untreated, metastatic, non-squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
556	Darvadstrocel for treating complex perianal fistulas in Crohn's disease	Autoimmune	Crohn's disease	N	N	-	Y	na	-
554	Tisagenlecleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people aged up to 25 years	Oncology	leukaemia	N	N	-	N	na	-
553	Pembrolizumab for adjuvant treatment of resected melanoma with high risk of recurrence	Oncology	melanoma	N	N	-	N	na	-
552	Liposomal cytarabine–daunorubicin for untreated acute myeloid leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
551	Lenvatinib for untreated advanced hepatocellular carcinoma	Oncology	HCC	N	N	-	N	na	-
550	Vandetanib for treating medullary thyroid cancer	Oncology	thyroid cancer	N	N	-	N	na	-
547	Tofacitinib for moderately to severely active ulcerative colitis	Autoimmune	UC	Y	Y	-	Y	na	-
546	Padeliporfin for untreated localised prostate cancer	Oncology	prostate cancer	N	N	-	Y	na	-
545	Gemtuzumab ozogamicin for untreated acute myeloid leukaemia	Oncology	leukaemia	N	N	-	N	na	-
544	Dabrafenib with trametinib for adjuvant treatment of resected BRAF V600 mutation-positive melanoma	Oncology	melanoma	N	N	-	Y	na	-
543	Tofacitinib for treating active psoriatic arthritis after inadequate response to DMARDs	Autoimmune	psoriasis	Y	Y	-	Y	na	-
542	Cabozantinib for untreated advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
541	Inotuzumab ozogamicin for treating relapsed or refractory B-cell acute lymphoblastic leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
540	Pembrolizumab for treating relapsed or refractory classical Hodgkin lymphoma	Oncology	lymphoma	N	N	-	Y	na	-
539	Lutetium (177Lu) oxodotreotide for treating unresectable or metastatic neuroendocrine tumours	Oncology	NETs	N (Novartis, AAA)	-	N	N (Novartis, AAA)	-	N
538	Dinutuximab beta for treating neuroblastoma	Oncology	neuroblastoma	N	N	-	N	na	-
537	Ixekizumab for treating active psoriatic arthritis after inadequate response to DMARDs	Autoimmune	psoriasis	Y	Y	-	Y	na	-
536	Alectinib for untreated ALK-positive advanced non-small-cell lung cancer	Oncology	NSCLC	N	N	-	Y	na	-
535	Lenvatinib and sorafenib for treating differentiated thyroid cancer after radioactive iodine	Oncology	thyroid cancer	N (Bayer, Eisai)	-	N	N (Bayer, Eisai)	-	N
534	Dupilumab for treating moderate to severe atopic dermatitis	Other	atopic dermatitis	N	N	-	N	na	-
533	Ocrelizumab for treating relapsing–remitting multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
531	Pembrolizumab for untreated PD-L1-positive metastatic non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
530	Nivolumab for treating locally advanced unresectable or metastatic urothelial cancer after platinum-containing chemotherapy	Oncology	urothelial cancer	N	N	-	N	na	-
529	Crizotinib for treating ROS1-positive advanced non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
528	Niraparib for maintenance treatment of relapsed, platinum-sensitive ovarian, fallopian tube and peritoneal cancer	Oncology	ovarian cancer	N	N	-	N	na	-
527	Beta interferons and glatiramer acetate for treating multiple sclerosis	Autoimmune	MS	N	-	N	N (Biogen, Teva, Merck)	-	N
526	Arsenic trioxide for treating acute promyelocytic leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
525	Atezolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy	Oncology	urothelial cancer	N	N	-	N	na	-
524	Brentuximab vedotin for treating CD30-positive Hodgkin lymphoma	Oncology	lymphoma	N	N	-	N	na	-
523	Midostaurin for untreated acute myeloid leukaemia	Oncology	leukaemia	N	N	-	N	na	-
522	Pembrolizumab for untreated PD-L1-positive locally advanced or metastatic urothelial cancer when cisplatin is unsuitable	Oncology	urothelial cancer	N	N	-	N	na	-
521	Guselkumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	N	Y	-	N	na	-
520	Atezolizumab for treating locally advanced or metastatic non-small-cell lung cancer after chemotherapy	Oncology	NSCLC	N	N	-	N	na	-
519	Pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy	Oncology	urothelial cancer	N	N	-	N	na	-
518	Tocilizumab for treating giant cell arteritis	Autoimmune	GCA	N	N	-	Y	na	-
517	Avelumab for treating metastatic Merkel cell carcinoma	Oncology	MCC	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
516	Cabozantinib for treating medullary thyroid cancer	Oncology	thyroid cancer	N (Sanofi), details unavailable (Ipsen)	-	N	N (Sanofi), na (Ipsen)	-	N
515	Eribulin for treating locally advanced or metastatic breast cancer after 1 chemotherapy regimen	Oncology	breast cancer	N	N	-	N	na	-
513	Obinutuzumab for untreated advanced follicular lymphoma	Oncology	lymphoma	N	N	-	N	na	-
512	Tivozanib for treating advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-
511	Brodalumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-
510	Daratumumab monotherapy for treating relapsed and refractory multiple myeloma	Oncology	MM	N	N	-	N	na	-
509	Pertuzumab with trastuzumab and docetaxel for treating HER2-positive breast cancer	Oncology	breast cancer	N	N	-	N	na	-
507	Sofosbuvir–velpatasvir–voxilaprevir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	Y	na	-
506	Lesinurad for treating chronic hyperuricaemia in people with gout	Other	gout	N	N	-	Y	na	-
505	Ixazomib with lenalidomide and dexamethasone for treating relapsed or refractory multiple myeloma	Oncology	MM	N	N	-	N	na	-
504	Pirfenidone for treating idiopathic pulmonary fibrosis	Autoimmune	IPF	N	N	-	N	na	-
503	Fulvestrant for untreated locally advanced or metastatic oestrogen-receptor positive breast cancer	Oncology	breast cancer	N	N	-	N	na	-
502	Ibrutinib for treating relapsed or refractory mantle cell lymphoma	Oncology	lymphoma	N	N	-	Y (scenario analysis)	na	-
500	Ceritinib for untreated ALK-positive non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
499	Glecaprevir–pibrentasvir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	Y	-	-
498	Lenvatinib with everolimus for previously treated advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-
497	Golimumab for treating non-radiographic axial spondyloarthritis	Autoimmune	spondyloarthritis	N	N	-	na	na	-
496	Ribociclib with an aromatase inhibitor for previously untreated, hormone receptor-positive, HER2-negative, locally advanced or metastatic breast cancer	Oncology	breast cancer	N	N	-	Y	na	na (DSU report)
495	Palbociclib with an aromatase inhibitor for previously untreated, hormone receptor-positive, HER2-negative, locally advanced or metastatic breast cancer	Oncology	breast cancer	N	N	-	Y	na	-
492	Atezolizumab for untreated PD-L1-positive locally advanced or metastatic urothelial cancer when cisplatin is unsuitable	Oncology	urothelial cancer	N	N	-	N	na	-
491	Ibrutinib for treating Waldenstrom's macroglobulinaemia	Oncology	WM	N	N	-	Y	na	-
490	Nivolumab for treating squamous cell carcinoma of the head and neck after platinum-based chemotherapy	Oncology	skin cancer	N	N	-	N	na	-
489	Vismodegib for treating basal cell carcinoma	Oncology	basal cell carcinoma	N	N	-	N	na	-
488	Regorafenib for previously treated unresectable or metastatic gastrointestinal stromal tumours	Oncology	GIST	N	N	-	N	na	-
487	Venetoclax for treating chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
485	Sarilumab for moderate to severe rheumatoid arthritis	Autoimmune	RA	Y	Y	-	Y	na	-
484	Nivolumab for previously treated non-squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
483	Nivolumab for previously treated squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
482	Immunosuppressive therapy for kidney transplant in children and young people	Other	immunosuppressive therapy	N	-	N	na	-	N
481	Immunosuppressive therapy for kidney transplant in adults	Other	immunosuppressive therapy	N	-	N	na	-	N
478	Brentuximab vedotin for treating relapsed or refractory systemic anaplastic large cell lymphoma	Oncology	lymphoma	N	N	-	N	na	-
476	Paclitaxel as albumin-bound nanoparticles with gemcitabine for untreated metastatic pancreatic cancer	Oncology	pancreatic cancer	N	N	-	N	na	-
475	Dimethyl fumarate for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-
472	Obinutuzumab with bendamustine for treating follicular lymphoma refractory to rituximab	Oncology	lymphoma	N	N	-	Y	na	-
471	Eluxadoline for treating irritable bowel syndrome with diarrhoea	Autoimmune	irritable bowel syndrome	N	N	-	N	na	-
466	Baricitinib for moderate to severe rheumatoid arthritis	Autoimmune	RA	Y	Y	-	Y	na	-
463	Cabozantinib for previously treated advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-
462	Nivolumab for treating relapsed or refractory classical Hodgkin lymphoma	Oncology	lymphoma	Y	Y	-	N (original report) -> Y (amendment)	na	-
460	Adalimumab and dexamethasone for treating non-infectious uveitis	Other	uveitis	N (AbbVie, Allergan)	-	N	na (AbbVie, Allergan)	-	N
458	Trastuzumab emtansine for treating HER2-positive advanced breast cancer after trastuzumab and a taxane	Oncology	breast cancer	N	N	-	N	na	-
457	Carfilzomib for previously treated multiple myeloma	Oncology	MM	-	-	-	-	-	-
456	Ustekinumab for moderately to severely active Crohn's disease after previous treatment	Autoimmune	Crohn's disease	N	N	-	Y	na	-
455	Adalimumab, etanercept and ustekinumab for treating plaque psoriasis in children and young people	Autoimmune	psoriasis	N	-	N	N	-	N
451	Ponatinib for treating chronic myeloid leukaemia and acute lymphoblastic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
450	Blinatumomab for previously treated Philadelphia-chromosome-negative acute lymphoblastic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
449	Everolimus and sunitinib for treating unresectable or metastatic neuroendocrine tumours in people with progressive disease	Oncology	NETs	N (Novartis, AAA, Pfizer)	-	N	N (Novartis, AAA, Pfizer), na (Pfizer)	-	N
448	Etelcalcetide for treating secondary hyperparathyroidism	Other	hyperparathyroidism	N	N	-	N	Y	
445	Certolizumab pegol and secukinumab for treating active psoriatic arthritis after inadequate response to DMARDs	Autoimmune	psoriasis	Y(UCB), N(Novartis)	-	Y	Y (UCB), N(Novartis)	-	Y
442	Ixekizumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
440	Pegylated liposomal irinotecan for treating pancreatic cancer after gemcitabine	Oncology	pancreatic cancer	N	N	-	Y	na	-
439	Cetuximab and panitumumab for previously untreated metastatic colorectal cancer	Oncology	colorectal cancer	N (Merck), details unavailable (Amgen)	-	N	Y (Merck), na (Amgen),	-	Y
433	Apremilast for treating active psoriatic arthritis	Autoimmune	psoriasis	Y	-	Y	Y	Y	-
432	Everolimus for advanced renal cell carcinoma after previous treatment	Oncology	RCC	N	N	-	N	na	-
430	Sofosbuvir–velpatasvir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	N	na	-
429	Ibrutinib for previously treated chronic lymphocytic leukaemia and untreated chronic lymphocytic leukaemia with 17p deletion or TP53 mutation	Oncology	leukaemia	N	N	-	Y	na	-
428	Pembrolizumab for treating PD-L1-positive non-small-cell lung cancer after chemotherapy	Oncology	NSCLC	N	N	-	N	na	-
427	Pomalidomide for multiple myeloma previously treated with lenalidomide and bortezomib	Oncology	MM	N	N	-	N	na	-
426	Dasatinib, nilotinib and imatinib for untreated chronic myeloid leukaemia	Oncology	leukaemia	Y (partial: in first-line setting, a of review TA251)	Y (partial: in first-line setting: a review of TA251)	-	N (cost-minimisation analysis), Y (TA 251: Novartis, BMS and AG)	na (DSU report)	-
425	Dasatinib, nilotinib and high-dose imatinib for treating imatinib-resistant or intolerant chronic myeloid leukaemia	Oncology	leukaemia	N	N	-	N	na (DSU report)	-
424	Pertuzumab for the neoadjuvant treatment of HER2-positive breast cancer	Oncology	breast cancer	N	N	-	N	na	-
423	Eribulin for treating locally advanced or metastatic breast cancer after 2 or more chemotherapy regimens	Oncology	breast cancer	N	N	-	N	na	-
422	Crizotinib for previously treated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
421	Everolimus with exemestane for treating advanced breast cancer after endocrine therapy	Oncology	breast cancer	N	N	-	N	na	-
420	Ticagrelor for preventing atherothrombotic events after myocardial infarction	Cardiovascular disease	TE	N	N	-	N	na	-
419	Apremilast for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	Y	Y	-	Y	na	-
418	Dapagliflozin in triple therapy for treating type 2 diabetes	Diabetes mellitus	DM	Y	Y	-	Y	na	-
417	Nivolumab for previously treated advanced renal cell carcinoma	Oncology	RCC	N	N	-	Y	na	-
416	Osimertinib for treating locally advanced or metastatic EGFR T790M mutation-positive non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
415	Certolizumab pegol for treating rheumatoid arthritis after inadequate response to a TNF-alpha inhibitor	Autoimmune	RA	Y	Y	-	Y	na	-
414	Cobimetinib in combination with vemurafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma	Oncology	melanoma	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
413	Elbasvir–grazoprevir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	N	na	-
412	Radium-223 dichloride for treating hormone-relapsed prostate cancer with bone metastases	Oncology	prostate cancer	N	N	-	N	na	-
411	Necitumumab for untreated advanced or metastatic squamous non-small-cell lung cancer	Oncology	NSCLC	N	N	-	Y	na	-
410	Talimogene laherparepvec for treating unresectable metastatic melanoma	Oncology	melanoma	N	N	-	Y	na	-
409	Aflibercept for treating visual impairment caused by macular oedema after branch retinal vein occlusion	Other	visual impairment	Y	Y	-	Y	na	-
408	Pegaspargase for treating acute lymphoblastic leukaemia	Oncology	leukaemia	Y	Y	-	Y	na	-
407	Secukinumab for active ankylosing spondylitis after treatment with non-steroidal anti-inflammatory drugs or TNF-alpha inhibitors	Autoimmune	spondyloarthritis	N	N	-	N	na	-
406	Crizotinib for untreated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
405	Trifluridine–tipiracil for previously treated metastatic colorectal cancer	Oncology	colorectal cancer	N	N	-	N	na	-
404	Degarelix for treating advanced hormone-dependent prostate cancer	Oncology	prostate cancer	N	N	-	Y	na	-
403	Ramucirumab for previously treated locally advanced or metastatic non-small-cell lung cancer	Oncology	NSCLC	N	N	-	Y	na	-
402	Pemetrexed maintenance treatment for non-squamous non-small-cell lung cancer after pemetrexed and cisplatin	Oncology	NSCLC	N	N	-	N	na	-
401	Bosutinib for previously treated chronic myeloid leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
400	Nivolumab in combination with ipilimumab for treating advanced melanoma	Oncology	melanoma	N	N	-	N	na	-
399	Azacitidine for treating acute myeloid leukaemia with more than 30% bone marrow blasts	Oncology	leukaemia	N	N	-	N	na	-
396	Trametinib in combination with dabrafenib for treating unresectable or metastatic melanoma	Oncology	melanoma	N	N	-	N	na	-
395	Ceritinib for previously treated anaplastic lymphoma kinase positive non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
393	Alirocumab for treating primary hypercholesterolaemia and mixed dyslipidaemia	Cardiovascular disease	dyslipidaemia	N	N	-	N	na	-
391	Cabazitaxel for hormone-relapsed metastatic prostate cancer treated with docetaxel	Oncology	prostate cancer	N	N	-	N	na	-
390	Canagliflozin, dapagliflozin and empagliflozin as monotherapies for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	Y (Janssen, AZ, Boehringer Ingelheim)	-	Y
389	Topotecan, pegylated liposomal doxorubicin hydrochloride, paclitaxel, trabectedin and gemcitabine for treating recurrent ovarian cancer	Oncology	ovarian cancer	N	N	-	N (Eli Lilly, PharmaMar)	-	N
388	Sacubitril valsartan for treating symptomatic chronic heart failure with reduced ejection fraction	Cardiovascular disease	HF	N	N	-	N	na	-
387	Abiraterone for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated	Oncology	prostate cancer	N	N	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
386	Ruxolitinib for treating disease-related splenomegaly or symptoms in adults with myelofibrosis	Oncology	myelofibrosis	N	N	-	Y	na	-
384	Nivolumab for treating advanced (unresectable or metastatic) melanoma	Oncology	melanoma	N	N	-	N	na	-
383	TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis	Autoimmune	spondyloarthritis	N (Pfizer, AbbVie, UCB, MSD)	-	N	Y (Pfizer), N (AbbVie, UCB, MSD)	-	Y
380	Panobinostat for treating multiple myeloma after at least 2 previous treatments	Oncology	MM	N	N	-	Y	na	-
378	Ramucirumab for treating advanced gastric cancer or gastro–oesophageal junction adenocarcinoma previously treated with chemotherapy	Oncology	gastric cancer	N	N	-	N	na	-
377	Enzalutamide for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated	Oncology	prostate cancer	Y	Y	-	Y	na	-
375	Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for rheumatoid arthritis not previously treated with DMARDs or after conventional DMARDs only have failed	Autoimmune	RA	Y (AbbVie, BMS, MSD*2, Pfizer, Roche, UCB)	-	Y	Y (AbbVie, BMS, MSD, Pfizer, Roche, UCB)	-	
374	Erlotinib and gefitinib for treating non-small-cell lung cancer that has progressed after prior chemotherapy	Oncology	NSCLC	N	-	N	na	-	N
373	Abatacept, adalimumab, etanercept and tocilizumab for treating juvenile idiopathic arthritis	Autoimmune	JIA	N (Roche, BMS, Pfizer, AbbVie)	-	N	N (Roche), na (BMS, Pfizer, AbbVie)	-	N
370	Bortezomib for previously untreated mantle cell lymphoma	Oncology	lymphoma	N	N	-	Y	na	-
367	Vortioxetine for treating major depressive episodes	Neurology/mental health	depression	N	N	-	Y	na	-
366	Pembrolizumab for advanced melanoma not previously treated with ipilimumab	Oncology	melanoma	N	N	-	N	na	-
365	Ombitasvir–paritaprevir–ritonavir with or without dasabuvir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	N	na	-
363	Ledipasvir–sofosbuvir for treating chronic hepatitis C	Infectious disease	hepatitis C	N	N	-	N	na	-
359	Idelalisib for treating chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
357	Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab	Oncology	melanoma	N	N	-	N	na	-
355	Edoxaban for preventing stroke and systemic embolism in people with non-valvular atrial fibrillation	Cardiovascular disease	TE	N	N	-	N	na	-
354	Edoxaban for treating and for preventing deep vein thrombosis and pulmonary embolism	Cardiovascular disease	TE	N	N	-	N	na	-
352	Vedolizumab for treating moderately to severely active Crohn's disease after prior therapy	Autoimmune	Crohn's disease	N	N	-	N	na	-
350	Secukinumab for treating moderate to severe plaque psoriasis	Autoimmune	psoriasis	N	N	-	N	na	-
347	Nintedanib for previously treated locally advanced, metastatic, or locally recurrent non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
346	Aflibercept for treating diabetic macular oedema	Other	DMO	N	N	-	N	na	-
345	Naloxegol for treating opioid-induced constipation	Other	constipation	N	N	-	N	na	-
343	Obinutuzumab in combination with chlorambucil for untreated chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
342	Vedolizumab for treating moderately to severely active ulcerative colitis	Autoimmune	UC	N	N	-	N	na	-
341	Apixaban for the treatment and secondary prevention of deep vein thrombosis and/or pulmonary embolism	Cardiovascular disease	TE	N	N	-	N	na	-
340	Ustekinumab for treating active psoriatic arthritis	Autoimmune	psoriasis	N	N	-	Y	na	-
336	Empagliflozin in combination therapy for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	N	na	-
335	Rivaroxaban for preventing adverse outcomes after acute management of acute coronary syndrome	Cardiovascular disease	ACS	N	N	-	N	na	-
333	Axitinib for treating advanced renal cell carcinoma after failure of prior systemic treatment	Oncology	RCC	N	N	-	N	na	-
329	Infliximab, adalimumab and golimumab for treating moderately to severely active ulcerative colitis after the failure of conventional therapy	Autoimmune	UC	N	-	N	N (AbbVie, MSD)	-	Y
327	Dabigatran etexilate for the treatment and secondary prevention of deep vein thrombosis and/or pulmonary embolism	Cardiovascular disease	TE	N	N	N	Y	na	-
326	Imatinib for the adjuvant treatment of gastrointestinal stromal tumours	Oncology	GIST	N	N	-	Y	na	-
325	Nalmefene for reducing alcohol consumption in people with alcohol dependence	Other	alcohol dependence	N	N	-	N	na	-
322	Lenalidomide for treating myelodysplastic syndromes associated with an isolated deletion 5q cytogenetic abnormality	Oncology	MDS	N	N	-	N	na	-
321	Dabrafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma	Oncology	melanoma	N	N	-	N	na	-
320	Dimethyl fumarate for treating relapsing-remitting multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
319	Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma	Oncology	melanoma	Y	Y	-	Y	na	-
316	Enzalutamide for metastatic hormone-relapsed prostate cancer previously treated with a docetaxel-containing regimen	Oncology	prostate cancer	N	N	-	N	na	-
315	Canagliflozin in combination therapy for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	Y	na	-
312	Alemtuzumab for treating relapsing-remitting multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
311	Bortezomib for induction therapy in multiple myeloma before high-dose chemotherapy and autologous stem cell transplantation	Oncology	MM	N	N	-	Y	na	-
310	Afatinib for treating epidermal growth factor receptor mutation-positive locally advanced or metastatic non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
308	Rituximab in combination with glucocorticoids for treating anti-neutrophil cytoplasmic antibody-associated vasculitis	Autoimmune	ANCA-associated vasculitis	Y	Y	-	Y	na	-
307	Aflibercept in combination with irinotecan and fluorouracil-based therapy for treating metastatic colorectal cancer that	Oncology	colorectal cancer	N	N	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
	has progressed following prior oxaliplatin-based chemotherapy								
306	Pixantrone monotherapy for treating multiply relapsed or refractory aggressive non-Hodgkin's B-cell lymphoma	Oncology	lymphoma	N	N	-	Y	na	-
305	Aflibercept for treating visual impairment caused by macular oedema secondary to central retinal vein occlusion	Other	visual impairment	N	N	-	N	na	-
303	Teriflunomide for treating relapsing–remitting multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
300	Peginterferon alfa and ribavirin for treating chronic hepatitis C in children and young people	Infectious disease	hepatitis C	N	N	-	N	-	N
298	Ranibizumab for treating choroidal neovascularisation associated with pathological myopia	Other	CNV	N	N	-	N	na	-
297	Ocriplasmin for treating vitreomacular traction	Other	VMT	N	N	-	N	na	-
294	Aflibercept solution for injection for treating wet age-related macular degeneration	Other	AMD	N	N	-	N	na	-
293	Eltrombopag for treating chronic immune (idiopathic) thrombocytopenic purpura	Autoimmune	ITP	Y	Y	-	Y	na	-
292	Aripiprazole for treating moderate to severe manic episodes in adolescents with bipolar I disorder	Neurology/mental health	bipolar disorder	Y	Y	-	Y	na	-
290	Mirabegron for treating symptoms of overactive bladder	Other	overactive bladder	N	N	-	Y	na	-
288	Dapagliflozin in combination therapy for treating type 2 diabetes	Diabetes mellitus	DM	N	N	-	Y	na	-
285	Bevacizumab in combination with gemcitabine and carboplatin for treating the first recurrence of platinum-sensitive advanced ovarian cancer	Oncology	ovarian cancer	N	N	-	N	na	-
284	Bevacizumab in combination with paclitaxel and carboplatin for first-line treatment of advanced ovarian cancer	Oncology	ovarian cancer	N	N	-	N (Eli Lilly, PharmaMar)	-	N
276	Colistimethate sodium and tobramycin dry powders for inhalation for treating pseudomonas lung infection in cystic fibrosis	Other	cystic fibrosis	N	N	-	N (Novartis)	-	N
275	Apixaban for preventing stroke and systemic embolism in people with nonvalvular atrial fibrillation	Cardiovascular disease	TE	N	N	-	N	na	-
274	Ranibizumab for treating diabetic macular oedema	Other	DMO	N	N	-	N	na	-
269	Vemurafenib for treating locally advanced or metastatic BRAF V600 mutation-positive malignant melanoma	Oncology	melanoma	N	N	-	N	na	-
268	Ipilimumab for previously treated advanced (unresectable or metastatic) melanoma	Oncology	melanoma	N	N	-	N	na	-
266	Mannitol dry powder for inhalation for treating cystic fibrosis	Other	cystic fibrosis	N	N	-	N	na	-
263	Bevacizumab in combination with capecitabine for the first-line treatment of metastatic breast cancer	Oncology	breast cancer	N	N	-	N	na	-
260	Botulinum toxin type A for the prevention of headaches in adults with chronic migraine	Neurology/mental health	migraine	N	N	-	N	na	-
259	Abiraterone for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen	Oncology	prostate cancer	N	N	-	N	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
257	Lapatinib or trastuzumab in combination with an aromatase inhibitor for the first-line treatment of metastatic hormone-receptor-positive breast cancer that overexpresses HER2	Oncology	breast cancer	N (Roche, GSK)	N	-	N (Roche, GSK)	-	N
256	Rivaroxaban for the prevention of stroke and systemic embolism in people with atrial fibrillation	Cardiovascular disease	TE	N	N	-	Y	na	-
254	Fingolimod for the treatment of highly active relapsing–remitting multiple sclerosis	Autoimmune	MS	N	N	-	N	na	-
249	Dabigatran etexilate for the prevention of stroke and systemic embolism in atrial fibrillation	Cardiovascular disease	TE	Y	Y	-	Y	na	-
247	Tocilizumab for the treatment of rheumatoid arthritis	Autoimmune	RA	N	N	-	Y	na	-
243	Rituximab for the first-line treatment of stage III-IV follicular lymphoma	Oncology	lymphoma	N	N	-	Y	na	-
242	Cetuximab, bevacizumab and panitumumab for the treatment of metastatic colorectal cancer after first-line chemotherapy: Cetuximab (monotherapy or combination chemotherapy), bevacizumab (in combination with non-oxaliplatin chemotherapy) and panitumumab (monotherapy) for the treatment of metastatic colorectal cancer after first-line chemotherapy	Oncology	colorectal cancer	N (Merck, Roche, Amgen)	-	N	N (Merck, Roche, Amgen)	-	N
239	Fulvestrant for the treatment of locally advanced or metastatic breast cancer	Oncology	breast cancer	N	N	-	N	na	-
238	Tocilizumab for the treatment of systemic juvenile idiopathic arthritis	Autoimmune	JIA	Y	Y	-	Y	na	-
228	Bortezomib and thalidomide for the first-line treatment of multiple myeloma	Oncology	MM	N (Janssen-Cilag, Celgene)	-	N	N (Janssen-Cilag, Celgene)	-	N
227	Erlotinib monotherapy for maintenance treatment of non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
226	Rituximab for the first-line maintenance treatment of follicular non-Hodgkin's lymphoma	Oncology	lymphoma	N	N	-	Y	na	-
225	Golimumab for the treatment of rheumatoid arthritis after the failure of previous disease-modifying anti-rheumatic drugs	Autoimmune	RA	Y	Y	-	Y	na	-
223	Cilostazol, naftidrofuryl oxalate, pentoxifylline and inositol nicotinate for the treatment of intermittent claudication in people with peripheral arterial disease	Cardiovascular disease	TE	N (Otsuka)	N	-	na (Otsuka)	-	N
220	Golimumab for the treatment of psoriatic arthritis	Autoimmune	psoriasis	N	N	-	N	N	-
217	Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease	Neurology/mental health	Alzheimer's disease	N (Lundbeck, Eisai-Pfizer)	-	N	N (Lundbeck, Eisai-Pfizer)	-	N
216	Bendamustine for the first-line treatment of chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
215	Pazopanib for the first-line treatment of advanced renal cell carcinoma	Oncology	RCC	N	N	-	N	na	-
214	Bevacizumab in combination with a taxane for the first-line treatment of metastatic breast cancer	Oncology	breast cancer	N	N	-	N	N	-
213	Aripiprazole for the treatment of schizophrenia in people aged 15 to 17 years	Neurology/mental health	schizophrenia	N	N	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
212	Bevacizumab in combination with oxaliplatin and either fluorouracil plus folinic acid or capecitabine for the treatment of metastatic colorectal cancer	Oncology	colorectal cancer	N	N	-	N	na	-
210	Clopidogrel and modified-release dipyridamole for the prevention of occlusive vascular events	Cardiovascular disease	TE	N(Boehringer-Ingelheim, Sanofi-BMS)	-	N	N (Boehringer-Ingelheim, Sanofi-BMS)	-	N
209	Imatinib for the treatment of unresectable and/or metastatic gastrointestinal stromal tumours	Oncology	GIST	N	N	-	na	Y	-
199	Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis	Autoimmune	psoriasis	N (Abbot, Schering-Plough, Wyeth)	-	N	N (Abbot, Schering-Plough, Wyeth)	-	N → Y (exploratory analysis)
197	Dronedarone for the treatment of non-permanent atrial fibrillation	Cardiovascular disease	AF	N	N		Y	na	-
195	Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor	Autoimmune	RA	Y (Abbot, Wyeth, Schering-Plough, Roche, BMS)	-	Y	Y (Abbot, Wyeth, Schering-Plough, Roche, BMS)	-	Y
193	Rituximab for the treatment of relapsed or refractory chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
190	Pemetrexed for the maintenance treatment of non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
187	Infliximab and adalimumab for the treatment of Crohn's disease	Autoimmune	Crohn's disease	N(Schering-Plough, Abbot)	-	N	N (Schering-Plough, Abbot)	-	N
183	Topotecan for the treatment of recurrent and stage IVB cervical cancer	Oncology	cervical cancer	N	N	-	N	na	-
181	Pemetrexed for the first-line treatment of non-small-cell lung cancer	Oncology	NSCLC	N	N	-	N	na	-
180	Ustekinumab for the treatment of adults with moderate to severe psoriasis	Autoimmune	psoriasis	N	N	-	N	na	-
178	Bevacizumab (first-line), sorafenib (first- and second-line), sunitinib (second-line) and temsirolimus (first-line) for the treatment of advanced and/or metastatic renal cell carcinoma	Oncology	RCC	N (Pfizer, Roche, Wyeth, Bayer)	N	-	N (Pfizer, Roche, Wyeth, Bayer)	-	N
174	Rituximab for the first-line treatment of chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	N	na	-
173	Tenofovir disoproxil for the treatment of chronic hepatitis B	Infectious disease	hepatitis B	Y	Y	-	Y	na	-
171	Lenalidomide for the treatment of multiple myeloma in people who have received at least 2 prior therapies	Oncology	MM	N	N	-	N	na	-
169	Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma	Oncology	RCC	N (Pfizer, Roche, Wyeth, Bayer)	N	-	N (Pfizer, Roche, Wyeth, Bayer)	-	N
164	Febuxostat for the management of hyperuricaemia in people with gout	Other	gout	N	N	-	N	na	-
161	Raloxifene and teriparatide for the secondary prevention of osteoporotic fragility fractures in postmenopausal women	Other	osteoporosis	N (Servier)	-	N	N (Servier)	-	N
154	Telbivudine for the treatment of chronic hepatitis B	Infectious disease	hepatitis B	Y	Y	-	Y	Y	-
153	Entecavir for the treatment of chronic hepatitis B	Infectious disease	hepatitis B	N	N	-	N	na	-
146	Adalimumab for the treatment of adults with psoriasis	Autoimmune	psoriasis	N	N	-	Y	na	-

TA	Title	Disease Area	Disease	Listed at least two treatment sequences for comparison			Employ a de novo model with treatment-sequencing structure¶		
				CS	ERG reports	AG reports	CS	ERG* reports	AG reports
138	Inhaled corticosteroids for the treatment of chronic asthma in adults and in children aged 12 years and over	Other	asthma	N	N	-	Y	na	-
137	Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma	Oncology	lymphoma	Y	Y	-	Y	N	-
134	Infliximab for the treatment of adults with psoriasis	Autoimmune	psoriasis	N	N	-	Y	na	-
121	Carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma	Other	glioma	N	N	-	Y	na	-
119	Fludarabine monotherapy for the first-line treatment of chronic lymphocytic leukaemia	Oncology	leukaemia	N	N	-	Y	na	-
118	Bevacizumab and cetuximab for the treatment of metastatic colorectal cancer	Oncology	colorectal cancer	N	N	-	N	na	-
116	Gemcitabine for the treatment of metastatic breast cancer	Oncology	breast cancer	N	N	-	N	na	-
103	Etanercept and efalizumab for the treatment of adults with psoriasis	Autoimmune	psoriasis	N (Wyeth, Serono)	Y	-	Details unavailable (Wyeth, Serono)	-	Y
101	Docetaxel for the treatment of hormone-refractory metastatic prostate cancer	Oncology	prostate cancer	N	N	-	N	na	-
100	Capecitabine and oxaliplatin in the adjuvant treatment of stage III (Dukes' C) colon cancer	Oncology	colorectal cancer	N	N	-	N	na	-
96	Adefovir dipivoxil and peginterferon alfa-2a for the treatment of chronic hepatitis B	Infectious disease	hepatitis B	N	N	-	N	na	-
82	Tacrolimus and pimecrolimus for atopic eczema	Other	eczema	N	N	-	N	na	-
70	Guidance on the use of imatinib for chronic myeloid leukaemia	Oncology	leukaemia	N	N	-	N	na	-
55	Guidance on the use of paclitaxel in the treatment of ovarian cancer	Oncology	ovarian cancer	N	N	-	N	na	-
52	Guidance on the use of drugs for early thrombolysis in the treatment of acute myocardial infarction	Cardiovascular disease	ACS	N	N	-	N	na	-
34	Guidance on the use of trastuzumab for the treatment of advanced breast cancer	Oncology	breast cancer	N	N	-	N	na	-
23	Guidance on the use of temozolomide for the treatment of recurrent malignant glioma (brain cancer)	Other	glioma	N	N	-	N	na	-

AAA: AAA Pharma, ACS: acute coronary syndrome (including myocardial infarction), ADPKD: autosomal dominant polycystic kidney disease, AG: Assessment Group reports, AMD: age-related macular degeneration, AZ: AstraZeneca, BMS: Bristol Myers Squibb, C. diff : Clostridium difficile, CNV: choroidal neovascularisation, COPD: chronic obstructive pulmonary disease, CS: company submissions, DM: diabetes mellitus, DMO: diabetic macular oedema, DSU: NICE Decision Support Unit, ERG: Evidence Review Group reports, FAD: Final Appraisal Determination document, GIST: gastrointestinal stromal tumour, GSK: GlaxoSmithKline, HCC: hepatocellular carcinoma, HF: heart failure: HNC: head and neck cancer, IPF: idiopathic pulmonary fibrosis, ITP: immune thrombocytopenia, JIA: juvenile arthritis, LSCD: limbal stem cell deficiency, MM: multiple myeloma, MS: multiple sclerosis, MSD: Merck Sharp & Dohme, NETs: neuroendocrine tumours, NSCLC: non-small cell lung cancer, RA: rheumatoid arthritis, RCC: renal cell carcinoma, SCLC: small cell lung cancer, SHE: TA: technology appraisals, TE: thromboembolism, UC: ulcerative colitis, UCB: Union Chimique Belge, WM: Waldenstrom macroglobulinemia.

Appendix 3.4 Details of treatment-sequence comparisons in Part B of the NICE Technology Appraisal (TA) systematic review in Chapter 3

For single and fast-track TAs (STA & FTA), treatment-sequencing comparisons were primarily extracted from CS reports, with relevant alternative sequences from ERG's sensitivity analyses also extracted. In the case of multiple TAs (MTA), sequencing comparisons from CS and AG reports were detailed separately, with the source for each comparison specified in parentheses beside after the TA number.

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
Oncology							
408 (CS)	Acute lymphoblastic leukaemia (ALL)	Yes	First-line and second-line in the model (newly diagnosed ALL)	life time horizon (5-year in decision tree)	Decision tree + Markov model	General treatment sequences in the decision tree model for adult (26-65 years) and paediatric patients (≤ 25 years): o **Pegaspargase** > Erwinase o Native E.coli asparaginase > Erwinase o Erwinase > **pegaspargase** o Erwinase > native E.coli asparaginase (Treatment switches occur in case of hypersensitivity to the first-line treatment)	Decision tree (5-year treatment phase): First-line treatment > (potential SCT in adults) > Second-line treatment > enter Markov model (experiencing OS and EFS outcomes) Survival extrapolation in Markov model (paediatric): EFS > Survival with relapse/secondary tumour (R/ST) > Death (OS) Survival extrapolation in Markov model (adult): Alive > Death (OS) Assumption: (1) Only patients in the EFS state continue to receive asparaginase treatments (2) Asparaginase treatment is ceased in adult patients once they receive a transplant (3) Pegaspargase, native asparaginase and Erwinase are equivalent in terms of OS and EFS. That is, all outcomes of interest are experienced during the treatment phase. EFS and OS are equivalent in adult patients
426 (CS/ERG) (CDF review of MTA 251 for dasatinib)	Chronic myeloid leukemia	No (cost-comparison analyses in TA 426) Yes (BMS, Novartis and AG model in TA 251)	First-line setting	Unspecified (likely to be life-time)	Unspecified (TA 426: cost-comparison analysis) (TA 251: likely to be Markov model or partitioned survival model)	Treatment sequences in AG report in TA 251: o **Dasatinib** --> nilotinib --> SCT/HU o Imatinib --> nilotinib --> SCT/HU o Nilotinib --> imatinib --> SCT/HU o **Dasatinib** --> imatinib --> SCT/HU Further exploratory treatment sequences suggested by the DSU in TA425 (dasatinib is available as later-line treatment since TA 241), no analysis were performed: o Nilotinib --> **dasatinib** --> SCT/HU o Imatinib --> **dasatinib** --> SCT/HU o Nilotinib --> **dasatinib** --> imatinib --> SCT/HU o **Dasatinib** --> nilotiinib --> imatinib --> SCT/HU o Imatinib --> nilotiinib --> **dasatinib** --> SCT/HU o Imatinib --> **dasatinib** --> nilotiinib --> SCT/HU o Nilotinib --> imatinib --> **dasatinib** --> SCT/HU o **Dasatinib** --> imatinib --> nilotiinib --> SCT/HU	TA 426: X (no economic model) The company performed scenario cost-comparison analysis without treatment sequences. Original TA 251: Lacking access to the original report
137 (CS)	Follicular non-Hodgkin's lymphoma	Yes	Second-line treatment: Rituximab use in either or both induction and maintenance phase in	Life-time (30 years)	Markov	2-arm model: assessing rituximab a maintenance treatment. Patients had previous complete or partial remission of at least a 4-week duration of CHOP +/- rituximab induction treatment. o **rituximab maintenance** o observation	4-arm model: PFS (induction phase)> PFS (maintenance phase) > PPS > death or

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
			treating relapse or refractory disease			4-arm model: assessing whether the use of rituximab as an induction therapy in addition to maintenance therapy is cost effective (i.e. induction treatment > maintenance treatment) o **R**-CHOP > **rituximab maintenance** o **R**-CHOP > observation (A proportion of patients were ineligible for maintenance after R-CHOP induction based on outcome of induction in the above two sequences) o CHOP > **rituximab maintenance** o CHOP > observation (A proportion of patients were ineligible for maintenance after CHOP induction based on outcome of induction in the above two sequences)	PFS (induction phase)> PFS (not in induction/maintenance phase) > PPS > death
462 (CS)	Hodgkin Lymphonma	No (Yes in scenario analysis)	First-line in relapsed or refractory HL (later-line in the treatment pathway: after chemotherapy and brentuximab with or without ASCT)	Life-time (40 years) Scenario analysis: 5-year, 10-year, 20-year	Semi-Markov model	Base-case (using pivotal trial survival data with sequence representing clinical practice, so it was not seen in the model structure): o **Nivolumab** -> BSC (chemotherapy, PAL, clinical trials) o Standard of care (survival curve: Cheah 2016; costs: bendamustine, BTX re-treatment, chemotherapy) > BSC Scenario analysis: o **Nivolumab** -> alloSCT> BSC o Standard of care > alloSCT > BSC	PFS -> PPS -> death (PFS and OS data implicitly include the effects of any subsequent treatment that may have been administered, the need to explicitly incorporate the effects of these subsequent treatments is negated) Amendment to model diagram: Progression free -> (scenario analysis: progression free with subsequent line) -> post-progression with subsequent line -> death
319 (CS)	Malignant melanoma	Yes	First-line (or second-line setting as salvage therapy in comparator arm)	Life-time (40 years) Scenario analysis: 10-, 20-, 30-year	Semi-Markov partitioned survival model	BRAF V600 mutation-genative pateints: o **Ipilimumab** > BSC > BSC o Dacarbazine > **Ipilimumab** > BSC BRAF V600 mutation-positive pateints: o **Ipilimumab** > vemurafenib > BSC o Vemurafenib > **Ipilimumab** > BSC o Dacarbazine > **Ipilimumab** > BSC Senario analysis: no active second-line treatments	First-line treatment > second-line treatment > Third-line treatment (> palliative care) > death Indirect transition to palliative care before death: a proportion of patients may receive 3 months palliative care before death from any lines of treatment.
377 (CS)	Prostate cancer	Yes	First-line treatment (for metastatic hormone resistant prostate cancer (mHRPC)	Life-time (10 years)	Markov model	Base-case: o **Enzalutamide** > docetaxel > palliative o BSC > docetaxel > **enzalutamide** > palliative o Abiraterone > docetaxel > palliative Company scenario analysis: o BSC > docetaxel > abieraterone > palliative ERG base-case: o **Enzalutamide** > docetaxel > abiraterone > palliative o Abiraterone > docetaxel > **enzalutam ide** > palliative o BSC > docetaxel > **enzalutamide** > palliative o BSC > docetaxel > abiraterone > palliative	Stable disease: first-line treatment > progressed disease (post-progression 1: second-line treatment > post progression 2: third-line treatment* > palliative care) > Death *only exists in the BSC first-line treatment arm
580 (CS)	Prostate cancer	Yes	First-line treatment in non-metastatic hormone-relapse prostate cancer	Life-time (20 years)	Partitioned survival (PFS/OS) + Markov model (Within PPS)	o **Enzalutamide** > ADT > ADT alone (60%), docetaxel (40%) > BSC o ADT > **Enzalutamide** (scenario analysis: abiraterone) > ADT alone (60%), docetaxel (40%) > BSC	nmHRPC > mHRPC (PD1(pre-chemo) > PD2(chemo) > PD3(post-chemo)) > Death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
Autoimmune							
308 (CS)	ANCA-associated vasculitis	Yes	First-line induction therapy (later-line in sensitivity analysis)	Life-time	Markov	Company's base-case of "all-patient", "treatment-naïve" and "recurrent disease" subgroups: o **Rituximab 1** (> **rituximab 2***) > cyclophosphamide (> AZA maintenance treatment) o Cyclophosphamide 1 > cyclophosphamide 2 (> AZA maintenance treatment) * The second course of RTX is only offered to patients who did not respond at all to the first course. Moreover, its efficacy is assumed to be lower. In this model, patients moved on to cyclophosphamide as the second-line induction therapy (in all treatment sequence comparators), if they did not achieve remission or relapse. Patients who achieve remission after a first-line rituximab induction received no further treatment until relapse, while patients achieved remission after first-line cyclophosphamide received azathioprine as maintenance therapy. ERG's proposed analysis of "all patient" & "treatment-naïve" subgroups: o cyclophosphamide > cyclophosphamide > supportive care o cyclophosphamide > **RTX** > cyclophosphamide > supportive care o cyclophosphamide > cyclophosphamide > **RTX** > supportive care o **RTX** > cyclophosphamide > cyclophosphamide > supportive care ERG's proposed analysis of the "recurrent disease" subgroup: o cyclophosphamide > supportive care o supportive care o cyclophosphamide > **RTX** > supportive care o **RTX** > cyclophosphamide > supportive care o **RTX** > supportive care	Non-remission (induction treatment first (+ second second) course)> complete remission > uncontrolled disease (treatment after release +/- maintenance treatment) > death
293 (CS)	Chronic Immune thrombocyto penic purpura	Yes	First-line (second- or later-line treatment in sensitivity analysis)	Life-time	Markov	Base-case (assumed patients receiving RTX prior to TPO-RA): o **Eltrombopag** > non TPO-RA pathway o Romiplostim > non TPO-RA pathway o Non TPO-RA pathway (azathioprine > mycophenolate mofetil > CYC > danazol > dapasone > cyclophosphamide > vincristine o> vinblastine > rescue as required) Alternative treatment sequences in sensitivity analyses: o RTX > non TPO-RA pathway o **Eltrombopag** > RTX > non TPO-RA pathway (in line with TA221) o Romiplostim > RTX > TPO-RA pathway (in line with TA221) o **Eltrombopag** o Romiplostim o RTX > non TPO-RA pathway o **Eltrombopag** > non TPO-RA pathway	(start a new treatment) Non-responder (cycle 1) > non-responder week (cycle 2) > non-responder week (cycle 3) > non-responder week (cycle 4) > responder > long-term non-responder > (go to a new treatment) A patient can go to the "responder" from any short-term non-responder state. Patients can die from general causes or from ITP-related death in any health state.

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						o Romiplostim > non TPO-RA pathway o Non TPO-RA pathway > **Eltrombopag** (before rescue treatment) o Non TPO-RA pathway > Romiplostim (before rescue treatment)	
238 **(CS)**	Juvenile Idiopathic Arthritis	Yes	First-line (after IR of NSAIDs/corticosteroids)	Life-time	Markov	Base-case 1 (MTX scenario): o **TCZ** > anakinra > ETN > ADA o MTX > anakinra > ETN > ADA Base-case 2 (anakinra scenario): o **TCZ** > ETN > ADA > ABA o anakinra > ETN > ADA > ABA Sensitivity analysis (entercept scenario): o **TCZ** > anakinra > ADA > ABA o ETN > anakinra > ADA > ABA Scenario analysis: o Use one treatment only o Use two treatments only o Use three treatments only Company's revised analysis 1: **TCZ** > INF INF > **TCZ** Infliximab anlone Company's revised analysis 2: **TCZ** > anakinra Anakinra > **TCZ** Anakinra alone ERG's scenario analysis including infliximab in the sequence: o **TCZ** > anakinra > INF > ABA o anakinra > INF > ADA > ABA o INF > ETN > ADA > ABA	First-line treatment (ACR response 30, 50, 70, 90, no response) > second-line treatment (ACR response 30, 50, 70, 90, no response) > third-line treatment (ACR response 30, 50, 70, 90, no response) > fourth-line treatment (ACR response 30, 50, 70, 90, no response) > death
103 **(AG)** **(Two CS from Wyeth and Serono did not compare treatment sequences, but only the AG report.)**	Psoriatic arthritis	Yes	First-line (after non-biologic systemic treatment)	Life-time	Markov	Base-case*: o **ETN 25 mg** o **ETN 25 mg continuous** o **ETN 50 mg** o **Efalizumab** o Supportive care Each treatment was compared to supportive care separately, and a threshold analysis testing at whether a treatment sequence including any other agent may be cost-effective at different level of willingness to pay (WTP) threshold. The AG claimed that their test provide potential ordering that patients trying different agents, but not strictly specify what treatment sequence is optimal. For example, in the base-case a single-line of supportive care was deemed to be the most cost-effective option when WTP threshold at 65000. However, the sequence including ETN 25 mg > supportive care becomes the most cost-effective option when the WTP threshold is above 70000. Below are treatment sequences identified to be the most cost-effective option at least once under different scenarios	First-line treatment > (yes/no response) > Second-line treatment > (yes/no response) > third-line treatment > (yes/no response) > fourth-line treatment > (yes/no response) > fifth-line treatment > (yes/no response) > sixth-line treatment > (yes/no response) > seventh-line treatment > (yes/no response) > eighth treatment > (yes/no response) > death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						Alternative scenario 1 (4th-quartile DLQI at baseline): o **ETN 25 mg** > **ETN 25 mg continuous** > **efalizumab** > **ETN 50 mg** > Supportive care o **ETN 25 mg** > **ETN 25 mg continuous** > **efalizumab** > Supportive care o **ETN 25 mg** > Supportive care o Supportive care Alternative scenario 2 (any DLQI at baseline, annual in-patient hospitalisation when not responding to therapy): o **ETN 25 mg** > **ETN 25 mg continuous** > efalizumab > Supportive care o **ETN 25 mg** > Supportive care o Supportive care Alternative scenario 3 (4th-quartile DLQI at baseline, annual in-patient hospitalisation when not responding to therapy): o **ETN 25 mg** > **ETN 25 mg continuous > efalizumab** > **ETN 50 mg** > Supportive care o **ETN 25 mg** > **ETN 25 mg continuous** > **efalizumab** > Supportive care o **ETN 25 mg** > Supportive care o Supportive care Alternative scenario 4 (comparison of biologics with other systemic therapies (patients with any baseline DLQI and assumption that non-responding patients are hospitalised for 21 days per year): o MTX > CYC > Funaderm > SC o MTX > CYC > Funaderm > **ETN 25 mg** > SC o MTX > CYC > Funadem > **ETN 25 mg** > **ETN 25 mg continuous** > **efalizumab** > SC o MTX > CYC > Funadem > **ETN 25 mg** > **ETN 25 mg continuous** > **efalizumab** > IFX > SC	
419 (CS, this submission was for a PAS review of TA368)	Psoriatic arthritis	Yes	First-line (after non-biologic systemic treatment)	10-year scenario analysis: 1-year, 5-year, 40-year	Markov	Base-case: o **APR** > ADA > ETN > BSC o ADA > ETN > BSC Scenario analysis 1 (alternative biologic treatment sequence): o **APR** > ADA > UST> BSC o ADA > UST > BSC Scenario analysis 2 (alternative biologic treatment sequence length, 1 biologic in the sequence): o **APR** > ADA > BSC o ADA > BSC Scenario analysis 3 (alternative biologic treatment sequence length, 3 biologics in the sequence): o **APR** > ADA > ETN > UST > BSC o ADA > ETN > UST > BSC Scenario analysis 4 (changing the position of apremilast, pre-biologic versus post-biologic): o **APR** > ADA > ETN > BSC	Model details unavailable (original TA 368 is no longer available)

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						o ADA > ETN > **APR** > BSC ERG's scenario analyses (each treatment versus BSC): o **APR** > BSC versus BSC o ADA > BSC versus BSC o ETN > BSC versus BSC o UST > BSC versus BSC	
433 (CS, PAS review of TA 372)	Psoriatic arthritis	Yes	First-line (after non-biologic systemic treatment)	40-year scenario analysis: 1-year, 5-year, 10-year	Markov	Base-case (APR strategy versus ADA strategy): o **APR** > ADA > ETN > BSC o ADA > ETN > GOL > BSC Scenario analysis 1 (APR versus ADA): o **APR** > BSC o ADA > BSC Scenario analysis 2 (comparison of 4-treatment strategies): o **APR** > ADA > ETN > GOL > BSC o ADA > ETN > GOL > IFX > BSC Scenario analysis 3 (Pair-wise comparisons of single-treatment strategy): o **APR** > BSC versus ADA > BSC o **APR** > BSC versus ETN > BSC o **APR** > BSC versus GOL > BSC o **APR** > BSC versus IFX > BSC Scenario analysis 4 (fully incremental analysis of single-treatment strategies): o **APR** o ADA o ETN o GOL o IFX o BSC Scenario analysis 5 (fully incremental analysis of 3-treatment strategies) (not an exhaustive list due to too many plausible treatment sequences: o **APR** > ADA > UST o ADA > UST > ETN o **APR** > ADA > ETN o ADA > UST > GOL o **APR** > ADA > GOL o **APR** > ETN > UST o ADA > ETN > UST o **APR** > ETN > ADA o **APR** > GOL > UST o ADA > GOL > UST o ETN > UST > ADA o ETN > ADA > UST o GOL > UST > ADA o GOL > ADA > UST o ADA > ETN > GOL o ADA > GOL > ETN o **APR** > ETN > GOL o ETN > UST > GOL	Model details unavailable (original TA 372 is no longer available)

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						o **APR** > GOL > ETN o GOL > UST > ETN o ETN > ADA > GOL o GOL > ADA > ETN o ETN > GOL > UST o GOL > ETN > UST o ETN > GOL > ADA o GOL > ETN > ADA o ADA > UST > IFX	
442 **(CS)**	Psoriatic arthritis	Yes	First-line (after non-biologic systemic treatment)	Life-time (45 year) scenario analysis: 10-year	Markov	Base-case: o **IXE** > UST 90 mg > IFX > BSC o ADA > UST 90 mg > IFX > BSC o ETN 50 mg > UST 90 mg > IFX > BSC o IFX > UST 90 mg > ADA > BSC o SEC > UST 90 mg > IFX > BSC o UST 90 mg > ADA > IFX > BSC o UST 45 mg > ADA > IFX > BSC Scenario analysis 1: o ADA > **IXE** > IFX > BSC o ADA > SEC > IFX > BSC o ADA > UST 45 mg > IFX > BSC o ADA > UST 90 mg > IFX > BSC Scenario analysis 2 o ETN 50 mg o ADA o UST 45 mg o **IXE Q2W** o UST 90 mg o IFX o SEC 300 mg Scenario analysis 3 o MTX o CYC o BSC o **IXE Q2W**	[First-line treatment (trial period month 1 > trial period month 2 > trial period month 3 > maintenance)]x > [Second-line treatment (trial period month 1 > trial period month 2 > trial period month 3 > maintenance)] > [Third-line treatment (trial period month 1 > trial period month 2 > trial period month 3 > maintenance)] > BSC > death Adapted from York model
445 **(AG)** **(The AG summary of CS were only extracted for UCB submission, but not Novartis (because Novartis submission did not comparing treatment sequences))**	Psoriatic arthritis	Yes	First-line in the model	Life-time (40 years)	Markov (York model: TA199)	Biologics-naïve (one prior cDMARD) o **SEC** > ETN > UST > BSC o **CZP** > ETN > UST > BSC o BSC Biologics-naïve (≥ 2 prior DMARDs) o **SEC** > UST > BSC o **CZP** > UST > BSC o ADA > UST > BSC o ETN > UST > BSC o IFX > UST > BSC o GOL > UST > BSC o BSC Biologic experienced or contraindicated o **SEC 300 mg** > BSC o UST > BSC BSC	[First-line treatment > (yes/no PASI 75 response)] > [Second-line treatment > (yes/no PSI 75 response)] > [Third-line treatment > (yes/no PSI 75 response)] > death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
445 (Extracted from AG summary of CS, UCB)	Psoriatic arthritis	Yes	First-line in the model	Life-time (50 years)	Markov model	Biologics-naïve (one prior cDMARD) o **CZP** >TNF> UST > Mix treatments o cDMARD >TNF> UST > Mix treatments Biologics-naïve (≥ 2 prior DMARDs) o **CZP** >UST>Mix treatments o ETN > UST > Mix treatments o IFX > UST > Mix treatments o ADA > UST > Mix treatments o GOL > UST > Mix treatments o SEC 150 mg Biologic-experienced o **CZP** > Mix treatments o SEC 300 mg > Mix treatments o UST > Mix treatments o Mixed treatments* *Mixed treatments: MTX, other cDMARDs, PAL	[First-line treatment > (yes/no PASI 75 response)] > [Second-line treatment > (yes/no PSI 75 response)] > [Third-line treatment > (yes/no PSI 75 response)] > death
475 (CS)	Plaque psoriasis	Yes	First-line treatment of patients who are eligible for systematic biologic interventions (later line in the whole treatment pathway)	10-year Scenario analysis: 20-year, life-time	Markov model	o **DMF** → ADA → UST → BSC o ADA → UST → BSC	Induction period (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death
511 (CS)	Plaque psoriasis	Yes	First-line treatment of patients who are eligible for systematic biologic interventions (later line in the whole treatment pathway)	Life-time (40 years) Scenario analysis: 10-year	Markov model	o **Brodalumab** > UST > SEC > BSC o ADA > UST > SEC > BSC o APR > UST > SEC > BSC o DMF > UST > SEC > BSC o ETN > UST > SEC > BSC o Infliximab > UST > SEC > BSC o IXE > UST > SEC > BSC o SEC > UST > ADA > BSC o UST > ADA > SEC > BSC	Induction period (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death
521 (CS)	Plaque psoriasis	No (cost-comparison analysis)	o Unspecified earlier lines of plaque psoriasis treatment o First-line in the cost-comparison analysis with treatment sequences	5-year Sensitivity analysis: 1-10 year	X	o **GUS**>ADA>ADA(50%)/UST(50%)>IFX o ADA>UST>IFX o UST>ADA>IFX	X (no economic model) The company performed scenario cost-comparison analysis with in different treatment sequences.
537 (CS)	Plaque psoriasis	Yes	o First-line treatment for patients who have experienced more than two lines of dDMARD treatments o Second-line TNFi in patient who are TNFi experienced	Life-time (40 years)	Markov model	Biologic-naïve: no psoriasis and mild-to-moderate psoriasis o **IXE Q4W** > UST > SEC > BSC o ADA > UST > SEC > BSC o APR > UST > SEC > BSC o CZP > UST > SEC > BSC o ETN > UST > SEC > BSC o Golimumab > UST > SEC > BSC o IFX > UST > SEC > BSC o SEC > UST > ADA > BSC Biologic-naïve: moderate to severe psoriasis o Same treatment sequences as those in "biologic-naïve: no psoriasis and moderate to severe" population. Changing	Induction period* (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death Trial period*

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						sequence containing **IXE Q4W** to **IXE Q2W+Q4W**, and SEC 300 150 mg to 300 mg <u>bMARD-experienced: no psoriasis and mild-to-moderate psoriasis</u> o **IXE** > BSC o UST > BSC o CZP > BSC o SEC 150 mg> BSC <u>bMARD-experienced: moderate to severe psoriasis</u> o **IXE** > BSC o UST > BSC o CZP > BSC o SEC 300 mg > BSC	
543 (CS)	Psoriatic arthritis	Yes	o First-line treatment for patients who have experienced more than two lines of cDMARD treatments o Second-line TNFi in patient who are TNFi experienced	Life-time (40 years)	Markov model	<u>Disease has not responded to at least 2 non-bDMARDs</u> o **TOF** > UST > BSC o APR > UST > BSC o CZP > UST > BSC o ETN > UST > BSC o GOL > UST > BSC o IFX > UST > BSC o SEC 188 mg > UST > BSC o BSC <u>Disease has not responded to non-bDMARDs and at least 1 TNFi</u> o **TOF** > BSC o SEC 300 mg > BSC o UST > BSC o BSC <u>1 TNFi contraindicated or not tolerated</u> o **TOF** > BSC o SEC 188 mg > BSC o UST > BSC o BSC	Induction period (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death
574 (CS)	Plaque psoriasis	Yes	First-line treatment of patients who are eligible for systematic biologic interventions (later line in the whole treatment pathway)	Life-time Scenario analysis: 10-year, 20-year, 1-year (trial)	Markov model	<u>Non-biologic therapy IR</u> o **CZP** > UST 90 mg > IFX > BSC > BSC o ADA > UST 90 mg > IFX > BSC > BSC o BROD > UST 90 mg > IFX > BSC > BSC o ETN > UST 90 mg > IFX > BSC > BSC o GUS > UST 90 mg > IFX > BSC > BSC o IXE > UST 90 mg > IFX > BSC > BSC o SEC > UST 90 mg > IFX > BSC > BSC o UST 45 mg > ADA > IFX > BSC > BSC o UST 90 mg > ADA > IFX > BSC > BSC	Induction period (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death
575 (CS)	Plaque psoriasis	Yes	First-line treatment of patients who are eligible for systematic biologic interventions (later line in the whole treatment pathway)	Life-time	Markov	o **Tildrakizumab** > UST > SEC > BSC o ADA> UST > SEC > BSC o UST > ADA > SEC > BSC o SEC> UST > ADA > BSC o ENT> UST > SEC > BSC o IXE> UST > SEC > BSC o BROD> UST > SEC > BSC o GUS> UST > SEC > BSC	Induction period (1st treatment) > maintenance period (1st treatment) > Induction period (2nd treatment) > maintenance period (2nd treatment) > Induction period (3rd treatment) > maintenance period (3rd treatment) > BSC > death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
195 (AG)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR at least one TNFi	Life-time	Individual sampling model (BRAM)	Base-case: o **ADA** > LEF > gold > CYC > AZA > PAL o **ETN** > LEF > gold > CYC > AZA > PAL o **IFX** > LEF > gold > CYC > AZA > PAL o **RTX** > LEF > gold > CYC > AZA > PAL o **ABA** > LEF > gold > CYC > AZA > PAL o LEF > gold > CYC > AZA > PAL Scenario analysis: o **ADA** o **ETN** o **INX** o **RTX** o **ABA** o DMARDs Pair-wise comparisons: o The AG conducted a series of pair-wise comparison of ICER between paired two treatment sequences. An exhaustive list is provided in the AG report.	[Start first-line treatment: on treatment (HAQ increase) > Quit treatment > Select next treatment] > [Start second-line treatment: on treatment (HAQ increase) > Quit treatment > Select next treatment] > [Start nth-line treatment: on treatment (HAQ increase) > Quit treatment > Select next treatment] > death Adapted from Birmingham Rheumatoid Arthritis Model (BRAM).
195 (Extracted from AG summary of CS: Abbot)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR to MTX, SSZ, HU and one TNFi	Life-time	DES	o Gold > LEF > CYC > rescue o **ADA**/ETN > LEF > CYC > rescue o IFX > Gold > LEF > CYC > IFX > Gold > LEF > CYC > rescue o ABA > Gold > LEF > CYC > IFX > Gold > LEF > CYC > rescue o **ADA**/ETN > RTX > gold > LEF > CYC > rescue	Model details unavailable
195 (Extracted from AG summary of CS: Wyeth)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR to ETN	Life-time	Markov	o **ETN**/IFX/ADA > DMARDs > salvage therapy o DMARDs > DMARDs > salvage therapy o RTX > DMARDs > salvage therapy	Model details unavailable
195 (Extracted from AG summary of CS: Schering-Plough)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR to two non-bDMARDs and one TNFi	Life-time	Patient-simulation	o ADA > DMARDs o ETN > DMARDs o **IFX** > DMARDs o ABA > DMARDs o RTX > DMARDs o ADA > RTX > DMARDs o ETN > RTX > DMARDs o **IFX** > RTX > DMARDs o DMARDs	Model details unavailable
195 (Extracted from AG summary of CS: Roche)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR a TNFi	Life-time	Patient-level simulation	o **RTX** > LEF > gold > CYC > PAL o ETN > LEF > gold > CYC > PAL o ADA > LEF > gold > CYC > PAL o IFX > LEF > gold > CYC > PAL o ABT > LEF > gold > CYC > PAL o LEF > gold > CYC > PAL	Model details unavailable
195 (AG summary of CS: BMS)	Rheumatoid arthritis	Yes	First-line treatment of patients who had an IR at least one TNFi	Life-time	Patient-level simulation	o **ABA** > IFX > LEF > gold > AZA > CYC > penicillamine > PAL o RTX > IFX > LEF > gold > AZA > CYC penicillamine > PAL o **ABA** > TNFi > LEF > gold > AZA > CYC > penicillamine > PAL o TNFi > TNFi > LEF > gold > AZA > CYC > penicillamine > PAL	Model details unavailable

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
225 (CS)	Rheumatoid arthritis	Yes	First-line treatment of patients who had progressed after cDMARDs	Life-time (45 years)	Markov	DMARD experienced o (MTX > SSZ + MTX >)* **GOL** + MTX > LEF > gold > AZA > CYC > PAL > PAL o (MTX > SSZ + MTX >)* TNFi + MTX > LEF > gold > AZA > CYC > PAL > PAL o (MTX > SSZ + MTX >)* MTX > LEF > gold > AZA > CYC > PAL > PAL TNFi experienced o (MTX > SSZ + MTX > TNFi)* > **GOL** + MTX > LEF > gold > AZA > CYC > PAL > PAL o (MTX > SSZ + MTX > TNFi)* > RTX + MTX > LEF > gold > AZA > CYC > PAL > PAL o (MTX > SSZ + MTX > TNFi)* > MTX > LEF > gold > AZA > CYC > PAL > PAL TNFi experienced – company sensitivity analysis o (MTX > SSZ + MTX > TNFi)* > **GOL** > RTX > LEF > gold > AZA > CYC > PAL > PAL o (MTX > SSZ + MTX > TNFi)* > MTX > RTX > LEF > gold > AZA > CYC > PAL > PAL * Previous treatment lines were only listed to demonstrate treatment history but not included as part of the treatment sequences in the model. Reference TNFis included in the model: ADA, IFX, ETN, RTX, CTZ	[Start first-line treatment: on treatment (baseline HAQ > ACR 20/50/no response > ACR 20/50/no response reassign)] > [Start second-line treatment: on treatment (baseline HAQ > ACR 20/50/no response > ACR 20/50/no response reassign)] > [Start nth-line treatment: on treatment (baseline HAQ > ACR 20/50/no response > ACR 20/50/no response reassign)] > death Adapted reference treatment sequence from the BRAM model.
375 (AG summary of CS, Abbvie)	Rheumatoid arthritis	Yes	o Population 1: adults with severe active RA not previously treated with cDMARDs o Population 2: adults with severe active RA that have been previously treated with cDMARDs but not bDMARDs; o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MTX.	Life-time	Individual patient simulation (DES)	Analysis 1 & 2: Population 3 & 2 in combination with MTX o LEF > SSZ > CYC > rescue o **ADA** + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o ETN + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o IFX + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o CZP + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o GOL + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o ABA + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o TCZ + MTX > > RTX + MTX > LEF > SSZ > CYC > rescue Analysis 3: Population 1 in combination with MTX o MTX > LEF > SSZ > CYC > rescue o **ADA** + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o ETN + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o IFX + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o GOL + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue o MTX + HCQ > **ADA** + MTX > RTX + MTX > TCZ + MTX > LEF > SSZ > CYC > rescue	Start sequence > initial response to treatment (ACR 50) > Time with treatment response & stay on treatment > switch to next treatment sequence

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						Analysis 4 & 5: Population 3 & 2, monotherapy o SSZ + HCQ > LEF > SSZ > CYC > rescue o **ADA** > LEF > SSZ > CYC > rescue o ETN > LEF > SSZ > CYC > rescue o CZP > LEF > SSZ > CYC > rescue o TCZ > LEF > SSZ > CYC > rescue Analysis 6: Population 1, monotherapy o SSZ + HCQ > LEF > SSZ > CYC > rescue o **ADA** > LEF > SSZ > CYC > rescue o ETN > LEF > SSZ > CYC > rescue o SSZ + HCQ > **ADA** > LEF > SSZ > CYC > rescue	
375 (AG summary of CS, BMS)	Rheumatoid arthritis	Yes	o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MTX. o General RA population who can receive MTX	Life-time	Individual patient model	Population 3 in combination with MTX and General RA Population who can receive MTX o LEF > gold > CYC > AZA > PAL o **ABA** SC + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o ADA + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o CZP + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o ETN + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o GOL + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o IFX + MTX> RTX + MTX > TCZ + MTX* > LEF > gold > CYC > AZA > PAL o TCZ + MTX > RTX + MTX > LEF > gold > CYC > AZA > PAL * TCZ + MTX would not be used if there was a DAS28 improvement of 1.2 or greater at six months	[Start first-line biologics (discontinue due to AE> DAS28 improves by at least 1.2?)] > [Start nth-line therapy (discontinue due to AE? > DAS28 improves by at least 1.2?)] > death The model was adapted from BRAM.
375 (AG summary of CS, MSD; MSD prepared their submission for two drugs: GOL, IFX)	Rheumatoid arthritis	Yes	o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MTX o General RA population who can receive MTX	Life-time (45 years) Sensitivity analysis: shorter timeframes	Markov	Population 3 in combination with MTX and General RA Population who can receive MTX o MTX > SSZ + MTX > **IFX** + MTX > RTX > LEF > gold > AZA > CYC > PAL o MTX > SSZ + MTX > **GOL** + MTX > RTX > LEF > gold > AZA > CYC > PAL o MTX > SSZ + MTX > other bDMARDs + MTX > RTX > LEF > gold > AZA > CYC > PAL o MTX > SSZ + MTX > MTX > RTX > LEF > gold > AZA > CYC > PAL	[Start first-line treatment: on treatment (baseline HAQ > ACR 20/50/70/no response > ACR 20/50/70/no response reassign)] > [Start second-line treatment: on treatment (baseline HAQ > ACR 20/50/70/no response > ACR 20/50/no response reassign)] > [Start nth-line treatment: on treatment (baseline HAQ > ACR 20/50/70/no response > ACR 20/50/70/no response reassign)] > death
375 (AG summary of CS, Pfizer)	Rheumatoid arthritis	Yes	o Population 1: adults with severe active RA not previously treated with cDMARDs o Population 2: adults with severe active RA that have been	Life-time Sensitivity analysis: shorter timeframes	DES	Population 1 & 2 & 3 in combination with MTX o **ETN** > RTX > TCZ > SSZ > LEF > PAL o ABA IV > RTX > TCZ > SSZ > LEF > PAL o ABA SC > RTX > TCZ > SSZ > LEF > PAL o CZP > RTX > TCZ > SSZ > LEF > PAL o ADA > RTX > TCZ > SSZ > LEF > PAL o IFX > RTX > TCZ > SSZ > LEF > PAL o TCZ > RTX > ABA IV> SSZ > LEF > PAL o GOL> RTX > TCZ > SSZ > LEF > PAL o cDMARD> RTX > TCZ > SSZ > LEF > PAL o Combo DMARD> RTX > TCZ > SSZ > LEF > PAL	[Start first-line treatment: on treatment (baseline in HAQ/no response)] > [Start second-line treatment: on treatment (baseline in HAQ/no response)] > [Start nth-line treatment: on treatment (baseline in HAQ/no response)]> death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
			previously treated with cDMARDs but not bDMARDs o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MT			Population 3 monotherapy o **ETN** > ADA > SSZ > LEF > PAL o ADA > **ETN** > SSZ > LEF > PAL o TCZ > **ETN** > SSZ > LEF > PAL o TCZ > ADA > SSZ > LEF > PAL o cDMARD > **ETN** > SSZ > LEF > PAL	
375 (AG summary of CS, Roche)	Rheumatoid arthritis	Yes	MTX intolerant or contraindicated RA population	Life-time	Individual patient model	o **TCZ** > CZP > ETN > ADA > PAL o CZP > ETN > ADA >PAL	Details unavailable
375 (AG summary of CS, UCB)	Rheumatoid arthritis	Yes	o Population 2: adults with severe active RA that have been previously treated with cDMARDs but not bDMARDs o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MTX.	Life-time (45 years) Sensitivity analysis: shorter timeframes	Markov	Population 3 in combination with MTX o **CZP** + MTX > follow-up intervention 1 o ADA + MTX > follow-up intervention 1 o ETN + MTX > follow-up intervention 1 o GOL + MTX > follow-up intervention 1 o TOZ + MTX > follow-up intervention 1 o IFX+ MTX > follow-up intervention 1 o ABA + MTX > follow-up intervention 1 Population 3 monotherapy o **CZP** > follow-up intervention 1 o ADA > follow-up intervention 1 o ETN > follow-up intervention 1 o TOZ > follow-up intervention 1 Population 2 in combination with MTX o **CZP** + MTX > follow-up intervention 1 o Placebo + MTX > follow-up intervention 1 Population 2 monotherapy o **CZP** + cDMARDs > follow-up intervention 1 o Placebo + cDMARDs > follow-up intervention 1 Follow-up intervention 1: RTX+MTX > (AZA, CYC, gold, HCQ, LEF, penicillamine, PAL) Follow-up intervention 2: MTX + SSZ, MTX + SSZ + HCQ, MTX + HCQ, MTX + LEF, SSZ + HCQ + CYC, penicillamine, PAL)	[Start first-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > [Start second-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > [Start nth-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > death
375 (AG)	Rheumatoid arthritis	Yes	o Population 1: adults with severe active RA not previously treated with cDMARDs o Population 2: adults with severe active RA that have been previously	Life-time	Individual-patient DES	Population 1 (for patients who could receive MTX) o MTX > intensive cDMARDs > MTX > non-biologic therapy o MTX > intensive cDMARDs > **bDMARDs*(excluding TCZ)** +MTX > RTX+MTX > TCZ + MTX > MTX > non-biologic therapy o MTX > intensive cDMARDs > **TCZ** +MTX > RTX+MTX > TCZ + MTX > MTX > non-biologic therapy	[Start first-line treatment: on treatment (good/moderate/no HAQ response) > continue treatment] >[Start second-line treatment: on treatment (good/moderate/no HAQ response) > continue treatment] > [Start nth-line treatment: on treatment (good/moderate/no HAQ response) > continue treatment] > death The ScHARR model.

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
			treated with cDMARDs but not bDMARDs o Population 3: adults with moderate to severe active RA that have been previously treated with cDMARDs only, including MTX			o **bDMARDs* (excluding ABA, CZP and TCZ)** + MTX > RTX + MTX > **TCZ** +MTX > MTX > non-biologic therapy <u>Population 1 (for patients who could not receive MTX)</u> o intensive cDMARDs > cDMARDs > intensive cDMARDs o intensive cDMARDs > **bDMARDs** > **bDMARDs (excluding TCZ)** > non-biologic therapy o **bDMARDs (excluding ABA, CZP and TCZ)** > **bDMARDs (excluding TCZ)** > non-biologic therapy <u>Population 2 & 3 (for patients who could receive MTX)</u> o MTX > non-biologic therapy o **ABA** IV + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **ABA** SC + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **ADA** + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **CZP** + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **ETN** + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **GOL** + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **IFX** + MTX > RTX + MTX > TCZ + MTX > MTX > non-biologic therapy o **TCZ** + MTX > RTX + MTX > MTX > non-biologic therapy <u>Population 2 & 3 (for patients who could not receive MTX)</u> o cDMARDs > non-biologic therapy o **bDMARDs** > **bDMARDs (excluding TCZ)** > cDMARDs > non-biologic therapy *** bDMARDs:** ABA, ADA, CZP, ETN, GOL, IFX, TCZ. The AG tested a series of treatment sequences combination for the broad sequence of bDMARDs > bDMARDs. This MTA did not have the intension to identify the most cost-effective treatment sequence.	
415 (CS)	Rheumatoid arthritis	Yes	First-line of second-line in TNFi-IR population	45 years Scenario analysis: 5 and 10 years	Markov	<u>Main comparator: RTX + MTX:</u> o **CZP** + MTX > RTX + MTX > TCZ SC + MTX > ABA + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL o RTX + MTX > TCZ SC + MTX > ABA + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL <u>Main comparator: other biologics in combination with MTX, for whom RTX is contraindicated or withdrawn</u> o **CZP** + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL o ABA + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL o ADA + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL	[Start first-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > [Start second-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > [Start nth-line treatment: on treatment (ACR 20/50/70 change or no response) > follow-up treatment] > death

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						○ ETN + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL ○ GOL + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL ○ IFX + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL ○ TCZ + MTX > MTX + HCQ + SSZ > LEF > gold > CYC > AZA > PAL Main comparator: other biologics monotherapy, for whom RTX is contraindicated or withdrawn ○ **CZP** > LEF > gold > CYC > AZA > PAL ○ ADA > LEF > gold > CYC > AZA > PAL ○ ETN > LEF > gold > CYC > AZA > PAL ○ TCZ > LEF > gold > CYC > AZA > PAL ERG exploratory analysis (main comparator: RTX + MTX): ○ **CZP** + MTX > RTX + MTX > TCZ SC + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL ○ RTX + MTX > **CZP** + MTX > TCZ SC + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL ○ **CZP** + MTX > TCZ + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL ○ **RTX** + MTX > TCZ + MTX > MTX + HCQ + SSC > non-biologic therapy > PAL	
466 (CS)	Rheumatoid arthritis	Yes	First-line treatment in each treatment sequence among patients with different treatment histories (i.e. had different lines of cDMARD and bDMARDs)	Life-time (45 years) Scenario analysis: 15-year, life-time	DES (based on AG model in TA375)	cDMARD-IR moderate population ○ **Baricitinib** > combination of cDMARDs > MTX > PAL ○ combination of cDMARDs > MTX > palliative care > None cDMARD-IR severe population ○ **Baricitinib** > RTX+MTX > TCZ+MTX > MTX > PAL ○ BDMARDs+MTX > RTX+MTX> TCZ+MTX > MTX > PAL ○ TCZ+MTX > RTX+MTX > ADA+MTX > MTX > PAL bDMARD-IR severe population (RTX eligible) ○ **Baricitinib** > TCZ+MTX > MTX > PAL ○ RTX+MTX > TCZ+MTX > MTX > PAL bDMARD-IR severe population (RTX ineligible) ○ **Baricitinib** > TCZ+MTX > MTX > PAL ○ bDMARD+MTX> TCZ+MTX > MTX > PAL ○ TCZ+MTX > ADA+MTX > MTX > PAL	Initial line treatment > Subsequent line treatment > 3rd line treatment > 4th line treatment > 5th line treatment > dead Total lines of treatment in the treatment sequence is depending on the subgroup specified in the TA (different scope of decision problem). Palliative can be a line of treatment
485 (CS)	Rheumatoid arthritis	Yes	First-line treatment in each treatment sequence among patients with different treatment histories (i.e. had different lines of cDMARD and bDMARDs)	Life-time (100 years) Scenario analysis: 10-year	Markov model	cDMARD+MTX IR: severe RA ○ **Sarilumab**+MTX > RTX+MTX > ABA IV + MTX > BSC ○ TCZ IV+MTX > RTX+MTX > ABA IV + MTX > BSC ○ TCZ SC+MTX > RTX+MTX > ABA IV + MTX > BSC ○ TNFi bundle> RTX+MTX > ABA IV + MTX > BSC ○ ABA IV+MTX > RTX+MTX > TCZ IV+MTX > BSC cDMARD+MTX IR: moderate RA or intoleratant of cDMARD+MTX ○ **Sarilumab** + MTX > BSC ○ BSC	Initial line treatment > Subsequent line treatment > 3rd line treatment > 4th line treatment > 5th line treatment > palliative care > dead Total lines of treatment in the treatment sequence is depending on the subgroup specified in the TA (different scope of decision problem)

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						cDMARD monotherapy-IR severe RA o **Sarilumab** > TNFi bundle > BSC o TCZ IV+MTX > TNFi bundle > BSC o TCZ SC+MTX > TNFi bundle > BSC o TNFi bundle > TNFi bundle > BSC cDMARD+MTX-IR (at least one TNFi, RTX+MTX intolerant): severe RA o **Sarilumab**+MTX > ABA IV + MTX > BSC o TCZ IV+MTX > ABA IV + MTX > BSC o TCZ SC+MTX > ABA IV + MTX > BSC o TNFi bundle > ABA IV + MTX > BSC o ABA IV+MTX > TCZ IV+MTX > BSC cDMARD+MTX-IR (at least one TNFi, RTX+MTX tolerant): severe RA o **Sarilumab** + MTX > ABA IV > BSC o RTX+MTX> ABA IV > BSC Severe RA: despite treatment with bMARDs o **Sarilumab** > BSC o TCZ IV+MTX > BSC o TCZ SC+MTX > BSC o TNFi bundle > BSC	
547 (CS)	Ulcerative colitis	Yes	o First-line treatment of patients who are eligible for systematic biologic interventions (later line in the whole treatment pathway) o Second-line treatment in those who had prior biologic exposure	Life-time	Markov model	Biologic-naïve o ADA > GOL > IFX > **TOF** > VED > ConT o ConT > ConT > ConT > ConT > ConT Biologic-prior-exposure o **TOF** > VED > ConT o ConT > ConT	Active UC (1st treatment) > remission (responder only, 1st treatment) > active UC (2nd treatment) > remission (responder only, 2nd treatment)> post-surgery (with our without complication) > dead
Cardiovascular							
249 (CS)	Prevention of stroke and systemic embolism in AF	Yes	o First-line or second-line treatment	Life-time Scenario analysis: 2, 10 and 15 years	Markov	Dabigatran as first-line setting: o **Dabigatran*** > aspirin plus clopidogrel, aspirin, or no treatment o Warfarin > aspirin plus clopidogrel, aspirin, or no treatment o Aspirin plus clopidogrel > aspirin or no treatment o Aspirin > no treatment Dabigatran as second-line setting o Warfarin > **dabigatran*** o Warfarin > aspirin o Warfarin > no treatment * Three different dosing regimen of dabigatran:	[On first-line treatment: Disability level: independent > disability level: moderate > disability level: dependent] > [On second-line treatment: Disability level: independent > disability level: moderate > disability level: dependent] > [Events: ischemic stroke/intracranial hemorrhage/other bleeds/other thrombotic events/non-clinical discontinuation] > [Off treatment: Disability level: independent > disability level: moderate > disability level: dependent]

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						○ dabigatran 150 mg BID for all eligible patients ○ dabigatran 110 mg BID for all eligible patients ○ dabigatran sequence: for patients less than 80 years old, started at 150 mg and switched to 110 mg at age 80; for patients at least 80 years old, started at 110 mg at baseline	
Neurology/mental health							
292 (CS)	Bipolar disorder	Yes	○ First-, second-, or third-line treatment	3 years whereby patient enter the model at aged 15 until adulthood	Markov	Base-case: ○ Strategy 1 (base-case): risperidone > quetiapine > olanzapine ○ Strategy 2: Risperidone > **aripiprazole** > quetiapine ○ Strategy 3: **Aripiprazole** > risperidone > quetiapine ○ Strategy 4: Risperidone > quetiapine > **aripiprazole** Scenario analysis: ○ Strategy 1 (base-case): risperidone > olanzapine > quetiapine ○ Strategy 2: Risperidone > **aripiprazole** > olanzapine ○ Strategy 3: **Aripiprazole** > risperidone > olanzapine ○ Strategy 4: Risperidone > olanzapine > **aripiprazole**	[First-line treatment : Acute phase week 1 > Acute phase week 2 > Acute phase week 3 > Sub-acute phase responder observation (5 cycles) > Maintenance phase: euthymic treated (4 cycles on average) > Maintenance phase: euthymic not treated)] > [(Second-line treatment : Acute phase week 1 > Acute phase week 2 > Acute phase week 3 > Sub-acute phase responder observation (5 cycles) > Maintenance phase: euthymic treated (4 cycles on average) > Maintenance phase: euthymic not treated)] > [(Third-line treatment : Acute phase week 1 > Acute phase week 2 > Acute phase week 3 > Sub-acute phase responder observation (5 cycles) > Maintenance phase: euthymic treated (4 cycles on average) > Maintenance phase: euthymic not treated)] > [Therapy resistance phase: Therapy resistance hospitalisation (5 cycles) > therapy resistance outpatient > euthymic treated > euthymic not treated] > death
Infectious disease							
154 (CS)	Chronic hepatitis B	Yes (sequence of drug treatments was not explicitly modelled, but subsequent transplantation)	○ First-line or salvage therapy in patients who are indicated for anti-viral therapy	Life-time (100 years) Scenario analysis: 70 years, 47 years	Transition state model	Seroconversion model (HBeAg-positive patients), in in with previous NICE assessment (patients who lost the e antigen were given further 6 months of treatment): ○ BSC > BSC > BSC ○ Lamivudine > BSC > BSC ○ **Telbivudine** > BSC > BSC ○ Adefovir > BSC > BSC ○ Lamivudine > Adefovir > BSC ○ **Telbivudine** > Adefovir > BSC ○ Adefovir > **Telbivudine** > BSC ○ Adefovir > Lamivudine > BSC ○ BSC (base-case) Viral load model (HBeAg positive and negative patients) (patients remain on treatment indefinitely unless they seroconvert or move to a different health state): ○ **Telbivudine** ○ Lamivudine	Seroconversion model Inactive carrier > cured or chronic hepatitis B > compensated cirrhosis > decompensated cirrhosis > hepatocellular carcinoma or [liver transplant year 1 > liver transplant year 2+] > chronic hepatitis B related death Viral load model, e antigen (HBeAg)-positive Compensated cirrhosis > chronic hepatitis decompensated cirrhosis > hepatocellular carcinoma or post liver transplant > death (patients may have positive or negative surface antigen in chronic hepatitis and compensated cirrhosis states) Viral load model, e antigen (HBeAg)-negative* Initial state > decompensated cirrhosis > hepatocellular carcinoma > death *Within each state, patients are subdivided into their viral load levels, serological marker status (both HBsAg and HBeAg positive or negative, or positive and negative HBsAg), and whether they have liver transplant or whether they are resistant to the drug.
173 (CS)	Chronic hepatitis B	Yes (sequence of drug treatments was not explicitly modelled, but subsequent transplantation)	○ First-line or salvage therapy in patients who are indicated for anti-viral therapy	Life-time	Markov model (time-dependent transition probabilities)	A total of 211 treatment sequences were compared. Twenty of them are listed: ○ BSC ○ Lamivudine > **tenofovir** ○ **Tenofovir** > lamivudine ○ **Tenofovir** > tenofovir + lamivudine ○ **Tenofovir** > **tenofovir** + lamivudine > entacavir	Half part of the model: Patients started with HBeAg+ HBeAg+ viral suppression (viral load < 300c/ml) > [[HBeAg seroconverted > HBsAg seroconverted] or HBeAg+ CHB (viral load > 300c/ml) > [HBeAg+ compensated cirrhosis (viral load > 300c/ml) or HBeAg+ compensated cirrhosis with detectable HBV

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
					Scenario analysis: 5-60 years	○ Lamivudine > BSC ○ Lamivudine > entacavir ○ Lamivudine > adefovir ○ Adefovir > lamivudine ○ Lamivudine > **tenofovir** + lamivudine ○ **Tenofovir** > BSC ○ **Tenofovir** > entacavir ○ Lamivudine > adenovir + lamivudine ○ Adefovir > **tenofovir** ○ Adefovir > **tenofovir** + lamivudine ○ Adefovir > **adefovir** + lamivudine ○ Entacavir > lamivudine ○ Entacavir > **tenofovir** ○ Adefovir + lamivudube > **tenofovir** + lamivudine	DNA] > HBeAg+ liver transplant > HBeAg+ post-liver transplant <u>Half part of the model: Patients started with HBeAg-</u> HBeAg- viral suppression (viral load < 300c/ml) > [[HBeAg seroconverted > HBsAg seroconverted] or HBeAg- CHB (viral load > 300c/ml) > [HBeAg+ compensated cirrhosis (viral load > 300c/ml) or HBeAg- compensated cirrhosis with detectable HBV DNA] > HBeAg- liver transplant > HBeAg- post-liver transplant Death can happen at any time while HBeAg+ HCC can happen in patients with positive HBeAg. Patients with HBeAg+ HCC can be eligible for liver transplant. It was assumed that patients can only develop HBeAg-negative chronic hepatitis B from HBeAg seroconverted disease state. A decision-tree model was used to model treatment resistance on an annual basis at the background of the Markov model. Within the first year, x proportion of patients might develop drug resistance to a treatment. Of those who did not develop drug resistance (1-x%), transition probability of the first-line treatment t will apply for the whole year. On the other hand, x% of patients will spend 1.5 months on the first-line treatment with transition probabilities of untreated patients (due to drug resistance), and are assumed to receive th a second-line treatment with corresponding transition probabilities for the rest of the year. Treatment strategies in the following years follows the same strategy. Patients receive the same treatment until they develop drug resistance and switch to the next-line of treatment.
Diabetes Mellitus							
418 (CS)	Diabetes Mellitus	Yes	○ First-line as a part of add-on therapy (patients that are inadequately controlled on dual therapy with either MET and a SU or MET and a DPP-4 inhibitor	Life-time (40 years) Scenario analysis: 5 years	DES (CARDIFF model: replicating UKPDS model of UKPDS 68 and UKPDS 82)	<u>Company base-case:</u> ○ **Dapagliflozin** + MET + SU > Insulin + MET > intensive insulin +MET ○ DPP-4 inhibitors* + MET + SU > Insulin + MET > intensive insulin +MET ○ Empagliflozin 10 mg + MET + SU > Insulin + MET > intensive insulin +MET ○ Empagliflozin 25 mg + MET + SU > Insulin + MET > intensive insulin +MET ○ Canagliflozin 100 mg + MET + SU > Insulin + MET > intensive insulin +MET ○ Canagliflozin 300 mg + MET + SU > Insulin + MET > intensive insulin +MET <u>ERG base-case:</u>	Initialize patient (set baseline demographics and risk factor profiles, such as HbA1c, SBP and weight) > [Treatment 1 > Treatment 2 > Treatment 3]*: determined by discontinuation rate > Events**including MI/stroke/CHF/IHD/Blindness/nephropathy/amputation /hypoglycemic event (symptomatic, severe, nocturnal)/CV death/Non-CV death * Treatment is intensified (moving onto next-line treatment) when patient's HbA1c is evolved n breaches the NICE intensification threshold of 7.5%. **Patients might experience Events when receiving treatments. The evolution of risk factors and patient baseline characteristics (e.g. age, gender, smoking), diabetic related complications and death can be modelled.

TA number (model source)	Disease	Treatment-sequencing model structure	Position of the new treatment in the treatment pathway	Time horizon	Model Type	Treatment sequences in the (base-case treatment sequences are listed or otherwise specified, new technology are labelled in bold)	Model health states/ model health state transitions (the longest in the model)
						○ **Dapagliflozin** + MET + SU > **Dapagliflozin** + MET + SU + insulin > **Dapagliflozin** + MET + SU + insulin + intensified insulin ○ DPP-4 inhibitors + MET + SU > DPP-4 inhibitors + MET + SU + insulin > DPP-4 inhibitors + MET + SU + insulin + intensified insulin ○ Empagliflozin 10 mg + MET + SU > Empagliflozin 10 mg + MET + SU + insulin > Empagliflozin 10 mg + MET + SU + insulin + intensified insulin ○ Empagliflozin 25 mg + MET + SU > Empagliflozin 25 mg + MET + SU + insulin > Empagliflozin 25 mg + MET + SU + insulin + intensified insulin ○ Canagliflozin 100 mg + MET + SU > Canagliflozin 100 mg + MET + SU + insulin > Canagliflozin 100 mg + MET + SU + insulin + intensified insulin ○ Canagliflozin 300 mg + MET + SU > Canagliflozin 300 mg + MET + SU + insulin > Canagliflozin 300 mg + MET + SU + insulin + intensified insulin *A basket of DPP-4 inhibitors comprising sitagliptin (71%), saxagliptin (10%), vildagliptin (3%), linagliptin (12%) and alogliptin (3%)	
Ophthalmology							
409 (CS)	Visual impairment caused by macular oedema after branch retinal vein occlusion	Yes	○ First-line or second-line	Life-time (35 years) Scenario analysis: 10 years	Markov	<u>First-line setting: comparison 1a</u> ○ **Aflibercept** > laser ○ Laser > ranibizumab <u>First-line setting: comparison 1b</u> ○ **Aflibercept** > laser ○ Laser > dexamethasone <u>First-line setting: comparison 1c</u> ○ **Aflibercept** > laser ○ Laser > **aflibercept** <u>Second-line setting: comparison 2a</u> ○ Laser > **aflibercept** ○ Laser > ranibizumab <u>Second-line setting: comparison 2b</u> ○ Laser > **aflibercept** ○ Laser > dexamethasone Patients will switch to a second-line treatment if treatment failure occurred within 0-6 of the first-lin treatment	[Efficacy phase Year 1: Initial first-line treatment (0-6 months) > continue first-line treatment or switch to second-line treatment (6-12 months)] > Maintenance phase Year 2-5: continue treatment that was used in the 6-12 months > Rest-of-life phase: off- treatment > death (can occur at any state) Maintenance phase and off-treatment phase can be divided into the following sub-states: visual acuity 1 > visual acuity 2 > visual acuity 3 > visual acuity 4 > visual acuity 5. Patients are only allowed to move to the next poorer level annually within the off-treatment phase.

ABA: abatacept, ADA: adalimumab, ADT: androgen deprivation therapy, AF: prevention of stroke and systemic, embolism in atrial fibrillation, AG: assessment group, alloSCT: allogeneic stem cell transplantation, APR: apremilast, ASCT: autologous stem cell transplant, AZA: azathioprine, BASDAI: Bath Ankylosing Spondylitis Disease Activity Index, BASFI: Bath Ankylosing Spondylitis Functional Index, BRAM: Birmingham Rheumatoid Arthritis Model, bDARMD: biologic DMARD, BID: twice a day, BMS: Bristol-Myers Squibb. BRAM: Birmingham Rheumatoid Arthritis Model, BROD: brodalumab, BSC: best supportive care, cDMARD: conventional DMARD, CHF: congestive heart failure, CHOP: cyclophosphamide, doxorubicin, vincristine, and prednisone, ConT : conventional therapy, CS: company submission, CZP: certolizumab pegol, CYC: cyclosporin, DAS28: disease activity score by 28 joints, DES: discrete event simulation, DLQI: Dermatology Life Quality Index, DMARD: disease-modifying antirheumatic drug, DMF: dimethyl fumarate, DPP-4: dipeptidyl peptidase 4, DSU: decision support unit, EFS: event-free survival, ERG: evidence review group, ETN: etanercept, GOL: golimumab, GUS: guselkumab, HbA1c: haemoglobin A1c, HBeAg : hepatitis B e-

antigen, HBsAg: hepatitis B surface antigen, HCQ: hydrixychlorine, HRPC: hormone-resistant prostate cancer, HU: hydroxyurea, IFX: infliximab, IHD: ischemic heart disease, IR: inadequate response, IV: intravenous injection, IXE: ixekizumab, LEF: leflunomide, MET: metformin, mHRPC: metastatic HRPC, MI: myocardial infarction, MTX: methotrexate, nmHRPC: non-metastatic HRPC, NSAIDs: nonsteroidal anti-inflammatory drugs, OS: overall survival, PAL: palliative care, PD: progressed disease, PFS: progression free survival, PPS: post-progression survival, Q2W: every two weeks, Q4W: every four weeks, RA: rheumatoid arthritis, R-CHOP: rituximab plus CHOP regimen, RTX: rituximab, SBP: systolic blood pressure, SC: subcutaneous injection, SCT: stem cell transplantation, SEC:
secukinumab, SSZ: sulfasalazine, SU: sulfonylurea TCZ: tocilizumab, TNF = tumour necrosis factor, TNFi = TNF inhibitor, TOF: tofacitinib, TPO-RAs: thrombopoeitin receptor agonists, UST: Ustekinumab, VED: vedolizumab

Appendix 4.1 PubMed search terms for Chapter 4 statistical methods review (March 13, 2022)

	PubMed search terms	Records
1	"treatment sequen*"[Title/Abstract] OR "therapy sequen*"[Title/Abstract] OR "optimal sequen*"[Title/Abstract] OR "optimum sequen*" [Title/Abstract] OR "sequential treatment*"[Title/Abstract] OR "sequential therap*"[Title/Abstract] OR "subsequent treatment*"[Title/Abstract] OR "subsequent therap*"[Title/Abstract] OR "treatment-switching" [Title/Abstract] OR "treatment-switching"[Title/Abstract] OR "therapy switching"[Title/Abstract] OR "therapy switching"[Title/Abstract] OR "clinical pathway*"[Title/Abstract] OR "treatment pathway*"[Title/Abstract] OR "treatment order*"[Title/Abstract] OR "optimal position*"[Title/Abstract] OR "optimal treatment position*"[Title/Abstract]	26,304
2	"dynamic treatment*"[Title/Abstract] OR "dynamic intervention*"[Title/Abstract]	476
3	"adaptive treatment*"[Title/Abstract] OR "adaptive intervention*"[Title/Abstract]	708
4	"treatment policy"[Title/Abstract] OR "treatment policies"[Title/Abstract]	2,390
5	"adapt*"[Title/Abstract] OR "dynamic*"[Title/Abstract] OR "regime*"[Title/Abstract]	1,779,500
6	#2 OR #3 OR (#4 AND #5)	1,386
7	"time-varying"[All Fields] OR "time-varying"[All Fields] OR "time-dependent"[All Fields] OR "time-dependent"[All Fields]	113,590
8	"causal*"[All Fields] OR "causality"[MeSH Terms] OR "causalit*"[All Fields]	1,047,710
9	"model*"[All Fields]	4,372,864
10	(#7 OR #8) AND #9	279,972
11	(#1 OR #6) AND #10	363

Appendix 4.2 EMBASE and MEDLINE search terms for Chapter 4 statistical methods review (March 13, 2022)

Through Ovid® EMBASE 1974 to 2022 March 1; and Ovid® MEDLINE and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations, Daily and Versions 1946 to March 11, 2022

	EMBASE/MEDLINE search terms	**Records**	
		EMBASE	**MEDLINE**
1	(treatment sequen* or therapy sequen* or optimal sequen* or optimum sequen* or sequential treatment* or sequential therap* or subsequent treatment* or subsequent therap* or treatment-switching or treatment-switching or therapy switching or therapy switching or clinical pathway* or treatment pathway* or treatment order* or optimal position* or optimal treatment position*).ti,ab,kw.	43,925	26,522
2	(dynamic treatment* or dynamic intervention*).ti,ab,kw.	588	478
3	(adaptive treatment* or adaptive intervention*).ti,ab,kw.	1,034	695
4	(treatment policy or treatment policies).ti,ab,kw.	3,148	2,381
5	(adapt* or dynamic* or regime*).ti,ab,kw.	2,093,948	1,747,204
6	2 or 3 or (4 and 5)	1,903	1,373
7	(time-varying or time-varying or time-dependent or time-dependent).af.	132,921	113,580
8	causal*.af.	173,643	144,445
9	causality.de.	4,438	18,083
10	causalit*.af.	30,520	37,505
11	8 or 9 or 10	173,643	144,445
12	model*.af.	5,541,489	4,371,055
13	(7 or 11) and 12	99,120	77,704
14	(1 or 6) and 13	249	172

Appendix 6.1 NICE technology appraisals of metastatic castration-resistant prostate cancer (mCRPC) treatments between 2006-2020

2006	2012	2014	2016	2018-2020
First-line therapy				
Docetaxel (TA 101)				
			Enzalutamide (TA377, Jan)	
			Abiratarone (TA387, Apr)	
Second- and subsequent-line therapy				
Docetaxel (TA 101)				
	Abiratarone (TA259, Jun, previously docetaxel-treated patients)			
		Enzalutaide (TA316, July, previously docetaxel-treated patients)		
			Carbazitaxel (TA391, May, docetaxel-treated)	

* The drugs appearing in the design of my Target Trial emulation studies in Chapter 7-8 are indicated by the yellow-highlighted cells in this table.

Appendix 6.2 NICE technology appraisals of metastatic renal cell carcinoma treatments between 2009-2020

2009-2012	2015	2016	2017	2018	2019	2020
First-line therapy						
Pazopanib (TA215 2011 Feb)						
Sunitinib (TA169 2009 March)						
~~Bevacizumab (not recommended, TA 178 2009 Aug)~~						
~~Sorafenib (not recommended, TA 178 2009 Aug)~~						
~~Temsirolimus (not recommended, TA 178 2009 Aug)~~						
				Tivozanib (TA512 2018 March)		
				Cabozantinib (TA 542 2018 Oct)		
					Nivolumab with ipilimumab (CDF, TA581 2019 May, intermediate- or poor-risk)	
						Avelumab with axitinib (CDF, TA645 2020 Sep)
						~~Pembrolizumab with axitinib (not recommended, TA 650 2020 Sep)~~
Second- and subsequent-line therapy						
~~Sorafenib (TA 178 2009 Aug, not recommended)~~						
~~Sunitinib (TA 178 2009 Aug, not recommended)~~						
~~Everolimus (TA219 not recommended~~, replaced by TA432 2017 Feb)						
	Axitinib (TA333 2015Feb, originally only after sunitinib)					
		Nivolumab (TA417 2016 Nov)				
			Everolimus (TA432 2017 Feb)			
			Cabozantinib (TA463 2017 Aug)			
				Lenvatinib with everolimus (TA498 2018 Jan)		

* The drugs appearing in the design of my Target Trial emulation studies in Chapter 7-8 are indicated by the yellow-highlighted cells in this table.
** The TAs (Treatment Arms) that have been crossed, meaning they have been terminated or replaced by the time of creating this table (January 2021).

Appendix 7.1 Specification for NCRAS data extraction

1. Cohort definition

- All patients who have been diagnosed with prostate cancer (C61) or renal cell carcinoma (C64) between 01/01/2011 and 31/12/2020.
- Limited to England.
- Prostate cancer: males. Kidney cancer: males and females.
- Dataset extraction:
  (1) Cancer Registration data, linked via tumourid for prostate cancer and renal cell carcinoma diagnoses as above.
  (2) Data extraction for the cohort from 2005 (6 years prior) for the above patients to the latest available data, from the following datasets: SACT, CWT, HES (APC), HES (OP), HES (A&E) and RTDS records, all linked at patient level (therefore comprising treatment data for the concurrent tumours below too). For the selected fields within each of the selected tables, as shown in the table below. Data is released in separate tables for each dataset.
  (3) A customized separate data table for concurrent tumour records: For the patients in the aforementioned cohort, the following fields from the Cancer Registry are required for all other cancer diagnoses (all ICD-10 C and D codes) diagnosed from 2009 to 2020: pseudonymised patientid, pseudonymised tumourid (ensure different pseudo tumourids to those used for cohort above), DIAGNOSISDATE1, DIAGNOSISDATE2, DIAGNOSISDATEBEST, DIAGNOSISDATEFLAG, SITE_ICD10_O2, SITE_ICD10_O2_3CHAR, MORPH_ICD10_O2, STAGE_BEST, T_BEST, N_BEST, M_BEST, GLEASON_PRIMARY, GLEASON_SECONDARY, GLEASON_TERTIARY, GLEASON_COMBINED.

2. Selected data tables (datasets) and fields (variables)

Data table (datasets within NCRAS)	Requested fields
Cancer Registry Data	PATIENTID (project specific pseudonymised)
	TUMOURID (project specific pseudonymised)
	SEX
	ETHNICITY
	ETHNICITYNAME
	AGE
	DIAGNOSISDATE1
	DIAGNOSISDATE2
	DIAGNOSISDATEBEST
	DIAGNOSISDATEFLAG
	BASISOFDIAGNOSIS
	SITE_ICD10_O2
	SITE_ICD10_O2_3CHAR
	MORPH_ICD10_O2
	BEHAVIOUR_ICD10_O2
	SITE_CODED
	SITE_CODED_DESC
	SITE_CODED_3CHAR
	CODING_SYSTEM

	CODING_SYSTEM_DESC
	MORPH_CODED
	BEHAVIOUR_CODED
	BEHAVIOUR_CODED_DESC
	HISTOLOGY_CODED
	HISTOLOGY_CODED_DESC
	GRADE
	TUMOURSIZE
	NODESEXCISED
	NODESINVOLVED
	TUMOURCOUNT
	BIGTUMOURCOUNT
	ROUTE_CODE
	FINAL_ROUTE
	STAGE_BEST
	T_BEST
	N_BEST
	M_BEST
	STAGE_BEST_SYSTEM
	T_IMG
	N_IMG
	M_IMG
	STAGE_IMG
	STAGE_IMG_SYSTEM
	T_PATH
	N_PATH
	M_PATH
	STAGE_PATH
	STAGE_PATH_SYSTEM
	STAGE_PATH_PRETREATED
	CHRL_TOT_27_03
	CHRL_TOT_78_06
	HES_LINKED
	GLEASON_PRIMARY
	GLEASON_SECONDARY
	GLEASON_TERTIARY
	GLEASON_COMBINED
	LATERALITY
	DCO
	VITALSTATUS
	VITALSTATUSDATE
	DEATHDATEBEST
	DEATHDATEFLAG
	EMBARKATION
	EMBARKATIONDATE
	DEATHCAUSECODE_1A
	DEATHCAUSECODE_1B
	DEATHCAUSECODE_1C
	DEATHCAUSECODE_2
	DEATHCAUSECODE_UNDERLYING
	DIAG_HOSP
	DIAG_HOSP_NAME
	FIRST_HOSP
	FIRST_HOSP_NAME
	FIRST_HOSP_DATE
	DIAG_TRUST
	DIAG_TRUST_NAME
	FIRST_TRUST
	FIRST_TRUST_NAME
	CCG_CODE
	CCG_NAME
	COUNTY_CODE
	COUNTY_NAME
	GOR_CODE
	GOR_NAME
	CTRY_CODE
	CTRY_NAME
	INCOME_QUINTILE (Income domain, based on version most appropriate for timing of tumour diagnosis)
	IMD_QUINTILE (Full IMD, based on version most appropriate for timing of tumour diagnosis)
	RT_FLAG
	CT_FLAG
	SG_FLAG
	EVENTID (project specific pseudonymised)
	NUMBER_OF_TUMOURS

	EVENTCODE
	EVENTDESC
	EVENTDATE
	PROVIDERCODE
	PROVIDERDESC
	TRUST_CODE
	TRUST_NAME
	PRACTITIONERCODE (Pseudonymised)
	WITHIN_SIX_MONTHS_FLAG
	SIX_MONTHS_AFTER_FLAG
	OPCS4_CODE
	OPCS4_NAME
	RADIOCODE
	RADIODESC
	IMAGINGCODE
	IMAGINGDESC
	LESIONSIZE
	CHEMO_ALL_DRUGS
	CHEMO_DRUG_GROUP
	MULTIFOCAL
	EXCISIONMARGIN
Systemic Anti-Cancer Therapy Dataset (SACT)	PATIENTID
	TUMOURID
	NHS_Number_Status
	Ethnicity
	Consultant_GMC_Code (pseudonymised)
	Consultant_Speciality_Code
	Organisation_Code_of_Provider
	Primary_Diagnosis
	Morphology_clean
	Stage_at_Start
	Programme_Number
	Regimen_Number
	Intent_of_Treatment
	Adjunctive therapy
	Analysis_Group
	Benchmark_Group
	Height_At_Start_of_Regimen
	Weight_At_Start_of_Regimen
	Performance_Status_at_Start_of_Regimen_Clean
	Comorbidity_Adjustment
	Date_Decision_To_Treat
	Start_Date_of_Regimen
	Clinical_Trial
	Chemo_Radiation
	Number_of_Cycles_Planned
	Cycle_Number
	Start_Date_of_Cycle
	Weight_At_Start_Of_Cycle
	Performance_Status_At_Start_Of_Cycle_Clean
	OPCS_Procurement_Code
	Drug_Group
	Actual_Dose_Per_Administration
	Administration_Route
	Administration_Date
	OPCS_Delivery_Code
	Date_of_Final_Treatment
	Regimen_Modification_Dose_Reduction
	Regimen_Modification_Time_Delay
	Regimen_Modification_Stopped_Early
	Regimen_Outcome_Summary
	regoutsum_cur_not_com_plan
	regoutsum_non_curat
	regoutsum_toxic
	regoutsum_cur_com_plan
Radiotherapy Dataset (RTDS)	PATIENTID (project specific pseudonymised)
	RADIOTHERAPYEPISODEID (project specific pseudonymised)
	APPTDATE
	DECISIONTOTREATDATE
	EARLIESTCLINAPPROPRIATEDATE
	RADIOTHERAPYPRIORITY
	TREATMENTSTARTDATE
	RADIOTHERAPYDIAGNOSISICD
	RADIOTHERAPYINTENT
	RTTREATMENTREGION

	RTTREATMENTANATOMICALSITE
	PRIMARYPROCEDUREOPCS
	PROCEDUREDATE
NDRS Linked HES A&E	PATIENTID (project specific pseudonymised)
	ethnos
	aeattendcat
	aeattenddisp
	arrivaldate (date YYYY-MM-DD)
	diag_n
	diag2_n
	diaga_n
	diags_n
	treat_n
	treat2_n
NDRS Linked HES APC	PATIENTID (project specific pseudonymised)
	startage
	ethnos
	admidate (date YYYY-MM-DD)
	elecdate (date YYYY-MM-DD)
	admimeth
	firstreg
	elecdur
	disdate (date YYYY-MM-DD)
	bedyear
	spelbgin
	epiend (date YYYY-MM-DD)
	epistart (date YYYY-MM-DD)
	speldur
	spelend
	epidur
	epiorder
	epitype
	diag_4n
	diag3_3n
	opertn_nn
	opdate_nn
	operstat
	intmanig
	mainspef
	tretspef
NDRS Linked HES OP	PATIENTID
	ethnos
	apptdate
	attended
	outcome
	priority
	diag_nn
	diag_4
	diag3
	opertn_nn
	opertn_01
	opertn3
	operstat
	mainspef
	tretspef
Cancer Waiting Times (Treatments Only)	PATIENTID (project specific pseudonymised)
	TUMOURID (project specific pseudonymised)
	TREAT_PERIOD_START
	SITE_ICD10
	METS_SITE
	LATERALITY
	TREAT_START
	WTA_TREAT
	WTA_TREAT_REASON
	CTE_TYPE
	MODALITY
	CLIN_TRIAL
	RADIO_PRIORITY

Appendix 8.1 Definition of the metastatic diagnosis date in the Flatiron metastatic prostate cancer database

"MetDiagnosisDate" in the Flatrion data was defined as the earliest confirmed date a patient developed distant metastatic prostate cancer, determined through manual abstraction by trained abstractors using the subsequent hierarchy:

(1) If a patient had distant metastases (M1) at initial prostate cancer diagnosis, the "MetDiagnosisDate" aligns with the initial prostate cancer diagnosis date “DiagnosisDate”.
(2) For patients diagnosed with Stage IV but with M0 or without initial distant metastases, the date they later developed metastasis is recorded as their "MetDiagnosisDate".
(3) Abstractors follow a hierarchy to determine the most accurate early date for metastatic confirmation by the treating physician:
    a. Biopsy specimen collection date from the pathology report
    b. Date reported by the physician for the biopsy
    c. Radiology scan date showing metastasis, later validated by a physician.
    d. Physician-stated metastatic diagnosis date.

Appendix 8.2 Definition of castration-resistant prostate cancer and the date of castration resistance in the Flatiron metastatic prostate cancer database

In the Flatiron data, the determination of CRPC diagnosis and the date of CRPC are as follows:

- Primary: Direct physician documentation (with the CRPC diagnosis date referred to by the clinician).
- Secondary: A documented PSA increase of at least 2.0 ng/ml during initial hormonal therapy, followed by another increase within 3 months (CRPC date is the second PSA rise's date).
- Tertiary: Physician documentation of rising PSA or PSA progression during the first hormonal therapy line accompanied by a treatment shift (CRPC date as specified by the physician).

Appendix 8.3 Details of diagnosis codes used in the study

Coding System	Codes*	Diagnosis Description
Chronic kidney disease with a creatinine clearance of < 30 ml/min		
ICD-9-CM	585.4x	Chronic kidney disease, Stage IV (severe)
	585.5x	Chronic kidney disease, Stage V
	585.6x	End stage renal disease
	585.9x	Chronic kidney disease, unspecified
ICD-10-CM	N18.4x	Chronic kidney disease, stage 4 (severe)
	N18.5x	Chronic kidney disease, stage 5
	N18.6x	End stage renal disease
	N18.9x	Chronic kidney disease, unspecified
Diagnosis of non-prostate primary malignancies (except non-melanomatous skin cancer)		
ICD-9-CM	140.x	Malignant neoplasm of lip
	141.x	Malignant neoplasm of tongue
	142.x	Malignant neoplasm of major salivary glands
	143.x	Malignant neoplasm of gum
	144.x	Malignant neoplasm of floor of mouth
	145.x	Malignant neoplasm of other and unspecified parts of mouth
	146.x	Malignant neoplasm of oropharynx
	147.x	Malignant neoplasm of nasopharynx
	148.x	Malignant neoplasm of hypopharynx
	149.x	Malignant neoplasm of other and ill-defined sites within the lip oral cavity and pharynx
	150.x	Malignant neoplasm of esophagus
	151.x	Malignant neoplasm of stomach
	152.x	Malignant neoplasm of small intestine including duodenum
	153.x	Malignant neoplasm of colon
	154.x	Malignant neoplasm of rectum rectosigmoid junction and anus
	155.x	Malignant neoplasm of liver and intrahepatic bile ducts
	156.x	Malignant neoplasm of gallbladder and extrahepatic bile ducts
	157.x	Malignant neoplasm of pancreas
	158.x	Malignant neoplasm of retroperitoneum and peritoneum
	159.x	Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum
	160.x	Malignant neoplasm of nasal cavities middle ear and accessory sinuses
	161.x	Malignant neoplasm of larynx
	162.x	Malignant neoplasm of trachea bronchus and lung
	163.x	Malignant neoplasm of pleura
	164.x	Malignant neoplasm of thymus heart and mediastinum
	165.x	Malignant neoplasm of other and ill-defined sites within the respiratory system and intrathoracic organs
	170.x	Malignant neoplasm of bone and articular cartilage
	171.x	Malignant neoplasm of connective and other soft tissue
	172.x	Malignant melanoma of skin
	174.x	Malignant neoplasm of female breast
	175.x	Malignant neoplasm of male breast
	176.x	Kaposi's sarcoma
	179.x	Malignant neoplasm of uterus, part unspecified
	180.x	Malignant neoplasm of cervix uteri
	181.x	Malignant neoplasm of placenta
	182.x	Malignant neoplasm of body of uterus
	183.x	Malignant neoplasm of ovary and other uterine adnexa

	184.x	Malignant neoplasm of other and unspecified female genital organs
	186.x	Malignant neoplasm of testis
	187.x	Malignant neoplasm of penis and other male genital organs
	188.x	Malignant neoplasm of bladder
	189.x	Malignant neoplasm of kidney and other and unspecified urinary organs
	190.x	Malignant neoplasm of eye
	191.x	Malignant neoplasm of brain
	192.x	Malignant neoplasm of other and unspecified parts of nervous system
	193.x	Malignant neoplasm of thyroid gland
	194.x	Malignant neoplasm of other endocrine glands and related structures
	195.x	Malignant neoplasm of other and ill-defined sites
	199.x	Malignant neoplasm without specification of site
	200.x	Lymphosarcoma and reticulosarcoma and other specified malignant tumors of lymphatic tissue
	201.x	Hodgkin's disease
	202.x	Other malignant neoplasms of lymphoid and histiocytic tissue
	203.x	Multiple myeloma and immunoproliferative neoplasms
	204.x	Lymphoid leukemia
	205.x	2Myeloid leukemia
	206.x	Monocytic leukemia
	207.x	Other specified leukemia
	208.x	Leukemia of unspecified cell type
	209.0x	Malignant carcinoid tumors of the small intestine
	209.1x	Malignant carcinoid tumors of the appendix, large intestine, and rectum
	209.2x	Malignant carcinoid tumors of other and unspecified sites
	209.3x	Malignant poorly differentiated neuroendocrine tumors
ICD-10-CM	C00.x	Malignant neoplasm of lip
	C01.x	Malignant neoplasm of base of tongue
	C02.x	Malignant neoplasm of other and unspecified parts of tongue
	C03.x	Malignant neoplasm of gum
	C04.x	Malignant neoplasm of floor of mouth
	C05.x	Malignant neoplasm of palate
	C06.x	Malignant neoplasm of other and unspecified parts of mouth
	C07.x	Malignant neoplasm of parotid gland
	C08.x	Malignant neoplasm of other and unspecified major salivary glands
	C09.x	Malignant neoplasm of tonsil
	C10.x	Malignant neoplasm of oropharynx
	C11.x	Malignant neoplasm of nasopharynx
	C12.x	Malignant neoplasm of pyriform sinus
	C13.x	Malignant neoplasm of hypopharynx
	C14.x	Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx
	C15.x	Malignant neoplasm of esophagus
	C16.x	Malignant neoplasm of stomach
	C17.x	Malignant neoplasm of small intestine
	C18.x	Malignant neoplasm of colon
	C19.x	Malignant neoplasm of rectosigmoid junction
	C20.x	Malignant neoplasm of rectum
	C21.x	Malignant neoplasm of anus and anal canal
	C22.x	Malignant neoplasm of liver and intrahepatic bile ducts
	C23.x	Malignant neoplasm of gallbladder
	C24.x	Malignant neoplasm of other and unspecified parts of biliary tract
	C25.x	Malignant neoplasm of pancreas

	C26.x	Malignant neoplasm of other and ill-defined digestive organs
	C30.x	Malignant neoplasm of nasal cavity and middle ear
	C31.x	Malignant neoplasm of accessory sinuses
	C32.x	Malignant neoplasm of larynx
	C33.x	Malignant neoplasm of trachea
	C34.x	Malignant neoplasm of bronchus and lung
	C37.x	Malignant neoplasm of thymus
	C38.x	Malignant neoplasm of heart, mediastinum and pleura
	C39.x	Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathoracic organs
	C40.x	Malignant neoplasm of bone and articular cartilage of limbs
	C41.x	Malignant neoplasm of bone and articular cartilage of other and unspecified sites
	C43.x	Malignant melanoma of skin
	C4A.x	Merkel cell carcinoma
	C45.x	Mesothelioma
	C46.x	Kaposi's sarcoma
	C47.x	Malignant neoplasm of peripheral nerves and autonomic nervous system
	C48.x	Malignant neoplasm of retroperitoneum and peritoneum
	C49.x	Malignant neoplasm of other connective and soft tissue
	C50.x	Malignant neoplasm of breast
	C51.x	Malignant neoplasm of vulva
	C52.x	Malignant neoplasm of vagina
	C53.x	Malignant neoplasm of cervix uteri
	C54.x	Malignant neoplasm of corpus uteri
	C55.x	Malignant neoplasm of uterus, part unspecified
	C56.x	Malignant neoplasm of ovary
	C57.x	Malignant neoplasm of other and unspecified female genital organs
	C58.x	Malignant neoplasm of placenta
	C60.x	Malignant neoplasm of penis
	C62.x	Malignant neoplasm of testis
	C63.x	Malignant neoplasm of other and unspecified male genital organs
	C64.x	Malignant neoplasm of kidney, except renal pelvis
	C65.x	Malignant neoplasm of renal pelvis
	C66.x	Malignant neoplasm of ureter
	C67.x	Malignant neoplasm of bladder
	C68.x	Malignant neoplasm of other and unspecified urinary organs
	C69.x	Malignant neoplasm of eye and adnexa
	C70.x	Malignant neoplasm of meninges
	C71.x	Malignant neoplasm of brain
	C72.x	Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system
	C73.x	Malignant neoplasm of thyroid gland
	C74.x	Malignant neoplasm of adrenal gland
	C75.x	Malignant neoplasm of other endocrine glands and related structures
	C76.x	Malignant neoplasm of other and ill-defined sites
	C7A.x	Malignant neuroendocrine tumors
	C80.x	Malignant neoplasm without specification of site
	C81.x	Hodgkin lymphoma
	C82.x	Follicular lymphoma
	C83.x	Non-follicular lymphoma
	C84.x	Mature T/NK-cell lymphomas
	C85.x	Other specified and unspecified types of non-Hodgkin lymphoma
	C86.x	Other specified types of T/NK-cell lymphoma

	C88.x	Malignant immunoproliferative diseases and certain other B-cell lymphomas
	C90.x	Multiple myeloma and malignant plasma cell neoplasms
	C91.x	Lymphoid leukemia
	C92.x	Myeloid leukemia
	C93.x	Monocytic leukemia
	C94.x	Other leukemias of specified cell type
	C95.x	Leukemia of unspecified cell type
	C96.x	Other and unspecified malignant neoplasms of lymphoid, hematopoietic and related tissue
Diagnosis of brain metastasis		
ICD-9-CM	198.3x	Secondary malignant neoplasm of brain and spinal cord
	198.4x	Secondary malignant neoplasm of other parts of nervous
ICD-10-CM	C79.3x	Secondary malignant neoplasm of brain and cerebral meninges
Diagnosis of seizure and cerebrovascular events		
ICD-9-CM	345.x	Epilepsy and recurrent seizures
	430.x	Subarachnoid hemorrhage
	431.x	Intracerebral hemorrhage
	432.x	Other and unspecified intracranial hemorrhage
	433.x	Occlusion and stenosis of precerebral arteries
	434.x	Occlusion of cerebral arteries
	463.x	Acute, but ill-defined, cerebrovascular disease
	437.x	Other and ill-defined cerebrovascular disease
	438.x	Late effects of cerebrovascular disease
ICD-10-CM	G40.x	Epilepsy and recurrent seizures
	I60.x	Nontraumatic subarachnoid hemorrhage
	I61.x	Nontraumatic intracerebral hemorrhage
	I62.x	Other and unspecified nontraumatic intracranial hemorrhage
	I63.x	Cerebral infarction
	I65.x	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
	I66.x	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
	I67.x	Other cerebrovascular diseases
	I68.x	Cerebrovascular disorders in diseases classified elsewhere
	I69.x	Sequelae of cerebrovascular disease

*The letter "x" signifies that all codes with a prefix preceding "x" are eligible.

Appendix 8.4 Prior treatments that were not allowed in the GUTG-001 Analogue Target Trial

abiraterone, enzalutamide, ketoconazole, apalutamide (ARN-509), galeterone (TOK-001), orteronel (TAK-700), samarium Sm 153 lexidronam, radium-223, sipuleucel-T, strontium-89, vipivotide, any other radiopharmaceutical drug, darolutamide, larotrectinib, olaparib, relugolix, rucaparib, talazoparib, clinical study drug

Appendix 8.5 R code for identifying the Target Trial cohort in Prostate Cancer Case Study 1 (PC1)

```
################################################################################
###   Cohort Identification                                                  ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                    ###
###   Data: Flatiron prostate cancer datasets                                ###
###                                                                          ###
###   University of Sheffield                                                ###
###   Date created: Dec 20, 2022                                             ###
###   Created Git & renv project: Aug 01, 2023                               ###
###   Applied GUTG-001 criteria: Sep 18, 2023                                ###
###   Applied unrestricted full mCRPC naive cohort: Sep 20, 2023             ###
###   Re-do all analysis using updated Flatiron data (Sep 2023): Oct 3, 2023 ###
################################################################################

## House keeping

# Load packages and functions
renv::status()
`%notin%` <- Negate(`%in%`)

if (!require(dplyr)) install.packages("dplyr")
library(dplyr)
library(lubridate)
library(stringr)


# load raw data using "00_Data Inspection"


#############################################################################################
###
###  Inclusion criteria
###  (The number of criteria were taken from the original GUTG-001 protocol)
###  https://classic.clinicaltrials.gov/ProvidedDocs/57/NCT02125357/Prot_SAP_000.pdf
###
#############################################################################################

##***** 1.  Willing and able to provide informed consent --> all patients gave consent to be treated

##***** 5.  Evidence of metastatic disease on bone scan or CT scan
#**** ==> See Flatiron: Knowledge centre how meta-diagnosis were defined in d_diag_metPC (i.e., the Met PC dataset)

# All 21711 patients in the Met PC dataset has a metastatic date
test <- rd_diag_metPC[rd_diag_metPC$MetDiagnosisDate %in% ""] # 0 rows (all patients in the dataset developed metastatic prostate cancer at some time point)

# Check diagnosis year of metastasis
test <- merge(rd_diag_metPC[ , c("DiagnosisDate", "MetDiagnosisDate", "PatientID")], rd_LOT, by = "PatientID", all.y = TRUE) %>%
  mutate_at(vars(DiagnosisDate, MetDiagnosisDate, StartDate, EndDate), as.Date, format="%Y-%m-%d") # 36350 (all raw LOT data)
summary(as.factor(year(test$MetDiagnosisDate)))
# year 2013-20233 each year approx. 3000-4000 pts with less patients in year 2022-2023 (closer to data cut-off)

# Select all patients from the Met prostate cancer dataset (their met diagnosis is either their first PC diagnosis or later)
# Create an initial dataset to store patient records for the Unrestricted full mCRPC cohort (each row = records for one patient)
d_TT_cohort_full_mCRPC <- rd_diag_metPC # 21711 (the GUTG-001 cohort will be a subset of full_mCRPC)
# Create an initial dataset to store patient records for the GUTG-001 Analogue cohort (each row = records for one patient)
d_TT_cohort_GUTG001 <- rd_diag_metPC # 21711


##***** 6.  Evidence of biochemical or imaging progression in the setting of surgical or
#   medical castration. Progressive disease for study entry is defined by one of
#   the following three criteria:
#    a. PSA progression: minimum of two rising PSA values from a baseline
#       measurement with an interval of ≥ 1 week between each
#       measurement. Minimum PSA at screening visit is > 2.0 ug/L
#    b. Soft tissue or visceral disease progression (see Appendix B for
#       definition of measurable disease as per RECIST 1.1 criteria)
#    c. Bone progression: ≥ 2 new lesions on bone scan

#**** Limit Cohort Met Patients to those who ever developed CPRC and a CPRC date
#**** Assign a mCPRC date to each mCRPC patient based on the later date among MetaDiagnosisDate and CRPCDate
```

```
# Unrestricted full mCRPC cohort
d_TT_cohort_full_mCRPC <- subset(d_TT_cohort_full_mCRPC, IsCRPC == "Yes") # 13233
d_TT_cohort_full_mCRPC <- d_TT_cohort_full_mCRPC[d_TT_cohort_full_mCRPC$CRPCDate %notin% "", ] %>%
  mutate_at(vars(DiagnosisDate, MetDiagnosisDate, CRPCDate), as.Date, format="%Y-%m-%d")
# 13233 --> 13016 rows, roughly 1.6% (dropping 217 rows)
d_TT_cohort_full_mCRPC$MCRPCDate <- apply(d_TT_cohort_full_mCRPC[ , c("MetDiagnosisDate", "CRPCDate")], 1, max)
d_TT_cohort_full_mCRPC$MCRPCDate <- as.Date(d_TT_cohort_full_mCRPC$MCRPCDate, format="%Y-%m-%d")
# Relocate the column of MCRPC date so it is easier to read
d_TT_cohort_full_mCRPC <- d_TT_cohort_full_mCRPC %>%
  relocate(CRPCDate, MCRPCDate, .after = MetDiagnosisDate)
sum(is.na(d_TT_cohort_full_mCRPC$MCRPCDate)) # no missing MCRPCdate
summary(as.factor(year(d_TT_cohort_full_mCRPC$MCRPCDate))) # Check MCRPC year 2013-2023

# GUTG-001 cohort: The same as the unrestricted mCRPC cohort at this stage
d_TT_cohort_GUTG001 <- d_TT_cohort_full_mCRPC # 13016

# For patients who had CRPC in Flatiron, some may have progressed from other stage e.g. nmCRPC or HSPC, nmHSPC
# Some patients in the data may have mHSPC, but did not progress to CRPC


#*******************************************************************************
#****** The selection process for unrestricted full mCRPC patients is complete.
#****** The following procedure is focusing on aligning patients with
#****** the remaining criteria specified in GUTG-001.
#*******************************************************************************

####******ADDITIONAL CRITERIA from Flatiron: deleting patients having "Line Zero"
####***** Rationale: Line info from these patients are incomplete

# See how many met patients have received treatment from the cohort
test <- rd_LOT %>% # 16248 (74.84% patients (out of 21711) have received treatments for prostate cancer)
  select(PatientID) %>%
    distinct(PatientID)

# Delete patients who have Line Zero in the LOT table
# --> hope this will solve the problem of patients having mCRPC treatment after their mCRPC date (CRPCDate or metDiagDate, whichever is the
latest)
v_ID_LineZero <- dplyr::filter(rd_LOT, LineName %in% "Line Zero") %>%
  select(PatientID) %>%
  distinct(PatientID)
# 1400 rows (1400 patients had Line Zero, in full met dataset (met be less in mCRPC dataset))
test <- merge(rd_LOT, v_ID_LineZero, by = "PatientID", all.y = TRUE)

# Testing whether any treatment interval of patients having LineZero is less than zero
test <- test %>%
  arrange(PatientID, LineNumber) %>%
  group_by(PatientID) %>%
  mutate(StartDate_lag = dplyr::lag(StartDate, n = 1, default = NA)) %>%
  mutate_at(vars(StartDate, StartDate_lag, EndDate), as.Date, format="%Y-%m-%d") %>%
  mutate(Gap_treat_interval = StartDate - StartDate_lag)
summary(as.numeric(test$Gap_treat_interval))
# All treatment intervals of patients having Line Zero (n = 1400) are above zero (28~3571)
# Line Zero are always before labelled treatments --> representing unknown previous treatment


### For the purpose of mimicking GUTG-001 trial --> deleting Patients having Line Zero
### We are uncertain whether these patients had abiraterone or enzalutamide as prior treatment (should be excluded)
### Also, it's harder to define time zero for these patients

# For GUTOG-001 cohort
d_TT_cohort_GUTG001 <- anti_join(d_TT_cohort_GUTG001, v_ID_LineZero, by = "PatientID")
# 11616 (remove 1400 patients that have LineZero)

# Note Dr. Philani's (Flatiron) replication code on 20231012 also deleted patients who doesn't have any LOT data (on top of those who had
incomplete treatment history)
# so his cohort came to a number of 10166 (this will be executed as part of in my next step: excluding patients who haven't initiated any treatment
since mCRPC as I can assure I am not completely deleting patients based on future events)

# Check if all Line zero have been deleted in the GUTG001 cohort
test <- merge(rd_LOT, d_TT_cohort_GUTG001, by = "PatientID") %>%
  arrange(PatientID, LineNumber) # 25798 LOTs (inner join)
summary(as.factor(test$LineNumber)) # no Line zero, but 1932 NA (potentially mHSPC)
summary(as.factor(test$LineSetting)) # HSPC 7, mCPRC 23141, mHSPC 1925, nmCRPC 725 (no LineSetting = NA) ### NOTE PATIENTS WITH mHSPC
can be used to study mHSPC treatments
```

```
sum(is.na(test$LineName)) # 0 unknown LineName


####******ADDITIONAL CRITERIA: Exclude Patients who did not initiate treatment following their mCRPC diagnosis.
####***** Rationale: These patients did not survive until "randomisation" (which is similar to the timing of receiving first-line treatment)

#**** Logic check
#**** Check how many patient's StartDtae for the first-line mCRPC treatment is before mCRPC date

# Check how many patients out of the cohort full_mCRPC cohort ever had PC treatment
test <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "DiagnosisDate", "MetDiagnosisDate", "CRPCDate", "MCRPCDate")], rd_LOT, by =
"PatientID", all.x = TRUE) %>%
  mutate_at(vars(DiagnosisDate, MetDiagnosisDate, CRPCDate, MCRPCDate, StartDate, EndDate), as.Date, format="%Y-%m-%d") %>%
  group_by(PatientID) %>%
  slice(which.min(LineNumber)) # 10007 out of 11616 patients had treatment in one of the setttings: mCRPC, mHSPC, nmCPRC, nmHSPC

test <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "DiagnosisDate", "MetDiagnosisDate", "CRPCDate", "MCRPCDate")], rd_LOT, by =
"PatientID", all.x = TRUE) %>%
  mutate_at(vars(DiagnosisDate, MetDiagnosisDate, CRPCDate, MCRPCDate, StartDate, EndDate), as.Date, format="%Y-%m-%d") %>%
  dplyr::filter(LineSetting %notin% "mCRPC") %>%
  group_by(PatientID) %>%
  slice(which.min(LineNumber)) # 464 out of 11616 patients had treatments of non mCRPC

test <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "DiagnosisDate", "MetDiagnosisDate", "CRPCDate", "MCRPCDate")], rd_LOT, by =
"PatientID", all.x = TRUE) %>%
  mutate_at(vars(DiagnosisDate, MetDiagnosisDate, CRPCDate, MCRPCDate, StartDate, EndDate), as.Date, format="%Y-%m-%d") %>%
  dplyr::filter(LineSetting %in% "mCRPC") %>%
  group_by(PatientID) %>%
  slice(which.min(LineNumber)) # 9938 out of 11616 patients had treatments of mCRPC


# Check how many patient's first-line mCRPC treatment is before mCRPC date
# Find the first mCRPC LOT for each patient
summary(test$LineNumber) # line 1-8: median = 1, Q3 = 1, max = 8 (meaning that first-line mCRPC  = x-line of CRPC)
summary(as.factor(test$LineNumber))
# n = 1 (first-line mCRPC = first-line of CRPC): 9543
# n = 2 (first-line mCPRC = second-line of CRPC), 260

# Gap from diagnosis of mCRPC to the first-treatment of mCRPC
test$gap_mCPRCDate_1stmCRPCtreatStart <- test$StartDate - test$MCRPCDate
test$gap_mCPRCDate_1stmCRPCtreatEnd <- test$EndDate - test$MCRPCDate
summary(as.numeric(test$gap_mCPRCDate_1stmCRPCtreatStart)) # median = 16, min = -2382~max 2598

# Identify patients who had "LineSetting = mCRPC" treatment before the mCRPC date (out of 7422 who had mCRPC treatments)
length(as.numeric(test$gap_mCPRCDate_1stmCRPCtreatStart[test$gap_mCPRCDate_1stmCRPCtreatStart < 0])) # 2313 patients
test2 <- test[test$gap_mCPRCDate_1stmCRPCtreatStart < 0, ]
summary(as.numeric(test2$gap_mCPRCDate_1stmCRPCtreatStart))
# Among these 2313 patients:
# date mCRPC date to the start date of the first mCRPC treatment: : -2382~-1: median -152.0, Q1: -341, Q3 -33
summary(as.numeric(test2$gap_mCPRCDate_1stmCRPCtreatEnd))
# date mCRPC date to the end date of the first mCRPC treatment: : 0~2430: Q1 32, median 89, Q3: 217


### Among current GUTG-001 eligible patients (n = 13016, including those that had LineZero) who ever received mCRPC treatments (n = 9938),
### 2313 (23.3%) received their "first mCRPC treatment" (LineSetting = "mCRPC" & LineNumber = the minimum number within "mCRPC") before
their mCRPC date

### Please see examples: IDs in the test2 dataset

### Update: A treatment regimen (LOT) is labeled as for "mCRPC" if ANY EPISODE of this regimen commences after the mCRPC date.
### Check: For all patients who began mCRPC treatment prior to their mCRPC date, all such treatments concluded by or on the mCRPC date, which
is reasonable.
#### However, I'll need to consult with the oncologist regarding classifying these treatments as "mCRPC" treatments for my study, especially since
some concluded just a week post the mCRPC date.
#### It's plausible that the subsequent regimen is the more appropriate treatment designated for mCRPC.

# Potential Scenarios:
#  1. Treatment regimen begins just prior to the mCRPC date and concludes much later:
#  This suggests the mCRPC treatment might have commenced before the official mCRPC documentation.

# 2. Treatment regimen starts long before the mCRPC date and finishes shortly after:
#  non-mCPRC treamtent was halted once mCRPC was documented.

# 3. Treatment regimen initiates shortly before the mCRPC date and ends shortly after:
```

```
#  ??? need to consult oncologists

# 4. Treatment regimen begins a long time before the mCRPC date and concludes much later:
#  ??? need to consult oncologists

# Additional Considerations:
# Establishing a grace period is essential.
# Understanding how Drug Episodes (and LOT) were defined is crucial (e.g. longer end date due to refill dates?)


#**** Current solution 1:  Check patient's first treatment which StartDate is on the date of mCRPC or after, label the date of first mCRPC treatment, and relabel the LOT
#### (assuming that all treatments across mCRPC or before were for non-mCRPC: caveat: some start right before mCRPC date, the date of mCPRC could be mis-classified)
# Create new numbering system for user-defined mCRPC treatments (only label of treatments starting after mCRPC diagnosis)
d_LOT_treat_after_mCRPC <- merge(rd_LOT, d_TT_cohort_GUTG001, by = "PatientID") %>%
  mutate_at(vars(StartDate, EndDate), as.Date, format="%Y-%m-%d") %>% # 20828
  filter(StartDate >= MCRPCDate) %>%
  arrange(PatientID, StartDate) %>%
  group_by(PatientID) %>%
  mutate(LineNumber_mCRPC_UserDef = seq_along(PatientID)) %>%
  select(colnames(rd_LOT), LineNumber_mCRPC_UserDef)
summary(as.factor(d_LOT_treat_after_mCRPC$LineName[d_LOT_treat_after_mCRPC$LineNumber_mCRPC_UserDef == 1])) # First-line 3036 abiraterone, 2915 Enzalutamide
length(unique(d_LOT_treat_after_mCRPC)) # 9224 patients ever had treatments starting after mCRPC diagnosis

# merge the relabelled mCRPC treatment line to all LOT for GUTG-001 patients
d_LOT_GUTG001 <- rd_LOT %>%
  mutate_at(vars(StartDate, EndDate), as.Date, format="%Y-%m-%d") %>%
  filter(PatientID %in% unique(d_LOT_treat_after_mCRPC$PatientID)) %>% # limited LOT to those who had treatments after mCRPC diagnosis
  left_join(d_LOT_treat_after_mCRPC, by = colnames(rd_LOT)) %>%
  arrange(PatientID, StartDate) # 24575 LOTs (including pre-treatment before mCRPC) for those 9224 patients

# For GUTOG-001 cohort
d_TT_cohort_GUTG001 <- d_TT_cohort_GUTG001[d_TT_cohort_GUTG001$PatientID %in% d_LOT_GUTG001$PatientID, ] # 9224 (2392/11616, 20.1% patients did not commence treatment after mCRPC diagnosis)



####******ADDITIONAL CRITERIA: Include only patients who instated abiraterone or enzalutamide following mCPRC treatment (time zero: start of treatment)
# Find ID that started abi or enza after mCRPC diagnosis
v_ID_1stmCRPCtreat_abi_enza <- d_LOT_GUTG001[d_LOT_GUTG001$LineNumber_mCRPC_UserDef %in% 1 &
                        d_LOT_GUTG001$LineName %in% c("Abiraterone", "Enzalutamide"), ] %>%
                        select(PatientID)
d_TT_cohort_GUTG001 <- d_TT_cohort_GUTG001[d_TT_cohort_GUTG001$PatientID %in% v_ID_1stmCRPCtreat_abi_enza[ ,"PatientID"], ] # 5951

# Limit LOT info to only patients starting with abi or enza patients as well
d_LOT_GUTG001 <- d_LOT_GUTG001[d_LOT_GUTG001$PatientID %in% v_ID_1stmCRPCtreat_abi_enza[ ,"PatientID"], ] # 24575 --> 14490
summary(as.factor(d_LOT_GUTG001[d_LOT_GUTG001$LineNumber_mCRPC_UserDef %in% 1, "LineName"])) # First-line mCRPC: abi: 3036; enza 2915



##### Assign the date of start of treatment for all patients (Baseline Date in the Trial: Day 1 == Time zero)
d_TT_cohort_GUTG001 <- filter(d_LOT_GUTG001, LineNumber_mCRPC_UserDef %in% 1) %>%
  select(PatientID, StartDate) %>%
  rename(TimeZero = StartDate) %>%
  inner_join(d_TT_cohort_GUTG001, by = "PatientID") # 5951

# Check the date of first-line treatment in relation to other dates
summary(as.numeric(d_TT_cohort_GUTG001$TimeZero - d_TT_cohort_GUTG001$MCRPCDate)) # Median: 32, Mean: 104, Q1: 8, Q3: 101, min: 0, max: 2382
summary(as.numeric(d_TT_cohort_GUTG001$TimeZero - d_TT_cohort_GUTG001$CRPCDate)) # Median: 49, Mean: 189, Q1: 13, Q3: 173, min: 0, max: 6754
summary(as.numeric(d_TT_cohort_GUTG001$TimeZero - d_TT_cohort_GUTG001$MetDiagnosisDate)) # Median: 340, Mean: 461, Q1: 105, Q3: 635, min: 0, max: 3317
summary(as.numeric(d_TT_cohort_GUTG001$TimeZero - d_TT_cohort_GUTG001$DiagnosisDate)) # Median: 1372, Mean: 2405, Q1: 525, Q3: 3676, min: 10, max: 15982: NA 74

####****** Other thought: excluding patients who still have ongoing treatment at the date mCRPC (having treatment LOT spanning across the mCRPC date (selection bias????))
####****** these patients as we cannot be sure whether these are treatments prior to mCRPC or for CRPC (NEED TO DO SCENARIO ANLAYSIS)
```

```
####****** ADDITIONAL CRITERIA TO ENSURE similar data-cut-off point for mature survival analysis
# GUTG-001: enrollment between 2014-10-21 and 2016-12-13
# patients were enrolled, of whom 101 were assigned to each group.
# The median follow-up at time of data cutoff (May 31, 2018) was  30·7  months  (IQR  25·1–36·2).
# Last enrollment to data-cut-off is roughly 17 months (The shortest period from enrollment to data-cut-off)

#**** Limit Patients to those who initiated their first treatment after MCRPC date no later than 2022-03-31 (2022-03-31~2023-08-31 is roughly 17 months)
# Exclude Patient's time-zero is later than 2022-03-31
d_TT_cohort_GUTG001 <- filter(d_TT_cohort_GUTG001, d_TT_cohort_GUTG001$TimeZero <= as.Date("2022-03-31")) # 5288
summary(as.factor(year(test$TimeZero)))

# Limit LOT info to only patients starting with abi or enza patients as well
d_LOT_GUTG001 <- d_LOT_GUTG001[d_LOT_GUTG001$PatientID %in% d_TT_cohort_GUTG001[ ,"PatientID"], ] # 13314 rows
summary(as.factor(d_LOT_GUTG001[d_LOT_GUTG001$LineNumber_mCRPC_UserDef %in% 1, "LineName"])) # First-line mCRPC: abi: 2717; enza 2571
summary(as.factor(d_LOT_GUTG001[d_LOT_GUTG001$LineNumber_mCRPC_UserDef %in% 2, "LineName"])) # Second-line mCRPC enza 818, doce 640, abi 618
summary(as.factor(year(d_LOT_GUTG001[d_LOT_GUTG001$LineNumber_mCRPC_UserDef %in% 1, "StartDate"])))



####************************************************************
####******Entering trial inclusion and exclusion criteria
####************************************************************
# Use LOINC for lab test relevant criteria
# https://loinc.org/2857-1
v_LOINC_all <- data.frame(sort(unique((rd_lab$LOINC)))) # 300 LOINCs, confirmed that these LOINC table only contains numbers

# Lab result norm: ref mayo clinic
# https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595


### Now we know the "Time Zero" in the study, so we can assess the age
# 2.  Adult males ≥ 18 years age
summary(rd_demo$BirthYear) # 1936-1986 --> all above 18 years old by 2013 (no need to exclude any patient)


# 3.  History of adenocarcinoma of the prostate diagnosed histologically without
#     evidence of neuroendocrine or small cell differentiation, or if patient does not
#     have pathology of adenocarcinoma of the prostate, patient has metastatic
#     disease typical of prostate cancer (i.e., involving bone or pelvic lymph nodes
#     or para-aortic lymph nodes) AND a serum concentration of PSA that is rising
#     and >20ng/mL at the time of when the patient was clinically diagnosed with
#     prostate cancer.

#****** Per Dr. Pezaro: all mCRPC patients can be included. This criterion is not so clinically relevant

# Check histology at original diagnosis among CRPC patients
summary(as.factor(d_TT_cohort_GUTG001$Histology))
# 10757 adenocarcinoma, 2259 unspecified

# N: unknown 7299, N0 3042, N1 1841, NX 834
# M: unknown 2687, M0 5159, M1 3752, M1a 110, M1b 986, M1c 322
# Stage info among patients with unspecified histology (see NCCN PC Guideline Ver1. 2023):
# **** ==> Check with Flatiron: any methods to check bone or pelvic lymph nodes or para-aortic lymph nodes (updated: not using this criteria)
summary(as.factor(d_TT_cohort_GUTG001$NStage))
summary(as.factor(d_TT_cohort_GUTG001$MStage))

# 7.  ECOG performance status 0-2 (see Appendix C)
# According to Flatiron, might not be able to do it for every patients (missing data) (include 0-2)
# Using Baseline ECOG data Exclude patients who have an ECOG > 2 within 30 days prior to TimeZero
d_ECOG_timezero_GUTG001 <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                         rd_baselineECOG[ , c("PatientID", "ECOGSource", "LineStartDate", "ECOGValue", "ECOGDate")],
                         by = "PatientID") %>%
  mutate_at(vars(LineStartDate, ECOGDate), as.Date, format="%Y-%m-%d") %>%
  filter(LineStartDate <= TimeZero) %>%
  group_by(PatientID) %>%
  slice(which.max(LineStartDate)) # 5288
d_ECOG_timezero_GUTG001$Gap_ECOGDate_TimeZero <- d_ECOG_timezero_GUTG001$ECOGDate - d_ECOG_timezero_GUTG001$TimeZero
summary(as.numeric(d_ECOG_timezero_GUTG001$Gap_ECOGDate_TimeZero)) # within 30 days before Time Zero or within a week afterwards
summary(as.factor(d_ECOG_timezero_GUTG001$ECOGValue)) # 3: 154, 4: 6, others unknown
summary(as.factor(d_ECOG_timezero_GUTG001$ECOGSource)) # extracted: 2307; strctured: 2981
```

```
v_ID_ECOG_poor <- d_ECOG_timezero_GUTG001 %>%
 filter(Gap_ECOGDate_TimeZero <= 0) %>% # 5288 --> 3187 (delete those using ECOG value after TimeZero)
 filter(ECOGValue %notin% "Unknown") %>%
 filter(ECOGValue > 2) %>%
 select(PatientID) %>%
 ungroup() # 144


# Exclude Patient's who has ECOG > 2
d_TT_cohort_GUTG001 <- anti_join(d_TT_cohort_GUTG001, v_ID_ECOG_poor, by = "PatientID")
# 5144


#########################################################################################################################
###########NOTE: FOR THE CRITERIA BELOW, patients are flagged and inclusion and exclusion criteria will be done at the end  ##############
#########################################################################################################################

# 4.  Prior surgical orchiectomy or if on LHRH agonist/antagonist then
#    testosterone < 1.7 nmol/L at screening visit (patients must maintain LHRH
#    agonist/antagonist therapy for duration of study treatment if not surgically
#    castrated)

# According to Dr. Pezaro, all patients should have undergone surgical castration or be under the effects of chemical castration.
# Therefore, we only need to check patient's testosterone level

### Exclude patients whose latest testosterone level is greater than 1.7 nmol/L (49.03 ng/dL) recorded within the 30 days preceding or on the date of TimeZero.
# Assume those who didn't have extreme values (no test) have remained within reasonable level (castrated)

# GUTG-001 cohort
# Testosterone level at screening visits
v_LabComponent_testosterone <- unique(grep("testosterone", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_testosterone))

# Create a table of all testosterone tests
d_LONIC_testosterone <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_testosterone,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
 arrange(LOINC, LabComponent, TestBaseName, Test)
# Create a vecTor containing all LOINC for testosterone
v_LOINC_testerone <- unique(d_LONIC_testosterone$LOINC) # 2986-8

v_ID_ADT_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                       rd_lab[rd_lab$LOINC %in% v_LOINC_testerone, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                       by = "PatientID",
                       all.y = TRUE) %>%
 mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
 filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # 1139 rows for testosterone within 30 days before diagnosis,test date: Day -30~-1 & Day 1

# Check none cleaned results
summary(as.factor(v_ID_ADT_fail$TestResult)) # there are < none number values, e.g. < or Test not performed
test <- v_ID_ADT_fail %>%
 distinct(TestResult) %>%
 arrange(TestResult)
# None numeric results: Note_Comment, Pending, Test not performed, the greatest is < 23 (which is still < 49.03, so doesn't matter ==> just have to look at cleaned results)

v_ID_ADT_fail <- v_ID_ADT_fail %>%
 filter(!is.na(TestResultCleaned)) %>% # excluded those with NA results: no results or non-numeric results
 group_by(PatientID) %>%
 slice(which.max(TestDate))  # 618 patients had testosterone test within 30 days of TimeZero
summary(v_ID_ADT_fail$TestResultCleaned)

# Find patients whose latest testosterone was >= 49.03 ng/dL
v_ID_ADT_fail <- v_ID_ADT_fail %>%
 filter(TestResultCleaned >= 49.03) %>%
 select(PatientID) # 37
v_ID_ADT_fail$f_timezero_ADT_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_ADT_fail, by = "PatientID", all.x = TRUE) %>%
 mutate(f_timezero_ADT_fail = ifelse(is.na(f_timezero_ADT_fail), 0, f_timezero_ADT_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_ADT_fail)) # 37
```

```
# 9.  Adequate organ function defined as:
# a.  Absolute neutrophil count ≥ 1.5 x 10 9 /L, platelet count ≥ 100 x 10 9 /L and hemoglobin ≥ 80 g/L
# b.  Creatinine clearance ≥ 30 ml/min (calculated by Cockcroft-Gault formula, see Appendix D)
# c.  Serum potassium > than lower limit of normal range
# d.  Total bilirubin ≤ 1.5 x upper limit of normal (ULN) except for
#     patients with known Gilbert's syndrome (direct bilirubin ≤ 1.5 x ULN)
# e.  Alanine aminotransferase (ALT) and aspartate aminotransferase (AST) ≤ 5 x ULN


#### Create LOINC lists for this criteria: not using patient's ICD-9 as they might not be as accurate

## Absolute neutrophil count
# Neutrophil level at screening visits: WBC x total neutrophils (segmented neutrophils% + segmented bands%) x 10 = ANC
v_LabComponent_abs_neutrophil <- unique(grep("neutrophil", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_abs_neutrophil))
# The most relevant ones is neutrophil count (absolute) --> can be calculated with WBC and others though
v_LabComponent_abs_neutrophil <- "Neutrophil count (absolute)"
# Create a table of absolute neutrophil - all tests had units of 10^9/L
d_LONIC_abs_neutrophil <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_abs_neutrophil,
                          c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
# Create a vecTor containing all LOINC for absolute neutrophil
v_LOINC_abs_neutrophil <- unique(d_LONIC_abs_neutrophil$LOINC) # 26499-4, 751-8, 753-4


## Platelet count
v_LabComponent_platelet <- unique(grep("platelet", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_platelet))
# Create a table of platelet - all tests had units of 10^9/L
d_LONIC_platelet <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_platelet,
                           c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
              arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_platelet <- unique(d_LONIC_platelet$LOINC) # 26515-7, 49497-1, 777-3, 778-1


## hemoglobin
v_LabComponent_Hb <- unique(grep("hemoglobin", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_Hb)) # one of them is HbA1c
v_LabComponent_Hb <- "Hemoglobin, whole blood"

# Create a table of Hb - all tests had units of g/dL
d_LONIC_Hb <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_Hb,
                        c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_Hb <- unique(d_LONIC_Hb$LOINC) # "20509-6" "718-7"



## Creatinine clearance
# Trial definition: Calculated creatinine clearance (Cockcroft-Gault formula):
# N x (140-Age) x weight in kg ÷ Serum Creatinine in µmol/L
# * For males N=1.23; for females N=1.04
v_LabComponent_Scr <- unique(grep("creatinine", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_Scr))
v_LabComponent_Scr <- "Creatinine, serum"
# Create a table of Scr - all tests had units of mg/dL
d_LONIC_Scr <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_Scr,
                     c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_Scr <- unique(d_LONIC_Scr$LOINC) # "2160-0"

## Potassium
v_LabComponent_K <- unique(grep("potassium", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_K))
v_LabComponent_K <- "Potassium [Moles/?volume] in Serum or Plasma"
## unit nmol/L
d_LONIC_K <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_K,
                       c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_K <- unique(d_LONIC_K$LOINC) # "2823-3"
```

```
## Bilirubin
v_LabComponent_tbil <- unique(grep("bilirubin", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_tbil))
#*** currently using total bilirubin, but need to accomondate those with Gilbert syndrome (using direct bilirubin)
v_LabComponent_tbil <- "Bilirubin (Total), serum"
v_LabComponent_dbil <- "Bilirubin (Direct, Conjugated), serum"
# unit mg/dL
# total bilirubin
d_LONIC_tbil <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_tbil,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_tbil <- unique(d_LONIC_tbil$LOINC) # "1975-2"  "42719-5"

# direct bilirubin
d_LONIC_dbil <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_dbil,
                        c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_dbil <- unique(d_LONIC_dbil$LOINC) # "15152-2" "1968-7"


## ALT
v_LabComponent_ALT <- unique(grep("alanine aminotransferase", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_ALT))
## unit U/L
d_LONIC_ALT <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_ALT,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_ALT <- unique(d_LONIC_ALT$LOINC) # "1742-6" "1743-4" "1744-2"

## AST
v_LabComponent_AST <- unique(grep("aspartate aminotransferase", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_AST))
## unit U/L
d_LONIC_AST <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_AST,
                        c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
  arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_AST <- unique(d_LONIC_AST$LOINC) # "1920-8"  "30239-8"



## Finding ID of patients who have these tests
# 9a.  Absolute neutrophil count < 1.5 x 10^9 /L, platelet count < 100 x 10 9 /L and hemoglobin < 80 g/L
### absolute neutrophil
v_ID_abs_neutrophil_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                      rd_lab[rd_lab$LOINC %in% v_LOINC_abs_neutrophil, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                      by = "PatientID",
                      all.y = TRUE) %>%
  mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before mCPRC
# 4237

# Check none cleaned results
summary(as.factor(v_ID_abs_neutrophil_fail$TestResult))
test <- v_ID_abs_neutrophil_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# None numeric results: Test not performed ==> so just have to look at cleaned results)

v_ID_abs_neutrophil_fail <- v_ID_abs_neutrophil_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
  slice(which.max(TestDate))  # patients who had tests within 30 days of TimeZero
# 2816

# Find patients whose max absolute neutrophil count < 1.5 x 10^9 /L and exclude them
v_ID_abs_neutrophil_fail <- v_ID_abs_neutrophil_fail %>%
  filter(TestResultCleaned < 1.5) %>% # unit count*10^9/L
  select(PatientID) # 25
v_ID_abs_neutrophil_fail$f_timezero_abs_neutrophil_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_abs_neutrophil_fail, by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_abs_neutrophil_fail = ifelse(is.na(f_timezero_abs_neutrophil_fail), 0, f_timezero_abs_neutrophil_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_abs_neutrophil_fail)) # 25
```

```
### Plt
v_ID_platelet_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                        rd_lab[rd_lab$LOINC %in% v_LOINC_platelet, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                        by = "PatientID",
                        all.y = TRUE) %>%
  mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before TimeZero
# 5166

# Check none cleaned results
summary(as.factor(v_ID_platelet_fail$TestResult))
test <- v_ID_platelet_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# None numeric results: Acceptable, Decreased, Not reportable, Note_Comment, Test not performed ==> so just have to look at cleaned results)

v_ID_platelet_fail <- v_ID_platelet_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
  slice(which.max(TestDate))
# 3164

# Find patients whose max Platelet count < 100 x 10^9 /L and exclude them
v_ID_platelet_fail <- v_ID_platelet_fail %>%
  filter(TestResultCleaned < 100) %>% # unit count*10^9/L
  select(PatientID) # 58
v_ID_platelet_fail$f_timezero_platelet_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_platelet_fail, by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_platelet_fail  = ifelse(is.na(f_timezero_platelet_fail ), 0, f_timezero_platelet_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_platelet_fail)) # 58


### Hb
v_ID_Hb_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                      rd_lab[rd_lab$LOINC %in% v_LOINC_Hb, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                      by = "PatientID",
                      all.y = TRUE) %>%
  mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before TimeZero
# 5090

# Check none cleaned results
summary(as.factor(v_ID_Hb_fail$TestResult))
test <- v_ID_Hb_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# None numeric results: Test not performed ==> so just have to look at cleaned results)

v_ID_Hb_fail <- v_ID_Hb_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
  slice(which.max(TestDate))
# 3571

# Find patients whose max Hb < 80 g/L (8 g/dL) and exclude them
v_ID_Hb_fail <- v_ID_Hb_fail %>%
  filter(TestResultCleaned < 8) %>% # unit count*10^9/dL
  select(PatientID) # 46
v_ID_Hb_fail$f_timezero_Hb_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_Hb_fail, by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_Hb_fail  = ifelse(is.na(f_timezero_Hb_fail), 0, f_timezero_Hb_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_Hb_fail)) # 46

# Combined factors for neutropenia
v_ID_neutropenia <- rbind(v_ID_abs_neutrophil_fail, v_ID_platelet_fail, v_ID_Hb_fail) %>%
  select(PatientID) %>%
  distinct()
# 116
```

```
# Flag these combined factor in the GUTG cohort
d_TT_cohort_GUTG001 <- d_TT_cohort_GUTG001 %>%
  mutate(f_timezero_neutropenia = pmax(f_timezero_abs_neutrophil_fail, f_timezero_platelet_fail, f_timezero_Hb_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_neutropenia)) # 116



# 9b.  Creatinine clearance ≥ 30 ml/min (calculated by Cockcroft-Gault formula, see Appendix D)
# Find patients whose max CrCl < 30 ml/min and exclude them
# # Trial definition: Calculated creatinine clearance (Cockcroft-Gault formula):
# N x (140-Age) x weight in kg ÷ Serum Creatinine in µmol/L
# * For males N=1.23; for females N=1.04
# need weight ==> use CKD diagnosis for this

# Use diagnosis instead: CKD diagnosis within half-year before mCRPC (chronic disease assessment --> using a year)
# (assume patient come to hospital at least very half year ==> in line with other chronic disease survey in the trial)
# diagnosis table
v_ID_CKD_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 5) %in% c("N18.4", "N18.5", "N18.6", "N18.9")| # icd-10 (checked, no overlap icd-9 codes)
                  substr(DiagnosisCode, 1, 5) %in% c("585.4", "585.5", "585.6", "585.9")) # icd-9 # 1537 rows
summary(as.factor(v_ID_CKD_diag$DiagnosisCode))
# v_ID_AKI_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 3) %in% c("N17")| # icd-10 (checked, no overlap icd-9 codes
#                   substr(DiagnosisCode, 1, 5) %in% c("585.5", "585.7", "585.9")) # icd-9 # 1431 rows
# summary(as.factor(v_ID_AKI_diag$DiagnosisCode))
v_ID_CKD_diag <-  merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_CKD_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # active CKD within a year before TimeZero
  distinct(PatientID) # 75 patients
v_ID_CKD_diag$f_timezero_CKD_diag <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_CKD_diag, by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_CKD_diag  = ifelse(is.na(f_timezero_CKD_diag), 0, f_timezero_CKD_diag))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_CKD_diag)) # 75

# v_ID_AKI_diag <-  merge(d_TT_cohort_GUTG001[ , c("PatientID", "MCRPCDate")], v_ID_AKI_diag) %>%
#  mutate_at(vars(DiagnosisDate, MCRPCDate), as.Date, format="%Y-%m-%d") %>%
#  filter(DiagnosisDate >= (MCRPCDate-30) & DiagnosisDate <= MCRPCDate) %>% # active AKI within a month before MCRPDC
#  distinct(PatientID) # 3 patients

# v_ID_renal_fail <- unique(rbind(v_ID_CKD_diag, v_ID_AKI_diag)) # 37



# 9c.  Serum potassium > than lower limit of normal range (3.5 to 5.2 mEq/L) = (3.5 to 5.2 mmol/L)
# ==> Dr. Pezaro says these number flunctuates and likely to be retested in clinical practice ==> consider now exclude these
### K+
v_ID_K_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "MCRPCDate", "TimeZero")],
                rd_lab[rd_lab$LOINC %in% v_LOINC_K, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                by = "PatientID",
                all.y = TRUE) %>%
  mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before mCPRC
# 4547

# Check none cleaned results
summary(as.factor(v_ID_K_fail$TestResult))
test <- v_ID_K_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# None numeric results: Test not performed, Note_Comment, Pending ==> so just have to look at cleaned results)

v_ID_K_fail <- v_ID_K_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
  slice(which.max(TestDate))
# 2400

# Find patients whose min K < 3.5 mEq/L
v_ID_K_fail <- v_ID_K_fail %>%
  filter(TestResultCleaned < 3.5) %>% # unit mmol/L
  select(PatientID) # 55
v_ID_K_fail$f_timezero_K_fail <- 1

# Flag these patients in the GUTG cohort
```

```
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_K_fail, by = "PatientID", all.x = TRUE)  %>%
  mutate(f_timezero_K_fail  = ifelse(is.na(f_timezero_K_fail), 0, f_timezero_K_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_K_fail)) # 55


# 9d.  Total bilirubin ≤ 1.5 x upper limit of normal (ULN) except for
#     patients with known Gilbert's syndrome (direct bilirubin ≤ 1.5 x ULN)
# Normal bilirubin: 0.1 to 1.2 mg/dL (1.71 to 20.5 µmol/L)
v_ID_tbil_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                rd_lab[rd_lab$LOINC %in% v_LOINC_tbil, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                by = "PatientID",
                all.y = TRUE) %>%
  mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before Time Zero
# 4574

# Check none cleaned results
summary(as.factor(v_ID_tbil_fail$TestResult))
test <- v_ID_tbil_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# None numeric results: greatest: < 0.3, Negative, Test not performed, Note_Comment, Pending ==> so just have to look at cleaned results)

v_ID_tbil_fail <- v_ID_tbil_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
  slice(which.max(TestDate))
# 3423

# Find patients whose max total bilirubin > 1.8 mg/dL
v_ID_tbil_fail <- v_ID_tbil_fail %>%
  filter(TestResultCleaned > 1.8) %>% # unit mg/dL
  select(PatientID) # 15

# Gilbert's syndrome (chronic condition, so use half year)
v_ID_GilbertSyndrome_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 4) %in% c("E80.4")| # icd-10 (checked no overlap ICD-9 codes)
                      substr(DiagnosisCode, 1, 5) %in% c("277.4")) # icd-9 # 5 rows
summary(as.factor(v_ID_GilbertSyndrome_diag$DiagnosisCode))

v_ID_GilbertSyndrome_diag <-  merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_GilbertSyndrome_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>%
  distinct(PatientID) # 0 patients (no patients in the GUTG-001 cohort had Gilbert syndrome!)

# Flag patients
v_ID_tbil_fail$f_timezero_tbil_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_tbil_fail, by = "PatientID", all.x = TRUE)  %>%
  mutate(f_timezero_tbil_fail  = ifelse(is.na(f_timezero_tbil_fail), 0, f_timezero_tbil_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_tbil_fail)) # 15

# 9e.  Alanine aminotransferase (ALT) and aspartate aminotransferase (AST) ≤ 5 x ULN
# normal range; a normal AST level for adults is: 8 to 48 IU/L. A normal ALT level for adults is 7 to 55 IU/L.

## ALT
v_ID_ALT_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                rd_lab[rd_lab$LOINC %in% v_LOINC_ALT, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                by = "PatientID",
                all.y = TRUE) %>%
  mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
  filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before mCPRC
# 4617

# Check none cleaned results
summary(as.factor(v_ID_ALT_fail$TestResult))
test <- v_ID_ALT_fail %>%
  distinct(TestResult) %>%
  arrange(TestResult)
# all numbered results

v_ID_ALT_fail <- v_ID_ALT_fail %>%
  filter(!is.na(TestResultCleaned)) %>%
  group_by(PatientID) %>%
```

```
 slice(which.max(TestDate))
# 3446

# Find patients whose min ALT > 275 IU/L (upper limit: 55 U/L, 5 times upper limit: 275 IU/L)
v_ID_ALT_fail <- v_ID_ALT_fail %>%
 filter(TestResultCleaned > 275) %>% # unit U/L = IU/L (for ALT)
 select(PatientID) # 1

# Flag patients
v_ID_ALT_fail$f_timezero_ALT_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_ALT_fail, by = "PatientID", all.x = TRUE) %>%
 mutate(f_timezero_ALT_fail = ifelse(is.na(f_timezero_ALT_fail), 0, f_timezero_ALT_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_ALT_fail)) # 1

## AST
v_ID_AST_fail <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")],
                rd_lab[rd_lab$LOINC %in% v_LOINC_AST, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestResult")],
                by = "PatientID",
                all.y = TRUE) %>%
 mutate_at(vars(TestDate), as.Date, format="%Y-%m-%d") %>%
 filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before TimeZero
# 4634

# Check none cleaned results
summary(as.factor(v_ID_AST_fail$TestResult))
test <- v_ID_AST_fail %>%
 distinct(TestResult) %>%
 arrange(TestResult)
# None numbered results: greatest:<10, Pending, Test not performed ==> so can use cleaned results directly

v_ID_AST_fail <- v_ID_AST_fail %>%
 filter(!is.na(TestResultCleaned)) %>%
 group_by(PatientID) %>%
 slice(which.max(TestDate))
# 3496

# Find patients whose min AST > 240 IU/L (norm: 48 U/L, 5 times norm: 240 IU/L)
v_ID_AST_fail <- v_ID_AST_fail %>%
 filter(TestResultCleaned > 240) %>% # unit U/L = IU/L (for ALT)
 select(PatientID) # 1
# Flag patients
v_ID_AST_fail$f_timezero_AST_fail <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_AST_fail, by = "PatientID", all.x = TRUE) %>%
 mutate(f_timezero_AST_fail = ifelse(is.na(f_timezero_AST_fail), 0, f_timezero_AST_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_AST_fail)) # 1

# Flag a summary of poor liver function patients
v_ID_liver_fail <- rbind(v_ID_ALT_fail, v_ID_AST_fail) %>%
 select(PatientID) %>%
 distinct()
# 1 obs

# Flag these combined factor in the GUTG cohort
d_TT_cohort_GUTG001 <- d_TT_cohort_GUTG001 %>%
 mutate(f_timezero_liver_fail = pmax(f_timezero_ALT_fail, f_timezero_AST_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_liver_fail)) # 1


### Exclude these patients from the GUTG-001 cohort due to inadequate organ function

# Flag all patients with inadequate organ function
d_TT_cohort_GUTG001 <- d_TT_cohort_GUTG001 %>%
 mutate(f_timezero_inadequate_oragan = pmax(f_timezero_neutropenia, f_timezero_CKD_diag,
                                f_timezero_K_fail, f_timezero_tbil_fail,
                                f_timezero_liver_fail))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_inadequate_oragan)) # 249


# 8.  Eligible for treatment with either abiraterone acetate or enzalutamide as per
# standard of care guidelines
```

```
# Assume all patients at time zero receiving abieraterone and enzalutamide are as eligible as in the trial


# 10. Able to swallow study drug and comply with study requirements including
# provision of peripheral blood samples at specified time points for correlative
# studies

# cannot be tested ==> patients receiving chemo might not be able to swallow pills

# 11. Recovery from all prior treatment-related toxicity to grade ≤ 2 (as per CTCAE 4.0)

# ==> assuming that patients who were eligible for treatments has recovered (Dr. Pezaro said it should highly correlate with ECOG anyways)


##############################################################################################
###
### Exclusion criteria
### (The number of criteria were taken from the original GUTG-001 protocol)
###  https://classic.clinicaltrials.gov/ProvidedDocs/57/NCT02125357/Prot_SAP_000.pdf
###
##############################################################################################

## 1. Severe concurrent illness or co-morbid disease that would make the subject
## unsuitable for enrollment (excluding those inadequate organ functions)

## 5.  Active concurrent malignancy (with the exception of non-melanomatous skin
## cancer) (need to augment this)


# Find people having active concurrent malignancy within 30 days of MCRPC (prior) other than C61 (not counting secondary and lymph ones or
uncertain)
# https://www.icd10data.com/ICD10CM/Codes/C00-D49/C00-C14

# ICD-10: concurrent cancers
# Malignant neoplasms of lip, oral cavity and pharynx C00-C14
# Malignant neoplasms of digestive organs C15-C26
# Malignant neoplasms of respiratory and intrathoracic organs C30-C39
# Malignant neoplasms of bone and articular cartilage C40-C41
# Melanoma and other malignant neoplasms of skin C43-C44 & 4A
# Malignant neoplasms of mesothelial and soft tissue C45-C49
# Malignant neoplasms of breast C50-C50
# Malignant neoplasms of female genital organs C51-C58
# Malignant neoplasms of male genital organs C60-C63
# Malignant neoplasms of urinary tract C64-C68
# Malignant neoplasms of eye, brain and other parts of central nervous system C69-C72
# (C71: Malignant of brain, C72: Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system)
# Malignant neoplasms of thyroid and other endocrine glands C73-C75
# Malignant neuroendocrine tumors C7A-C7A
# Secondary neuroendocrine tumors C7B-C7B
# Malignant neoplasms of ill-defined, other secondary and unspecified sites C76-C80
# Malignant neoplasms of lymphoid, hematopoietic and related tissue: C81-96
# D ==> non-malignant tumours or unknown behavior tumour
# In situ neoplasms: D00-D09
# Neoplasms of uncertain behavior, polycythemia vera and myelodysplastic syndromes: D37-D48
# Benign neuroendocrine tumors: D3A-D3A
# Neoplasms of unspecified behavior: D49-D49


# Originally found a lot of: ICD-10 C79.51: secondary malignant neoplasm of bone C79.51 (now not going to exclude C79)
# Create a vector that holds all the ICD10 code for concurrent cancer
v_concurrent_primary_cancer_ICD10_3CHR <- c("00", "01", "02", "03", "04", "05",
                                            "06", "07", "08", "09",
                                            seq(10, 26, by = 1),
                                            seq(30, 34, by = 1),
                                            seq(37, 41, by = 1),
                                            "43",
                                            # except C44 other and unspecified malignant neoplasm of skin (non-melanoma skin cancer)
                                            "4A", # c4A: Merkel cell carcinoma
                                            seq(45, 58, by = 1),
                                            "60",
                                            # except C61 prostate cancer
                                            # ???? Malignant neoplasm of other and ill-defined sites: C76 (cannot rule out that they are meta due to prostate cancer)
                                            seq(62, 76, by = 1),
```

```
                    "7A",
                    # except: Secondary neuroendocrine tumors C7B,
                    # except: Secondary and unspecified malignant neoplasm of lymph nodes: C77 (cannot rule out that they are meta due to
prostate cancer)
                    # except: Secondary malignant neoplasm of respiratory and digestive organs: C78
                    # except: Secondary malignant neoplasm of other and unspecified sites: C79
                    # ?????? C80.1 Malignant (primary) neoplasm, unspecified
                    seq(80, 86, by = 1),
                    "88",
                    seq(90, 96, by = 1))
                    # D sections: neoplasm, non-malignant tumours
v_concurrent_primary_cancer_ICD10_3CHR  <- paste("C", v_concurrent_primary_cancer_ICD10_3CHR, sep="")

# ICD-9: concurrent cancers
# https://www.aapc.com/codes/icd9-codes-range/18/
# http://www.icd9data.com/2015/Volume1/default.htm (2015 ver)

# NEOPLASMS (140-239)
# MALIGNANT NEOPLASM OF LIP, ORAL CAVITY, AND PHARYNX (140-149)
# MALIGNANT NEOPLASM OF DIGESTIVE ORGANS AND PERITONEUM (150-159)
# MALIGNANT NEOPLASM OF RESPIRATORY AND INTRATHORACIC ORGANS (160-165)
# MALIGNANT NEOPLASM OF BONE, CONNECTIVE TISSUE, SKIN, AND BREAST (170-176)
# MALIGNANT NEOPLASM OF GENITOURINARY ORGANS (179-189)
# MALIGNANT NEOPLASM OF OTHER AND UNSPECIFIED SITES (190-199)
# MALIGNANT NEOPLASM OF LYMPHATIC AND HEMATOPOIETIC TISSUE (200-208)
# MALIGNANT NEUROENDOCRINE TUMORS (209)
# BENIGN NEOPLASMS (210-229)
# CARCINOMA IN SITU (230-234)
# NEOPLASMS OF UNCERTAIN BEHAVIOR (235-238)
# NEOPLASMS OF UNSPECIFIED NATURE (239)

# Create vectors (3 character and 4 character) that holds all the ICD9 code for concurrent cancer
v_concurrent_primary_cancer_ICD9_3CHR <- c(seq(140, 165, by = 1),
                        # except: 173 Other malignant neoplasm of skin (non melanoma)
                        seq(170, 172, by = 1),
                        # 185 malignant neoplasm of the prostate
                        seq(174, 176, by = 1), # There are no 177.x and 178.x
                        seq(179, 184, by = 1),
                        # except: Secondary and unspecified malignant neoplasm of lymph nodes: 196
                        # except: Secondary malignant neoplasm of respiratory and digestive systems: 197
                        # except: Secondary malignant neoplasm of other specified sites: 198
                        seq(186, 195, by = 1),
                        # ????: Malignant neoplasm without specification of site: 199
                        # There are more detailed codes within 209: MALIGNANT NEUROENDOCRINE TUMORS (209) (see 4CHR)
                        seq(199, 208, by = 1)
                        # 210-239: benign or non malignant tumours
                        )
v_concurrent_primary_cancer_ICD9_3CHR
v_concurrent_primary_cancer_ICD9_4CHR <- c(# 209.0 Malignant carcinoid tumors of the small intestine
                         # 209.1 Malignant carcinoid tumors of the appendix, large intestine, and rectum
                         # 209.2 Malignant carcinoid tumors of other and unspecified sites
                         # 209.3 Malignant poorly differentiated neuroendocrine tumors
                        "209.0", "209.1","209.2","209.3")
                         # except:
                         # 209.4 Benign carcinoid tumors of the small intestine
                         # 209.5 Benign carcinoid tumors of the appendix, large intestine, and rectum
                         # 209.6 Benign carcinoid tumors of other and unspecified sites
                         # 209.7 Secondary neuroendocrine tumors
v_concurrent_primary_cancer_ICD9_4CHR


# Find concurrent cancer diags (primary cancers)
v_ID_concurrent_cancer_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 3) %in% c(v_concurrent_primary_cancer_ICD10_3CHR,
                                                v_concurrent_primary_cancer_ICD9_3CHR) | # icd-10, icd-9: 3 character concurrent cancers
                                substr(DiagnosisCode, 1, 5) %in% c(v_concurrent_primary_cancer_ICD9_4CHR))
summary(as.factor(v_ID_concurrent_cancer_diag$DiagnosisCode)) # 17653

# Most common: concurrent cancer (assume they are active cuz the diag appear within half year)
# C67.9: Malignant neoplasm of bladder, unspecified
# C80.1: Malignant (primary) neoplasm, unspecified
# C199.1: Other malignant neoplasm without specification of site
# C90.00: Multiple myeloma not having achieved remission
# C91.10: Chronic lymphocytic leukemia of B-cell type not having achieved remission
# 188.9: Malignant neoplasm of bladder, part unspecified
```

```
# C18.9: Malignant neoplasm of colon, unspecified

# Originally found a lot of: secondary neoplasm, that that's almost a must because of the metastatic diagnosis
# ICD-10 C79.51: secondary malignant neoplasm of bone C79.51
# ICD-9 198.5: diagnosis of bone metastasis
# These are not popping out after adjusting codes


# Find Diagnosis containing these primary cancer diagnosis within 182 days prior to Time Zero
v_ID_concurrent_cancer_diag <-  merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_concurrent_cancer_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # tumor concurrent cancer prior to Timezero
  distinct(PatientID) # 132
v_ID_concurrent_cancer_diag$f_timezero_concurrent_cancer <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_concurrent_cancer_diag , by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_concurrent_cancer = ifelse(is.na(f_timezero_concurrent_cancer), 0, f_timezero_concurrent_cancer))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_concurrent_cancer)) # 132



## 7.  Brain metastases or active epidural disease
#     (treated epidural disease is permitted)

# brain meta codes: ICD-10
v_brain_meta_ICD10_4CHR <- c("C79.3") # Dr. Pezaro: ICD-10-CM: C79.31, C79.32 (Malignant neoplasm of spinal cord C72, already included in
primary concurrent cancer)

# brain meta codes: ICD-9
# Dr. Pezaro: ICD-9: 198.3 Secondary malignant neoplasm of brain and spinal cord
# I found: 198.4 Secondary malignant neoplasm of other parts of nervous (not necessary spinal cord meta though)
# (ICD-9 198 Secondary malignant neoplasm of other specified sites was not excluded)
v_brain_meta_ICD9_4CHR <- c("198.3", "198.4")

# Find concurrent cancer diags (primary cancers)
v_ID_brain_meta_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 5) %in% c(v_brain_meta_ICD10_4CHR,
                                                v_brain_meta_ICD9_4CHR)) # icd-10, icd-9: 4 character meta (and a dot)
summary(as.factor(v_ID_brain_meta_diag$DiagnosisCode)) # 1103

# Find Diagnosis containing these brain/spinal meta within 182 days prior to Time Zero
v_ID_brain_meta_diag <-  merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_brain_meta_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # tumor concurrent cancer prior to Timezero
  distinct(PatientID) # 11
v_ID_brain_meta_diag$f_timezero_brain_meta <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_brain_meta_diag , by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_brain_meta = ifelse(is.na(f_timezero_brain_meta), 0, f_timezero_brain_meta))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_brain_meta)) # 11



## 2.  Prior therapy with CYP17 inhibitors (including abiraterone acetate, TAK-
# 700, TOK-001 and ketoconazole), enzalutamide or other experimental anti-
# androgens (e.g. ARN-509 (apalutamide), TOK-001)

## 6.  Wide-field radiotherapy or radioisotopes such as Strontium-89 or Radium-
# 223 ≤ 28 days prior to starting study drug (limited-field palliative
#                                 radiotherapy for 1-5 fractions is permitted)

# 3.  Prior systemic chemotherapy for mCRPC

# Update the d_LOT again
d_LOT_GUTG001 <- d_LOT_GUTG001[d_LOT_GUTG001$PatientID %in% d_TT_cohort_GUTG001[ ,"PatientID"], ] # 13085

# Find treatment records to the user-defined first-line mCRPC treatment
d_LOT_GUTG001_treat_before_TimeZero <- d_LOT_GUTG001 %>%
    filter(is.na(LineNumber_mCRPC_UserDef)) %>% # Select rows without user-defined mCPRC treatment line
    select(-RegimenClass, -IsMaintenanceTherapy, -EnhancedCohort) %>%
    mutate_at(vars(StartDate, EndDate), as.Date, format="%Y-%m-%d")
# 1350 rows
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$LineName))
```

```
# 323 rows abiraterone, the others are mostly a combination of abiraterone with other drugs
# 49 rows Sipuleucel-T, 13 rows Radium-223, 2 rows Samarium Sm 153 Lexidronam
# Some had Clinical Study Drug


# There are also quite some systematic treatment for other type of cancer: check if excluding patient with the previous criteria would decrease these
# Probabily need to label other systematic treatment as well (though not typical for prostate cancer)
test <- d_TT_cohort_GUTG001 %>%
  filter(pmax(f_timezero_ADT_fail, f_timezero_inadequate_oragan,
         f_timezero_concurrent_cancer, f_timezero_brain_meta) == 0) # 4740 patients left
test <- d_LOT_GUTG001_treat_before_TimeZero [d_LOT_GUTG001_treat_before_TimeZero$PatientID %in% test[ ,"PatientID"], ] # 1207
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$LineName))


# There are also quite some clinical study drug
# --> could it be possible that the regimen is just extended because it's a combination of drugs that other drugs been added on?
# It's related to Flatiron's definition (technically some of them could actually receive abiraterone later but counted as the treatment for non-mCRPC)
# However Time Zero is defind as the next-line of treamtent anyways

# Create a list of previous treatments that need to be flagged
summary(as.factor(rd_treat_ABemit$DrugName))
summary(as.factor(rd_treat_oral$DrugName))
v_treat_not_allowed_pre_TimeZero <- c("Abiraterone", "Enzalutamide", "Ketoconazole", "Apalutamide",
                                      "Galeterone", "Orteronel", # these drugs don't seem to be used in the us but still included
                                      "Clinical Study Drug",
                                      "Samarium Sm 153 Lexidronam", "Radium-223", "Sipuleucel-T", "strontium-89",
                                      "Vipivotide", "Radiopharmaceutical", # not listed in the protocol but was available in the US later)
                                      "Darolutamide", "Larotrectinib", "Olaparib", "Relugolix", "Rucaparib", "Talazoparib") # approved after the trial
v_treat_allowed_pre_TimeZero_docetaxel <- c("Docetaxel")


# flag if regimen contain not allowed treatment, docetaxel or other systematic treatment
d_LOT_GUTG001_treat_before_TimeZero <- d_LOT_GUTG001_treat_before_TimeZero %>%
  rowwise() %>%
  mutate(f_pretreat_treat_not_allowed = max(str_detect(LineName, v_treat_not_allowed_pre_TimeZero))) %>%
  mutate(f_pretreat_treat_allowed_docetaxel = ifelse(LineName %in% v_treat_allowed_pre_TimeZero_docetaxel, 1, 0)) %>%
  mutate(f_pretreat_treat_other_systemic = ifelse(pmax(f_pretreat_treat_not_allowed, f_pretreat_treat_allowed_docetaxel) == 0, 1, 0)) %>%
  left_join(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero", "MCRPCDate")], by = "PatientID")
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$f_pretreat_treat_not_allowed)) # 0: 645; 1: 705
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$f_pretreat_treat_allowed_docetaxel)) # 0: 790; 1: 560
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$f_pretreat_treat_other_systemic))# 0: 1265; 1: 85
summary(as.factor(d_LOT_GUTG001_treat_before_TimeZero$LineSetting))
# pre-treatment for the setting based on Flatiron's definition: HSPC: 1, mCRPC: 451: mHSCP: 740; nmCRPC: 158

# Check unflagged patients, what type of systematic treatments they have
test <- test.data [test.data$f_pretreat_treat_not_allowed %in% 0 & test.data$f_pretreat_treat_allowed_docetaxel %in% 0, ]
summary(as.factor(test$LineName))


# Flag these treatment history in GUTG cohort
v_ID_treat_pre_TimeZero <- d_LOT_GUTG001_treat_before_TimeZero %>%
  select(PatientID, f_pretreat_treat_not_allowed, f_pretreat_treat_allowed_docetaxel, f_pretreat_treat_other_systemic) %>%
  group_by(PatientID) %>%
  summarise_at(c("f_pretreat_treat_not_allowed", "f_pretreat_treat_allowed_docetaxel", "f_pretreat_treat_other_systemic"), sum, na.rm = TRUE)
%>%
  mutate(f_pretreat_lines = f_pretreat_treat_not_allowed + f_pretreat_treat_allowed_docetaxel + f_pretreat_treat_other_systemic)
summary(as.factor(v_ID_treat_pre_TimeZero$f_pretreat_treat_not_allowed)) # 0: 550, 1: 574; 2: 48; 3: 9; 4: 2
summary(as.factor(v_ID_treat_pre_TimeZero$f_pretreat_treat_allowed_docetaxel)) # 0: 629; 1: 548; 2: 6 (some patients had two lines of docetaxel as monotherapy, perhaps because  re-initiating the same treatment more than 90 days later was considered a new line of therapy)
summary(as.factor(v_ID_treat_pre_TimeZero$f_pretreat_treat_other_systemic)) # 0: 1116; 1: 56; 2: 7; 3: 1; 4: 3
summary(as.factor(v_ID_treat_pre_TimeZero$f_pretreat_lines)) # 1: 1046; 2: 116; 3: 14; 4: 6; 6: 1
# Among 1183 treatments that had pre-treatments prior to mCRPCDate: 1046 had only 1 line; others had more than one line
# 633 patients had un-allowed treatments, the rest had only docetaxel or other systematic treatments


# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(data.frame(PatientID = d_TT_cohort_GUTG001[ , "PatientID"]), v_ID_treat_pre_TimeZero , by = "PatientID", all.x = TRUE) %>%
                replace(is.na(.), 0) %>%
                inner_join(d_TT_cohort_GUTG001, by = "PatientID") %>%
                relocate(f_pretreat_treat_not_allowed,f_pretreat_treat_allowed_docetaxel,
                     f_pretreat_treat_other_systemic, f_pretreat_lines,
                     .after = last_col()) # relocate these columns to the end of the table
```

```
length(d_TT_cohort_GUTG001$f_pretreat_treat_not_allowed[d_TT_cohort_GUTG001$f_pretreat_treat_not_allowed >= 1]) # 633 patients had pre-treatments that were not allowed



## 10. History of seizure or seizure disorder, or history of any cerebrovascular event

# ICD-10:
# G40  Epilepsy and recurrent seizures
# G45  Transient cerebral ischemic attacks and related syndromes (TIA is probably not significant enough?)
# I60-I69  Cerebrovascular diseases

# ICD-9
# 345 Epilepsy and recurrent seizures
# 430-438 Cerebrovascular Disease (435 Transient cerebral ischemia)


# Create a vector that holds all the ICD10 code for seizure
v_seizure_ICD10_3CHR <- c("G40")

# Create a vector that holds all the ICD9 code for seizure
v_seizure_ICD9_3CHR <- c("345")

# Find seizure diags
v_ID_seizure_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 3) %in% c(v_seizure_ICD10_3CHR, v_seizure_ICD9_3CHR)) # icd-10, icd-9
summary(as.factor(v_ID_seizure_diag$DiagnosisCode)) # 350

# Find Diagnosis containing these seizure diag within 182 days prior to Time Zero
v_ID_seizure_diag <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_seizure_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # seizure prior to TimeZero
  distinct(PatientID) # 4
v_ID_seizure_diag$f_timezero_seizure <- 1


# Create a vector that holds all the ICD10 code for cerebrovascular event (not including G45 as TIA is not specific)
v_cerebrovasc_event_ICD10_3CHR <- c(seq(60, 63, by = 1),
                        seq(65, 69, by = 1))
v_cerebrovasc_event_ICD10_3CHR <- paste("I", v_cerebrovasc_event_ICD10_3CHR, sep="")

# Create a vector that holds all the ICD9 code for cerebrovascular event (not including 435 TIA is not specific)
v_cerebrovasc_event_ICD9_3CHR <- c(seq(430, 434, by = 1),
                        seq(436, 438, by = 1))

# Find cerebral event diag
v_ID_cerebrovas_event_diag <- filter(rd_diag, substr(DiagnosisCode, 1, 3) %in% c(v_cerebrovasc_event_ICD10_3CHR,
v_cerebrovasc_event_ICD9_3CHR)) # icd-10, icd-9
summary(as.factor(v_ID_cerebrovas_event_diag$DiagnosisCode)) # 2619

# Find Diagnosis containing these seizure diag within 182 days prior to Time Zero
v_ID_cerebrovas_event_diag <- merge(d_TT_cohort_GUTG001[ , c("PatientID", "TimeZero")], v_ID_cerebrovas_event_diag) %>%
  mutate_at(vars(DiagnosisDate), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # Active recent cerebroevent prior to Timezero
  distinct(PatientID) # 35
v_ID_cerebrovas_event_diag$f_timezero_cerebrovas_event <- 1

# Flag these patients in the GUTG cohort
d_TT_cohort_GUTG001 <- merge(d_TT_cohort_GUTG001, v_ID_seizure_diag, by = "PatientID", all.x = TRUE) %>%
  mutate(f_timezero_seizure = ifelse(is.na(f_timezero_seizure), 0, f_timezero_seizure)) %>%
  left_join(v_ID_cerebrovas_event_diag, by = "PatientID") %>%
  mutate(f_timezero_cerebrovas_event = ifelse(is.na(f_timezero_cerebrovas_event), 0, f_timezero_cerebrovas_event)) %>%
  mutate(f_timezero_seizure_cerebrovas = pmax(f_timezero_seizure, f_timezero_cerebrovas_event))
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_seizure)) # 4
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_cerebrovas_event)) # 35
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_seizure_cerebrovas)) #38



## 11. Gastrointestinal disorder affecting absorption

# ==> Assume that patients having gastrointestinal disorder won't be prescribed with abiraterone and enzalutamide
# ==> not using ICD-9/10 diangosis as these are not particular enough to indicate severe GI disorder interfering absorption
```

```
## 12. Major surgery within 4 weeks of starting study treatment

# ==> There are no procedure codes in Flatiron
# ==> Using ICD-9-10 surgery diagnosis might detected non-active surgery record
# ==> Therefore, we assume all patients did not have major surgery if their ECOG > 2


## 4.  Life expectancy < 6 months
# ==> assume all patients had life expectancy > 6 because we cannot select patient based on their actual survival, which will create selection bias
# ==> Importantly, patients having ECOG > 2 are likely to be healthier with better expectancy

## 8.  Use of herbal products that may lower PSA level (e.g. saw palmetto)
# ==> Limitation: unknown in the database

## 9.  Contraindication to prednisone therapy including poorly controlled DM
# ambiguous criteria

# ==> unknown whether patient's DM is poorly controlled (HbA1C > 10?)
# Assume that patients who were contradicted to prednisolone won't take abiraterone anyways
# but because those that have contradiction to prednisone are often prescribed enzalutamide (per Dr. Pezaro)
# Dr. Pezaro suggested isolated DM as a confounder and adjust it at baseline
# Still, naturally there could be some unmeasured confounders (patients that were prescribed enzalutamide due to other contradiction to prednisone)


#########################################################################################################################
####################EXECUTING EXCLUSION CRITERIA USING FLAGS#############################################
#########################################################################################################################

# Select patients that did not meet any exclusion criteria
d_TT_cohort_GUTG001_final <- d_TT_cohort_GUTG001 %>%
  filter(pmax(f_timezero_ADT_fail, f_timezero_inadequate_oragan,
          f_timezero_concurrent_cancer, f_timezero_brain_meta,
          f_pretreat_treat_not_allowed, f_timezero_seizure_cerebrovas) == 0)
# 4144 patients

# Check the detail numbers of excluded patients from each detailed criteria
# ADT fail
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_ADT_fail)) # 37
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_inadequate_oragan)) # 249
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_concurrent_cancer)) # 132
summary(as.factor(d_TT_cohort_GUTG001$f_pretreat_treat_not_allowed)) # 633
summary(as.factor(d_TT_cohort_GUTG001$f_timezero_seizure_cerebrovas)) # 38


# Test the updated d_LOT and see the first-line treatment
d_LOT_GUTG001_final <- d_LOT_GUTG001[d_LOT_GUTG001$PatientID %in% d_TT_cohort_GUTG001_final[ ,"PatientID"], ] # 10007
summary(as.factor(d_LOT_GUTG001_final[d_LOT_GUTG001_final$LineNumber_mCRPC_UserDef %in% 1, "LineName"])) # First-line mCRPC: abi: 2172; enza 1972
summary(as.factor(d_LOT_GUTG001_final[d_LOT_GUTG001_final$LineNumber_mCRPC_UserDef %in% 2, "LineName"])) # Second-line mCRPC: enza: 716; abi: 537: doce: 447


# Save these cohorts: manually move them to "derived_data"
# write.csv(d_TT_cohort_GUTG001_final, "out\\d_TT_cohort_GUTG001_final_20231102.csv", row.names=FALSE) # patients excluded with the most strict criteria
# write.csv(d_LOT_GUTG001_final, "out\\d_LOT_GUTG001_final_20231102.csv", row.names=FALSE)# LOTs for patients excluded with the most strict criteria

# save identified cohort data
# save(d_TT_cohort_GUTG001_final, file = "derived_data\\d_TT_cohort_GUTG001_final.RData") # final use
# save(d_LOT_GUTG001_final, file = "derived_data\\d_LOT_GUTG001_final.RData") # final use
```

Appendix 8.6 R code for defining patient baseline characteristics in Prostate Cancer Case Study 1 (PC1)

```
################################################################################
###   Cohort Characteristics                                                  ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                     ###
###   Data: Flatiron prostate cancer datasets                                 ###
###                                                                           ###
###   University of Sheffield                                                 ###
###   Date created: Nov 2, 2023                                               ###
###   Created Git & renv project: Aug 01, 2023                                ###
################################################################################

## House keeping

# Load packages and functions
renv::status()
renv::snapshot()
`%notin%` <- Negate(`%in%`)

library(dplyr)
library(lubridate)
library(table1)

################################################
#
# GUTG-001 Analogue
#
################################################

##### Read files (from cohort identification)

# d_TT_cohort_GUTG001_final <- read.csv("derived_data\\d_TT_cohort_GUTG001_final_20231102.csv") # patients excluded with the strictest criteria
# d_LOT_GUTG001_final <- read.csv("derived_data\\d_LOT_GUTG001_final_20231102.csv")# LOTs for patients excluded with the strictest criteria

##### Baseline characteristics

# Merge with demographic and first-line therapy info
d_TT_cohort_GUTG001_final <- merge(d_TT_cohort_GUTG001_final, rd_demo, by = "PatientID", all.x = TRUE) %>%
  select(-f_timezero_ADT_fail, -f_timezero_abs_neutrophil_fail,
       -f_timezero_platelet_fail, -f_timezero_Hb_fail,
       -f_timezero_neutropenia, -f_timezero_K_fail,
       -f_timezero_tbil_fail, -f_timezero_ALT_fail,
       -f_timezero_AST_fail, -f_timezero_liver_fail,
       -f_timezero_concurrent_cancer, -f_timezero_brain_meta,
       -f_timezero_seizure, -f_timezero_cerebrovas_event,
       -f_timezero_seizure_cerebrovas, -f_timezero_CKD_diag,
       -f_timezero_inadequate_oragan, -f_pretreat_treat_not_allowed,
       -Gender) %>%
    inner_join(d_LOT_GUTG001_final[d_LOT_GUTG001_final$LineNumber_mCRPC_UserDef %in% 1,
                         c("PatientID", "LineName")], by = "PatientID")  %>%
    rename(FirstLine = LineName) %>%
    mutate(Age_TimeZero = as.numeric(year(TimeZero)) - BirthYear)

# Age by Group
tapply(d_TT_cohort_GUTG001_final$Age, d_TT_cohort_GUTG001_final$FirstLine, summary)
summary(d_TT_cohort_GUTG001_final$Age)
# Ethnicity by Group
tapply(as.factor(d_TT_cohort_GUTG001_final$Race), d_TT_cohort_GUTG001_final$FirstLine, summary)

# State by Group
tapply(as.factor(d_TT_cohort_GUTG001_final$MStage), d_TT_cohort_GUTG001_final$FirstLine, summary)



### Hb
d_demo_GUTG001_Hb <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                           rd_lab[rd_lab$LOINC %in% v_LOINC_Hb, c("PatientID", "TestDate", "LOINC", "TestResultCleaned")],
                           by = "PatientID",
                           all.y = TRUE) %>%
  mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
  dplyr::filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before Time Zero
# 3851
```

```
d_demo_GUTG001_Hb <- d_demo_GUTG001_Hb %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 group_by(PatientID) %>%
 slice(which.max(TestDate))
# 2809
summary(as.factor(d_demo_GUTG001_Hb$FirstLine)) # only 1495 (abi) and 1314 patient had records (enza)
tapply(d_demo_GUTG001_Hb$TestResultCleaned, d_demo_GUTG001_Hb$FirstLine,
    function(x) format(summary(x)))

### ECOG
d_demo_GUTG001_ECOG <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                      rd_baselineECOG[ , c("PatientID", "ECOGSource", "LineStartDate", "ECOGValue", "ECOGDate")],
                      by = "PatientID") %>%
 mutate_at(vars(LineStartDate, ECOGDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
 filter(LineStartDate <= TimeZero) %>%
 group_by(PatientID) %>%
 slice(which.max(LineStartDate)) %>%
 ungroup() %>%
 mutate(Gap_ECOGDate_TimeZero = ECOGDate - TimeZero) %>%
 filter(Gap_ECOGDate_TimeZero <= 0)
# 2424
summary(as.factor(d_demo_GUTG001_ECOG$FirstLine))
tapply(as.factor(d_demo_GUTG001_ECOG$ECOGValue), d_demo_GUTG001_ECOG$FirstLine,
    function(x) format(summary(x)))

### Previous treatment
tapply(as.factor(d_TT_cohort_GUTG001_final$f_pretreat_treat_allowed_docetaxel), d_TT_cohort_GUTG001_final$FirstLine,
    function(x) format(summary(x))) # 234 (abi), 232 (enza)
tapply(as.factor(d_TT_cohort_GUTG001_final$f_pretreat_treat_other_systemic), d_TT_cohort_GUTG001_final$FirstLine,
    function(x) format(summary(x))) # 30 (abi), 16 (enza)
tapply(as.factor(d_TT_cohort_GUTG001_final$f_pretreat_lines), d_TT_cohort_GUTG001_final$FirstLine,
    function(x) format(summary(x))) # 259 (abi), 247 (enza)

### PSA
# Check Labs for self-defined PSA: 300 unique labs (based on data inspection)
# Use LOINC to find those PSA ones
# https://loinc.org/2857-1
v_LOINC_all <- data.frame(sort(unique((rd_lab$LOINC)))) # confirmed that these LOINC table only contains numbers

# check the LOINC number of prostate sensitive antigen
# Create a vector of all LabComponent that can be mapped to prostate
v_LabComponent_PSA <- unique(grep("prostate", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_PSA))
# Therefore a different test name for the same LOINC, and there are LOINC for total PSA (2857-1, 35741-8), free PSA(10886-0), and ratio of free PSA
(12841-3)
# Create a table of all PSA tests
d_LONIC_PSA <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_PSA,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
 arrange(LOINC, LabComponent, TestBaseName, Test)
# Create a vecor containing all LOINC for total PSA
v_LOINC_PSA_all <- unique(d_LONIC_PSA$LOINC)
v_LOINC_PSA_t <- c("2857-1", "35741-8")

d_demo_GUTG001_PSA_t <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                     rd_lab[rd_lab$LOINC %in% v_LOINC_PSA_t, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestUnitsCleaned")],
                     by = "PatientID",
                     all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before Time Zero
# 2098
d_demo_GUTG001_PSA_t <- d_demo_GUTG001_PSA_t %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 group_by(PatientID) %>%
 slice(which.max(TestDate))
# 1701
summary(as.factor(d_demo_GUTG001_PSA_t$FirstLine)) # only 965 (abi) and 736 patient had records (enza)
tapply(d_demo_GUTG001_PSA_t$TestResultCleaned, d_demo_GUTG001_PSA_t$FirstLine,
    function(x) format(summary(x)))

# Check very extreme values
test <- d_demo_GUTG001_PSA_t %>%
 filter(TestResultCleaned > 2800) # only 10 pts, and have the same test units
```

```
head(d_demo_GUTG001_PSA_t)


### ALP
# check the LOINC number of ALP
# Create a vector of all LabComponent that can be mapped to prostate
v_LabComponent_ALP <- unique(grep("alkaline", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_ALP)) # only Alkaline phosphatase (ALP)
#-  Create a table of all ALP tests: only 6768-6 (one has unit U/L, the other doesn't)
d_LONIC_ALP <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_ALP,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
 arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_ALP <- unique(d_LONIC_ALP$LOINC)

d_demo_GUTG001_ALP <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                      rd_lab[rd_lab$LOINC %in% v_LOINC_ALP, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestUnitsCleaned")],
                      by = "PatientID",
                      all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before Time Zero
# 3572

d_demo_GUTG001_ALP <- d_demo_GUTG001_ALP %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 group_by(PatientID) %>%
 slice(which.max(TestDate))
# 2745
summary(as.factor(d_demo_GUTG001_ALP$FirstLine)) # only 1471 (abi) and 1274 patient had records (enza)
tapply(d_demo_GUTG001_ALP$TestResultCleaned, d_demo_GUTG001_ALP$FirstLine,
    function(x) format(summary(x)))

# Check very extreme values: normal value 30-130 IU/L
# https://www.nbt.nhs.uk/severn-pathology/requesting/test-information/alp
test <- d_demo_GUTG001_ALP %>%
 filter(TestResultCleaned > 195) # 503 pts, over 1.5 times of normal value
head(d_demo_GUTG001_ALP)


### LDH
# check the LOINC number of LDH
# Create a vector of all LabComponent that can be mapped to prostate
v_LabComponent_LDH <- unique(grep("lactate", rd_lab$LabComponent, ignore.case = TRUE, value=TRUE))
summary(as.factor(v_LabComponent_LDH)) # only Lactate dehydrogenase (LDH), serum
#-  Create a table of all LDH tests: 14804-9, 2532-0
d_LONIC_LDH <- unique(rd_lab[rd_lab$LabComponent %in% v_LabComponent_LDH,
                      c("LOINC", "Test", "LabComponent", "TestBaseName", "LabSource", "TestUnitsCleaned")]) %>%
 arrange(LOINC, LabComponent, TestBaseName, Test)
v_LOINC_LDH <- unique(d_LONIC_LDH$LOINC)

d_demo_GUTG001_LDH <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                    rd_lab[rd_lab$LOINC %in% v_LOINC_LDH, c("PatientID", "TestDate", "LOINC", "TestResultCleaned", "TestUnitsCleaned")],
                    by = "PatientID",
                    all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate >= (TimeZero-30) & TestDate <= TimeZero) # records within 30 days before Time Zero
# 558

d_demo_GUTG001_LDH <- d_demo_GUTG001_LDH %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 group_by(PatientID) %>%
 slice(which.max(TestDate))
# 456
summary(as.factor(d_demo_GUTG001_LDH$FirstLine)) # only 229 (abi) and 227 patient had records (enza)
tapply(d_demo_GUTG001_LDH$TestResultCleaned, d_demo_GUTG001_LDH$FirstLine,
    function(x) format(summary(x)))

# Check very extreme values: normal value 135-225 IU/L
# https://www.yorkhospitals.nhs.uk/our-services/a-z-of-services/lab-med/test-directory/clinical-biochemistry/lactate-dehydrogenase/
test <- d_demo_GUTG001_LDH %>%
 filter(TestResultCleaned > 450) # 30 pts, over 2 times of normal value
head(d_demo_GUTG001_LDH)
```

```
# DM
# brain meta codes: ICD-10
# E10  Type 1 diabetes mellitus
# E11  Type 2 diabetes mellitus
# E13  Other specified diabetes mellitus
v_DM_ICD10_3CHR <- c("E10", "E11", "E13")

# brain meta codes: ICD-9
# 250 Diabetes mellitus
v_DM_ICD9_3CHR <- c("250")

# Find concurrent cancer diags (primary cancers)
d_demo_GUTG001_DM <- filter(rd_diag, substr(DiagnosisCode, 1, 3) %in% c(v_DM_ICD10_3CHR,
                                          v_DM_ICD9_3CHR)) # icd-10, icd-9: 4 character meta (and a dot)
summary(as.factor(d_demo_GUTG001_DM$DiagnosisCode)) # 16770

# Find Diagnosis containing these brain/spinal meta within 182 days prior to Time Zero
d_demo_GUTG001_DM <-  merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "TimeZero")], d_demo_GUTG001_DM) %>%
  mutate_at(vars(DiagnosisDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero) %>% # DM within 182 days prior to Timezero
  distinct(PatientID) # 191
d_demo_GUTG001_DM$f_timezero_DM <- 1


### Gleason score (at initial diagnosis)


### SES: social economic status: each pt has one record (is it measured before they started treatment? can they change?)
d_demo_GUTG001_SES <-  d_TT_cohort_GUTG001_final %>%
  select(PatientID) %>%
  left_join(rd_socialdeterm, by = "PatientID") %>%
  mutate(f_SES = case_when(SESIndex2015_2019 == "" ~ "Unknown / Not documented",
                    TRUE ~ SESIndex2015_2019)) %>%
  select(PatientID, f_SES)
# 4144
summary(as.factor(d_demo_GUTG001_SES$f_SES)) # 4144: 448 missing, others from 1~5


### Comorbidity Score
# Adapted from Comorbidity Score for claims data from the
# Division of Pharmacoepidemiology and Pharmacoeconomics Department of Medicine, Harvard Medical School
# https://www.drugepi.org/dope/software#Combined1
# ref: https://pubmed.ncbi.nlm.nih.gov/21208778/
# ref: https://pubmed.ncbi.nlm.nih.gov/29087983/

## Create a list of ICD codes to identify comorbidities
#- Alcohol Abuse
v_Comorbid_Alcohol_abuse_ICD9_4CHR <- c("291.1", "291.2", "291.5", "291.8", "291.9", "303.9", "305.0", "V11.3")

v_Comorbid_Alcohol_abuse_ICD10_3CHR <- c("E52", "F10", "T51")
v_Comorbid_Alcohol_abuse_ICD10_4CHR <- c("G62.1", "I42.6", "K29.2", "K70.0", "K70.3", "K70.9", "Z65.8", "Z71.4")

#- Any Tumour
# No need to examine "Any Tumour" (prostate cancer is already tumour)

#- Cardiac arrythmia
v_Comorbid_Cardiac_arrhythmias_ICD9_4CHR <- c("426.2", "426.3", "426.4",
                            "426.6", "426.7", "426.8",
                            "427.0", "427.2", "427.6", "427.9",
                            "785.0", "V45.0", "V53.3")
v_Comorbid_Cardiac_arrhythmias_ICD9_5CHR <- c("426.10", "426.11", "426.13",
                            "426.50", "426.51", "426.52", "426.53",
                            "427.31")

v_Comorbid_Cardiac_arrhythmias_ICD10_3CHR <- c("I44", "I47", "I48", "I49")
v_Comorbid_Cardiac_arrhythmias_ICD10_4CHR <- c("I44.0", "I44.1",
                             "I44.3", "I44.5", "I44.6", "I44.7",
                             "I45.0", "I45.1", "I45.2",
                             "I45.4", "I45.5", "I45.6", "I45.7", "I45.8", "I45.9",
                             "R00.0", "R00.1", "R00.8",
                             "T82.1", "Z45.0", "Z95.0", "Z95.9")
v_Comorbid_Cardiac_arrhythmias_ICD10_6CHR <- c("Z95.810", "Z95.818")
```

```
#- Chronic pulmonary disease
v_Comorbid_Chronic_pulmonary_disease_ICD9_3CHR <- c("491", "492", "493", "494", "496")
v_Comorbid_Chronic_pulmonary_disease_ICD9_4CHR <- c("415.0", "416.8", "416.9")

v_Comorbid_Chronic_pulmonary_disease_ICD10_3CHR <- c("J40", "J41", "J42", "J43", "J44", "J45", "J47",
                                        "J60", "J61", "J62", "J63", "J64", "J65", "J66", "J67")
v_Comorbid_Chronic_pulmonary_disease_ICD10_4CHR <- c("I26.0", "I27.2", "I27.8", "I27.9",
                                         "J68.4", "J70.1", "J70.3")

# Coagulopathy
v_Comorbid_Coagulopathy_ICD9_3CHR <- c("286")
v_Comorbid_Coagulopathy_ICD9_4CHR <- c("287.1", "287.3", "287.4", "287.5")

v_Comorbid_Coagulopathy_ICD10_3CHR <- c("D65", "D66", "D67", "D68")
v_Comorbid_Coagulopathy_ICD10_4CHR <- c("D69.1", "D69.3", "D69.4", "D69.5", "D69.6")

# Complicated_diabetes
v_Comorbid_Complicated_DM_ICD9_4CHR <- c("250.4", "250.5", "250.6", "250.7")
v_Comorbid_Complicated_DM_ICD9_5CHR <- c("250.90", "250.91", "250.92", "250.93")

v_Comorbid_Complicated_DM_ICD10_4CHR <- c("E10.2", "E10.3", "E10.4", "E10.5",
                        "E10.6", "E10.7", "E10.8",
                        "E12.2", "E12.3", "E12.4", "E12.5",
                        "E12.6", "E12.7", "E12.8",
                        "E13.2", "E13.3", "E13.4", "E13.5",
                        "E13.6", "E13.7", "E13.8")

# CHF
v_Comorbid_CHF_ICD9_3CHR <- c("425", "428")
v_Comorbid_CHF_ICD9_4CHR <- c("429.3")
v_Comorbid_CHF_ICD9_5CHR <- c("402.01", "402.11", "402.91")

v_Comorbid_CHF_ICD10_3CHR <- c("I42", "I43", "I50")
v_Comorbid_CHF_ICD10_4CHR <- c("I09.9", "I11.0","I13.0", "I13.2", "I25.5",
                       "I51.7", "P29.0")
v_Comorbid_CHF_ICD10_5CHR <- c("A18.84")

# Deficiency_anemia
v_Comorbid_Deficiency_anemia_ICD9_4CHR <- c("280.1", "280.8", "280.9", "285.9")

v_Comorbid_Deficiency_anemia_ICD10_3CHR <- c("D51", "D52", "D53")
v_Comorbid_Deficiency_anemia_ICD10_4CHR <- c("D50.1", "D50.8", "D50.9", "D64.9")

# Dementia
v_Comorbid_Dementia_ICD9_3CHR <- c("290")
v_Comorbid_Dementia_ICD9_4CHR <- c("331.0", "331.1", "331.2")

v_Comorbid_Dementia_ICD10_3CHR <- c("F01", "F02", "F03", "F05",
                           "G30")
v_Comorbid_Dementia_ICD10_4CHR <- c("G31.1")
v_Comorbid_Dementia_ICD10_5CHR <- c("G31.01", "G31.09")

# Fluid and electrolyte disorders
v_Comorbid_Fluid_Electro_ICD9_3CHR <- c("276")

v_Comorbid_Fluid_Electro_ICD10_3CHR <- c("E86", "E87")
v_Comorbid_Fluid_Electro_ICD10_4CHR <- c("E22.2")

# HIV
v_Comorbid_HIV_ICD9_3CHR <- c("042", "043", "044")

v_Comorbid_HIV_ICD10_3CHR <- c("B20")

# Hemiplegia
v_Comorbid_Hemiplegia_ICD9_3CHR <- c("342", "344")

v_Comorbid_Hemiplegia_ICD10_3CHR <- c("G81", "G82", "G83")
v_Comorbid_Hemiplegia_ICD10_4CHR <- c("G04.1", "G04.1", "G11.4",
                             "G80.1", "G80.2")

# HTN
v_Comorbid_HTN_ICD9_4CHR <- c("401.1", "401.9")
v_Comorbid_HTN_ICD9_5CHR <- c("402.10", "402.90", "404.10", "404.9",
                      "405.11", "405.19", "405.91", "405.99")
```

```
v_Comorbid_HTN_ICD10_3CHR <- c("I10", "I11", "I12", "I13", "I15")
v_Comorbid_HTN_ICD10_4CHR <- c("N26.2")

# Liver disease
v_Comorbid_LiverD_ICD9_4CHR <- c("456.0", "456.1",
                        "571.0", "571.2", "571.3", "571.4",
                        "571.5", "571.6", "571.8", "571.9",
                        "572.3", "572.8", "V42.7")
v_Comorbid_LiverD_ICD9_5CHR <- c("070.32", "070.33", "070.54",
                        "456.20", "456.21")

v_Comorbid_LiverD_ICD10_3CHR <- c("B18", "I85", "K70", "K73", "K74")
v_Comorbid_LiverD_ICD10_4CHR <- c("I86.4",
                        "K71.1", "K71.3", "K71.4", "K71.5", "K71.7",
                        "K72.1", "K72.9", "K75.4",
                        "K76.0", "K76.2", "K76.3", "K76.4", "K76.5", "K76.6",
                        "K76.7", "K76.8", "K76.9", "Z94.4")
v_Comorbid_LiverD_ICD10_5CHR <- c("K75.81", "Z48.23")

# Metastatic Cancer (all pts = 1)

# Peripheral vascular disease
v_Comorbid_PeripheralVasc_ICD9_3CHR <- c("440")
v_Comorbid_PeripheralVasc_ICD9_4CHR <- c("441.2", "441.4", "441.7", "441.9",
                           "443.1", "443.2", "443.8", "443.9",
                           "447.1", "557.1", "557.9", "V43.4")

v_Comorbid_PeripheralVasc_ICD10_3CHR <- c("I70", "I71", "I79")
v_Comorbid_PeripheralVasc_ICD10_4CHR <- c("I67.0", "I73.1", "I73.8", "I73.9",
                            "I77.1",
                            "K55.1", "K55.8", "K55.9",
                            "Z95.9")
v_Comorbid_PeripheralVasc_ICD10_5CHR <- c("E08.51", "E08.52", "E09.51", "E09.52",
                            "E10.51", "E10.52", "E11.51",
                            "E13.51", "E11.52",
                            "I77.71", "I77.72", "I77.73", "I77.74",
                            "I77.79",
                            "Z95.82")

# Psychosis
v_Comorbid_Psychosis_ICD9_3CHR <- c("295", "296", "297", "298")
v_Comorbid_Psychosis_ICD9_4CHR <- c("299.1") # original Harvard code has 299.11 but this was not found in ICD-9 dictionary

v_Comorbid_Psychosis_ICD10_3CHR <- c("F20", "F22", "F23", "F24", "F25",
                          "F28", "F29",
                          "F30", "F31", "F32", "F33", "F39")
v_Comorbid_Psychosis_ICD10_4CHR <- c("F34.8", "F34.9", "F84.3")
v_Comorbid_Psychosis_ICD10_5CHR <- c("F44.89")

# Pulmonary circulation disorders
v_Comorbid_PulmonaryCirc_ICD9_3CHR <- c("416")
v_Comorbid_PulmonaryCirc_ICD9_4CHR <- c("417.9")

v_Comorbid_PulmonaryCirc_ICD10_3CHR <- c("I26", "I27")
v_Comorbid_PulmonaryCirc_ICD10_4CHR <- c("I28.0", "I28.8", "I28.9")

# Renal Failure
v_Comorbid_RenalFail_ICD9_3CHR <- c("585", "586")
v_Comorbid_RenalFail_ICD9_4CHR <- c("V42.0", "45.1", "V56.0", "V56.8")
v_Comorbid_RenalFail_ICD9_5CHR <- c("03.11", "403.91", "404.12", "404.92")

v_Comorbid_RenalFail_ICD10_3CHR <- c("I13", "N18", "N19")
v_Comorbid_RenalFail_ICD10_4CHR <- c("I12.0",
                          "N03.2", "N03.3", "N03.4", "N03.5",
                          "N03.6", "N03.7",
                          "N05.2", "N05.3", "N05.4", "N05.5",
                          "N05.6", "N05.7",
                          "N25.0",
                          "Z49.0", "Z94.0", "Z99.2")
v_Comorbid_RenalFail_ICD10_5CHR <- c(# Harvard code has Z39.32 - there doesn't seem to be Z39.32?
                          "Z48.22", "Z49.31",
                          "Z91.15")
# Weight loss
```

```
v_Comorbid_WeightLoss_ICD9_3CHR <- c("260", "261", "262", "263")

v_Comorbid_WeightLoss_ICD10_3CHR <- c("E40", "E41", "E42", "E43",
                              "E44", "E45", "E46",
                              "R64")
v_Comorbid_WeightLoss_ICD10_4CHR <- c("E64.0", "R63.4")


# Put all Comorbid codes into a list (73 vectors)
l_Comorbid_ICDs <- mget(ls(pattern = "^v_Comorbid_.*ICD"))
rm(list = ls(pattern = "^v_Comorbid_.*ICD"))

# Create a vector of comorbid_names
v_comorbid_names <- sapply(names(l_Comorbid_ICDs), function(x) {
  match <- gsub("^v_Comorbid_([A-Za-z_]+)_ICD.*$", "\\1", x)
  if (match == x) {
    return(NA)  # Return NA if no match is found
  } else {
    return(match)
  }
})
v_comorbid_names <- unique(v_comorbid_names)
# 18 names: except any tumour and metastatsis

# Create a function to label these diagnosis data with flags of comorbidities
label_comorbid <- function(data){
  for(comorbid in v_comorbid_names){

    f_comorbidities_name <- paste("Comorbid_", comorbid, sep = "")

    data <- data %>%
      mutate(!!sym(f_comorbidities_name) :=
               case_when(
                 # ICD-9: 3 character (check if such ICD code exist first)
                 {
                   icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD9_3CHR$", sep = ""), names(l_Comorbid_ICDs))
                   if (length(icd_index) > 0) {
                     icd_list <- l_Comorbid_ICDs[[icd_index]]
                     DiagnosisCodeSystem == "ICD-9-CM" & substr(DiagnosisCode, 1, 3) %in% icd_list
                   } else FALSE
                 } ~ 1,
                 # ICD-9: 4 character, including comma
                 {
                   icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD9_4CHR$", sep = ""), names(l_Comorbid_ICDs))
                   if (length(icd_index) > 0) {
                     icd_list <- l_Comorbid_ICDs[[icd_index]]
                     DiagnosisCodeSystem == "ICD-9-CM" & substr(DiagnosisCode, 1, 5) %in% icd_list
                   } else FALSE
                 } ~ 1,
                 # ICD-9: 5 character, including comma
                 {
                   icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD9_5CHR$", sep = ""), names(l_Comorbid_ICDs))
                   if (length(icd_index) > 0) {
                     icd_list <- l_Comorbid_ICDs[[icd_index]]
                     DiagnosisCodeSystem == "ICD-9-CM" & substr(DiagnosisCode, 1, 6) %in% icd_list
                   } else FALSE
                 } ~ 1,
                 # ICD-9: 6 character, including comma
                 {
                   icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD9_6CHR$", sep = ""), names(l_Comorbid_ICDs))
                   if (length(icd_index) > 0) {
                     icd_list <- l_Comorbid_ICDs[[icd_index]]
                     DiagnosisCodeSystem == "ICD-9-CM" & substr(DiagnosisCode, 1, 7) %in% icd_list
                   } else FALSE
                 } ~ 1,

                 # ICD-10: 3 character
                 {
                   icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD10_3CHR$", sep = ""), names(l_Comorbid_ICDs))
                   if (length(icd_index) > 0) {
                     icd_list <- l_Comorbid_ICDs[[icd_index]]
                     DiagnosisCodeSystem == "ICD-10-CM" & substr(DiagnosisCode, 1, 3) %in% icd_list
                   } else FALSE
                 } ~ 1,
```

```
            # ICD-10: 4 character, including comma
            {
              icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD10_4CHR$", sep = ""), names(l_Comorbid_ICDs))
              if (length(icd_index) > 0) {
                icd_list <- l_Comorbid_ICDs[[icd_index]]
                DiagnosisCodeSystem == "ICD-10-CM" & substr(DiagnosisCode, 1, 5) %in% icd_list
              } else FALSE
            } ~ 1,
            # ICD-10: 5 character, including comma
            {
              icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD10_5CHR$", sep = ""), names(l_Comorbid_ICDs))
              if (length(icd_index) > 0) {
                icd_list <- l_Comorbid_ICDs[[icd_index]]
                DiagnosisCodeSystem == "ICD-10-CM" & substr(DiagnosisCode, 1, 6) %in% icd_list
              } else FALSE
            } ~ 1,
            # ICD-10: 6 character, including comma
            {
              icd_index <- grep(paste("^v_Comorbid_", comorbid, "_ICD10_6CHR$", sep = ""), names(l_Comorbid_ICDs))
              if (length(icd_index) > 0) {
                icd_list <- l_Comorbid_ICDs[[icd_index]]
                DiagnosisCodeSystem == "ICD-10-CM" & substr(DiagnosisCode, 1, 7) %in% icd_list
              } else FALSE
            } ~ 1,
            TRUE ~ 0
          )
    )
  }
  return(data)
}


# Find co-morbid Diagnosis within 182 days prior to Time Zero
d_demo_GUTG001_CoMorbid <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "TimeZero")], rd_diag, by = "PatientID") %>%
  select(-PracticeID) %>%
  mutate_at(vars(DiagnosisDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate >= (TimeZero-182) & DiagnosisDate <= TimeZero)

# Summaries each patient's morbidity flag using the function created
d_demo_GUTG001_CoMorbid <- label_comorbid(d_demo_GUTG001_CoMorbid) %>%
  select(-TimeZero, -DiagnosisDate, -DiagnosisCode,
         -DiagnosisDescription, -DiagnosisCodeSystem) %>%
  group_by(PatientID) %>%
  summarise_all(max, na.rm = TRUE) %>%
  rename_with(~ paste0(., "_TimeZero"), -PatientID)

# Label every patient's cormorbidity score, adding any tumour and metastatsis = 1
d_demo_GUTG001_CoMorbid <- d_TT_cohort_GUTG001_final[ , "PatientID", drop = FALSE] %>%
  left_join(d_demo_GUTG001_CoMorbid, by = "PatientID") %>%
  mutate(across(everything(), ~ ifelse(is.na(.), 0, .))) %>%
  mutate(Comorbid_AnyTumour_TimeZero = 1,
         Comorbid_Metastatic_cancer_TimeZero = 1) %>%
  # Combined comorbidity score conditions and weights for a Medicare population (Gagne 2011)
  # Not necessarily suitable for other population
  mutate(Comorbid_Score_TimeZero =
         5*Comorbid_Metastatic_cancer_TimeZero +
         2*Comorbid_CHF_TimeZero +
         2*Comorbid_Dementia_TimeZero +
         2*Comorbid_RenalFail_TimeZero +
         2*Comorbid_WeightLoss_TimeZero +
          Comorbid_AnyTumour_TimeZero +
          Comorbid_Alcohol_abuse_TimeZero +
          Comorbid_Cardiac_arrhythmias_TimeZero +
          Comorbid_Chronic_pulmonary_disease_TimeZero +
          Comorbid_Coagulopathy_TimeZero +
          Comorbid_Complicated_DM_TimeZero +
          Comorbid_Deficiency_anemia_TimeZero +
          Comorbid_Fluid_Electro_TimeZero +
          Comorbid_Hemiplegia_TimeZero +
          Comorbid_LiverD_TimeZero +
          Comorbid_PeripheralVasc_TimeZero +
          Comorbid_Psychosis_TimeZero +
          Comorbid_PulmonaryCirc_TimeZero +
```

```
        (-1 * Comorbid_HIV_TimeZero) + # There is are patients having HIV but without other non-any tumour, non-meta comorbid diagnosis within
182 days
        (-1 * Comorbid_HTN_TimeZero) # not sure why HTN os -1
       )

summary(as.factor(d_demo_GUTG001_CoMorbid$Comorbid_Score_TimeZero)) # ranging from 5~13

####################################
# Prepare Table 1
####################################

# Factor the basic variables that
# we're interested in
table1_TimeZero <- d_TT_cohort_GUTG001_final %>%
  mutate_at(vars(TimeZero, MCRPCDate,
            MetDiagnosisDate, DiagnosisDate, CRPCDate), as.Date, format="%Y-%m-%d") %>%
  mutate(Gap_TimeZero_mCRPC = as.numeric(TimeZero - MCRPCDate),
         Gap_TimeZero_mPC = as.numeric(TimeZero - MetDiagnosisDate),
         Gap_TimeZero_PC = as.numeric(TimeZero - DiagnosisDate),
         Gap_TimeZero_CRPC = as.numeric(TimeZero - CRPCDate)) %>%
  left_join(d_demo_GUTG001_PSA_t[ , c("PatientID", "TestResultCleaned")], by = "PatientID") %>%
  rename(PSA_TimeZero = TestResultCleaned)  %>%
  left_join(d_demo_GUTG001_Hb[ , c("PatientID", "TestResultCleaned")], by = "PatientID") %>%
  rename(Hb_TimeZero = TestResultCleaned) %>%
  mutate(Hb_TimeZero = Hb_TimeZero*10) %>% # Change unit to be in line with Khalaf 2019 %>%
  left_join(d_demo_GUTG001_ALP[ , c("PatientID", "TestResultCleaned")], by = "PatientID") %>%
  rename(ALP_RULN_TimeZero = TestResultCleaned)  %>%
  mutate(ALP_RULN_TimeZero = ALP_RULN_TimeZero/130) %>% # relative to ULN
  left_join(d_demo_GUTG001_LDH[ , c("PatientID", "TestResultCleaned")], by = "PatientID") %>%
  rename(LDH_RULN_TimeZero = TestResultCleaned)  %>% # relative to ULN
  mutate(LDH_RULN_TimeZero = LDH_RULN_TimeZero/225) %>% # relative to ULN
    left_join(d_demo_GUTG001_ECOG[ , c("PatientID", "ECOGValue")], by = "PatientID") %>%
  rename(ECOG_TimeZero = ECOGValue) %>%
  left_join(d_demo_GUTG001_DM[ , c("PatientID", "f_timezero_DM")], by = "PatientID") %>%
  mutate(f_timezero_DM = ifelse(is.na(f_timezero_DM), 0, f_timezero_DM)) %>%
  rename(DM_TimeZero = f_timezero_DM) %>%
  left_join(d_demo_GUTG001_SES[ , c("PatientID", "f_SES")], by = "PatientID") %>%
  left_join(d_demo_GUTG001_CoMorbid[ , c("PatientID", "Comorbid_Score_TimeZero")], by = "PatientID")

# Race relabel
table1_TimeZero <- table1_TimeZero %>%
  mutate(Race = case_when(
    Race == "" ~ "Unknown / Not documented",
    Race == "Hispanic or Latino" ~ "Other Race",
    TRUE ~ Race)) %>%
  mutate(Race = factor(Race, levels = c("Asian", "Black or African American", "White", "Other Race", "Unknown / Not documented")))

# Previous treatment relabel
table1_TimeZero <- table1_TimeZero %>%
  mutate(f_pretreat_lines = case_when(
    f_pretreat_lines > 1 ~ "> 1",
    TRUE ~ as.character(f_pretreat_lines))) %>%
  mutate(f_pretreat_lines = factor(f_pretreat_lines, levels = c("0", "1", "> 1")))

table1_TimeZero$f_pretreat_treat_allowed_docetaxel <- factor(
  ifelse(table1_TimeZero$f_pretreat_treat_allowed_docetaxel > 0, "Yes", "No"),
  levels = c("Yes", "No"))

table1_TimeZero$f_pretreat_treat_other_systemic <- factor(
  ifelse(table1_TimeZero$f_pretreat_treat_other_systemic > 0, "Yes", "No"),
  levels = c("Yes", "No"))

# DM relabel
table1_TimeZero$DM_TimeZero <-
  factor(table1_TimeZero$DM_TimeZero, levels=c(1,0),
         labels=c("Yes", "No"))

# Relabel Group stage
table1_TimeZero <- table1_TimeZero %>%
  mutate(GroupStage = case_when(
    GroupStage %in% c("IIA", "IIB", "IIC") ~ "II",
    GroupStage %in% c("IIIA", "IIIB", "IIIC") ~ "III",
    GroupStage %in% c("IVA", "IVB") ~ "IV",
    TRUE ~ GroupStage))
```

```
# Relabel T stage
table1_TimeZero <- table1_TimeZero %>%
  mutate(TStage = case_when(
    TStage %in% c("T0", "T1", "T1a", "T1b", "T1c") ~ "T0/T1",
    TStage %in% c("T2a", "T2b", "T2c") ~ "T2",
    TStage %in% c("T3a", "T3b", "T3c") ~ "T3",
    TRUE ~ TStage))

# Relabel Gleason score
table1_TimeZero$GleasonScore <-
  factor(table1_TimeZero$GleasonScore ,
         levels=c("Less than or equal to 6",
                  "3 + 4 = 7",
                  "4 + 3 = 7",
                  "7 (when breakdown not available)",
                  "8",
                  "9",
                  "10",
                  "Unknown / Not documented"),
         labels=c("Low/very low risk: ≤6",
                  "Intermediate risk: 7",
                  "Intermediate risk: 7",
                  "Intermediate risk: 7",
                  "High/very high risk: 8-10",
                  "High/very high risk: 8-10",
                  "High/very high risk: 8-10",
                  "Unknown / Not documented"))

# Add calender time of treatment initation
table1_TimeZero <- table1_TimeZero %>%
  mutate(cal_time = as.factor(as.numeric(year(TimeZero))))


### Checking the relationship between diagnosis dates (a bit confusing --> decide not to include in table 1)
# table1_TimeZero$Gap_MetDiag_CRPC <- as.numeric(table1_TimeZero$MetDiagnosisDate - table1_TimeZero$CRPCDate)
# summary(table1_TimeZero$Gap_MetDiag_CRPC)
# table1_TimeZero$Gap_MCRPC_PCDiag <- as.numeric(table1_TimeZero$MCRPCDate - table1_TimeZero$DiagnosisDate)
# summary(table1_TimeZero$Gap_MCRPC_PCDiag)
# table1_TimeZero$Gap_MCRPC_MetDiag <- as.numeric(table1_TimeZero$MCRPCDate - table1_TimeZero$MetDiagnosisDate)
# summary(table1_TimeZero$Gap_MCRPC_MetDiag)
# table1_TimeZero$Gap_MCRPC_CRPC <- as.numeric(table1_TimeZero$MCRPCDate - table1_TimeZero$CRPCDate)
# summary(table1_TimeZero$Gap_MCRPC_CRPC)


label(table1_TimeZero$Age_TimeZero) <- "Age"
label(table1_TimeZero$PSA_TimeZero) <- "PSA (ng/mL)"
label(table1_TimeZero$ALP_RULN_TimeZero) <- "Alkaline phosphatase, relative to ULN"
label(table1_TimeZero$LDH_RULN_TimeZero) <- "Lactate dehydrogenase, relative to ULN"
label(table1_TimeZero$Hb_TimeZero) <- "Hb (g/L)"
label(table1_TimeZero$ECOG_TimeZero) <- "ECOG"
label(table1_TimeZero$f_pretreat_lines) <- "Previous treatment lines for mHSPC, nmCPRC and/or nmHSPC"
label(table1_TimeZero$f_pretreat_treat_allowed_docetaxel) <- "Previous treatment of docetaxel"
label(table1_TimeZero$f_pretreat_treat_other_systemic) <- "Previous other systematic treamtents"
label(table1_TimeZero$GroupStage) <- "Group Stage (initial PC diagnosis)"
label(table1_TimeZero$TStage) <- "T Stage (initial PC diagnosis)"
label(table1_TimeZero$NStage) <- "N Stage (initial PC diagnosis)"
label(table1_TimeZero$MStage) <- "M Stage (initial PC diagnosis)"
label(table1_TimeZero$GleasonScore) <- "Gleason Score (initial PC diagnosis)"
label(table1_TimeZero$DM_TimeZero) <- "Diabetes mellitus"
label(table1_TimeZero$f_SES) <- "SES quantiles"
label(table1_TimeZero$Gap_TimeZero_mCRPC) <- "Days since mCRPC diagnosis"
label(table1_TimeZero$Gap_TimeZero_CRPC) <- "Days since castration resistance"
label(table1_TimeZero$Gap_TimeZero_mPC) <- "Days since confirmation of metastasis"
label(table1_TimeZero$Gap_TimeZero_PC) <- "Days since initial PC diagnosis"
label(table1_TimeZero$Comorbid_Score_TimeZero) <- "Combined comorbidity score"
label(table1_TimeZero$cal_time) <- "Year of initial first-Line mCRPC Treatment"


table1(~ Age_TimeZero + PSA_TimeZero +
       ALP_RULN_TimeZero + LDH_RULN_TimeZero +
       Hb_TimeZero +
       ECOG_TimeZero +
       f_pretreat_lines +
```

```
       f_pretreat_treat_allowed_docetaxel +
       f_pretreat_treat_other_systemic +
       GleasonScore + GroupStage +
       TStage + NStage + MStage +
       Comorbid_Score_TimeZero +
       DM_TimeZero + Race +
       f_SES +
       Gap_TimeZero_mCRPC + Gap_TimeZero_mPC + Gap_TimeZero_CRPC +
       Gap_TimeZero_PC +
       cal_time
       | FirstLine, data = table1_TimeZero)

pvalue <- function(x, ...) {
 # Construct vectors of data y, and groups (strata) g
 y <- unlist(x)
 g <- factor(rep(1:length(x), times=sapply(x, length)))
 if (is.numeric(y)) {
  # For numeric variables, perform a standard 2-sample t-test
  p <- t.test(y ~ g)$p.value
 } else {
  # For categorical variables, perform a chi-squared test of independence
  p <- chisq.test(table(y, g))$p.value
 }
 # Format the p-value, using an HTML entity for the less-than sign.
 # The initial empty string places the output on the line below the variable label.
 c("", sub("<", "&lt;", format.pval(p, digits=3, eps=0.001)))
}

# Create a function to compute p-value for continuous (t-test) and categorical vars (chi-square)
# add p-value
table1(~ Age_TimeZero + PSA_TimeZero +
       ALP_RULN_TimeZero + LDH_RULN_TimeZero +
       Hb_TimeZero +
       ECOG_TimeZero +
       f_pretreat_lines +
       f_pretreat_treat_allowed_docetaxel +
       f_pretreat_treat_other_systemic +
       GleasonScore + GroupStage +
       TStage + NStage + MStage +
       Comorbid_Score_TimeZero +
       DM_TimeZero + Race +
       f_SES +
       Gap_TimeZero_mCRPC + Gap_TimeZero_mPC + Gap_TimeZero_CRPC +
       Gap_TimeZero_PC +
       cal_time
      | FirstLine, data = table1_TimeZero, overall = F, extra.col = list(`P-value`=pvalue))


#####################################
# Prepare Table 1 for crossover
#####################################

# Use data from (already time-varying covaraites labelled long dataset 04_Target Trial Emulation_GUTG001)

# Find patient characteristics at crossover (use the the characteristics a day before starting second-line treatment)
table1_Crossover <- d_tvary_GUTG001_PD %>%
 dplyr::filter(LineNumber_mCRPC_UserDef == 2) %>%
 dplyr::select(PatientID, Gap_PD_TimeZero) %>%
 rename(time = Gap_PD_TimeZero) %>%
 mutate(time = as.numeric(time)-1) %>%
 inner_join(TTE_GUTG001_DTR_IPW, by = c("PatientID", "time")) # 2418

#- relabel table1 at cross-over
table1_Crossover$DM_tvary <-
 factor(table1_Crossover$DM_tvary, levels=c(1,0),
       labels=c("Yes", "No"))

# Update Age
table1_Crossover <- table1_Crossover %>%
 mutate(Age_Crossover = Age_TimeZero + (time/365))

# Change unknown labels PSA, ALP, LDH, Hb to NA, and as numeric variables for characteristic table
table1_Crossover <- table1_Crossover %>%
 mutate(across(c(PSA_tvary, ALP_RULN_tvary, LDH_RULN_tvary, Hb_tvary),
```

```
            ~ as.numeric(replace(., . == "Unknown", NA)))) %>%
  mutate(across(c(ECOG_tvary),
            ~ replace(., . == "Unknown", NA)))

table1::label(table1_Crossover$Age_Crossover) <- "Age"
table1::label(table1_Crossover$PSA_tvary) <- "PSA (ng/mL)"
table1::label(table1_Crossover$ALP_RULN_tvary) <- "Alkaline phosphatase, relative to ULN"
table1::label(table1_Crossover$LDH_RULN_tvary) <- "Lactate dehydrogenase, relative to ULN"
table1::label(table1_Crossover$Hb_tvary) <- "Hb (g/L)"
table1::label(table1_Crossover$ECOG_tvary) <- "ECOG"
table1::label(table1_Crossover$Comorbid_Score_tvary) <- "Combined comorbidity score"
table1::label(table1_Crossover$DM_tvary) <- "Diabetes mellitus"
table1::label(table1_Crossover$time) <- "Time since the treatment initiation"

gc()
# table 1 at crossover (n = 2418)
table1(~ Age_Crossover +
         PSA_tvary +
         ALP_RULN_tvary + LDH_RULN_tvary +
         Hb_tvary +
         ECOG_tvary +
         Comorbid_Score_tvary +
         DM_tvary +
         time
       | FirstLine, data = table1_Crossover)

# table 1 at crossover with p-value
table1(~ Age_Crossover +
         PSA_tvary +
         ALP_RULN_tvary + LDH_RULN_tvary +
         Hb_tvary +
         ECOG_tvary +
         Comorbid_Score_tvary +
         DM_tvary +
         time
       | FirstLine, data = table1_Crossover, overall = F, extra.col = list(`P-value`=pvalue))
```

Appendix 8.7 R code for extracting time-varying characteristics during patient follow-up in Prostate Cancer Case Study 1 (PC1)

```
################################################################################
###   Defining time-varying covariates                                       ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                    ###
###   Data: Flatiron prostate cancer datasets                                ###
###                                                                          ###
###   University of Sheffield                                                ###
###   Date created: Nov 2, 2023                                              ###
###   Created Git & renv project: Aug 01, 2023                               ###
################################################################################

## House keeping

# Load packages and functions
renv::status()
`%notin%` <- Negate(`%in%`)

library(dplyr)
library(lubridate)
library(tidyverse)


##############################################
#
# GUTG-001 Analogue: Time-varying covariates
#
##############################################


##### Read files
# use d_TT_cohort_GUTG001_final, d_LOT_GUTG001_final clearned from 02_02_Cohort Characteristics
# use table1_TimeZero

# Baseline covariate (confirmed with Carmel):
# age, ECOG performance status, Hb,
# diabetes, prior treatments,
# whether patient was initially diagnosed with metastatic disease,
# stage at original diagnosis,
# (comorbidity score), (PSA prior to treatment), (Gleason score), (ethnicity)
# Others: institution, calender year?

# Time-varying covariates
# age (expressed as time with daily interval), ECOG performance status,
# Hb, diabetes,
# progression status (whether patient switched treatment),
# prior treatments,
# (comorbidity score), (PSA prior to treatment), (Gleason score)

## ECOG from the ECOG baseline table, keep only those after TimeZero and delte line ECOG value unknown
# Delete LineStart Date info from the Baseline Ecog table (redefine in my own analysis)
d_tvary_GUTG001_ECOG <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                      rd_baselineECOG[ , c("PatientID", "ECOGSource", "LineStartDate", "ECOGValue", "ECOGDate")],
                      by = "PatientID") %>%
  mutate_at(vars(LineStartDate, ECOGDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  filter(ECOGDate > TimeZero) %>%
  mutate(Gap_ECOGDate_TimeZero = ECOGDate - TimeZero) %>%
  filter(ECOGValue %notin% "Unknown") %>%
  select(-LineStartDate) # 4478


temp <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                       rd_ECOG[ , c("PatientID", "EcogValue", "EcogDate")],
                       by = "PatientID") %>%
  mutate_at(vars(EcogDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  group_by(PatientID, FirstLine, TimeZero, EcogDate) %>%
  slice(which.min(EcogValue)) %>% # if there are multiple Ecog on the same date, then select the smallest
  filter(EcogDate > TimeZero) %>%
  mutate(Gap_ECOGDate_TimeZero = EcogDate - TimeZero) %>%
  rename(ECOGValue = EcogValue) %>%
  rename(ECOGDate = EcogDate) %>%
  mutate(ECOGSource = "Structured") # 58272
```

```
# merge unstructured and structured data
# delete complete duplicates, and pick extracted data over structured data if any on the same date
d_tvary_GUTG001_ECOG <- rbind (d_tvary_GUTG001_ECOG, temp) %>% # 62750
 distinct() %>% # 58906
 group_by(PatientID, ECOGDate) %>%
 arrange(PatientID, ECOGDate, ECOGSource) %>%
 slice(1) %>%
 ungroup() # 58896


## Lab results

#- Hb
# keep source rather than derived, and then within the remaining, keep the highest if multiple tests per day
d_tvary_GUTG001_Hb <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                  rd_lab[rd_lab$LOINC %in% v_LOINC_Hb, c("PatientID", "TestDate",
                                        "TestResultCleaned", "LabSource")],
                  by = "PatientID",
                  all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate > TimeZero) %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 mutate(Gap_LabDate_TimeZero = TestDate - TimeZero) %>%
 arrange(PatientID, TestDate, desc(LabSource), desc(TestResultCleaned)) %>%
 group_by(PatientID, TestDate) %>%
 slice(1) %>%
 ungroup() # 31470

#- PSA
# keep source over derived, and the smallest if multiple tests per day
d_tvary_GUTG001_PSA <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                   rd_lab[rd_lab$LOINC %in% v_LOINC_PSA_t, c("PatientID", "TestDate", "TestResultCleaned", "LabSource")],
                   by = "PatientID",
                   all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate > TimeZero) %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 mutate(Gap_LabDate_TimeZero = TestDate - TimeZero) %>%
 arrange(PatientID, TestDate, desc(LabSource), TestResultCleaned) %>%
 group_by(PatientID, TestDate) %>%
 slice(1) %>%
 ungroup() # 31470


#- ALP
# keep source over derived, and the smallest if multiple tests per day
d_tvary_GUTG001_ALP <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                   rd_lab[rd_lab$LOINC %in% v_LOINC_ALP, c("PatientID", "TestDate", "TestResultCleaned", "LabSource")],
                   by = "PatientID",
                   all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate > TimeZero) %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 mutate(Gap_LabDate_TimeZero = TestDate - TimeZero) %>%
 arrange(PatientID, TestDate, desc(LabSource), TestResultCleaned) %>%
 group_by(PatientID, TestDate) %>%
 slice(1) %>%
 ungroup() # 71456

#- LDH
# keep source over derived, and the smallest if multiple tests per day
d_tvary_GUTG001_LDH <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "FirstLine", "TimeZero")],
                   rd_lab[rd_lab$LOINC %in% v_LOINC_LDH, c("PatientID", "TestDate", "TestResultCleaned", "LabSource")],
                   by = "PatientID",
                   all.y = TRUE) %>%
 mutate_at(vars(TimeZero, TestDate), as.Date, format="%Y-%m-%d") %>%
 dplyr::filter(TestDate > TimeZero) %>%
 dplyr::filter(!is.na(TestDate)) %>%
 dplyr::filter(!is.na(TestResultCleaned)) %>%
 mutate(Gap_LabDate_TimeZero = TestDate - TimeZero) %>%
 arrange(PatientID, TestDate, desc(LabSource), TestResultCleaned) %>%
```

```
  group_by(PatientID, TestDate) %>%
  slice(1) %>%
  ungroup() # 9683 (not suprising, Carmel said it is not a specific indicator)


# Diagnosis

#- DM
# find diagnosis after Time Zero
d_tvary_GUTG001_DM <- filter(rd_diag, substr(DiagnosisCode, 1, 3)
                             %in% c(v_DM_ICD10_3CHR,v_DM_ICD9_3CHR)) %>%
  select(PatientID, DiagnosisDate) %>%
  inner_join(d_TT_cohort_GUTG001_final[ , c("PatientID", "TimeZero")],
             by = "PatientID") %>%
  mutate_at(vars(DiagnosisDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate > TimeZero) %>%
  group_by(PatientID) %>%
  slice(which.min(DiagnosisDate)) %>%
  ungroup() # 245 (more than baseline, (1) patient develop DM later, or (2) 182 is not long enough to capture)


#- Progression date (start receiving subsequent treatment)
# Define the date of treatment-switching as the date of progression disease
# (indicating following dynamic strategy but progressed)
d_tvary_GUTG001_PD <- d_LOT_GUTG001_final %>%
  filter(!is.na(LineNumber_mCRPC_UserDef),
         LineNumber_mCRPC_UserDef > 1) %>%
  left_join(d_TT_cohort_GUTG001_final[ , c("PatientID", "TimeZero")]) %>%
  mutate(PDDate = StartDate) %>%
  mutate_at(vars(PDDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  mutate(Gap_PD_TimeZero = PDDate - TimeZero) %>%
  select(PatientID, PDDate, LineNumber_mCRPC_UserDef, Gap_PD_TimeZero)
# 5336


# Co-morbidity

# Find co-morbid Diagnosis within 182 days prior to Time Zero
d_tvary_GUTG001_CoMorbid <- merge(d_TT_cohort_GUTG001_final[ , c("PatientID", "TimeZero")], rd_diag, by = "PatientID") %>%
  select(-PracticeID) %>%
  mutate_at(vars(DiagnosisDate, TimeZero), as.Date, format="%Y-%m-%d") %>%
  filter(DiagnosisDate > TimeZero)
# 81935

# Summaries each patient's morbidity flag using the function created
# label time-varying diagnosis with 0 as ., so they can be replace with base line co-morbidity later
# (as each row represent only one type of co-morbidity update)
# Only update when there are additional 1
d_tvary_GUTG001_CoMorbid <- label_comorbid(d_tvary_GUTG001_CoMorbid) %>%
  select(-DiagnosisCode, -DiagnosisDescription, -DiagnosisCodeSystem) %>%
  filter(rowSums(select(., -PatientID, -DiagnosisDate, -TimeZero)) != 0) %>% # Only keep rows where row sum does not equal to 0
  mutate(Gap_Comorbid_TimeZero = as.numeric(DiagnosisDate -TimeZero)) %>% # calculate gap to TimeZero
  select(-DiagnosisDate, -TimeZero) %>%
  mutate(across(everything(), ~ifelse(. == 0, NA, .))) %>%
  arrange(PatientID, Gap_Comorbid_TimeZero) %>%
  rename_with(~ifelse(. %in% c("PatientID", "Gap_Comorbid_TimeZero"), ., paste0(., "_tvary"))) # rename comorbidity columns with _tvary (time
updated comorbid)
# 9430: some patients had multiple records on the same date for different diagnosis Gap_Comorbid_TimeZero


# Summerise records to per patient, per day (e.g. combining records of any updated comorbid on the same day)
# Summarise_at is faster tham summary(across())
# Custom max function that returns NA for all NA values
# (otherwise for some comorbid it would return -InF), and take long > 3 mins
safe_max <- function(x, na.rm = TRUE) {
  if (all(is.na(x))) NA else max(x, na.rm = na.rm)
}
variable_names <- paste("Comorbid_", unique(v_comorbid_names), "_tvary", sep = "")
d_tvary_GUTG001_CoMorbid <- d_tvary_GUTG001_CoMorbid %>%
  group_by(PatientID, Gap_Comorbid_TimeZero) %>%
  summarise_at(vars(all_of(variable_names)), safe_max, na.rm = TRUE) %>%
  ungroup() %>%
  mutate(across(everything(), ~ ifelse(. == -Inf, NA, .)))
# 6724
```

```
# Add baseline detailed comorbidities for all patients
# Noted that AnyTumour and Metastaitc_cancer are only in baseline and both are labelled as 1 for every patient
d_tvary_GUTG001_CoMorbid <- d_demo_GUTG001_CoMorbid %>%
  select(-Comorbid_Score_TimeZero) %>%
  rename_with(~str_replace(., "_TimeZero$", "_tvary")) %>%
  mutate(Gap_Comorbid_TimeZero = 0) %>%
  bind_rows(d_tvary_GUTG001_CoMorbid) %>%
  arrange(PatientID, Gap_Comorbid_TimeZero)
# 10868

# In summary, the d_tvary_GUTG001_CoMorbid table contains each detailed comorbidity flag at baseline and any updated new comorbidity
```

Appendix 8.8 R code for finalising the final cleaned "wide" baselined (TTE_GUTG001_base) and "long" time-varying datasets (TTE_GUTG001_DTR_IPW) for Prostate Cancer Case Study 1 (PC1)

```
###Project Info ##############################################################################
###   Target Trial Emulation – Final TTE dataset derivation                                  ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                                    ###
###   Data: Flatiron prostate cancer datasets                                                ###
###                                                                                          ###
###   University of Sheffield                                                                ###
###   Date created: Nov 16, 2023                                                             ###
##############################################################################################

## House keeping
# Load packages and functions
renv::status()
# renv::snapshot()

`%notin%` <- Negate(`%in%`)

# install backports when there are conflicts of tidyr
library(backports)
library(tidyr)
library(remotes)
# install.packages("tidyr")
library(tidyr)
library(dplyr)
library(lubridate)
library(ggpubr)
library(survival)
library(zoo)
library(data.table) # faster than data.frame
library(speedglm)
library(dplyr) # too avoid conflict, read dplyr again
library(boot)
library(scales)
library(date)
library(survminer) # use tidy R 1.3.0
library(splines)
library(sandwich)
library(lmtest)
# library(broom) # use tidy R 1.2.0
# library(haven)
# library(psych)
# library(flexsurv)
# library(cowplot)
# library(geepack)
# library(Hmisc)
# library(eha)
# library(rpsftm)
# library(rms)


# Disable printing results in scientific notation
options(scipen=100)

# Check data
head(table1_TimeZero)

# Set time point for estimation of risks (1440 days, 48 months)
K <- 1440 # 4 years
K_minus_1 = K-1 # time interval K-1~K
Time <- data.frame(time = seq(0, K, by = 1)) %>%
  mutate(tstart = time) %>% # label time interval for survival analysis
  mutate(tstop = time + 1) # 1440 (intervals): from  tstart = 0, tstop 1, to tstart = 1339, tstop = 1440

# Check size and number of objects
# sizes <- sapply(ls(), function(x) object.size(get(x)))
# length(sizes)
# sizes <- sort(sizes, decreasing = TRUE)
# sizes
gc()
```

```
#############################################################################################################
#  I. Clean dataset, add outcome, censoring indicator, and follow-up time to cohort characterstics                                  ####
#############################################################################################################

#***Define Administrative Censoring Date based on "Structural Activity + Grace period" ####
# This part is adapted from Dr. Philani Mpofu's Aug 31, 2023 code (Flatiron's standard way of finding last structural activity)

# Craete a function to identify the last date for each patient from a dataset
f_LastDate <- function(ids = d_TT_cohort_GUTG001_final$PatientID,
                       file = rd_diag,
                       DateName = DiagnosisDate) {

  DateName <- enquo(DateName) # Create quosure (column in dplyr can be problematic in R)

  last_date <- file %>%
    dplyr::filter(PatientID %in% ids) %>%
    dplyr::select(PatientID, !!DateName) %>% # Use !! to unquote the quosure
    dplyr::mutate_at(vars(!!DateName), as.Date, format="%Y-%m-%d") %>%
    dplyr::filter(!is.na(!!DateName)) %>%
    group_by(PatientID) %>%
    slice(which.max(!!DateName)) %>%
    rename(max_date = !!DateName) %>%
    ungroup()

  return(last_date)
}

# define the subset of ids to extract the last dates
ids <- d_TT_cohort_GUTG001_final$PatientID

#- last diagnosis date
last_diag <- f_LastDate(ids, rd_diag, DiagnosisDate)
# 4143 patients have a last diagnosis date

#- last visit date
last_visit <- f_LastDate(ids, rd_visit, VisitDate)
# 4144 patients have a last visit date

# crate a list to store these dates
last_dates <- list(last_diag = last_diag, last_visit = last_visit)
rm(last_diag, last_visit)

#- last lab date
last_lab <- f_LastDate(ids, rd_lab, TestDate) # 4074
last_dates[["last_lab"]] <- last_lab
rm(last_lab)

#- last lab result date
last_labresult <- f_LastDate(ids, rd_lab, ResultDate) # 4024
last_dates[["last_labresult"]] <- last_labresult
rm(last_labresult)

#- last vital date
last_vital <- f_LastDate(ids, rd_vital, TestDate) # 4142
last_dates[["last_vital"]] <- last_vital
rm(last_vital)

#- last vital result date
last_vitalresult <- f_LastDate(ids, rd_vital, ResultDate) # 493
last_dates[["last_vitalresult"]] <- last_vitalresult
rm(last_vitalresult)

#- last med order date
last_medorder <- f_LastDate(ids, rd_med_order, OrderedDate) # 4042
last_dates[["last_medorder"]] <- last_medorder
rm(last_medorder)

#- last medication administration date
last_medadmin <- f_LastDate(ids, rd_med_admin, AdministeredDate) # 3720
last_dates[["last_medadmin"]] <- last_medadmin
rm(last_medadmin)

#- last ecog date
last_ecog <- f_LastDate(ids, rd_ECOG, EcogDate) # 3404
```

```
last_dates[["last_ecog"]] <- last_ecog
rm(last_ecog)

#- last baseline ecog date (including ML abstraction)
last_baseecog <- f_LastDate(ids, rd_baselineECOG, ECOGDate) # 3270
last_dates[["last_baseecog"]] <- last_baseecog
rm(last_baseecog)


#- last tele med date
last_visit_tele <- f_LastDate(ids, rd_visit_tele, VisitDate) # 825
last_dates[["last_visit_tele"]] <- last_visit_tele
rm(last_visit_tele)

#- last enhanced oral date (LOT was from med order, administration and enhanced oral)
last_enchenced_oral <- f_LastDate(ids, rd_treat_oral, EndDate) # 4144
last_dates[["last_enchenced_oral"]] <- last_enchenced_oral
rm(last_enchenced_oral)


#- last enhanced AB emitters date (LOT was from med order, administration and enhanced oral)
last_enchenced_ABemit <- f_LastDate(ids, rd_treat_ABemit, AdministrationDate) # 485
last_dates[["last_enchenced_ABemit"]] <- last_enchenced_ABemit
rm(last_enchenced_ABemit)

#- last primary treatment date (priory treatment date prior to mPC)
last_enchenced_primary <- f_LastDate(ids, rd_treat_primary, TreatmentDate) # 1912
last_dates[["last_enchenced_primary"]] <- last_enchenced_primary
rm(last_enchenced_primary)

#- last NGS test date
last_test_NGS <- f_LastDate(ids, rd_test_NGS, NGSTestDate) # 980
last_dates[["last_test_NGS"]] <- last_test_NGS
rm(last_test_NGS)

#- last enhanced ADT date (LOT was from med order, administration and enhanced oral)
last_enchenced_ADT <- f_LastDate(ids, rd_treat_ADT, EndDate) # 400
last_dates[["last_enchenced_ADT"]] <- last_enchenced_ADT
rm(last_enchenced_ADT)

#- last enhanced Provenge date (LOT was from med order, administration and enhanced oral)
last_enchenced_sipT <- f_LastDate(ids, rd_treat_sipT, StartDate) # 212
last_dates[["last_enchenced_sipT"]] <- last_enchenced_sipT
rm(last_enchenced_sipT)

#- last biomarker specimen collection date (new dataset)
last_biomarker_collect <- f_LastDate(ids, rd_biomarkers, SpecimenCollectedDate) # 1186
last_dates[["last_biomarker_collect"]] <- last_biomarker_collect
rm(last_biomarker_collect)

#- last biomarker specimen received date (new dataset)
last_biomarker_received <- f_LastDate(ids, rd_biomarkers, SpecimenReceivedDate) # 1156
last_dates[["last_biomarker_received"]] <- last_biomarker_received
rm(last_biomarker_received)

#- last biomarker specimen result date (new dataset)
last_biomarker_result <- f_LastDate(ids, rd_biomarkers, ResultDate) # 1227
last_dates[["last_biomarker_result"]] <- last_biomarker_result
rm(last_biomarker_result)

#-- derive the last activity date by taking the maximum of the activity dates (from 20 different dates)
d_last_activity <- do.call("rbind", last_dates) %>%
  group_by(PatientID) %>%
  slice(which.max(max_date)) %>%
  rename(DateAdminCen_lastactive = max_date) # 4144: every one has a last activity date



#***Define the censoring date due to deviating from treatment strategy ####
# Abiraterone Group: second line not being enzalutamide
# Enatalumide Group: second line not being abiraterone

#- Label C_DTRdeviation (censoring due to treatment deviation)
```

```
# C_DTRdeviation = 1 (treatment-switching date), when the initiation of a non-protocol defined second-line treatment (on date k) at the interval of
tstart = k
d_DTRdeviation <- d_LOT_GUTG001_final %>%
  filter(LineNumber_mCRPC_UserDef == 2) %>%
  left_join(table1_TimeZero[ , c("PatientID", "FirstLine", "TimeZero")], by = "PatientID") %>%
  dplyr::select(PatientID, TimeZero, StartDate, FirstLine, LineName) %>%
  mutate_at(vars(TimeZero, StartDate), as.Date, format="%Y-%m-%d") %>%
  mutate(time = as.numeric(StartDate - TimeZero)) %>%
  rename(SecondLine = LineName) %>%
  # Flagging patients who were censored due to deviation
  mutate(Cenor_DTRdev2 = ifelse((FirstLine == "Abiraterone" & SecondLine == "Enzalutamide")|
                          (FirstLine == "Enzalutamide" & SecondLine == "Abiraterone"),
                           0, 1)) %>%
  # label the date of second-line deviation as the date of initiating a non-protocol second-line
  rename(DateCen_DTRdev2 = StartDate)
# 2468 patients had a second-line record
# (some patients did not receive second-line but died before, while some were censored/lost-to-follow-up)
tapply(as.factor(d_DTRdeviation$Cenor_DTRdev2), d_DTRdeviation$FirstLine, summary)
# 1: 1319 (received non-protocol second-line: Abi group 693, Enza 626)
# 0: 1149 (received protocol-based second-line: Abi 655; Enza 494
# (more patients deviating protocol in the enza group, that's probably why the survival curve look far different from those in the trial?))

tapply(as.factor(d_DTRdeviation$SecondLine), d_DTRdeviation$FirstLine, summary)


#***Label outcome (death), administrative censoring date and censoring due to DTR deviation ####
TTE_GUTG001_base <- table1_TimeZero %>%
  dplyr::select(-MCRPCDate, -DiagnosisDate, -MetDiagnosisDate, -CRPCDate, -Ethnicity,
       -Histology, -IsCRPC, -PSADiagnosis, -PSAMetDiagnosis, -State) %>%
  # Insert death date: data only detailed to which month, so use 15th of that month for each patient
  left_join(rd_death) %>%
  mutate(DateOfDeath = as.Date(paste(DateOfDeath, "-15", sep=""), format="%Y-%m-%d")) %>%
  # Define Admin censoring Scenario 1: 2023-08-31 (data cutoff)
  mutate(DateAdminCen_dcutoff = as.Date("2023-08-31", format="%Y-%m-%d")) %>%
  # Define Admin censoring Scenario 2: date of the last structural activity for each patient
  left_join(d_last_activity) %>%
  # Define Admin censoring Scenario 3: censoring due to alive > observation timeframe K (it may vary but the default is 1440 days (4 years) to be
align with GUTG-001)
  mutate(DateAdminCen_maxK = TimeZero + K) %>%
  # Define Censoring due to deviating from DTR
  left_join(d_DTRdeviation[d_DTRdeviation$Cenor_DTRdev2 == 1, c("PatientID", "DateCen_DTRdev2")]) %>%
  # comparison of the last structural active date to (1) death and (2) date of censoring due to treatment deviation
  mutate(Gap_lastactive_death = as.numeric(DateOfDeath - DateAdminCen_lastactive),
       Gap_DTRDev2_lastactive = as.numeric(DateAdminCen_lastactive - DateCen_DTRdev2))

# Check how many patients had a death date in the database
sum(!is.na(TTE_GUTG001_base$DateOfDeath)) # 2887


#***Label outcome (death), administrative censoring + grace period date and censoring due to DTR deviation ####
TTE_GUTG001_base <- TTE_GUTG001_base %>%
  mutate(DateAdminCen_lastactive_grace182 =  DateAdminCen_lastactive + 182) %>%
  rowwise() %>%
  # Define the Date of (Y) for the ITT analysis as the minimum date of event
  # (death, or staying alive at the last structural activity date + grace period, or staying alive on day K)
  mutate(DateY_FirstLine = min(DateOfDeath, DateAdminCen_lastactive_grace182,
                      DateAdminCen_dcutoff, DateAdminCen_maxK, na.rm = T),
       # Label the Type of Y with the order of death, date of censoring
       TypeY_FirstLine = case_when(
         DateY_FirstLine == DateOfDeath ~ "DateOfDeath",
         TRUE ~ "DateOfAdminCensor"
       )) %>%
  # Define the Date of (Y) for the DTR analysis as the minimum date of event
  # (death, or staying alive at the last structural activity date + grace period, or staying alive on day K.
  # or staying alive on the date of switching to a non protocol allowed second-line treatment)
  mutate(DateY_DTR = min(DateOfDeath, DateAdminCen_lastactive_grace182,
                   DateAdminCen_dcutoff, DateAdminCen_maxK,
                   DateCen_DTRdev2, na.rm = T),
       # Label the Type of Y with the order of death, date of admin censoring, date of deviation
       TypeY_DTR = case_when(
         DateY_DTR == DateOfDeath ~ "DateOfDeath",
         # Noted that structural activity should always be after deviation of treatment
         # However structural activity + grace period might happen after or on the date of deviation of treatment
         # Therefore label Date of Censoring due to treatment deviation first if DateCen_DTRdev = DateAdminCen_lastactive_grace182
```

```
        DateY_DTR == DateCen_DTRdev2 ~ "DateOf2LTreatDeviation",
        TRUE ~ "DateOfAdminCensor"
      )) %>%
 # Define Y for the ITT analysis
 mutate(Y_FirstLine =  case_when(
       TypeY_FirstLine == "DateOfDeath" ~ 1,
       TRUE ~ 0 )) %>%
 # Define Y for DTR per-protocol analysis
 mutate(Y_DTR = case_when(
       TypeY_DTR == "DateOfDeath" ~ 1,
       TRUE ~ 0 )) %>%
 # Define Administrative censoring for ITT analysis
 mutate(C_Admin_FirstLine = case_when(
    TypeY_FirstLine == "DateOfAdminCensor" ~ 1,
    TRUE ~ 0 )) %>%
 # Define Administrative censoring for DTR analysis
 mutate(C_Admin_DTR = case_when(
  TypeY_DTR == "DateOfAdminCensor" ~ 1,
  TRUE ~ 0 )) %>%
 # Define Treatment deviation censoring for DTR analysis
 mutate(C_DTRdevL2 = case_when(
  TypeY_DTR == "DateOf2LTreatDeviation" ~ 1,
  TRUE ~ 0
 ))



#***Label Outcome and Censoring label for ITT & DTR analysis####


#***Label SurvTime for ITT analysis and DTR (the earliest of death or censoring dates) ####
# Define a function to label survival times
label_SurvTime <- function(dataset) {
 dataset %>%
   mutate(SurvTime_FirstLine = case_when(
     # For patients who had a death date, then they only survived until the date of death
     TypeY_FirstLine  == "DateOfDeath" ~ as.numeric(DateY_FirstLine - TimeZero),
     # For patients who had administrative censoring, we only know they remained alive on the last date of observation
     # They technically could have survived until the censoring date + 1 day, so use censoring date + 1 for outcome calculation (which is unknown)
     # Keep max survival time as K for now, but in the long data, create an additional row for interval K~K+1 and label death as NA (for admin
censoring weighting?)
     TRUE ~ as.numeric(DateY_FirstLine - TimeZero)
   )) %>%
   mutate(SurvTime_DTR = case_when(
     # For patients who had a death date, then they only survived until the date of death
     TypeY_DTR  == "DateOfDeath" ~ as.numeric(DateY_DTR - TimeZero),
     # For patients who had a known date for initiating a second-line that is not allowed in protocol
     # We censor them at the interval when they start receiving non-protocol second-line
     # In the long format, that row will have a weight of 0, as we shouldn't observe the outcome at the interval when they start receiving non-
protocol treatment
     # For patients who had administrative censoring, we only know they remained alive on the last date of observation
     # They technically could have survived until the censoring date + 1 day, so use censoring date + 1 for outcome calculation (which is unknown)
     # We will censoring them in the next interval after the were confirmed remaining alive
     # Keep max survival time as K for now, but in the long data, create an additional row for interval K~K+1 (for admin censoring weighting?)
     TypeY_DTR  == "DateOf2LTreatDeviation" ~ as.numeric(DateY_DTR - TimeZero),
     TRUE ~ as.numeric(DateY_DTR - TimeZero)
   # Live up until the interval of treatent switching
   ))
}

# Apply the function to both all baseline datasets
TTE_GUTG001_base <- label_SurvTime(TTE_GUTG001_base)

#***Give the groups numeric label: Abi as active group = 1 enza = 0 (inline with Khalaf 2018) ####
TTE_GUTG001_base <- TTE_GUTG001_base %>%
 mutate(FirstLine_num = ifelse(FirstLine %in% "Abiraterone", 1, 0)) # label the last interval y = 1 (if death)


#***Organising baseline variables (relabel) for analysis: cannot have NA for regression models ####

# reduce the level of variables
#- Relabel MStage
TTE_GUTG001_base$MStage  <- ifelse(TTE_GUTG001_base$MStage %in% c("M1a", "M1b", "M1c"), "M1",
                          TTE_GUTG001_base$MStage)
```

```
# Need to transform any missing into a categorical variable:
# therefore, all numeric variables with missing values  ==> transform them into categories

# create a function to relabel PSA, creat new variable called PSA_cat
relabel_PSA <- function(data, column) {
  new_col_name <- paste(column, "cat", sep = "_")
  data <- data %>%
    mutate(!!sym(new_col_name) := case_when(
      is.na(.data[[column]]) ~ "Unknown",
      .data[[column]] < 4 ~ "< 4",
      .data[[column]] < 10 ~ "4 ≤ PSA < 10",
      .data[[column]] < 20 ~ "10 ≤ PSA < 20",
      .data[[column]] < 100 ~ "20 ≤ PSA < 100",
      TRUE ~ "> 100"))
  return(data)
}

#- Relabel PSA
TTE_GUTG001_base <- relabel_PSA(TTE_GUTG001_base, "PSA_TimeZero")

# create a function to relabel Hb
relabel_Hb <- function(data, column) {
  new_col_name <- paste(column, "cat", sep = "_")
  data <- data %>%
    mutate(!!sym(new_col_name) := case_when(
      is.na(.data[[column]]) ~ "Unknown",
      .data[[column]] < 120 ~ "Hb < 120",
      .data[[column]] < 140 ~ "120 ≤ Hb < 140",
      TRUE ~ "Hb ≥ 140"))
  return(data)
}

#- Relabel Hb
TTE_GUTG001_base <- relabel_Hb(TTE_GUTG001_base, "Hb_TimeZero")

#- Relabel ECOG
TTE_GUTG001_base$ECOG_TimeZero <- ifelse(is.na(TTE_GUTG001_base$ECOG_TimeZero), "Unknown",
                                         TTE_GUTG001_base$ECOG_TimeZero)

#- Relabel DM
TTE_GUTG001_base$DM_TimeZero <- ifelse(TTE_GUTG001_base$DM_TimeZero == "Yes", 1, 0)

#- Relabel Gap days between diagnosis and time zero
quantile(TTE_GUTG001_base$Gap_TimeZero_mCRPC) # quartile 0, 8, 31, 93, 2382
quantile(TTE_GUTG001_base$Gap_TimeZero_mCRPC[TTE_GUTG001_base$FirstLine == "Abiraterone"]) # quartile 0, 8, 31, 93, 1779
quantile(TTE_GUTG001_base$Gap_TimeZero_mCRPC[TTE_GUTG001_base$FirstLine == "Enzalutamide"]) # quartile 0, 7, 31, 96, 2382

quantile(TTE_GUTG001_base$Gap_TimeZero_mPC) # quartile 0, 101, 335, 636, 3317 (signifiacnt difference between groups)
quantile(TTE_GUTG001_base$Gap_TimeZero_mPC[TTE_GUTG001_base$FirstLine == "Abiraterone"]) # quartile 0, 94, 321, 587, 3317
quantile(TTE_GUTG001_base$Gap_TimeZero_mPC[TTE_GUTG001_base$FirstLine == "Enzalutamide"]) # quartile 0, 108, 351, 692, 3004

quantile(TTE_GUTG001_base$Gap_TimeZero_CRPC) # quartile 0, 12, 43, 147, 6754
quantile(TTE_GUTG001_base$Gap_TimeZero_CRPC[TTE_GUTG001_base$FirstLine == "Abiraterone"]) # quartile 0, 12, 45, 150, 6754
quantile(TTE_GUTG001_base$Gap_TimeZero_CRPC[TTE_GUTG001_base$FirstLine == "Enzalutamide"]) # quartile 0, 11, 41, 144, 5166

relabel_Gap_Diag <- function(data) {
  data <- data %>%
    # mCRPC
    mutate(Gap_TimeZero_mCRPC_cat = case_when(
      Gap_TimeZero_mCRPC < 30 ~ "< 30 days",
      Gap_TimeZero_mCRPC <= 90 ~ "31-90 days",
      Gap_TimeZero_mCRPC <= 182 ~ "91-182 days",
      Gap_TimeZero_mCRPC <= 365 ~ "182-365 days",
      TRUE ~ "> 365 days")) %>%

    # mPC
    mutate(Gap_TimeZero_mPC_cat = case_when(
      Gap_TimeZero_mPC <= 30 ~ "≤ 30 days",
      Gap_TimeZero_mPC <= 90 ~ "31-90 days",
      Gap_TimeZero_mPC <= 182 ~ "91-182 days",
      Gap_TimeZero_mPC <= 365 ~ "182-365 days",
      Gap_TimeZero_mPC <= 730 ~ "366-730 days",
      TRUE ~ "> 730 days")) %>%
```

```
   # CRPC
   mutate(Gap_TimeZero_CRPC_cat = case_when(
    Gap_TimeZero_CRPC <= 7 ~ "≤ 7 days",
    Gap_TimeZero_CRPC <= 30 ~ "8-30 days",
    Gap_TimeZero_CRPC <= 60 ~ "31-60 days",
    Gap_TimeZero_CRPC <= 90 ~ "61-90 days",
    Gap_TimeZero_CRPC <= 182 ~ "91-182 days",
    Gap_TimeZero_CRPC <= 365 ~ "182-365 days",
    TRUE ~ "> 365 days"))

 return(data)
}

#- Convert treatment number into factor
TTE_GUTG001_base$FirstLine_num <- as.factor(TTE_GUTG001_base$FirstLine_num)


##########################################
#   Wide to Long survival data for TTE ####
##########################################


#***Create survival interval by days for the ITT analysis (only baseline covariates) ####
TTE_GUTG001_ITT <- merge(Time, TTE_GUTG001_base) %>%
 # Keep only records within observed time
      # For patients who had an outcome of death, keep only rows when their survival time >= tstop (as death on date n is labelled at interval n-1~n)
 filter((TypeY_FirstLine  == "DateOfDeath" & SurvTime_FirstLine >= tstop) |
      # For patients who had an outcome of admin censoring, keep only rows when their survival time +1 >= tstop (as death on date n is labelled at
interval n-1~n)
      # This is because the admin censoring time has taken into account that they might still be alive a day after they were observed alive
      # i.e., max tstop = K + 1
      (TypeY_FirstLine  == "DateOfAdminCensor" & SurvTime_FirstLine + 1 >= tstop)) %>%
 # label y = outcome for the right patient-observation interval for each patient
 mutate(y = case_when(
       # For death outcome, label y = 1 at the interval where tstop = survival time
       Y_FirstLine == 1 &
         SurvTime_FirstLine == tstop
       ~ 1,
      # For administrative censoring, label y = 0 at the last interval (where tstop = admin censoring survival, K-1~k), then K~K+1 as with outcome =
NA (we have baseline of K but no outcome at K+1)
      # Such rows (y = NA) won't be used for outcome modelling, but purely for censoring weighting models
      Y_FirstLine == 0 &
        SurvTime_FirstLine + 1 == tstop
       ~ NA,
       TRUE ~ 0)
      )
# 2963088 obs-intervals
# Check how many death labelled interval: smaller than total death (some patient only died after 1440 days, or activity censor before death))
summary(as.factor(TTE_GUTG001_ITT$y)) # 2452, NA: 1692; other intervals: 0
gc()


#***Organising time-varying censoring variable for the ITT set####
# label c_admin = administrative censoring for the right patient-observation interval for each patient
TTE_GUTG001_ITT <- TTE_GUTG001_ITT %>%
 arrange(PatientID, tstop) %>% # sort by tstop
 # Label the censoring only at the interval of censoring
 mutate(c_admin = case_when(
   # For interval having death outcome, label c_admin = NA
   y == 1 ~ NA,
   # For interval having admin censoring, label c_admin = 1
   is.na(y) ~ 1,
   # For all other intervals, c_admin = 0
   TRUE ~ 0)) %>%
 dplyr::select(-C_Admin_DTR, -C_DTRdevL2, -Y_DTR, -SurvTime_DTR) # delete DTR set variables
# check if the number of individual with censoring matching those in the TTE base dataset
summary(as.factor(TTE_GUTG001_ITT$c_admin)) # 1692 intervals had admin censoring, other intervals no censoring 0 (including 2452 death
intervals = NA)



#***Create survival interval by days for the DTR_IPW analysis ####
# Create a function to label survival time of DTR dataset from Wide to Long (W2L)
```

```
# Time interval sheet have already been defind at the global environment
label_survival_W2L_DTR <- function(baseline_dataset) {
  merge(Time, baseline_dataset) %>%
  # Keep only records within observed time
  # For patients who had an outcome of death, keep only rows when their survival time >= tstop (as death on date n is labelled at interval n-1~n)
  filter((TypeY_DTR  == "DateOfDeath" & SurvTime_DTR >= tstop)|
         # For patients who had an outcome of admin censoring, keep only rows when their survival time >= tstop (as death on date n is labelled at
interval n-1~n)
         # This is because the admin censoring time has taken into account that they might still be alive a day after they were observed alive
         # i.e., max tstop = K + 1
         (TypeY_DTR  == "DateOfAdminCensor" & SurvTime_DTR + 1 >= tstop)|
         # For patients who had censoring due to protocol non-adherence
         # Keep their record only until survival time >= tstart
         # We only know that they are surviving in this interval and censoring at that interval
         # we don't want to use outcome of that interval
         (TypeY_DTR  == "DateOf2LTreatDeviation" & SurvTime_DTR >= tstart)
  ) %>%
  # label y = outcome for the right patient-observation interval for each patient
  mutate(y = case_when(
    # For death outcome, label y = 1 at the interval where tstop = survival time
    Y_DTR == 1 &
      SurvTime_DTR == tstop
    ~ 1,
    # For administrative censoring, label y = 0 at the last interval (where tstop = admin censoring survival, K-1~k), then K~K+1 as with outcome = NA
(we have baseline of K but no outcome at K+1)
    # Such rows (y = NA) won't be used for outcome modelling, but purely for censoring weighting models
    Y_DTR == 0 &
      TypeY_DTR  == "DateOfAdminCensor" &
      SurvTime_DTR + 1 == tstop # max tstop for admin censoring is K + 1
    ~ NA,
    # For censoring due to receiving non-protocol second-line, label y = NA at the last interval (where tstart = censoring due to treatment deviation)
    # need to change to 0 for IPCW-deviation modelling (require that line even the weight will be forced to 0 (meaning death wouldn't be counted))
    # Not use this line for non-adjusted PP modelling
    Y_DTR == 0 &
      TypeY_DTR  == "DateOf2LTreatDeviation" &
      SurvTime_DTR == tstart
    ~ NA,
    # all other intervals label y = 0
    TRUE ~ 0)
  )
}
gc()

# Label all DTR datasets
TTE_GUTG001_DTR_IPW <- label_survival_W2L_DTR(TTE_GUTG001_base)
# 2394119 obs-intervals
# Check how many death labelled interval: smaller than total death (some patient only died after 1440 days, or activity censor before death))
summary(as.factor(TTE_GUTG001_DTR_IPW$y)) # 1631, NA: 2513; other intervals: 0
gc()



#***Organising time-varying censoring variable for the DTR set####
# Create a function for labeling long data set censoring
label_censoring_long_DTR <- function(DTR_dataset) {

DTR_dataset %>%
  arrange(PatientID, tstart) %>% # sort by start
  # Label the admin censoring only at the interval of admin censoring
  mutate(c_admin = case_when(
    # For interval having death outcome, label c_admin = 0
    y == 1 ~ 0,
    # For interval having admin censoring, label c_admin = 1
    is.na(y) &
      TypeY_DTR  == "DateOfAdminCensor"
    ~ 1,
    # For interval having treatment deviation censoring, label c_admin = NA
    is.na(y) &
      TypeY_DTR  == "DateOf2LTreatDeviation"
    ~ NA,
    TRUE ~ 0)
    ) %>%
  # Label the treatment deviation censoring only at the interval of treatment deviation censoring
  mutate(c_treatdevL2 = case_when(
```

```
   # For interval having death outcome, label c_treatdevL2 = 0
   y == 1 ~ 0,
   # For interval having admin censoring, label c_treatdevL2 = NA
   is.na(y) &
    TypeY_DTR  == "DateOfAdminCensor"
    ~ NA,
   # For interval having treatment deviation censoring, label c_treatdevL2 = 1
   is.na(y) &
    TypeY_DTR  == "DateOf2LTreatDeviation"
   ~ 1,
   TRUE ~ 0)
   ) %>%
  dplyr::select(-C_Admin_FirstLine, -Y_FirstLine) # delete ITT set variables

}

# Label all DTR datasets
TTE_GUTG001_DTR_IPW <- label_censoring_long_DTR(TTE_GUTG001_DTR_IPW)


#***Label time-varying variables for DRT_IPW analysis set ####
#- DM
length(TTE_GUTG001_base$PatientID[TTE_GUTG001_base$DM_TimeZero == 1]) # baseline DM 191 pts
length(unique(d_tvary_GUTG001_DM$PatientID)) # 245 patients have DM diagnosis after baseline

# In total 379 patients ever have DM in the dataset (note baseline DM using 182 within TimeZero, could be mis-labeled)
# Some might really developed DM later though, once having DM diagnosis, then keep having DM
test <- unique(c(TTE_GUTG001_base$PatientID[TTE_GUTG001_base$DM_TimeZero == 1],
        d_tvary_GUTG001_DM$PatientID))

# Find patients who developed DM later (no DM at baseline)
temp <- d_tvary_GUTG001_DM %>%
  mutate(tstart = as.numeric(DiagnosisDate - TimeZero)) %>%
  mutate(DM_tvary = 1) %>%
  dplyr::select(-DiagnosisDate, -TimeZero) %>%
  filter(PatientID %notin% unique(TTE_GUTG001_base$PatientID[TTE_GUTG001_base$DM_TimeZero == 1]))
# 188
summary(temp$tstart) # some developed DM relatively late (> 1440 days, only 140 pts developed DM later but < 1440 days)

# Label all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who developed DM later
  mutate(DM_tvary = ifelse(tstart == 0, DM_TimeZero, DM_tvary)) %>% # label everyone's baseline DM status
  arrange(PatientID, time) %>% # arrange rows before fill in missing DM_tvary values
  mutate(DM_tvary = na.locf(DM_tvary)) # label everyone's DM base on baseline DM until it changes (use zoo package)


#- ECOG
# Find patients who had ECOG after Time Zero
temp <- d_tvary_GUTG001_ECOG %>%
  mutate(tstart = as.numeric(Gap_ECOGDate_TimeZero),
       ECOG_tvary = ECOGValue) %>%
  dplyr::select(PatientID, tstart, ECOG_tvary) # 58896

# Label for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who had updated ECOG
  mutate(ECOG_tvary = ifelse(tstart == 0, ECOG_TimeZero, ECOG_tvary)) %>% # label everyone's baseline ECOG status
  arrange(PatientID, time) %>% # arrange rows before fill in missing ECOG_tvary values
  mutate(ECOG_tvary = na.locf(ECOG_tvary)) # label everyone's ECOG base on baseline ECOG until it changes (use zoo package)


#- Hb
# Find patients who had Hb after Time Zero
temp <- d_tvary_GUTG001_Hb %>%
  mutate(tstart = as.numeric(Gap_LabDate_TimeZero),
       Hb_tvary = TestResultCleaned*10) %>% # change test unit to g/L
  dplyr::select(PatientID, tstart, Hb_tvary) # 84466

# Label for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who had updated Hb
  mutate(Hb_tvary = ifelse(tstart == 0, Hb_TimeZero, Hb_tvary))  # label everyone's baseline
```

```
# relabel Hb for all DTR datasets
TTE_GUTG001_DTR_IPW <- relabel_Hb(TTE_GUTG001_DTR_IPW, "Hb_tvary")

# Change all unknown Hb_tvary_cat at tstart != in to NA, so that the zoo pakcage can be used
TTE_GUTG001_DTR_IPW$Hb_tvary_cat[TTE_GUTG001_DTR_IPW$tstart != 0 & TTE_GUTG001_DTR_IPW$Hb_tvary_cat == "Unknown"] <- NA

# Changed those with original value at time zero with NA as unknown, so that the zoo package can be used (not carrying the wrong value to the
next patient)
TTE_GUTG001_DTR_IPW$Hb_tvary[TTE_GUTG001_DTR_IPW$tstart == 0 & is.na(TTE_GUTG001_DTR_IPW$Hb_tvary)] <- "Unknown"

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary_cat))
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary_cat[TTE_GUTG001_DTR_IPW$tstart == 0]))
summary(as.factor(TTE_GUTG001_base$Hb_TimeZero_cat))

# label everyone's time-varying Hb_cat
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  arrange(PatientID, time) %>% # arrange rows before fill in missing tvary values
  mutate(Hb_tvary = na.locf(Hb_tvary)) %>% # label everyone's value base on baseline value until it changes (use zoo package)
  mutate(Hb_tvary_cat = na.locf(Hb_tvary_cat)) # label everyone's value base on baseline value until it changes (use zoo package)

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary_cat))
summary(as.factor(TTE_GUTG001_DTR_IPW$Hb_tvary_cat[TTE_GUTG001_DTR_IPW$tstart == 0]))


#- PSA
# Find patients who had PSA after Time Zero
temp <- d_tvary_GUTG001_PSA %>%
  mutate(tstart = as.numeric(Gap_LabDate_TimeZero),
         PSA_tvary = TestResultCleaned) %>% # change test unit to g/L
  dplyr::select(PatientID, tstart, PSA_tvary) # 31316

# Label for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who had updated PSA
  mutate(PSA_tvary = ifelse(tstart == 0, PSA_TimeZero, PSA_tvary))

# relabel PSA for all DTR datasets
TTE_GUTG001_DTR_IPW <- relabel_PSA(TTE_GUTG001_DTR_IPW, "PSA_tvary")

# Change all unknown values at tstart != in to NA, so that the zoo pakcage can be used
TTE_GUTG001_DTR_IPW$PSA_tvary_cat[TTE_GUTG001_DTR_IPW$tstart != 0 & TTE_GUTG001_DTR_IPW$PSA_tvary_cat == "Unknown"] <- NA

# Changed those with original value at time zero with NA as unknown, so that the zoo package can be used (not carrying the wrong value to the
next patient)
TTE_GUTG001_DTR_IPW$PSA_tvary[TTE_GUTG001_DTR_IPW$tstart == 0 & is.na(TTE_GUTG001_DTR_IPW$PSA_tvary)] <- "Unknown"

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary_cat))
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary_cat[TTE_GUTG001_DTR_IPW$tstart == 0]))
summary(as.factor(TTE_GUTG001_base$PSA_TimeZero_cat))

# label everyone's time-varying PSA_cat
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  arrange(PatientID, time) %>% # arrange rows before fill in missing tvary values
  mutate(PSA_tvary = na.locf(PSA_tvary)) %>%
  mutate(PSA_tvary_cat = na.locf(PSA_tvary_cat)) # label everyone's value base on baseline value until it changes (use zoo package)

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary_cat))
summary(as.factor(TTE_GUTG001_DTR_IPW$PSA_tvary_cat[TTE_GUTG001_DTR_IPW$tstart == 0]))



#- ALP (only for the purpose of table 1 cross-over characteristics). Therefore no need to estabalisehd categorical variable
# Find patients who had PSA after Time Zero
temp <- d_tvary_GUTG001_ALP %>%
  mutate(tstart = as.numeric(Gap_LabDate_TimeZero),
```

```
      ALP_RULN_tvary = TestResultCleaned/130) %>% # relative to ULN
  dplyr::select(PatientID, tstart, ALP_RULN_tvary) # 71456
summary(temp$ALP_RULN_tvary)

# Label for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who had updated PSA
  mutate(ALP_RULN_tvary = ifelse(tstart == 0, ALP_RULN_TimeZero, ALP_RULN_tvary))

# Changed those with original value at time zero with NA as unknown, so that the zoo package can be used (not carrying the wrong value to the
next patient)
TTE_GUTG001_DTR_IPW$ALP_RULN_tvary[TTE_GUTG001_DTR_IPW$tstart == 0 & is.na(TTE_GUTG001_DTR_IPW$ALP_RULN_tvary)] <- "Unknown"

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$ALP_RULN_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$ALP_RULN_tvary[TTE_GUTG001_DTR_IPW$tstart == 0]))
summary(TTE_GUTG001_base$ALP_RULN_TimeZero)  # 1399 unknown at baseline

TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  arrange(PatientID, time) %>% # arrange rows before fill in missing tvary values
  mutate(ALP_RULN_tvary = na.locf(ALP_RULN_tvary)) # label everyone's value base on baseline value until it changes (use zoo package)

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$ALP_RULN_tvary))



#- LDH (only for the purpose of table 1 cross-over characteristics)
# Find patients who had PSA after Time Zero
temp <- d_tvary_GUTG001_LDH %>%
  mutate(tstart = as.numeric(Gap_LabDate_TimeZero),
      LDH_RULN_tvary = TestResultCleaned/225) %>% # relative to ULN
  dplyr::select(PatientID, tstart, LDH_RULN_tvary) # 9683
summary(temp$LDH_RULN_tvary)

# Label for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add info of patients who had updated PSA
  mutate(LDH_RULN_tvary = ifelse(tstart == 0, LDH_RULN_TimeZero, LDH_RULN_tvary))
summary(TTE_GUTG001_DTR_IPW$LDH_RULN_TimeZero)

# Changed those with original value at time zero with NA as unknown, so that the zoo package can be used (not carrying the wrong value to the
next patient)
TTE_GUTG001_DTR_IPW$LDH_RULN_tvary[TTE_GUTG001_DTR_IPW$tstart == 0 & is.na(TTE_GUTG001_DTR_IPW$LDH_RULN_tvary)] <-
"Unknown"

# Check
summary(as.factor(TTE_GUTG001_DTR_IPW$LDH_RULN_tvary))
summary(as.factor(TTE_GUTG001_DTR_IPW$LDH_RULN_tvary[TTE_GUTG001_DTR_IPW$tstart == 0]))
summary(TTE_GUTG001_base$LDH_RULN_TimeZero)  # 3688 unknown at baseline

# relative to ULN
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  arrange(PatientID, time) %>% # arrange rows before fill in missing tvary values
  mutate(LDH_RULN_tvary = na.locf(LDH_RULN_tvary)) # label everyone's value base on baseline value until it changes (use zoo package)
summary(as.factor(TTE_GUTG001_DTR_IPW$LDH_RULN_tvary))



#- PD status (based on the initiation of next line treatment)
# (initiation of a x-line of treatment, meaning that on day x, pt already experienced disease progression)
summary(as.factor(d_tvary_GUTG001_PD$LineNumber_mCRPC_UserDef)) # max 10 times PD (initiation of line 2-11)
# Record line 2-6 PD

# create a function to label multiple linse of PD
label_PD <- function(TTEdata, PDdata, init_line_number) {

  pd_column_name <- paste0("PD", init_line_number - 1, "_tvary")

  temp <- PDdata %>%
    filter(LineNumber_mCRPC_UserDef == init_line_number) %>%
    mutate(tstart = as.numeric(Gap_PD_TimeZero)) %>%
    mutate(!!pd_column_name := 1) %>% # PDx-1 = progression after first-line (initiation of line x)
    dplyr::select(PatientID, tstart, !!rlang::sym(pd_column_name))
```

```
  TTEdata <- TTEdata %>%
    left_join(temp) %>% # add info of patients who had PD
    mutate(!!pd_column_name := ifelse(tstart == 0, 0, !!sym(pd_column_name))) %>%  # PD always started as 0
    arrange(PatientID, time) %>% # arrange rows before fill in missing values
    mutate(!!pd_column_name := na.locf(!!sym(pd_column_name))) # label pt's PD when it changes

  return(TTEdata)
}

# Label lines for all DTR datasets
# 1. TTE_GUTG001_DTR_IPW
# initiation of x-line
for (init_line_number in 2:11) {
  TTE_GUTG001_DTR_IPW <- label_PD(TTE_GUTG001_DTR_IPW, d_tvary_GUTG001_PD, init_line_number)
  summary(as.factor(TTE_GUTG001_DTR_IPW$PD2_tvary))
}

# PD during treatment on line x (initiation date of line x+1)
for (i in 1:10) {
  pd_column_name <- paste0("PD", i, "_tvary")
  print(paste("Summary for", pd_column_name, ":"))
  print(summary(as.factor(TTE_GUTG001_DTR_IPW[[pd_column_name]])))
}
# PD9 and PD10 seem to happen beyond censoring or the end of K (study timeframe, so no label 1)

# add a column of PD1_now to the DTR_IPW data
temp <- d_tvary_GUTG001_PD %>%
  filter(LineNumber_mCRPC_UserDef == 2) %>%
  mutate(tstart = as.numeric(Gap_PD_TimeZero)) %>%
  dplyr::select(PatientID, tstart) %>%
  mutate(PD1_Now_tvary = 1) # 2468 (distinct record)

# add a column of PD1_Recent to the DTR_IPW data
temp2 <- temp %>%
  mutate(tstart = tstart + 29) %>%
  rename(PD1_Recent30D_tvary = PD1_Now_tvary) # 2468

# add a column of PD1_Recent, after 30 days (PD1 Recent = 0) (append)
temp3 <- temp2 %>%
  mutate(tstart = tstart + 1) %>%
  mutate(PD1_Recent30D_tvary = 0)

temp2 <- rbind(temp2, temp3) %>%
  arrange(PatientID, tstart) # 4936

TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp) %>% # add PD1_now info of patients who had PD1 at that interval
  mutate(PD1_Now_tvary := case_when(
    PD1_Now_tvary == 1 ~ 1,
    TRUE ~ 0
  ))

# add a column of PD1_Recent30D_tvary to the DTR_IPW data
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  left_join(temp2, by = c("PatientID", "tstart"))

# add PD1_now info of patients who had PD1 at that interval
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  mutate(PD1_Recent30D_tvary := case_when(
    PD1_Now_tvary == 1 ~ 1,
    TRUE ~ PD1_Recent30D_tvary
  )) %>%
  mutate(PD1_Recent30D_tvary = ifelse(tstart == 0, 0, PD1_Recent30D_tvary)) %>%
  arrange(PatientID, tstart) %>% # arrange rows before fill in missing tvary values
  mutate(PD1_Recent30D_tvary = na.locf(PD1_Recent30D_tvary)) # label everyone's value base on baseline value until it changes (use zoo package)
summary(as.factor(TTE_GUTG001_DTR_IPW$PD1_Recent30D_tvary))
# 35123 intervals of recent PD (1.47%), 2358996
```

```
#- Comorbidity Score (time-varying)
# Comorbidity score time-varying table has been cleaned, everyone has score for each seperate comorbid 0 or 1 (n = 4144),
# with any additional comorbidity incidence (new diagnosis) for each comorbid item labeled as 1 at each time point
# therefore the Comorbidity_tvary table can be directly merged with the long dataset

# Change Gap_Comorbid_TimeZero to tstart for merging with the long TTE dataset
temp <- d_tvary_GUTG001_CoMorbid %>%
  mutate(tstart = as.numeric(Gap_Comorbid_TimeZero)) %>%
  dplyr::select(-Gap_Comorbid_TimeZero)

variable_names <- c(paste("Comorbid_", c(unique(v_comorbid_names), "AnyTumour", "Metastatic_cancer"), "_tvary", sep = ""))

# label all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
   left_join(temp, by = c("PatientID", "tstart")) %>%
   arrange(PatientID, time) %>% # arrange rows before fill in missing tvary values
   mutate(across(all_of(variable_names), ~na.locf(.))) # label everyone's value base on baseline value until it changes (use zoo package)

# Make a function to label every patient's time-varying cormorbidity score, adding any tumour and metastatsis = 1
label_comorbid_tvary <- function(TTEdata) {
TTEdata <- TTEdata %>%
   # Combined comorbidity score conditions and weights for a Medicare population (Gagne 2011)
   # Not necessarily suitable for other population
   mutate(Comorbid_Score_tvary =
          5*Comorbid_Metastatic_cancer_tvary +
          2*Comorbid_CHF_tvary +
          2*Comorbid_Dementia_tvary +
          2*Comorbid_RenalFail_tvary +
          2*Comorbid_WeightLoss_tvary +
          Comorbid_AnyTumour_tvary +
          Comorbid_Alcohol_abuse_tvary +
          Comorbid_Cardiac_arrhythmias_tvary +
          Comorbid_Chronic_pulmonary_disease_tvary +
          Comorbid_Coagulopathy_tvary +
          Comorbid_Complicated_DM_tvary +
          Comorbid_Deficiency_anemia_tvary +
          Comorbid_Fluid_Electro_tvary +
          Comorbid_Hemiplegia_tvary +
          Comorbid_LiverD_tvary +
          Comorbid_PeripheralVasc_tvary +
          Comorbid_Psychosis_tvary +
          Comorbid_PulmonaryCirc_tvary +
          (-1 * Comorbid_HIV_tvary) + # There is are patients having HIV but without other non-any tumour, non-meta comorbid diagnosis within 182
days
          (-1 * Comorbid_HTN_tvary) # not sure why HTN os -1
   )
 return(TTEdata)
}
gc()

# label all DTR datasets
TTE_GUTG001_DTR_IPW <- label_comorbid_tvary(TTE_GUTG001_DTR_IPW)
summary(as.factor(TTE_GUTG001_DTR_IPW$Comorbid_Score_tvary)) # ranging from 5~19

# Delete individual time-varying comorbidity item for clarity for all DTR datasets
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  dplyr::select(-variable_names)



#***Label "Strategy" for each DTR dataset
TTE_GUTG001_DTR_IPW <- TTE_GUTG001_DTR_IPW %>%
  mutate(Seq_Strategy = case_when(
    FirstLine == "Abiraterone" ~ "Abiraterone -> Enzalutamide", # 1210088 intervals
    TRUE ~ "Enzalutamdie -> Abiraterone" # 1184031 intervals
  ))
summary(as.factor(TTE_GUTG001_DTR_IPW$Seq_Strategy))


# save ready-to-analyse data
# save(TTE_GUTG001_base, file = "derived_data\\TTE_GUTG001_base.RData")
# save(TTE_GUTG001_ITT, file = "derived_data\\TTE_GUTG001_ITT.RData")
# save(TTE_GUTG001_DTR_IPW, file = "derived_data\\TTE_GUTG001_DTR_IPW.RData")
```

Appendix 8.9 Reconstructing the data and findings from published Kaplan-Meier survival curves in GUTG-001

```
################################################################################
###   Digitisation of Published survival curves  & fundings                  ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                     ###
###                                                                           ###
###   University of Sheffield                                                 ###
###   Date created: Nov 19, 2023                                              ###
###   Git & renv project created on Aug 01, 2023                              ###
################################################################################

# Install & load the library of survHE package for using the digitise function
# survHE requires pre-installation of Rcpp, flexurv, survival pakcages

install.packages("Rcpp")
install.packages("flexsurv")
install.packages("survival")
install.packages("survHE")

library(Rcpp)
library(flexsurv)
library(survival)
library(survHE)

# check the help file of survHE package
help(survHE)
help("fit.models")
# check the help file of digitise function from the survHE package
help("digitise")
help("surv_inp")

# Gianluca's reply on data format for digitise
# https://github.com/giabaio/survHE/issues/24
# https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-9


# Digitising the OS of the Khalaf paper (GUTG-001)
# https://pubmed.ncbi.nlm.nih.gov/31727538/

# pre-modification to the original data (avoid manually key-in data to prevent mistakes): max survival should be 1 not 100
# use coarse file cuz too many points might violate the monotonicity, not compatabile with codes derived from guyot et al.
digizeit_Abi <- read.csv("digitise/Surv_Khalaf_Digitise_OS_GroupA_Abi_coarse.csv", header = TRUE, row.names = NULL) # 93 rows, 2 vars

# create a new column (indicating serial of observations) to digizeit in front of the original columns
serial_num <- data.frame(serial = 1:length(digizeit_Abi$Time))
digizeit_Abi <- cbind(serial_num, digizeit_Abi)


nrisk_Abi <- read.table("digitise/nrisk_Khalaf2019_OS_Abi_coarse.txt", header = TRUE, row.names = NULL)

head(digizeit_Abi)
head(nrisk_Abi)

# save it as a new txt file  (digitise need txt)
# https://github.com/giabaio/survHE/issues/31
write.table(digizeit_Abi, file = "digitise/Surv_Khalaf_Digitise_OS_GroupA_Abi_coarse.txt", sep = "\t",
            row.names = FALSE)


# For enza arm

digizeit_Enza <- read.csv("digitise/Surv_Khalaf_Digitise_OS_GroupB_Enza_coarse2.csv", header = TRUE, row.names = NULL) # 133 rows, 2 vars (use
the second one, firsto one is a bit unprecise)

# create a new column (indicating serial of observations) to digizeit in front of the original columns
serial_num <- data.frame(serial = 1:length(digizeit_Enza$Time))
digizeit_Enza <- cbind(serial_num, digizeit_Enza)


nrisk_Enza <- read.table("digitise/nrisk_Khalaf2019_OS_Enza_coarse2.txt", header = TRUE, row.names = NULL)
```

```
# save it as a new txt file  (digitise need txt)
write.table(digizeit_Enza, file = "digitise/Surv_Khalaf_Digitise_OS_GroupB_Enza_coarse2.txt", sep = "\t",
          row.names = FALSE)


# rerun digitise analysis with the new file: no need to define output because there is default name
digitise (surv_inp = "digitise/Surv_Khalaf_Digitise_OS_GroupB_Enza_coarse2.txt",
        nrisk_inp = "digitise/nrisk_Khalaf2019_OS_Enza_coarse2.txt")
# Kaplan Meier data written to file: .KMdata.txt
# IPD data written to file: .IPDdata.txt


Khalaf.plots <- list()
Khalaf.plots [[1]] <- ggplot() +
  geom_line(data = digizeit_Abi, aes(x = Time, y = Survival * 100), group = "Group A (Abi -> Enza)", color = lancet_colors[1]) +
  geom_point(data = digizeit_Abi, aes(x = Time, y = Survival * 100), group = "Group A (Abi -> Enza)", color = lancet_colors[1]) +
  geom_line(data = digizeit_Enza, aes(x = Time, y = Survival * 100), group = "Group B (Enza -> Abi)", color = lancet_colors[2]) +
  geom_point(data = digizeit_Enza, aes(x = Time, y = Survival * 100), group = "Group B (Enza -> Abi)", color = lancet_colors[2]) +
  scale_x_continuous(breaks = seq(0, 45, by = 9), name = "Time from start of first-line therapy (months)") +
  scale_y_continuous(breaks = seq(0, 100, by = 20), name = "Overall Survival (%)", limits = c(0, 100)) +
  theme_minimal() +
  labs(title = "Overall survival (Khalaf et al. 2019)", subtitle = "Group A=abiraterone plus prednisone followed by enzalutamide
  \nGroup B=enzalutamide followed by abiraterone plus prednisone.") +
  scale_color_manual(values = lancet_colors, labels = c("Group A (Abi -> Enza)", "Group B (Enza -> Abi)")) +
  guides(color = guide_legend(title = "Treatment Groups"))


# Read the table of KM data:) for checking results :D
KMdata_Khalaf_Abi <- read.table("digitise/KMdata_Khalaf_GroupA.txt", header = TRUE, row.names = NULL) # Proxy of each censor point using
sandwich-ish strategy (my understanding)
IPDdata_Khalaf_Abi <- read.table("digitise/IPDdata_Khalaf_GroupA.txt", header = TRUE, row.names = NULL) # proxy of the time-to-event data for
every patients (patient number = n.risk at time 0)

# Read the table of KM data:) for checking results :D
KMdata_Khalaf_Enza <- read.table("digitise/KMdata_Khalaf_GroupB2.txt", header = TRUE, row.names = NULL) # Proxy of each censor point using
sandwich-ish strategy (my understanding)
IPDdata_Khalaf_Enza <- read.table("digitise/IPDdata_Khalaf_GroupB2.txt", header = TRUE, row.names = NULL) # proxy of the time-to-event data for
every patients (patient number = n.risk at time 0)


# Add a group variable
IPDdata_Khalaf_Abi$Group <- 0 # numbering for for plot plotting purpose
IPDdata_Khalaf_Enza$Group <- 1

# Combine the datasets
IPDdata_Khalaf_combined <- rbind(IPDdata_Khalaf_Abi, IPDdata_Khalaf_Enza)

# Fit Kaplan-Meier model to combined data
fit.km.Khalaf2019 <- survfit(Surv(time, event) ~ Group, data = IPDdata_Khalaf_combined)

summary(fit.km.Khalaf2019)$table

# Compute median follow up
quantile(with(IPDdata_Khalaf_combined, Surv(time, event)), probs = 0.5) # 27.31 (24.26, 34.05)
summary(IPDdata_Khalaf_combined$time)

# Plot
Khalaf.plots [[2]] <- ggsurvplot(
    fit_Khalaf ,
    conf.int = TRUE,
    pval = TRUE,
    conf.int.alpha = 0.1,
    censor = TRUE,
    censor.shape = 3,  # Shape of censoring points (default is '+')
    censor.size = 3,   # Increased size of censoring points
    risk.table = "nrisk_cumcensor",
    fontsize = 4,
    linetype = "strata",
    break.x.by = 9,
    xlim = c(0, 45),
```

```
    xlab = "Time from start of first-line therapy (months)",
    title = "Overall Survival - reproduction from Khalaf et al.",
    subtitle = "Reconstruct KM curve/IPD data using Guyot et al. digitisation method \nGroup A = abiraterone followed by enzalutamide, Group B =
enzalutamide followed by abiraterone",
    surv.median.line = "hv",
    ggtheme = theme_bw(),
    palette = lancet_colors,
    legend.title = "Randomisation Group",
    legend.labs = c("Group A", "Group B")
  )
Khalaf.plots [[2]]


# Combine and arrange the plots with custom layout
combined_plot <- grid.arrange(
  Khalaf.plots [[1]],
  Khalaf.plots[[2]]$plot,
  Khalaf.plots[[2]]$table,
  layout_matrix = rbind(
    c(1, 2),
    c(1, 2),
    c(1, 2),
    c(1, 2),
    c(1, 3)
  )
)



#### Fit a Cox model to the digitise curve####
TTE_KM_Khalaf2019_wide <- rbind(IPDdata_Khalaf_Enza, IPDdata_Khalaf_Abi) %>%
  mutate(FirstLine_num = case_when(
    Group == 0 ~ 1,
    TRUE ~ 0
  )) %>%
  rename(Y_Death = event,
         SurvTime = time) %>%
  select(-arm, -Group)


fit.cox.Khalaf2019 <- coxph(Surv(SurvTime, Y_Death) ~ FirstLine_num,
                            data = TTE_KM_Khalaf2019_wide) # default method = "efron"
summary(fit.cox.Khalaf2019) # HR: 0.783 (0.535, 1.147)
AIC(fit.cox.Khalaf2019) # 39657.08
BIC(fit.cox.Khalaf2019) # 39708.69
logLik(fit.cox.Khalaf2019) # -19824.54 (df=1)
```

Appendix 8.10 R code for survival analyses in Prostate Cancer Case Study 1 (PC1)

```
###Project Info ##############################################################################
###   Target Trial Emulation – Final survival analyses                                      ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                                   ###
###   Data: Flatiron prostate cancer datasets                                               ###
###                                                                                         ###
###   University of Sheffield                                                               ###
###   Date created: Nov 16, 2023                                                            ###
##############################################################################################
# Part of this code has adapted and extended insights from the CAUSALab courses at Harvard University as well as
# the causal inference course at the Private University for Health Sciences and Health Technology in Hall in Tirol (UMiT)

## House keeping
# Load packages and functions
renv::status()
# renv::snapshot()

`%notin%` <- Negate(`%in%`)

# install backports when there are conflicts of tidyr
library(backports)
library(tidyr)
library(remotes)
# install.packages("tidyr")
library(tidyr)
library(dplyr)
library(lubridate)
library(ggpubr)
library(survival)
library(zoo)
library(data.table) # faster than data.frame
library(speedglm)
library(dplyr) # too avoid conflict, read dplyr again
library(boot)
library(scales)
library(date)
library(survminer) # use tidy R 1.3.0
library(splines)
library(sandwich)
library(lmtest)
# library(broom) # use tidy R 1.2.0
# library(haven)
# library(psych)
# library(flexsurv)
# library(cowplot)
# library(geepack)
# library(Hmisc)
# library(eha)
# library(rpsftm)
# library(rms)

#############################################
#   TTE ANALYSES STARTING FROM HERE         ####
#############################################

##############################################################################################
### 1a. Estimating Effect in a Target Trial (Unadjusted ITT)                  ##################
##############################################################################################

#***1a.1. Non parametric estimators for time-to-event outcome: KM for comparison of risks ####

#### KM estimator stratified by study arms ####

# Use the baseline wide dataset (Y_FirstLine has been trimmed based on the max observation time K)
fit.km  <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                   conf.type = "log-log", # maintain constant hazard ratio assumption
                   data = TTE_GUTG001_base)
summary(TTE_GUTG001_base$SurvTime_FirstLine)

# Review KM survival estimates, and estimates at time = K (default = 1440 days)
fit.km
# Review median survival time
summary(fit.km)$table
```

```
summary(fit.km, times = K) # slight difference when setting SurvTime at K or K + 1 for admin censoring where patients still alive at K

# Out put median follow-up time
summary(TTE_GUTG001_base$SurvTime_FirstLine, na.rm = TRUE) # 643 (346.8, 1061.5)


####Construct KM curves ####
kmplots <- list()

kmplots[[1]] <- ggsurvplot(fit.km,
                    conf.int = TRUE,
                    pval = TRUE,
                    censor = FALSE, # don't include tick marks for events/censorings
                    # risk.table = TRUE, # Add risk table
                    risk.table = "nrisk_cumcensor", # seem to be weird
                    fontsize = 3,
                    risk.table.col = "strata", # Change risk table color by groups
                    linetype = "strata", # Change line type by groups
                    break.x.by = 180,
                    xlab = "Days", # label x-axis
                    title = "Unadjusted survival of GUTG-001 Analogue mCPRC patients in Flatiron data",
                    subtitle = "Included all patients who started with Abiraterone or Enzalutamide (max: 4-year follow-up)",
                    legend.title = "mCRPC treatment",
                    legend.labs = c("First-line: Enzalutamide", "First-line: Abiraterone"),
                    surv.median.line = "hv", # Specify median survival
                    ggtheme = theme_bw(), # Change ggplot2 theme
                    palette = c("#E7B800", "#2E9FDF", "orchid2"))
kmplots[[1]]



#***1a.2. Cox proportional hazards model (Semi-parametric) for comparison ####
#### Fit a Cox model ####
fit.cox <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
            data = TTE_GUTG001_base) # default method = "efron"
levels(TTE_GUTG001_base$FirstLine_num)
summary(fit.cox)
AIC(fit.cox) # 39657.08
BIC(fit.cox) # 39708.69
logLik(fit.cox) # -19824.54 (df=1)

#### Calculate HR####
results.unadj.cox.hr.K <- list(
  Hazard_Ratio = exp(coef(fit.cox)), # Calculate the hazard ratio
  CI95_Lower = exp(confint(fit.cox)[, 1]), # Calculate 95% confidence intervals from the model's coefficients
  CI95_Upper = exp(confint(fit.cox)[, 2]) # Exponentiate to get the 95% confidence intervals for the hazard ratios
)
# Print the results
print(results.unadj.cox.hr.K)

#### Plot HR ####
# library(rms)
hazard.ratio.plot(as.numeric(TTE_GUTG001_base_reduced$FirstLine_num),
              Surv(TTE_GUTG001_base_reduced$SurvTime_FirstLine, TTE_GUTG001_base_reduced$Y_FirstLine),
              e=20,
              legendloc='ll',
              antilog = TRUE )
unadj.coxplots[["HR_plot"]] <- recordPlot()
unadj.coxplots[["HR_plot"]]

# ggplot version
# ggplot(km_GUTG001Analogue_data, aes(x = log(time), y = cloglog_surv, color = strata)) +
#  geom_line() +
#  scale_x_log10() + # Apply log scale to x-axis
#  labs(x = "Time (log scale)", y = "log(-log S(t))", color = "Strata") +
#  theme_minimal(



##############################################################################################
### 1b. Estimating Unadjusted Effect in a Target Trial (PP)                          ############
##############################################################################################
```

```
#***1b.1. Non-parametric KM estimators: unadjusted per-protocol effect ####
#### KM estimator stratified by study arms ####


# Use the baseline wide dataset (Y_FirstLine has been trimmed based on the max observation time K)
fit.km.pp  <- survfit(Surv(SurvTime_DTR, Y_DTR) ~ FirstLine_num,
              conf.type = "log-log",
              data = TTE_GUTG001_base)


# Review KM survival estimates, and estimates at time = K (default = 1440 days)
summary(fit.km.pp)
summary(fit.km.pp, times = K)


summary(fit.km.pp)$table
 summary(TTE_GUTG001_base$SurvTime_DTR)

# Out put median follow-up time
summary(TTE_GUTG001_base$SurvTime_FirstLine, na.rm = TRUE) # 643 (346.8, 1061.5)



####Construct KM curves ####
km.pp.plots <- list()
km.pp.plots[[1]] <- ggsurvplot(fit.km.pp,
                    conf.int = TRUE,
                    pval = TRUE,
                    censor = FALSE, # don't include tick marks for events/censorings
                    # risk.table = TRUE, # Add risk table
                    risk.table = "nrisk_cumcensor", # seem to be weird
                    fontsize = 3,
                    risk.table.col = "strata", # Change risk table color by groups
                    linetype = "strata", # Change line type by groups
                    break.x.by = 180,
                    xlab = "Days", # label x-axis
                    title = "Unadjusted per-protocol survival of GUTG-001 Analogue mCPRC patients in Flatiron data",
                    subtitle = "Included all patients who initiated treatment with Abiraterone or Enzalutamide. \nPatients were censored when they
began receiving a second-line treatment not stipulated in the protocol (max: 4-year follow-up)",
                    legend.title = "mCRPC treatment",
                    legend.labs = c("Group B: Enza-Abi", "Group A: Abi-Enza"),
                    surv.median.line = "hv", # Specify median survival
                    ggtheme = theme_bw(), # Change ggplot2 theme
                    palette = c("#E7B800", "#2E9FDF", "orchid2"))
km.pp.plots[[1]]


#***1b.2. Cox estimators: Unadjusted per-protocol effect ####

fit.pp.cox <- coxph(Surv(SurvTime_DTR, Y_DTR) ~ FirstLine_num,
                 data = TTE_GUTG001_base_reduced) # default method = "efron"

summary(fit.pp.cox)
AIC(fit.pp.cox) # 24672.21
BIC(fit.pp.cox) # 24677.61
logLik(fit.pp.cox) # -12335.11 (df=1)



##############################################
### 1c & 1d. Unadjusted As-treated patients     #######
##############################################


# 1c & 1d.1  For As-treated (limited to second-Line treatment recipients)####

# Mimicking Terada 2017 & Matsubara 2018

# As-treated (limited to second-Line treatment recipients)
# Selecting only patients being able to received second-line (immortal time bias)
temp <- TTE_GUTG001_DTR_IPW_reduced$PatientID[TTE_GUTG001_DTR_IPW_reduced$PD1_Now_tvary == 1 &
TTE_GUTG001_DTR_IPW_reduced$c_treatdevL2 == 0] # 1132 patients

options(digits=7)
```

```
summary(as.factor(TTE_GUTG001_base_reduced$FirstLine_num[TTE_GUTG001_base_reduced$PatientID %in% temp]))
fit.km.at.2L <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
              data = TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %in% temp, ])
fit.km.at.2L

# Out put median follow-up time
summary(TTE_GUTG001_base_reduced$SurvTime_FirstLine[TTE_GUTG001_base_reduced$PatientID %in% temp], na.rm = TRUE) # 812 (528.5,
1172.5)

# n events median 0.95LCL 0.95UCL
# FirstLine_num=0 487   312   897    836    988
# FirstLine_num=1 645   392   957    920   1042

fit.cox.at.2L <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
          data = TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %in% temp, ]) # default method = "efron"
summary(fit.cox.at.2L)
AIC(fit.cox.at.2L)
BIC(fit.cox.at.2L)
logLik(fit.cox.at.2L)
# 0.92 (0.80-1.07)

# 5.2 For As-treated (including patients only received first-line)####

# As-treated
# Selecting all patients who did not violated second-line (conditioned on post-treatment covariates)
temp2 <- TTE_GUTG001_DTR_IPW_reduced$PatientID[TTE_GUTG001_DTR_IPW_reduced$c_treatdevL2 == 1] # 1132 patients

options(digits=7)

summary(as.factor(TTE_GUTG001_base_reduced$FirstLine_num[TTE_GUTG001_base_reduced$PatientID %notin% temp2])) # enza 1358, abi: 1500
fit.km.at <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                data = TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %notin% temp2, ])
fit.km.at

# Out put median follow-up time
summary(TTE_GUTG001_base_reduced$SurvTime_FirstLine[TTE_GUTG001_base_reduced$PatientID %notin% temp2], na.rm = TRUE)
# 594 (285.2, 1024.8)

fit.cox.at <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                data = TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %notin% temp2, ]) # default method = "efron"
summary(fit.cox.at)
AIC(fit.cox.at)
BIC(fit.cox.at)
logLik(fit.cox.at)
# 1.14 (1.03-1.25)


##############################################
### 2a. Baseline adjusted ITT patients       #######
##############################################

# Note: need to run the derivation of IPTW in 2b. IPTW adjusted pp analysis before running this the 2a.section
#***2a.1. Fit weighted KM####
options(digits=5)
fit.iptw.b.km.itt <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                  data = TTE_GUTG001_base_reduced,
                  weights = w.treat_baseline.s)
fit.iptw.b.km.itt
summary(fit.iptw.b.km.itt)$table

# Out put median follow-up time
weighted_followup <- TTE_GUTG001_base_reduced$SurvTime_FirstLine * TTE_GUTG001_base_reduced$w.treat_baseline.s
summary(weighted_followup) # 618.928 (322.968-1012.514)

#***2a.2. Fit weighted Cox####
options(digits=7)
fit.iptw.b.cox.itt <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                data = TTE_GUTG001_base_reduced,
                weights = w.treat_baseline.s) # default method = "efron"

summary(fit.iptw.b.cox.itt) # 1.11 (1.02-1.20)
AIC(fit.iptw.b.cox.itt) # 37929.09
BIC(fit.iptw.b.cox.itt) # 37934.89
```

```
logLik(fit.iptw.b.cox.itt) # -18963.54 (df=1)

# HR
exp(coefficients(fit.iptw.b.cox.itt))

# Naive 95% CI
exp(confint(fit.iptw.b.cox.itt))

# Sandwich robust 95% CI
coeftest(fit.iptw.b.cox.itt, vcov = sandwich)
exp(confint(coeftest(fit.iptw.b.cox.itt, vcov = sandwich))) # 1.02-1.20



################################################################################
### 2b. Baselin eadjusted (IPW) - PP                          ##############################
################################################################################

library(speedglm)
gc()

# reduce dataset
TTE_GUTG001_base_reduced <- TTE_GUTG001_base %>%
  dplyr::select(PatientID, FirstLine_num, Age_TimeZero, cal_time, # keep essential variables
        GroupStage, MStage, GleasonScore, DM_TimeZero,
        f_pretreat_treat_allowed_docetaxel, f_pretreat_treat_other_systemic, f_SES,
        Race, Gap_TimeZero_mCRPC_cat, Gap_TimeZero_mPC_cat,
        PSA_TimeZero_cat, Hb_TimeZero_cat, ECOG_TimeZero, DM_TimeZero,
        Comorbid_Score_TimeZero,
        TypeY_FirstLine, TypeY_DTR, # keep outcome variables
        Y_FirstLine, Y_DTR,
        SurvTime_FirstLine, SurvTime_DTR) # 4144 rows, 23 vars


#***2b.1: Fit a IPW (treatment) to estimate the denominator of the nonstabilised weights ####

# non-saturated model, high-dimensional

# IPW treatment model only considering treatment as a point intervention
# For per-protocol population (all)
iptw.b.pp.denom <- speedglm(FirstLine_num ~ # Propensity for Strategy 1 Abi-Enza (first-line being abiraterone) treatment

                    ## Time-fixed covariates: baseline
                    Age_TimeZero + # Omitted quadratic age term I(Age_TimeZero^2) due to computational constraints
                    factor(cal_time) + # Calendar year of treatment start influences treatment strategy (availability treatments change)
                    factor(GroupStage) + factor(MStage) + # Group stage encompasses N, M stage information, but missingness in GroupStage is larger
                    factor(GleasonScore) +  # Gleason Score
                    factor(f_pretreat_treat_allowed_docetaxel) + factor(f_pretreat_treat_other_systemic) + # Prior docetaxel/other treatment
                    factor(f_SES) +
                    factor(Race) + # Race, considering potential grouping of Asians
                    # Diagnosis to Treatment Initiation Gap
                    factor(Gap_TimeZero_mCRPC_cat) +
                    factor(Gap_TimeZero_mPC_cat) +

                    ## Baseline values of time-varying covariates
                    factor(PSA_TimeZero_cat) + factor(Hb_TimeZero_cat) + factor(ECOG_TimeZero) +
                    Comorbid_Score_TimeZero + factor(DM_TimeZero),

                  data = TTE_GUTG001_base_reduced, # Use only baseline data
                  family = binomial(link="logit"))
# 0.094 secs
summary(iptw.b.pp.denom)

# Add IPTW_baseline propensity to TTE_GUTG001_base_reduced data
TTE_GUTG001_base_reduced$p.treat_baseline.denom <- predict(iptw.b.pp.denom, TTE_GUTG001_base_reduced, type="response")


#***2b.2: Fit a IPW (treatment) to estimate the numerator of the stabilised weights ####
# Specify a logistic regression model that models treatment strategy group as a function of only an intercept
# (use glm as speedglm might not converge for this, 4.2 MB)
iptw.b.pp.num <- glm(FirstLine_num ~ 1, family = binomial(link="logit"), data = TTE_GUTG001_base_reduced) # error can occur with speedglm here
summary(iptw.b.pp.num)

# Add numerator to TTE_GUTG001_base_reduced data
```

```
TTE_GUTG001_base_reduced$p.treat_baseline.num <- predict(iptw.b.pp.num, type="response")


#***2b.3: Calculate nonstablised IPTW (baseline) ####
# Calculate nonstablaised weights (based on baseline treatment)
TTE_GUTG001_base_reduced$w.treat_baseline.ns <-
  ifelse(TTE_GUTG001_base_reduced$FirstLine_num == 1,
     1/TTE_GUTG001_base_reduced$p.treat_baseline.denom,
     1/(1 - TTE_GUTG001_base_reduced$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced$w.treat_baseline.ns) # 1.086-10.425
sd(TTE_GUTG001_base_reduced$w.treat_baseline.ns) # SD: 0.618

# Check propensity
ipw.plots <- list()
ipw.plots[["p.treat_baseline.denom"]] <- ggplot(TTE_GUTG001_base_reduced, aes(x = p.treat_baseline.denom,
                                                fill = as.factor(FirstLine_num))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 100) +
  theme_minimal() +
  scale_fill_manual(values = c("maroon", "skyblue"),
                labels = c("Enza-Abi", "Abi-Enza"),
                name = "mCPRC treatment strategy") +
  ggtitle("Propensity Score Overlap (baseline treatemnt)")
ipw.plots[["p.treat_baseline.denom"]]


# check nostabalised weights
ipw.plots[["w.treat_baseline.ns"]]  <- ggplot(TTE_GUTG001_base_reduced, aes(x = w.treat_baseline.ns,
  fill = as.factor(FirstLine_num))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 500) +
  theme_minimal() +
  scale_fill_manual(values = c("red", "blue"),
                labels = c("Enza-Abi", "Abi-Enza"),
                name = "mCPRC treatment strategy") +
  ggtitle("Distribution of baseline treatemnt weights (non-stablised)")
ipw.plots[["w.treat_baseline.ns"]]

#***2b.4: Calculate stablised IPTW (baseline) ####

# Compute stablised weights
TTE_GUTG001_base_reduced$w.treat_baseline.s <-
  ifelse(TTE_GUTG001_base_reduced$FirstLine_num == 1,
       TTE_GUTG001_base_reduced$p.treat_baseline.num/TTE_GUTG001_base_reduced$p.treat_baseline.denom,
       (1 - TTE_GUTG001_base_reduced$p.treat_baseline.num)/(1 - TTE_GUTG001_base_reduced$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced$w.treat_baseline.s) # 0.5682~5.0164, mean = 1.0006
sd(TTE_GUTG001_base_reduced$w.treat_baseline.s) # 0.3004998

# check stabalised weights
ipw.plots[["w.treat_baseline.s"]]  <- ggplot(TTE_GUTG001_base_reduced, aes(x = w.treat_baseline.s,
                                                fill = as.factor(FirstLine_num))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 500) +
  theme_minimal() +
  scale_fill_manual(values = c("red", "blue"),
                labels = c("Enza-Abi", "Abi-Enza"),
                name = "mCPRC treatment strategy") +
  ggtitle("Distribution of baseline treatemnt weights (stablised)")


ggarrange(
  ipw.plots[["p.treat_baseline.denom"]],  # Top graph
  ggarrange(
    ipw.plots[["w.treat_baseline.ns"]],
    ipw.plots[["w.treat_baseline.s"]],
    ncol = 2, nrow = 1  # 2 bottom graphs side by side
  ),
  ncol = 1, nrow = 2,    # overall layout: 1 column, 2 rows
  heights = c(1, 1)     # equal heights for top and bottom
)

gc()
```

```
#***2b.5: MSM with stabalised weights ####

# reduce dataset: use Per-protocol (DTR dataset)
# Add treatment base-line weight
TTE_GUTG001_DTR_IPW_reduced <- TTE_GUTG001_DTR_IPW %>%
  dplyr::select(PatientID, FirstLine_num, Age_TimeZero, cal_time, # keep essential variables
            GroupStage, MStage, GleasonScore, DM_TimeZero,
            f_pretreat_treat_allowed_docetaxel, f_pretreat_treat_other_systemic, f_SES,
            Race, Gap_TimeZero_mCRPC_cat, Gap_TimeZero_mPC_cat,
            PSA_TimeZero_cat, Hb_TimeZero_cat, ECOG_TimeZero, Comorbid_Score_TimeZero, DM_TimeZero,
            PSA_tvary_cat, Hb_tvary_cat, ECOG_tvary, Comorbid_Score_tvary, DM_tvary,
            PD1_tvary,
            PD1_Now_tvary, PD1_Recent30D_tvary, # all patients all only remained switched on the date of first-progress/prior to progression #
Should I add progression now indicator?
            time,y,
            c_admin, c_treatdevL2) %>%
  left_join(TTE_GUTG001_base_reduced[ , c("PatientID",
                                "p.treat_baseline.denom", "p.treat_baseline.num",
                                "w.treat_baseline.ns", "w.treat_baseline.s")],
          by = "PatientID") %>%
  mutate(y = as.integer(y)) # need to be integer for outcome models

# data.table faster
TTE_GUTG001_DTR_IPW_reduced <- data.table(TTE_GUTG001_DTR_IPW_reduced)
# 2394119
gc()


#***2b.6. Fit weighted KM####
library(sandwich)
options(digits=5)
fit.iptw.b.km <- survfit(Surv(SurvTime_DTR, Y_DTR) ~ FirstLine_num,
                      data = TTE_GUTG001_base_reduced,
                      weights = w.treat_baseline.s)
fit.iptw.b.km

# Review KM survival estimates, and estimates at time = K (default = 1440 days)
summary(fit.iptw.b.km)
summary(fit.iptw.b.km, times = K)

summary(fit.iptw.b.km)$table

# Estimates at K days (default = 1440 days)
results.iptw.b.km.K <- compute_km_estimates(fit.iptw.b.km, K)
# Print results
results.iptw.b.km.K

# Out put median follow-up time
weighted_followup <- TTE_GUTG001_base_reduced$SurvTime_DTR * TTE_GUTG001_base_reduced$w.treat_baseline.s
summary(weighted_followup) # 444.791 (202.917-841.439)


####Construct KM curves ####
options(digits=2)
kmplots[["fit.iptw.b.km"]] <- ggsurvplot(fit.iptw.b.km,
                    conf.int = TRUE,
                    pval = TRUE,
                    censor = FALSE, # don't include tick marks for events/censorings
                    # risk.table = TRUE, # Add risk table
                    risk.table = "nrisk_cumcensor", # seem to be weird
                    fontsize = 3,
                    risk.table.col = "strata", # Change risk table color by groups
                    linetype = "strata", # Change line type by groups
                    break.x.by = 180,
                    xlab = "Days", # label x-axis
                    title = "Baseline adjusted survival of GUTG-001 Analogue mCPRC patients in Flatiron data (IPTW weighted KM)",
                    subtitle = "Included all patients who started with Abiraterone or Enzalutamide (max: 4-year follow-up)",
                    legend.title = "mCRPC treatment",
                    legend.labs = c("First-line: Enzalutamide", "First-line: Abiraterone"),
                    surv.median.line = "hv", # Specify median survival
                    ggtheme = theme_bw(), # Change ggplot2 theme
                    palette = c("#E7B800", "#2E9FDF", "orchid2"))
kmplots[["fit.iptw.b.km"]]
```

```
#***2b.7. Fit weighted Cox####
options(digits=7)
fit.iptw.b.cox <- coxph(Surv(SurvTime_DTR, Y_DTR) ~ FirstLine_num,
                  data = TTE_GUTG001_base_reduced,
                  weights = w.treat_baseline.s) # default method = "efron"

summary(fit.iptw.b.cox)
AIC(fit.iptw.b.cox) # 24505.77
BIC(fit.iptw.b.cox) # 24511.17
logLik(fit.iptw.b.cox) # -12251.89 (df=1)

# HR
exp(coefficients(fit.iptw.b.cox))

# Naive 95% CI
exp(confint(fit.iptw.b.cox))

# Sandwich robust 95% CI
library("sandwich")
library("lmtest")
coeftest(fit.iptw.b.cox, vcov = sandwich)
exp(confint(coeftest(fit.iptw.b.cox, vcov = sandwich)))

# Extract absolute risks at K
iptw.b.cox0 <- compute_risks_cox(fit.iptw.b.cox, "FirstLine_num", 0, 0:K_minus_1)
iptw.b.cox1 <- compute_risks_cox(fit.iptw.b.cox, "FirstLine_num", 1, 0:K_minus_1)


# Use Cox estimates to compute rd/rr estimates and 95% CI
results.iptw.b.cox.K <- compute_cox_estimates(fit.iptw.b.cox, iptw.b.cox0, iptw.b.cox1, "FirstLine_num", K_minus_1) # outcome at K is from the
time interval K-1~K
results.iptw.b.cox.K



#################################################
### 2c & 2d. Baseline adjusted AT patients      #######
#################################################

# For As-treated (limited to second-Line treatment recipients)####

# 2d.1 Fit a IPW (treatment) to estimate the denominator of the nonstabilised weights ####
TTE_GUTG001_base_reduced_at2L <- TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %in% temp, ]

iptw.b.at.2L.denom <- speedglm(FirstLine_num ~ # Propensity for Strategy 1 Abi-Enza (first-line being abiraterone) treatment

                    ## Time-fixed covariates: baseline
                    Age_TimeZero + # Omitted quadratic age term I(Age_TimeZero^2) due to computational constraints
                    factor(cal_time) + # Calendar year of treatment start influences treatment strategy (availability treatments change)
                    factor(GroupStage) + factor(MStage) + # Group stage encompasses N, M stage information, but missingness in GroupStage is larger
                    factor(GleasonScore) +  # Gleason Score
                    factor(f_pretreat_treat_allowed_docetaxel) + factor(f_pretreat_treat_other_systemic) + # Prior docetaxel/other treatment
                    factor(f_SES) +
                    factor(Race) + # Race, considering potential grouping of Asians
                    # Diagnosis to Treatment Initiation Gap
                    factor(Gap_TimeZero_mCRPC_cat) +
                    factor(Gap_TimeZero_mPC_cat) +

                    ## Baseline values of time-varying covariates
                    factor(PSA_TimeZero_cat) + factor(Hb_TimeZero_cat) + factor(ECOG_TimeZero) +
                    Comorbid_Score_TimeZero + factor(DM_TimeZero),

                  data = TTE_GUTG001_base_reduced_at2L, # Use only baseline data
                  family = binomial(link="logit"))
# 0.094 secs
summary(iptw.b.at.2L.denom)

# Add IPTW_baseline propensity to TTE_GUTG001_base_reduced data
TTE_GUTG001_base_reduced_at2L$p.treat_baseline.denom <- predict(iptw.b.at.2L.denom, TTE_GUTG001_base_reduced_at2L, type="response")


#***2d. 2 Fit a IPW (treatment) to estimate the numerator of the stabilised weights ####
```

```
iptw.b.at.2L.num <- glm(FirstLine_num ~ 1, family = binomial(link="logit"), data = TTE_GUTG001_base_reduced_at2L) # error can occur with
speedglm here
summary(iptw.b.at.2L.num)

# Add numerator to TTE_GUTG001_base_reduced data
TTE_GUTG001_base_reduced_at2L$p.treat_baseline.num <- predict(iptw.b.at.2L.num, type="response")


#***2d. 3: Calculate nonstablised IPTW (baseline) ####
# Calculate nonstablaised weights (based on baseline treatment)
TTE_GUTG001_base_reduced_at2L$w.treat_baseline.ns <-
  ifelse(TTE_GUTG001_base_reduced_at2L$FirstLine_num == 1,
       1/TTE_GUTG001_base_reduced_at2L$p.treat_baseline.denom,
       1/(1 - TTE_GUTG001_base_reduced_at2L$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced_at2L$w.treat_baseline.ns) # 1.799-9.150
sd(TTE_GUTG001_base_reduced_at2L$w.treat_baseline.ns) # SD: 0.788

#***Calculate stablised IPTW (baseline) ####

# Compute stablised weights
TTE_GUTG001_base_reduced_at2L$w.treat_baseline.s <-
  ifelse(TTE_GUTG001_base_reduced_at2L$FirstLine_num == 1,
       TTE_GUTG001_base_reduced_at2L$p.treat_baseline.num/TTE_GUTG001_base_reduced_at2L$p.treat_baseline.denom,
       (1 - TTE_GUTG001_base_reduced_at2L$p.treat_baseline.num)/(1 - TTE_GUTG001_base_reduced_at2L$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced_at2L$w.treat_baseline.s)
sd(TTE_GUTG001_base_reduced_at2L$w.treat_baseline.s)

#***2d. 4. Fit weighted KM/Cox####

# As-treated (limited to second-Line treatment recipients)
# Selecting only patients being able to received second-line (immortal time bias)
length(TTE_GUTG001_base_reduced_at2L$PatientID)
options(digits=7)
fit.iptw.b.km.at.2L <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                 data = TTE_GUTG001_base_reduced_at2L,
                 weights = w.treat_baseline.s)
fit.iptw.b.km.at.2L

# Out put median follow-up time
weighted_followup <- TTE_GUTG001_base_reduced_at2L$SurvTime_FirstLine * TTE_GUTG001_base_reduced_at2L$w.treat_baseline.s
summary(weighted_followup)

fit.iptw.b.cox.at.2L <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                 data = TTE_GUTG001_base_reduced_at2L,
                 weights = w.treat_baseline.s) # default method = "efron"
summary(fit.iptw.b.cox.at.2L)
AIC(fit.iptw.b.cox.at.2L) # 8951.784
BIC(fit.iptw.b.cox.at.2L) # 8956.341
logLik(fit.iptw.b.cox.at.2L)  # -4474.892 (df=1)
# 0.97 (0.82-1.14)

# Sandwich robust 95% CI
coeftest(fit.iptw.b.cox.at.2L, vcov = sandwich)
exp(confint(coeftest(fit.iptw.b.cox.at.2L, vcov = sandwich))) # 0.82-1.14


# 2c For As-treated (including patients only received first-line)####

# 2c.1 Fit a IPW (treatment) to estimate the denominator of the nonstabilised weights ####
TTE_GUTG001_base_reduced_at <- TTE_GUTG001_base_reduced[TTE_GUTG001_base_reduced$PatientID %notin% temp2, ]
summary(TTE_GUTG001_base_reduced_at$FirstLine_num)

iptw.b.at.denom <- speedglm(FirstLine_num ~ # Propensity for Strategy 1 Abi-Enza (first-line being abiraterone) treatment

                    ## Time-fixed covariates: baseline
                    Age_TimeZero + # Omitted quadratic age term I(Age_TimeZero^2) due to computational constraints
                    factor(cal_time) + # Calendar year of treatment start influences treatment strategy (availability treatments change)
                    factor(GroupStage) + factor(MStage) + # Group stage encompasses N, M stage information, but missingness in GroupStage is
larger
                    factor(GleasonScore) +  # Gleason Score
```

```
                    factor(f_pretreat_treat_allowed_docetaxel) + factor(f_pretreat_treat_other_systemic) + # Prior docetaxel/other treatment
                    factor(f_SES) +
                    factor(Race) + # Race, considering potential grouping of Asians
                    # Diagnosis to Treatment Initiation Gap
                    factor(Gap_TimeZero_mCRPC_cat) +
                    factor(Gap_TimeZero_mPC_cat) +

                    ## Baseline values of time-varying covariates
                    factor(PSA_TimeZero_cat) + factor(Hb_TimeZero_cat) + factor(ECOG_TimeZero) +
                    Comorbid_Score_TimeZero + factor(DM_TimeZero),

                  data = TTE_GUTG001_base_reduced_at, # Use only baseline data
                  family = binomial(link="logit"))
# 0.094 secs
summary(iptw.b.at.denom)

# Add IPTW_baseline propensity to TTE_GUTG001_base_reduced data
TTE_GUTG001_base_reduced_at$p.treat_baseline.denom <- predict(iptw.b.at.denom, TTE_GUTG001_base_reduced_at, type="response")


#***2c.2 Fit a IPW (treatment) to estimate the numerator of the stabilised weights ####
iptw.b.at.num <- glm(FirstLine_num ~ 1, family = binomial(link="logit"), data = TTE_GUTG001_base_reduced_at) # error can occur with speedglm
here
summary(iptw.b.at.num)

# Add numerator to TTE_GUTG001_base_reduced data
TTE_GUTG001_base_reduced_at$p.treat_baseline.num <- predict(iptw.b.at.num, type="response")


#***2c.3: Calculate nonstablised IPTW (baseline) ####
# Calculate nonstablaised weights (based on baseline treatment)
TTE_GUTG001_base_reduced_at$w.treat_baseline.ns <-
  ifelse(TTE_GUTG001_base_reduced_at$FirstLine_num == 1,
       1/TTE_GUTG001_base_reduced_at$p.treat_baseline.denom,
       1/(1 - TTE_GUTG001_base_reduced_at$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced_at$w.treat_baseline.ns) # 1.884
sd(TTE_GUTG001_base_reduced_at$w.treat_baseline.ns) # SD: 0.679

#***Calculate stablised IPTW (baseline) ####

# Compute stablised weights
TTE_GUTG001_base_reduced_at$w.treat_baseline.s <-
  ifelse(TTE_GUTG001_base_reduced_at$FirstLine_num == 1,
       TTE_GUTG001_base_reduced_at$p.treat_baseline.num/TTE_GUTG001_base_reduced_at$p.treat_baseline.denom,
       (1 - TTE_GUTG001_base_reduced_at$p.treat_baseline.num)/(1 - TTE_GUTG001_base_reduced_at$p.treat_baseline.denom))

# Summary of weights
summary(TTE_GUTG001_base_reduced_at$w.treat_baseline.s)  # 0.9399
sd(TTE_GUTG001_base_reduced_at$w.treat_baseline.s) # 0.328

#***2c.4. Fit weighted KM/Cox####

# As-treated (limited to second-Line treatment recipients)
# Selecting only patients being able to received second-line (immortal time bias)
length(TTE_GUTG001_base_reduced_at$PatientID)
options(digits=7)
fit.iptw.b.km.at <- survfit(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                    data = TTE_GUTG001_base_reduced_at,
                    weights = w.treat_baseline.s)
fit.iptw.b.km.at

# Out put median follow-up time
weighted_followup <- TTE_GUTG001_base_reduced_at$SurvTime_FirstLine * TTE_GUTG001_base_reduced_at$w.treat_baseline.s
summary(weighted_followup) # 569.6 (270.0-972.6)

fit.iptw.b.cox.at <- coxph(Surv(SurvTime_FirstLine, Y_FirstLine) ~ FirstLine_num,
                   data = TTE_GUTG001_base_reduced_at,
                   weights = w.treat_baseline.s) # default method = "efron"
summary(fit.iptw.b.cox.at)
AIC(fit.iptw.b.cox.at) # 23907.89
BIC(fit.iptw.b.cox.at) # 23913.29
logLik(fit.iptw.b.cox.at)  # -11952.94 (df=1)
```

```
# 1.14 (1.03-1..27)

# Sandwich robust 95% CI
coeftest(fit.iptw.b.cox.at, vcov = sandwich)
exp(confint(coeftest(fit.iptw.b.cox.at, vcov = sandwich))) # 1.03-1.27



#########################################################################################################################################
### 3. Adjusted for baseline (IPW) & censoring due to treatment deviation (IPCW) - Hypothetical perfect per-protocol effect                ####
#########################################################################################################################################



# Create a variable called TimeCrossover (time of switching to a second-line treatment (previously called PD1_Now_tvary))
TTE_GUTG001_DTR_IPW_reduced$TimeCrossover <- TTE_GUTG001_DTR_IPW_reduced$PD1_Now_tvary

# Spline modelss
# Fit Treamtent censoring model for each treatment arm separately, assuming that time-varying covariate has different effect on censoring for
different treatment arms
# Use a new data
TTE_GUTG001_DTR_IPW_Cross_cspline <- TTE_GUTG001_DTR_IPW_reduced

columns_to_remove <- c("p.treatcen.denom", "w.treatcen.temp.ns", "w.treatcen.ns", "p.treatcen.num",
                       "w.treatcen.temp.s", "w.treatcen.s", "w_overall_ns", "w_overall_s")

# Remove the columns
TTE_GUTG001_DTR_IPW_Cross_cspline[, (columns_to_remove) := NULL]
TTE_GUTG001_DTR_IPW_Cross_cspline <- data.table(TTE_GUTG001_DTR_IPW_Cross_cspline)

#***3.1: Fit a IPCW to estimate the denominator of the stabilised weights#####

# Enza-Abi arm0
# generate spline based on censoring time in arm 0
library(Hmisc)
library(rms)
gc()
spline_knots_TreatCen_arm0 <- rcspline.eval(TTE_GUTG001_DTR_IPW_Cross_cspline$time[TTE_GUTG001_DTR_IPW_Cross_cspline$c_treatdevL2
== 1 &
                                                    TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 0], knots.only = T,pc = T)
spline_knots_TreatCen_arm0 # 56, 162.3, 295.5, 512.4, 1001,4


# Set display digits
options(digits=7)
gc()
start_time <- Sys.time()
options(warn = -1)
ipcw.tdev3.pp.CO.denom_arm0 <- glm( # speedglm might not converge
  c_treatdevL2 == 1 ~ # modelling the probability of switching

    ## Time-fixed covariates
    Age_TimeZero + factor(cal_time) +
    factor(GroupStage) + factor(MStage) +
    factor(GleasonScore) +
    factor(f_pretreat_treat_allowed_docetaxel) + factor(f_pretreat_treat_other_systemic) +
    factor(f_SES) + factor(Race) +
    factor(Gap_TimeZero_mCRPC_cat) + factor(Gap_TimeZero_mPC_cat) +

    ## Baseline values of time-varying covariates
    factor(PSA_TimeZero_cat) + factor(Hb_TimeZero_cat) + factor(ECOG_TimeZero) +
    Comorbid_Score_TimeZero + factor(DM_TimeZero) +

    ## Time-updated values of time-varying covariates
    factor(PSA_tvary_cat) + factor(Hb_tvary_cat) + factor(ECOG_tvary) +
    Comorbid_Score_tvary + factor(DM_tvary) +

    # Not including an exact progression time due to no perfect timing proxy in the database.
    # (1) Assumeing other time-varying covariates reflect prognosis related to progression at the time of switching
    #     ==> Caveat: Possible residual confounding not captured by the identified time-varying variables.
    # (2) Treatment switches in GUTG-0001 trial may also occur due to factors other than progression, like treatment toxicity, which should might be
captured through ECOG/Comorbidity score.

    # Time
```

```
   rcs(time, spline_knots_TreatCen_arm0),

  data = TTE_GUTG001_DTR_IPW_Cross_cspline[TimeCrossover == 1 & FirstLine_num == 0], # Use only CrossOver baseline interval from each
patient if any (i.e, patients who never switched to a second-line treamtent have prob of switching = 0)

  family = binomial(link="logit"))
options(warn = 0)
Sys.time() - start_time
summary(ipcw.tdev3.pp.CO.denom_arm0) # 3.5 sec


# Abi-Enza arm 1
# generate spline for arm 1
spline_knots_TreatCen_arm1 <- rcspline.eval(TTE_GUTG001_DTR_IPW_Cross_cspline$time[TTE_GUTG001_DTR_IPW_Cross_cspline$c_treatdevL2
== 1 &
                                                    TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 1],knots.only = T,pc = T)
spline_knots_TreatCen_arm1 # 54, 127.5, 226.5, 416.4, 919.6

gc()
start_time <- Sys.time()
options(warn = -1)
ipcw.tdev3.pp.CO.denom_arm1 <- glm( # speedglm might not converge
  c_treatdevL2 == 1 ~ # modelling the probability of switching

    ## Time-fixed covariates
    Age_TimeZero + factor(cal_time) +
    factor(GroupStage) + factor(MStage) +
    factor(GleasonScore) +
    factor(f_pretreat_treat_allowed_docetaxel) + factor(f_pretreat_treat_other_systemic) +
    factor(f_SES) + factor(Race) +
    factor(Gap_TimeZero_mCRPC_cat) + factor(Gap_TimeZero_mPC_cat) +

    ## Baseline values of time-varying covariates
    factor(PSA_TimeZero_cat) + factor(Hb_TimeZero_cat) + factor(ECOG_TimeZero) +
    Comorbid_Score_TimeZero + factor(DM_TimeZero) +

    ## Time-updated values of time-varying covariates
    factor(PSA_tvary_cat) + factor(Hb_tvary_cat) + factor(ECOG_tvary) +
    Comorbid_Score_tvary + factor(DM_tvary) +

    # Time
    rcs(time, spline_knots_TreatCen_arm1),

  data = TTE_GUTG001_DTR_IPW_Cross_cspline[TimeCrossover == 1 & FirstLine_num == 1], # Use only the CrossOver baseline interval from each
patient if any (i.e, patients who never switched to a second-line treatment will have prob of switching = 0)
  family = binomial(link="logit"))
options(warn = 0)
Sys.time() - start_time
summary(ipcw.tdev3.pp.CO.denom_arm1) # 0.12 sec


#***3.2: Fit a IPCW to estimate the numerator of the stabilised weights#####
# arm 0
gc()
start_time <- Sys.time()
options(warn = -1)
ipcw.tdev3.pp.CO.num_arm0 <- speedglm(
  c_treatdevL2 == 1 ~
    rcs(time, spline_knots_TreatCen_arm0), # modelling the probability of switching given time interval
  data = TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(c_treatdevL2) & FirstLine_num == 0], # all rows except the additional row for admin
censoring weighting
  family = binomial(link="logit"))
options(warn = 0)
Sys.time() - start_time
# 6.3 secs
summary(ipcw.tdev3.pp.CO.num_arm0)

# arm 1
gc()
start_time <- Sys.time()
options(warn = -1)
ipcw.tdev3.pp.CO.num_arm1 <- speedglm(
  c_treatdevL2 == 1 ~
    rcs(time, spline_knots_TreatCen_arm1), # modelling the probability of switching given time interval
```

```
  data = TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(c_treatdevL2) & FirstLine_num == 1], # all rows except the additional row for admin
censoring weighting
  family = binomial(link="logit"))
options(warn = 0)
Sys.time() - start_time
# 7.2 secs
summary(ipcw.tdev3.pp.CO.num_arm1)

# Set the number of digits to display

#***3.3 Calculate nonstablised deviation censoring weights #####
gc()

# Set display digits
options(digits = 10)

# Arm 0
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 0, p.treatcen.denom := predict(ipcw.tdev3.pp.CO.denom_arm0, .SD, type = "response")]
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 0 & TimeCrossover == 0, p.treatcen.denom:= 0] # Force all non-crossover baseline rows
to have a censoring weight = 0
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$p.treatcen.denom[TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 0 &
TTE_GUTG001_DTR_IPW_Cross_cspline$TimeCrossover == 1])
# Weight at cross-over baseline: median: 0.56942330, min: 0.05741535, max: 0.99999942

# Arm 1
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 1, p.treatcen.denom := predict(ipcw.tdev3.pp.CO.denom_arm1, .SD, type = "response")]
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 1 & TimeCrossover == 0, p.treatcen.denom:= 0] # Force all non-crossover baseline rows
to have a censoring weight = 0
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$p.treatcen.denom[TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 1 &
TTE_GUTG001_DTR_IPW_Cross_cspline$TimeCrossover == 1])
# Weight at cross-over baseline: median: 0.5083406504, min 0.0000011964, max: 0.9773441602


# Set display digits
options(digits = 20)

# nonstablised weight at each interval
TTE_GUTG001_DTR_IPW_Cross_cspline[, w.treatcen.temp.ns := 1 / (1 - p.treatcen.denom)] # weight for patients remained uncensored at the time
of cross-over
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w.treatcen.temp.ns[TTE_GUTG001_DTR_IPW_Cross_cspline$TimeCrossover == 1])
# median 2.1499153065092300, min 1.0000011964258400 max 1710292.0496234854217619 (including those who censored)

# Taking the cumulative product of the weights for each patient
setorder(TTE_GUTG001_DTR_IPW_Cross_cspline, PatientID, time) # arrange data before doing cumulative product
TTE_GUTG001_DTR_IPW_Cross_cspline[, w.treatcen.ns:=cumprod(w.treatcen.temp.ns), by = PatientID]

# If censored, then that line of weight = 0
TTE_GUTG001_DTR_IPW_Cross_cspline[c_treatdevL2 == 1, w.treatcen.ns:= 0]

summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w.treatcen.ns[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)])
# median: 1.0000000000000000; max: 10.5695611279343709


#***3.4 Calculate stabalised deviation censoring weights #####

# arm 0
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 0, p.treatcen.num := predict(ipcw.tdev3.pp.CO.num_arm0, .SD, type = "response")]
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$p.treatcen.num[TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 0 &
!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)]) # prob of switching given time for all non-admin censoring rows
# median 0.00048114473457106747, max: 00.00089470602271352409

# Arm 1
TTE_GUTG001_DTR_IPW_Cross_cspline[FirstLine_num == 1, p.treatcen.num := predict(ipcw.tdev3.pp.CO.num_arm1, .SD, type = "response")]
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$p.treatcen.num[TTE_GUTG001_DTR_IPW_Cross_cspline$FirstLine_num == 1 &
!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)]) # prob of switching given time for all non-admin censoring rows
# median 0.00048565200649136847, max 0.00119607559536391117

# stabalised weight at each interval
TTE_GUTG001_DTR_IPW_Cross_cspline[, w.treatcen.temp.s := (1 - p.treatcen.num) / (1 - p.treatcen.denom)] # weight for patients remained
uncensored at each time interval
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w.treatcen.temp.s[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)])
# median 0.99951781739044065, min 0.99880392440463606 max 10.56057687794355537

# Taking the cumulative product of the weights for each id
```

```
TTE_GUTG001_DTR_IPW_Cross_cspline[, w.treatcen.s:=cumprod(w.treatcen.temp.s), by = PatientID]

# If censored, then weight = 0
TTE_GUTG001_DTR_IPW_Cross_cspline[c_treatdevL2 == 1, w.treatcen.s:=0]
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w.treatcen.s[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)])
# median: 0.86268259967874439, max: 9.75164816080160790


# Ref: Harvard 2023 ACA5
#***3.5 Calculate Overall weights #####
# Combining weights: non-stabalised
TTE_GUTG001_DTR_IPW_Cross_cspline[, w_overall_ns:=w.treat_baseline.ns*w.treatcen.ns]
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w_overall_ns[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)])
# median: 2.0402396450963853,  max 35.0769540901478010


# Combining weights: stabalised
TTE_GUTG001_DTR_IPW_Cross_cspline[, w_overall_s:=w.treat_baseline.s*w.treatcen.s]
summary(TTE_GUTG001_DTR_IPW_Cross_cspline$w_overall_s[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y)])
# median:  0.82424314227787387, max: 13.14245719335347751

# Check who have extreme weights: check weights by the order and keep only records with weights over 5 (for stablised weights)
test <- TTE_GUTG001_DTR_IPW_Cross_cspline[w_overall_s > 5][order(-w_overall_s)] # 6503

# Check weights (don't use geom_point, it will crash with this size of database)
iptw.b.ipcwCO3.plots <- list()
iptw.b.ipcwCO3.plots[["p.treatcen.denom"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ],
aes(x = time, y = p.treatcen.denom)) +
  geom_hex(bins = 50) +
  scale_fill_gradientn(colors = c("lightblue", "blue", "darkblue")) +  # Custom gradient
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("p.treatcen.denom") +
  ggtitle("Probability of censoring due to treatment deviation given \ngiven time-vary covariates over time by group \n0 = Enza-Abi, 1 = Abi-Enza
(cubic spline time, Crossover baseline approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))  # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["p.treatcen.denom"]]


gc()
# numerator
iptw.b.ipcwCO3.plots[["p.treatcen.num"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ],
aes(x = time, y = p.treatcen.num)) +
  geom_hex(bins = 50) +
  scale_fill_gradient(low = "lightgreen", high = "darkgreen") +
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("p.treatcen.num") +
  ggtitle("Probability of censoring due to treatment deviation \ngiven time \n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover baseline
approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["p.treatcen.num"]]

ggarrange(
  iptw.b.ipcwCO3.plots[["p.treatcen.denom"]],
  iptw.b.ipcwCO3.plots[["p.treatcen.num"]],
  ncol = 1, nrow = 2,    # overall layout: 1 column, 2 rows
  heights = c(1, 1)      # equal heights for top and bottom
)

# check nostabalised weights for remaining uncernsored
iptw.b.ipcwCO3.plots[["w.treatcen.temp.ns"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y),
], aes(x = time, y = w.treatcen.temp.ns)) +
  geom_hex(bins = 50) +
  scale_fill_gradient(low = "#EAA9A8", high = "#800000") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("w.treatcen.temp.ns") +
  ggtitle("Nonstabalised weights for patients remaind uncensored by Group\n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover baseline
approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
```

```
iptw.b.ipcwCO3.plots[["w.treatcen.temp.ns"]]

iptw.b.ipcwCO3.plots[["w.treatcen.ns"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ], aes(x
= time, y = w.treatcen.ns)) +
  geom_hex(bins = 50) +
  scale_fill_gradient(low = "#EAA9A8", high = "#800000") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("w.treatcen.ns") +
  ggtitle("Cumulative nonstabalised weights for patients remaind uncensored by Group\n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover
baseline approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["w.treatcen.ns"]]

# check stabalised weights

iptw.b.ipcwCO3.plots[["w.treatcen.temp.s"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ],
aes(x = time, y = w.treatcen.temp.s)) +
  geom_hex(bins = 50) +
  scale_fill_gradient(low = "#D7BDE2", high = "#4A235A") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("w.treatcen.temp.s") +
  ggtitle("Stabalised weights for patients remaind uncensored by Group \n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover baseline
approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["w.treatcen.temp.s"]]

iptw.b.ipcwCO3.plots[["w.treatcen.s"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ], aes(x
= time, y = w.treatcen.s)) +
  geom_hex(bins = 75) +
  scale_fill_gradient(low = "#D7BDE2", high = "#4A235A") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("w.treatcen.s") +
  ggtitle("Cumulative stabalised weights for patients remaind uncensored by Group\n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover
baseline approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["w.treatcen.s"]]

ggarrange(
  iptw.b.ipcwCO3.plots[["w.treatcen.temp.ns"]],
  iptw.b.ipcwCO3.plots[["w.treatcen.ns"]],
  iptw.b.ipcwCO3.plots[["w.treatcen.temp.s"]],
  iptw.b.ipcwCO3.plots[["w.treatcen.s"]],
  ncol = 2, nrow = 2,    # overall layout: 1 column, 2 rows
  heights = c(1, 1)      # equal heights for top and bottom
)


# check overall weights (nonstablised)
iptw.b.ipcwCO3.plots[["w_overall_ns"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ], aes(x
= time, y = w_overall_ns)) +
  geom_hex(bins = 500) +  # Reduced number of bins and added alpha for transparency
  scale_fill_gradient(low = "#D3BDBD", high = "#5E4747") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
  xlab("Time") +
  ylab("w_overall_ns") +
  ggtitle("Cumulative overall non-stabalised weights for patients remaind uncensored by Group \n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time,
Crossover baseline approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["w_overall_ns"]]

# check overall weights (stablised)
iptw.b.ipcwCO3.plots[["w_overall_s"]] <- ggplot(TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ], aes(x =
time, y = w_overall_s)) +
  geom_hex(bins = 500) +  # Reduced number of bins and added alpha for transparency
  scale_fill_gradient(low = "#808080", high = "#1A1A1A") +  # Light to dark color scale
  facet_wrap(~ FirstLine_num) +  # Facet by the grouping variable
  theme_minimal() +
```

```
  xlab("Time") +
  ylab("w_overall_s") +
  ggtitle("Cumulative overall stabalised weights for patients remaind uncensored by Group \n0 = Enza-Abi, 1 = Abi-Enza (cubic spline time, Crossover
baseline approach)") +
  scale_x_continuous(breaks = seq(0, 1440, by = 180))   # X-axis labels every 90 days
iptw.b.ipcwCO3.plots[["w_overall_s"]]

ggarrange(
  iptw.b.ipcwCO3.plots[["w_overall_ns"]],
  iptw.b.ipcwCO3.plots[["w_overall_s"]],
  ncol = 1, nrow = 2,    # overall layout: 1 column, 2 rows
  heights = c(1, 1)      # equal heights for top and bottom
)
gc()


# save ipw.plots and remove from global environment (7 GB)
# save(iptw.b.ipcwCO3.plots, file = "derived_data\\iptw.b.ipcwCO3.plots.RData")
# rm(iptw.b.ipcwCO3.plots)
gc()

#***3.5 Outcome Model #####
# generate spline for outcome
spline_knots_TreatCen_y <- rcspline.eval(TTE_GUTG001_DTR_IPW_Cross_cspline$time[TTE_GUTG001_DTR_IPW_Cross_cspline$y == 1],knots.only
= T,pc = T)
spline_knots_TreatCen_y # 55.5, 213.25, 411, 693, 1190.5



#### Fit a Cox model using stabalised weight ####
fit.iptw.b.ipcwCO.s3.Cox <- coxph(Surv(time, time + 1, y) ~ FirstLine_num,
                          data = TTE_GUTG001_DTR_IPW_Cross_cspline[!is.na(TTE_GUTG001_DTR_IPW_Cross_cspline$y), ],
                          weights = w_overall_s,
                          cluster = PatientID) # default method = "efron"
# Set display digits
options(digits=7)
summary(fit.iptw.b.ipcwCO.s3.Cox)
AIC(fit.iptw.b.ipcwCO.s3.Cox) # 26697.83
BIC(fit.iptw.b.ipcwCO.s3.Cox) # 26703.23
logLik(fit.iptw.b.ipcwCO.s3.Cox) # -13347.92 (df=1)

# Sandwich robust variance (using this instead of the internal robust estimator robust = TRUE within the coxph function)
# More appropriate if there are complex clustering
coeftest(fit.iptw.b.ipcwCO.s3.Cox, vcov = sandwich, cluster = ~PatientID)
exp(confint(coeftest(fit.iptw.b.ipcwCO.s3.Cox, vcov = sandwich, cluster = ~PatientID)))
#     2.5 %   97.5 %
#   [1,] 0.9551065 1.203951

gc()
# Standerdised difference
HR1 <- 0.79 # HR1
CI1_lower <- 0.54
CI1_upper <- 1.16

HR2 <- 1.07 # HR2
CI2_lower <- 0.96
CI2_upper <- 1.20

# Calculate the standard errors
SE1 <- (log(CI1_upper) - log(CI1_lower)) / (2 * 1.96)
SE2 <- (log(CI2_upper) - log(CI2_lower)) / (2 * 1.96)

# Calculate the standardised difference
std_diff <- (log(HR1) - log(HR2)) / sqrt(SE1^2 + SE2^2)

std_diff
# -1.493095


options(digits=20)



########################################
```

```
# Plot comparison with KM Khalaf
########################################

# extract layer data from fit.Khalaf (extract from Khalaf study), and check unique strata
km_Khalaf_data_new <- tidy(fit_Khalaf) %>% # 161
  mutate(strata = case_when(strata == "Group=0" ~ "Abi-Enza (KM - Khalaf 2019)", # Not that for coloring convenience Group 0 = Group A
(abiraterone in the Khalaf data)
                    TRUE ~ "Enza-Abi (KM - Khalaf 2019)")) %>%
  mutate(time = time*30) %>% # change time scale to days to align with my study
  add_row(tibble( # add row at time zero for both strata as it doesn't have data at 0
    time = rep(0, 2), n.risk = NA, n.event = 0, n.censor = NA,
    estimate = 1, std.error = 0, conf.high = 1, conf.low = 1,
    strata = c("Abi-Enza (KM - Khalaf 2019)", "Enza-Abi (KM - Khalaf 2019)"))) # 163


# extract layer data from fit.Khalaf (extract from Khalaf study), and check unique strata
km_iptw.b.ipcwCO.s3_data <- tidy(fit.iptw.b.ipcwCO.s3.km) %>%
  mutate(strata = case_when(strata == "FirstLine_num=1" ~ "Abi-Enza (IPTW*IPCW weighted KM)",
                    TRUE ~ "Enza-Abi (IPTW*IPCW weighted KM)")) %>%
  add_row(tibble( # add row at time zero for both strata as it doesn't have data at 0
    time = rep(0, 2), n.risk = NA, n.event = 0, n.censor = NA,
    estimate = 1, std.error = 0, conf.high = 1, conf.low = 1,
    strata = c("Abi-Enza (IPTW*IPCW weighted KM)", "Enza-Abi (IPTW*IPCW weighted KM)")))

# Add Khalaf Km to IPTW*IPC weighted KM

test <- ggplot() +
  geom_line(data = km_iptw.b.ipcwCO.s3_data, aes(x = time, y = estimate, color = strata), linewidth = 1) +
  geom_ribbon(data = km_iptw.b.ipcwCO.s3_data, aes(x = time, ymin = conf.low, ymax = conf.high, fill = strata), alpha = 0.2) +
  geom_line(data = km_Khalaf_data_new , aes(x = time, y = estimate, color = strata), linewidth = 1) +
  geom_ribbon(data = km_Khalaf_data_new , aes(x = time, ymin = conf.low, ymax = conf.high, fill = strata), alpha = 0.2) +
  scale_color_manual(values = c(
    "Abi-Enza (IPTW*IPCW weighted KM)" = "#2E9FDF",
    "Enza-Abi (IPTW*IPCW weighted KM)" = "#E7B800",
    "Abi-Enza (KM - Khalaf 2019)" = "#0044CC",
    "Enza-Abi (KM - Khalaf 2019)" = "#8B0000"
  )) +
  scale_fill_manual(values = c(
    "Abi-Enza (IPTW*IPCW weighted KM)" = scales::alpha("#2E9FDF", 1),
    "Enza-Abi (IPTW*IPCW weighted KM)" = scales::alpha("#E7B800", 1),
    "Abi-Enza (KM - Khalaf 2019)" = scales::alpha("#0044CC", 0.5), # alpha doesn't seem to work here (adjust geom_ribbon)
    "Enza-Abi (KM - Khalaf 2019)" = scales::alpha("#8B0000", 0.5)
  )) +
  xlab("Days") + # label x axis
  ylab("Survival (%)") + # label y axis
  theme_minimal() + # set plot theme elements
  theme(axis.text = element_text(size = 14), legend.position = c(0.7, 0.8),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks=seq(0, 1440, by = 180)) +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = 0.1), labels = label_percent(scale = 100)) + # label survival it as percentage
  expand_limits(x = 0, y = 0) +
  font("xlab",size = 14)+
  font("ylab",size = 14)+
  font("legend.text",size = 10)+
  guides(fill = "none") + # Suppress the legend for shaded area
  guides(color = guide_legend("First-line mCRPC treatment")) +
  ggtitle("IPTW*IPCW adjusted of GUTG-001 Analogue mCPRC patients \nin Flatiron data (Weighted KM versus reconstructed KM from Khalaf
2019)") # replace a new title

test


test <- ggplot() +
  geom_line(data = filter(km_iptw.b.ipcwCO.s3_data, strata == "Abi-Enza (IPTW*IPCW weighted KM)"), aes(x = time, y = estimate, color = strata),
linewidth = 1) +
  geom_ribbon(data = filter(km_iptw.b.ipcwCO.s3_data, strata == "Abi-Enza (IPTW*IPCW weighted KM)"), aes(x = time, ymin = conf.low, ymax =
conf.high, fill = strata), alpha = 0.2) +
  geom_line(data = filter(km_Khalaf_data_new, strata == "Abi-Enza (KM - Khalaf 2019)"), aes(x = time, y = estimate, color = strata), linewidth = 1) +
  geom_ribbon(data = filter(km_Khalaf_data_new, strata == "Abi-Enza (KM - Khalaf 2019)"), aes(x = time, ymin = conf.low, ymax = conf.high, fill =
strata), alpha = 0.2) +
  scale_color_manual(values = c(
    "Abi-Enza (IPTW*IPCW weighted KM)" = "#2E9FDF",
    "Abi-Enza (KM - Khalaf 2019)" = "#0044CC"
  )) +
```

```
  scale_fill_manual(values = c(
    "Abi-Enza (IPTW*IPCW weighted KM)" = scales::alpha("#2E9FDF", 1),
    "Abi-Enza (KM - Khalaf 2019)" = scales::alpha("#0044CC", 0.5) # alpha doesn't seem to work here (adjust geom_ribbon)
  )) +
  xlab("Days") + # label x axis
  ylab("Survival (%)") + # label y axis
  theme_minimal() + # set plot theme elements
  theme(axis.text = element_text(size = 14), legend.position = c(0.7, 0.8),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks=seq(0, 1440, by = 180)) +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = 0.1), labels = label_percent(scale = 100)) + # label survival it as percentage
  expand_limits(x = 0, y = 0) +
  font("xlab",size = 14)+
  font("ylab",size = 14)+
  font("legend.text",size = 10)+
  guides(fill = "none") + # Suppress the legend for shaded area
  guides(color = guide_legend("First-line mCRPC treatment")) +
  ggtitle("IPTW*IPCW adjusted of GUTG-001 Analogue mCPRC patients \nin Flatiron data (Weighted KM versus reconstructed KM from Khalaf
2019)") # replace a new title
test


test2 <- ggplot() +
  geom_line(data = filter(km_iptw.b.ipcwCO.s3_data, strata == "Enza-Abi (IPTW*IPCW weighted KM)"), aes(x = time, y = estimate, color = strata),
linewidth = 1) +
  geom_ribbon(data = filter(km_iptw.b.ipcwCO.s3_data, strata == "Enza-Abi (IPTW*IPCW weighted KM)"), aes(x = time, ymin = conf.low, ymax =
conf.high, fill = strata), alpha = 0.2) +
  geom_line(data = filter(km_Khalaf_data_new, strata == "Enza-Abi (KM - Khalaf 2019)"), aes(x = time, y = estimate, color = strata), linewidth = 1) +
  geom_ribbon(data = filter(km_Khalaf_data_new, strata == "Enza-Abi (KM - Khalaf 2019)"), aes(x = time, ymin = conf.low, ymax = conf.high, fill =
strata), alpha = 0.2) +
  scale_color_manual(values = c(
    "Enza-Abi (IPTW*IPCW weighted KM)" = "#E7B800",
    "Enza-Abi (KM - Khalaf 2019)" = "#8B0000"
  )) +
  scale_fill_manual(values = c(
    "Enza-Abi (IPTW*IPCW weighted KM)" = scales::alpha("#E7B800", 1),
    "Enza-Abi (KM - Khalaf 2019)" = scales::alpha("#8B0000", 0.5)
  )) +
  xlab("Days") + # label x axis
  ylab("Survival (%)") + # label y axis
  theme_minimal() + # set plot theme elements
  theme(axis.text = element_text(size = 14), legend.position = c(0.7, 0.8),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks=seq(0, 1440, by = 180)) +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = 0.1), labels = label_percent(scale = 100)) + # label survival it as percentage
  expand_limits(x = 0, y = 0) +
  font("xlab",size = 14)+
  font("ylab",size = 14)+
  font("legend.text",size = 10)+
  guides(fill = "none") + # Suppress the legend for shaded area
  guides(color = guide_legend("First-line mCRPC treatment")) +
  ggtitle("IPTW*IPCW adjusted of GUTG-001 Analogue mCPRC patients \nin Flatiron data (Weighted KM versus reconstructed KM from Khalaf
2019)") # replace a new title
test2

ggarrange(test,test2,
          ncol = 2, nrow = 1)
```

Appendix 8.11 R code for ad-hoc restricted mean survival time analyses in Prostate Cancer Case Study 1 (PC1)

The RMST for the unweighted Kaplan-Meier (KM) analyses (Analysis 1a-1d in Section 8.3, Chapter 8) and the digitised curves from the GUTG-001 trial (Khalaf et al.[84]) was derived using the rmst2 package in R. However, no R packages support the calculation of RMST for weighted KM. Therefore, I wrote a function "perform_rmst_trapezoids_wKM" in R to calculate the RMST for each group by determining the area under the weighted KM curve for the remaining five weighted analyses: Analysis: 2a-2d baseline IPTW adjusted analysis and Analysis 3 IPTW*IPCW$_{texdev}$ adjusted analysis. Given the daily interval data used for analysis, this is an accurate method for estimating RMST.

The perform_rmst_trapezoids_wKM function involves the following steps:

1. Obtaining the survival probability at each time point from the weighted KM survival object in R for each group's survival curve
2. Using the trapezoid rule, the RMST is calculated by determining the survival probability between consecutive time points in the weighted KM, summing the area under the curve for each interval:

$$RMST = \sum_{1}^{\tau}\left(\frac{\text{survival probability at time point 1} + \text{survival probability at time point 2}}{2}\right) \times (time\ point\ 2 - time\ point\ 1),$$

$$\tau\ is\ the\ time\ limit\ for\ restricted\ mean\ survival$$

3. To calculate the 95% CI of the RMST for each group, I used the survival probability of the naïve 95% CI at each time point from the weighted KM (the same approach used for obtaining the median overall survival 95% CI in Table 8.10). For more precise CI estimates that account for uncertainties in the weight derivation models, bootstrapping may be necessary, as described in Section 8.3.4.4. However, this was not applied in the thesis due to the marginal differences expected given the study's sample size and the computational resources required (see Section 8.3.4.4).
4. The RMST differences between the two arms were assessed by measuring the point estimate difference in RMST between the treatment groups.

The R code for ad-hoc restricted mean survival time analyses and the perform_rmst_trapezoids_wKM function are detailed below.

```
###Project Info ##############################################################################
###   Target Trial Emulation – Final survival analyses – RMST calculation                  ###
###   Amy Chang's PhD Thesis - Treatment Sequence project                                  ###
###   Data: Flatiron prostate cancer datasets                                              ###
###                                                                                        ###
###   University of Sheffield                                                              ###
###   Date created: July, 2024                                                             ###
##############################################################################################

# Write a function to perform RMST for multiple endpoints at the same time, and can accomondate different data & endpoints
perform_rmst <- function(data, time_col, status_col, arm_col, taus) {
```

```
 # Check if columns exist
 if (!all(c(time_col, status_col, arm_col) %in% names(data))) {
  stop("One or more specified columns do not exist in the data frame.")
 }

 # Select and rename columns
 data <- data %>%
  rename(time = !!sym(time_col), status = !!sym(status_col), arm = !!sym(arm_col))

 # Initialize a list to store results
 results <- list()

 # Loop through each tau and perform RMST analysis
 for (tau in taus) {
  fit <- rmst2(time = data$time, status = data$status, arm = data$arm, tau = tau)
  results[[paste0("RMST_", tau, "months")]] <- fit
 }

 return(results)
}

# RMST for weighted KM: This version solves problems where initial survival probability is not 1 compared to version 1 (add time 0)
# Note: This function takes time input in days but outputs in months.
perform_rmst_trapezoids_wKM <- function(surv_object, tau_vec, title) {
 results <- list()

 # Function to add initial time 0 with survival probability 1
 add_initial_time <- function(times, surv, lower, upper) {
  if (times[1] != 0) {
   times <- c(0, times)
   surv <- c(1, surv)
   lower <- c(1, lower)
   upper <- c(1, upper)
  }
  return(list(times = times, surv = surv, lower = lower, upper = upper))
 }

 # Get the group indices: 0, number of record in arm 0, number of record in arm 0 + 1 (due to the cumulative sum of records)
 # Need to do this as each arm may have slightly different rows
 group_indices <- cumsum(c(0, surv_object$strata))

 for (tau in tau_vec) {
  # Extract KM times and survival probabilities for Arm 1 based on group indices
  km_times_1 <- surv_object$time[(group_indices[2]+1):group_indices[3]] # number of record in arm 0 + 1 to the last record (records for arm 1)
  km_surv_1 <- surv_object$surv[(group_indices[2]+1):group_indices[3]]

  # Extract KM times and survival probabilities for Arm 0 based on group indices
  km_times_0 <- surv_object$time[(group_indices[1]+1):group_indices[2]]  # 1 to the last record in arm 0 (the number of arm 0)
  km_surv_0 <- surv_object$surv[(group_indices[1]+1):group_indices[2]]

  # Extract the survival lower and upper limits from the KM object for Arm 1
  km_surv_ll_1 <- surv_object$lower[(group_indices[2]+1):group_indices[3]]
  km_surv_ul_1 <- surv_object$upper[(group_indices[2]+1):group_indices[3]]

  # Extract the survival lower and upper limits from the KM object for Arm 0
  km_surv_ll_0 <- surv_object$lower[(group_indices[1]+1):group_indices[2]]
  km_surv_ul_0 <- surv_object$upper[(group_indices[1]+1):group_indices[2]]

  # Add time 0 and survival probability 1 if not already present for both arms
  arm1_data <- add_initial_time(km_times_1, km_surv_1, km_surv_ll_1, km_surv_ul_1)
  km_times_1 <- arm1_data$times
  km_surv_1 <- arm1_data$surv
  km_surv_ll_1 <- arm1_data$lower
  km_surv_ul_1 <- arm1_data$upper

  arm0_data <- add_initial_time(km_times_0, km_surv_0, km_surv_ll_0, km_surv_ul_0)
  km_times_0 <- arm0_data$times
  km_surv_0 <- arm0_data$surv
  km_surv_ll_0 <- arm0_data$lower
  km_surv_ul_0 <- arm0_data$upper

  # Ensure the times and survival probabilities are in ascending order
  km_times_1 <- sort(km_times_1)
  km_surv_1 <- km_surv_1[order(km_times_1)]
```

```
km_surv_ll_1 <- km_surv_ll_1[order(km_times_1)]
km_surv_ul_1 <- km_surv_ul_1[order(km_times_1)]

km_times_0 <- sort(km_times_0)
km_surv_0 <- km_surv_0[order(km_times_0)]
km_surv_ll_0 <- km_surv_ll_0[order(km_times_0)]
km_surv_ul_0 <- km_surv_ul_0[order(km_times_0)]

# Filter times and survival probabilities up to tau for both arms
km_times_1 <- km_times_1[km_times_1 <= tau]
km_surv_1 <- km_surv_1[1:length(km_times_1)]
km_surv_ll_1 <- km_surv_ll_1[1:length(km_times_1)]
km_surv_ul_1 <- km_surv_ul_1[1:length(km_times_1)]

km_times_0 <- km_times_0[km_times_0 <= tau]
km_surv_0 <- km_surv_0[1:length(km_times_0)]
km_surv_ll_0 <- km_surv_ll_0[1:length(km_times_0)]
km_surv_ul_0 <- km_surv_ul_0[1:length(km_times_0)]

# Calculate the differences between consecutive time points
diffs_1 <- diff(km_times_1)
diffs_0 <- diff(km_times_0)

# Perform numerical integration using the trapezoidal rule for both arms
rmst_1 <- sum((km_surv_1[-length(km_surv_1)] + km_surv_1[-1]) * diffs_1 / 2)
rmst_0 <- sum((km_surv_0[-length(km_surv_0)] + km_surv_0[-1]) * diffs_0 / 2)

# Calculate the difference in RMST between the two arms
rmst_diff <- rmst_1 - rmst_0

# Calculate RMST lower limit for both arms
rmst_ll_1 <- sum((km_surv_ll_1[-length(km_surv_ll_1)] + km_surv_ll_1[-1]) * diffs_1 / 2)
rmst_ll_0 <- sum((km_surv_ll_0[-length(km_surv_ll_0)] + km_surv_ll_0[-1]) * diffs_0 / 2)

# Calculate RMST upper limit for both arms
rmst_ul_1 <- sum((km_surv_ul_1[-length(km_surv_ul_1)] + km_surv_ul_1[-1]) * diffs_1 / 2)
rmst_ul_0 <- sum((km_surv_ul_0[-length(km_surv_ul_0)] + km_surv_ul_0[-1]) * diffs_0 / 2)

# Calculate the 95% CI for the RMST difference (most conservative, lower of arm 1 - higher of arm 2; higher of arm 1 - lower of arm 2)
ci_lower_diff <- rmst_ll_1 - rmst_ul_0
ci_upper_diff <- rmst_ul_1 - rmst_ll_0

# Round all results to 2 decimal places and in months
rmst_1 <- round(rmst_1 / 30, 2)
rmst_0 <- round(rmst_0 / 30, 2)
rmst_diff <- round(rmst_diff / 30, 2)
rmst_ll_1 <- round(rmst_ll_1 / 30, 2)
rmst_ll_0 <- round(rmst_ll_0 / 30, 2)
rmst_ul_1 <- round(rmst_ul_1 / 30, 2)
rmst_ul_0 <- round(rmst_ul_0 / 30, 2)
ci_lower_diff <- round(ci_lower_diff / 30, 2)
ci_upper_diff <- round(ci_upper_diff / 30, 2)

# Store results in a list
results[[as.character(tau)]] <- list(
  rmst_1 = rmst_1,
  rmst_1_ll = rmst_ll_1,
  rmst_1_ul = rmst_ul_1,
  rmst_0 = rmst_0,
  rmst_0_ll = rmst_ll_0,
  rmst_0_ul = rmst_ul_0,
  rmst_diff = rmst_diff,
  ci_lower_diff = ci_lower_diff,
  ci_upper_diff = ci_upper_diff
)

# Print results for clarity
cat(paste("\nResults for tau =", tau, "days:\n"))
cat(paste("RMST for Arm 1 (95% CI): ", rmst_1, " (", rmst_ll_1, ", ", rmst_ul_1, ")\n", sep = ""))
cat(paste("RMST for Arm 0 (95% CI): ", rmst_0, " (", rmst_ll_0, ", ", rmst_ul_0, ")\n", sep = ""))
cat(paste("Difference in RMST (95% CI): ", rmst_diff, " (", ci_lower_diff, ", ", ci_upper_diff, ")\n", sep = ""))

# Prepare data for plotting trapezoids for RMST calculation
trapezoids_1 <- data.frame(
```

```
    time = rep(km_times_1, each = 2)[-1],
    surv = rep(km_surv_1, each = 2)[-1],
    group = "Arm 1"
   )

   trapezoids_0 <- data.frame(
    time = rep(km_times_0, each = 2)[-1],
    surv = rep(km_surv_0, each = 2)[-1],
    group = "Arm 0"
   )

   # Combine trapezoid data
   trapezoids <- rbind(trapezoids_1, trapezoids_0)

   # Prepare data for KM curves
   km_data_1 <- data.frame(time = km_times_1, surv = km_surv_1, group = "Arm 1")
   km_data_0 <- data.frame(time = km_times_0, surv = km_surv_0, group = "Arm 0")
   km_data <- rbind(km_data_1, km_data_0)

   # Plot the trapezoids and KM curves with different colors and add vertical lines
   p <- ggplot() +
    geom_step(data = km_data, aes(x = time, y = surv, color = group), size = 1) +
    geom_ribbon(data = trapezoids_1, aes(x = time, ymin = 0, ymax = surv, fill = group), fill = "blue", alpha = 0.2) +
    geom_ribbon(data = trapezoids_0, aes(x = time, ymin = 0, ymax = surv, fill = group), fill = "red", alpha = 0.2) +
    geom_vline(xintercept = c(360, 720, 1080, 1440), linetype = "dashed", color = "black") +
    annotate("text", x = c(360, 720, 1080, 1440), y = Inf, label = c("360 days", "720 days", "1080 days", "1440 days"), angle = 90, vjust = -0.5, hjust =
1.1) +
    scale_x_continuous(breaks = seq(0, 1440, by = 180)) +
    labs(title = paste("RMST Trapezoids and KM Curves for tau =", tau, "days:", title),
        x = "Time (days)",
        y = "Survival Probability") +
    theme_minimal()

   print(p)
  }

  return(results)
}

### Output RMST for thesis correction
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM2(fit.iptw.b.km, tau_values, "Baseline adjusted PP")
print(results)




################################################################################
### 1a. Estimating Effect in a Target Trial (Unadjusted ITT)                  ####################
################################################################################

# Calculated restricted mean survival time (RMST): 11 July 2024
# Make a month variable
TTE_GUTG001_base$SurvTime_FirstLine_month <- TTE_GUTG001_base$SurvTime_FirstLine/30

fit.km_RMST <- perform_rmst(TTE_GUTG001_base, "SurvTime_FirstLine_month", "Y_FirstLine", "FirstLine_num", c(12, 24, 36, 48))
fit.km_RMST



################################################################################################
### 1b. Estimating Unadjusted Effect in a Target Trial (PP)                                 ############
################################################################################################

# Calculated restricted mean survival time (RMST): 11 July 2024
# Make a month variable
TTE_GUTG001_base$SurvTime_DTR_month <- TTE_GUTG001_base$SurvTime_DTR/30

# See above for the function
fit.km.pp_RMST <- perform_rmst(TTE_GUTG001_base, "SurvTime_DTR_month", "Y_DTR", "FirstLine_num", c(12, 24, 36, 48))
fit.km.pp_RMST
```

```
##############################################
### 1c & 1d. Unadjusted As-treated patients     #######
##############################################

# 1c
# Make a month variable
# Use the dataset created below
TTE_GUTG001_base_reduced_at$SurvTime_FirstLine_month <- TTE_GUTG001_base_reduced_at$SurvTime_FirstLine/30

# See above for the function
fit.km.at_RMST <- perform_rmst(TTE_GUTG001_base_reduced_at, "SurvTime_FirstLine_month", "Y_FirstLine", "FirstLine_num", c(12, 24, 36, 48))
fit.km.at_RMST

# 1d
# Make a month variable
# Use the dataset created below
TTE_GUTG001_base_reduced_at2L$SurvTime_FirstLine_month <- TTE_GUTG001_base_reduced_at2L$SurvTime_FirstLine/30

# See above for the function
fit.km.at.2L_RMST <- perform_rmst(TTE_GUTG001_base_reduced_at2L, "SurvTime_FirstLine_month", "Y_FirstLine", "FirstLine_num", c(12, 24, 36,
48))
fit.km.at.2L_RMST



##############################################
### 2a. Baseline adjusted ITT patients              #######
##############################################

### Output RMST for thesis correction: see function created in the main TTE file
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM(fit.iptw.b.km.itt, tau_values, "Baseline adjusted ITT")
print(results)



#############################################################################
### 2b. Baselin eadjusted (IPW) - PP                                  ##############################
#############################################################################

### Output RMST for thesis correction: see function created in the main TTE file
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM(fit.iptw.b.km, tau_values, "Baseline adjusted PP")
print(results)



##################################################
### 2c & 2d. Baseline adjusted AT patients              #######
##################################################

# 2c
### Output RMST for thesis correction: see function created in the main TTE file
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM(fit.iptw.b.km.at, tau_values, "Baseline adjusted AT")
print(results)

# 2d
### Output RMST for thesis correction: see function created in the main TTE file
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM(fit.iptw.b.km.at.2L, tau_values, "Baseline adjusted AT, limited to second-line treatment recipients")
print(results)



#########################################################################################################################
### 3. Adjusted for baseline (IPW) & censoring due to treatment deviation (IPCW) - Hypothetical perfect per-protocol effect                  ####
#########################################################################################################################

### Output RMST for thesis correction: see function created in the main TTE file
tau_values <- c(360, 720, 1080, 1440)
results <- perform_rmst_trapezoids_wKM(fit.iptw.b.ipcwCO.s3.km, tau_values, "IPTW*IPCW adjusted PP")
print(results)
```