Treatment outcome modelling in anal cancer radiotherapy: utilising distributed learning across multiple international centres

Stelios Theophanous

University of Leeds

School of Medicine

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

November 2022

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 2 is based on jointly authored work which is being prepared for publication:

Theophanous S, Baldwin J, Saif A, Gilbert A, Scarsbrook A, Wheller B, Jones C, Samuel R, Appelt A, Lilley J. Development of a comprehensive institutional anal cancer data warehouse for real-world data analysis. In preparation.

Stelios Theophanous contributed to the conception of the study idea and to the formulation of the research question. He was responsible for the identification of relevant data items and data sources, for the generation of the data dictionary and for the specification of the structure of the data warehouse. He was also responsible for carrying out all manual data collection, evaluating the completeness and quality of the data, preparing the initial draft manuscript, revising the manuscript, and preparing the final version included in this thesis. John Lilley, Ane Appelt, Alexandra Gilbert and Bob Wheller contributed to the conception of the study idea, to the formulation of the research question and to the identification of relevant data items and data sources. Jack Baldwin and Ahmed Saif contributed to the generation of the data dictionary and to the specification of the data warehouse structure. They were also responsible for setting up a pipeline for automatic extraction of data and contributed to the evaluation of data completeness and quality. All co-authors reviewed the manuscript and provided feedback on written drafts. All co-authors read and approved the final manuscript included in this thesis.

Chapter 3 is based on work from the following jointly authored publication:

Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, Appelt A. Prognostic factors for patients with anal cancer treated with conformal radiotherapy—a systematic review. BMC Cancer 2022;22:607. Available from: https://doi.org/10.1186/s12885-022-09729-4.

Stelios Theophanous conceived the study concept and the research question. He was responsible for conducting the systematic review, including carrying out the literature search, screening the identified articles, selecting the relevant articles to be reviewed, extracting the data from these articles, analysing the data, evaluating the methodological quality of the articles, preparing the initial draft manuscript, revising the manuscript, and preparing the final version. Ane Appelt, Alexandra Gilbert and Ann Henry contributed to the conception of the study idea and to the formulation of the research question. For quality assurance purposes, Robert Samuel screened a subset of the articles, extracted the data from a subset of the articles, and evaluated the methodological quality of all articles. All co-authors reviewed the manuscript, provided feedback on written drafts, and read and approved the final manuscript.

Chapter 4 is based on work from the following jointly authored publication:

Theophanous S*, Choudhury A*, Lønne P-I* (joint first authors), Samuel R, Guren MG, Berbee M, Brown P, Lilley J, van Soest J, Dekker A, Gilbert A, Malinen E, Wee L, Appelt A. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study. Radiotherapy and Oncology 2021;159:183–9. Available from: https://doi.org/10.1016/j.radonc.2021.03.013.

Stelios Theophanous contributed to the formulation of the overarching research goals and aims, and to the design of the methodology. He was responsible for coordinating the planning and execution of research activity, for preparing the prospective study protocol and statistical analysis plan, for implementing the technical distributed learning infrastructure in Leeds Cancer Centre, for running the statistical analyses, for writing the draft manuscript, for responding to reviewers’ comments and for producing the final version of the manuscript. Ane Appelt and Leonard Wee conceived the study idea and set up the research collaboration. Ananya Choudhury and Leonard Wee were primarily responsible for the development and implementation of the distributed learning infrastructure. Alexandra Gilbert, Marianne Grønlie Guren, Maaike Berbee, Peter Brown and Robert Samuel contributed to data collection, and to the discussion of the clinical relevance of the study findings. All co-authors critically reviewed, edited, and approved the study protocol and statistical analysis plan. All co-authors critically reviewed the manuscript and provided feedback on written drafts. All co-authors read and approved the final version of the manuscript.

Chapter 5 is based on work from the following jointly authored publication:

Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Dennis K, Dewdney A, Gambacorta MA, Gilbert A, Guren MG, Holloway L, Rashmi J, Kochhar R, Mohamed AA, Muirhead R, Parés O, Raszewski Ł, Roy R, Scarsbrook A, Sebag-Montefiore D, Spezi E, Spindler KL, van Triest B, Vassiliou V, Malinen E, Wee L, Appelt A, on behalf of the atomCAT consortium. Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study. Diagnostic and Prognostic Research 2022;6:14. Available from: https://doi.org/10.1186/s41512-022-00128-8.

Stelios Theophanous contributed to the formulation of the overarching research goals and aims, to the design of the methodology, and to the recruitment of international centres to the atomCAT consortium. He was responsible for coordinating the planning and execution of all research activity, preparing the initial draft prospective study protocol and statistical analysis plan, writing the manuscript, responding to reviewers' comments, and producing the final version of the manuscript. Ane Appelt and Leonard Wee conceived the study idea. Ane Appelt, Leonard Wee and Eirik Malinen contributed to the recruitment of international centres to the atomCAT consortium. Alexandra Gilbert, Marianne Grønlie Guren, Maaike Berbee and Andrew Scarsbrook contributed to the development of the study data dictionary, critically reviewed the findings from the systematic review to select and prioritise factors for model inclusion and contributed to the discussion of the clinical relevance of the study findings. All co-authors critically reviewed, edited, and approved manuscript and provided feedback on written drafts. All co-authors read and approved the final version of the manuscript.

Chapter 6 is based on jointly authored work which is being prepared for publication:

Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Gambacorta MA, Gilbert A, Guren MG, Jadon R, Mohamed AA, Parés O, Raszewski Ł, Roy R, Scarsbrook A, Sebag-Montefiore D, Spezi E, Spindler KL, Deijen C, Vassiliou V, Malinen E, Wee L, Appelt A, on behalf of the atomCAT consortium. Prognostic models for anal cancer using distributed learning: the international multi-centre atomCAT2 study. In preparation.

Stelios Theophanous contributed to the formulation of the overarching research goals and aims, and to the design of the methodology. He was responsible for coordinating the planning, coordination, and execution of all research activity, for obtaining the required HRA and REC approvals, for implementing the technical distributed learning infrastructure in Leeds Cancer Centre, for collecting the local data and preparing the dataset for analysis and for supporting all other participating centres in setting up the infrastructure, in obtaining the required approvals and in preparing the local datasets. He was also responsible for running the distributed learning analysis, and all subsequent analyses, as well as for preparing the initial draft manuscript, revising the manuscript, and preparing the final version included in this thesis. Ane Appelt and Leonard Wee conceived the study idea. Ane Appelt contributed to obtaining legal and ethics study approvals, as well as to the coordination of the consortium. Ananya Choudhury and Leonard Wee contributed to the development of the distributed learning infrastructure. Alexandra Gilbert and Robert Samuel contributed to the local data collection for the Leeds dataset. Marianne Grønlie Guren, Per-Ivar Lønne, Maaike Berbee, Charlotte Deijen, Rajarshi Roy, Rashmi Jadon, Łukasz Raszewski, Maria Antonietta Gambacorta, Emiliano Spezi, Vassilios Vassiliou, Ahmed Allam Mohamed and Oriol Parés were responsible for the local data collection at each participating centre. Numerous additional people contributed to the local aspects of atomCAT2 in individual centres, including data collection, infrastructure setup and local coordination. Please see Appendix A for a full list of the atomCAT consortium members. Alexandra Gilbert, Marianne Grønlie Guren, Maaike Berbee, Andrew Scarsbrook and David Sebag-Montefiore contributed to the discussion of the clinical relevance of the study findings. Ane Appelt, Alexandra Gilbert, David Sebag-Montefiore, Marianne Grønlie Guren and Andre Dekker critically reviewed the manuscript and provided feedback on written drafts. All co-authors read and approved the final version of the manuscript included in this thesis.

Acknowledgements

I am deeply grateful to my supervisors Ane Appelt, Alexandra Gilbert, Ann Henry, and John Lilley for their continuous support throughout my PhD studies. Ane, I could not have completed this challenging journey without your insightful comments and suggestions. Thank you for guiding me so positively, for inspiring me and for always making me feel confident in my abilities.

I would like to offer my special thanks to Leonard Wee, Eirik Malinen, Per-Ivar Lønne and Ananya Choudhury for their assistance at every stage of the work on the atomCAT projects. I very much appreciate your hard work and support.

I would also like to express my thanks to Robert Samuel, Jack Baldwin, and Ahmed Saif for sharing their knowledge and expertise with me, which were key to the success of this research.

To my family; Eftychia, Fanos, Paris, grandmas, and grandads, without you I would not be who I am today. Thank you for motivating me, for always being there for me, and for always supporting my dreams.

I am grateful to my partner Esther, and to my friends Luke and Fee for their unwavering support along the entire journey.

Abstract

Anal cancer is a rare disease typically treated with concurrent chemoradiotherapy. Lack of understanding of prognostic factors renders options for treatment individualisation limited. Due to the rarity of the cancer, single-centre data are rarely sufficient for robust prognostic model development. Distributed learning enables the analysis of datasets from multiple centres without exchanging sensitive individual-level patient data. This thesis aimed to determine prognostic factors for patients treated for anal cancer with modern radiotherapy by using distributed learning to analyse real-world data across an international consortium.

To achieve this, a local anal cancer data warehouse was established, which includes data for 568 patients treated at Leeds Cancer Centre between 2013 and 2022. The literature was systematically reviewed to identify established prognostic factors for anal cancer outcomes after treatment with conformal radiotherapy. 19 studies were evaluated, and N stage, T stage, and sex were identified as the most prevalent clinical prognostic factors for the majority of outcomes explored.

The atomCAT1 three-centre proof-of-concept study was successful in demonstrating the value of distributed learning in outcome modelling for rare cancers. This study guided the expansion of the initial collaboration into an international consortium consisting of 14 radiotherapy treatment centres. Distributed learning was implemented for collaborative prognostic model development and validation across the atomCAT consortium. In the atomCAT2 study, the distributed learning analysis of data from 1,099 patients treated across 12 centres established nodal involvement, male sex, older age, and larger primary tumour size as prognostic for poorer overall survival; male sex, higher T stage, and larger primary tumour size as prognostic for poorer locoregional control; and nodal involvement and larger primary tumour size as prognostic for poorer freedom from distant metastasis. These results may guide the design of future clinical trials in anal cancer and may ultimately aid the personalisation of treatment for future patients.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Definition
%	Per cent
α/β	Alpha/beta ratio
5-FU	5-fluorouracil
AJCC	American Joint Committee on Cancer
APR	Abdominoperineal resection
ASCC	Anal squamous cell carcinoma
atomCAT	Anal cancer Treatment Outcome Modelling with Computer Aided Theragnostics
CFS	Colostomy-free survival
CI	Confidence interval
CORMAC	Core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer
CRT	Chemoradiotherapy
CT	Computed tomography
CTV	Clinical target volume
DFS	Disease-free survival
DL	Distributed learning
EBRT	External-beam radiotherapy
EQD2	Equivalent dose in 2Gy fractions (α/β=10Gy)
ETL	Extract Transform Load
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
GTV	Gross tumour volume
GWAS	Genome-wide association studies
HEI	Hemo-eosinophil inflammation
HIV	Human immunodeficiency virus
HPV	Human papilloma virus

HR	Hazard ratio
HRA	Health Research Authority
ICD10	International Classification of Diseases 10th revision
ID	Identification number
IG	Information governance
IKNL	Netherlands Comprehensive Cancer Organisation
IMRT	Intensity-modulated radiation therapy
IQR	Interquartile range
IRAS	Integrated Research Application System
IRB	Institutional review board
KNN	k-Nearest neighbour algorithm
LCC	Leeds Cancer Centre
LP	Linear predictor
LRF	Locoregional failure
MAR	Missing at random
MDT	Multidisciplinary Team
MDW	Medical Data Works
MFS	Metastasis-free survival
MICE	Multiple imputation by chained equations
MLCs	Multi-leaf collimators
MMC	Mitomycin C
MNAR	Missing not at random
MRI	Magnetic resonance imaging
MVA	Multivariable analysis
NIH	National Institutes of Health
NTCP	Normal tissue complication probability
OAR	Organs at risk
OS	Overall survival
PACS	Picture archiving and communication systems
PET	Positron emission tomography
PPI	Patient and public involvement

PPM	Patient Pathway Manager
PROGRESS	PROgnosis RESearch Strategy initiative
PROs	Patient-reported outcomes
PTV	Planning target volume
RCT	Randomised controlled trial
REC	Research Ethics Committee
ROC	Receiver Operating Characteristics
SCC	Squamous cell carcinoma
SIB	Simultaneous integrated boost
SSIS	SQL Server Integration Services
SUVmax	Maximum standardized uptake value
TCP	Tumour control probability
TNM	Tumour-node-metastasis
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UK	United Kingdom
UVA	Univariable analysis
Vantage6	priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange
VMAT	Volumetric-modulated arc therapy
NHS	National Health Service
IRAS	Integrated Research Application System
3D-CRT	Three-dimensional conformal radiotherapy

Approvals for use of patient data for research

An approval for the use of data for the purpose of the anal cancer data warehouse project described in Chapter 2 was obtained from the LeedsCAT (REC reference: 19/YH/0300, IRAS project ID: 255585) research governance board. This approval covered the use of clinical and radiotherapy data from patients treated for anal cancer in Leeds Cancer Centre between 2008 and 2022.

For the atomCAT1 work described in Chapter 4, each participating centre acquired a separate local approval for accessing and collecting patient data for research. In Leeds Cancer Centre, the study was approved by the LeedsCAT (REC reference: 19/YH/0300, IRAS project ID: 255585) research governance board. In MAASTRO clinic, IRB approval was obtained to extract patient data from electronic records (reference: P-0535). In Oslo, Regional Ethics Committee approval (reference number REK 2012/2274) was obtained for re-use of data from the ANCARAD trial (NCT01937780) via a study amendment, and the local data protection officer reviewed and approved the distributed learning infrastructure.

For the atomCAT2 work described in Chapter 6, each centre acquired a separate local approval for accessing and collecting patient data for research. Each local coordinating investigator has provided a copy of the letter confirming that use of data for research has been approved, including approval reference number, to the central study coordinator.

In Leeds Cancer Centre, the study was covered by the LeedsCAT approval referenced above. In Cardiff University, the study received CardiffCAT approval (REC reference: 19/WA/0119, IRAS project ID: 262568). Institutional Review Board (IRB) approval was obtained by Aachen University (reference: EK 478/21), by Greater Poland Cancer Centre (reference: KB 530/21), by NKI-AVL (reference: RBd21-166), and by MAASTRO (reference: P-0535). The Bank of Cyprus Oncology Centre received approval from the President of the National Committee of Bioethics (no reference provided), and the Champalimaud Foundation received approval from the President of the Ethics Committee (no reference provided). At Oslo University Hospital, Regional Ethics Committee approval was obtained (reference: 2012/2274). At Fondazione

Policlinico Universitario A. Gemelli IRCCS, local Medical Ethics Committee approval was obtained (reference: ID 4350).

For UK centres which could not gain local approval to access and collect patient data for the purpose of the project (Hull University Teaching Hospitals NHS Trust and Cambridge University Hospitals NHS Foundation Trust), a central project application sponsored by the University of Leeds and with Ane Appelt as primary investigator was submitted for review by the Health Research Authority (HRA) and the Research Ethics Committee (REC). The central project application received HRA and REC approval (IRAS project ID: 303103, REC reference: 22/WA/0081). See Appendix B for the HRA approval letter and Appendix C for the REC approval letter.

All Leeds data collected and used in this project were stored in access-controlled folders on Leeds Teaching Hospitals NHS Trust computers. For atomCAT1 and atomCAT2, no individual level patient data were shared between centres.

Chapter 1 - Introduction

1.1 Opening statement

This thesis focuses on the development and validation of outcome prediction models for anal cancer across multiple UK and European centres. The main aim is to demonstrate the feasibility of developing such models without sharing sensitive patient data, through the employment of a novel data analysis technique called distributed (or federated) learning. To achieve this, the study explores and addresses several research questions, which are fundamental to learning from every single patient treated for anal cancer and as a result improving outcomes for future patients.

Leeds Cancer Centre (LCC) is a large tertiary cancer centre treating thousands of patients each year. However, only a small percentage of these patients participate in clinical trials. This means that the routinely collected data for all patients treated are not contributing to the improvement of treatment of future patients. This is an issue, particularly for rare cancers such as anal cancer, where each centre will only treat a few patients each year (e.g. in Leeds 40-50 patients), and patient data are rarely shared nationally and internationally between centres. Existing solutions depend on data being pooled from multiple locations in large repositories and analysed centrally. This leads to data privacy and ethical concerns regarding individual-level patient data sharing, which hinder potential collaborations between centres. Such approaches are also technically challenging, especially when dealing with complex data, for example detailed radiotherapy, imaging, and biomarker data. There are thus significant barriers to learning from every patient treated for anal cancer.

To address these challenges, this thesis describes the evaluation of patient data availability at LCC and the establishment of a local anal cancer patient data warehouse. Subsequently, it investigates the availability of established factors which impact anal cancer patient outcomes after radiotherapy through a systematic review of the literature. The routine patient data from the local database and the identified prognostic factors can then feed into the development of distributed outcome models of increasing complexity. The formation of the atomCAT (Anal cancer Treatment Outcome Modelling with Computer Aided Theragnostics) consortium, which initially consisted of an initial pilot of three centres and was later extended to 14 centres from the UK and Europe, is central to this work. This collaboration renders the analysis of data from a large number of patients possible, which in turn allows us to advance our understanding of anal cancer in a real-world setting. Robust outcome prediction models may inform the design of future clinical trials and support shared decision-making regarding treatment of future patients.

1.1.1 Introduction chapter overview

This introductory chapter covers several aspects of anal cancer, including standard treatment in the UK, focusing mainly on the radiotherapy aspects of treatment. The anal cancer treatment pathway will be described in detail, beginning from the first stages of diagnosis, and ending at the follow-up procedures that are carried out once treatment is complete. The data-driven approaches used to identify factors affecting patient outcomes will then be discussed, emphasising the significance of prognostic research in this field. The need for more multicentre and international collaborations to carry out outcome modelling for anal cancer will be highlighted. Distributed learning will be discussed as an approach to promote further collaboration between centres, and details will be provided regarding its present and possible future applications, as well as its current limitations. Specifically, reasons why distributed learning may be of significant value to anal cancer prognostic research will be discussed. At the end, the overall aims of this research will be specified, and the specific aims of the following chapters of this thesis will be summarised.

1.2 Anal cancer and modern radiotherapy

1.2.1 Anal cancer epidemiology

Anal cancer is a rare disease comprising about 0.3% of all cancers [1]. Over the past few decades, a rise in incidence has been observed. In England, the incidence rate increased from 1.54 (95% CI, 1.42-1.66) to 2.31 (95% CI, 2.18-2.45) per 100,000 person-years between 2001 and 2017 [2]. Females are twice as likely to be diagnosed with anal cancer in Northern Europe (age standardised incidence rates of 0.79 per 100,000 person-years in males compared to 1.60 per 100,000 person-years in females) and in Western Europe (age standardised incidence rates of 0.96 per 100,000 person-years in males compared to 2.00 per 100,000 person-years in females) [1]. The median age of diagnosis of anal cancer in the UK is 67 years [2].

Apart from female sex, multiple other risk factors associated with the development of anal cancer have been identified. These include human papilloma virus (HPV) infection, human immunodeficiency virus (HIV) infection, multiple sexual partners, receptive anal intercourse, history of cervical, vulvar or vaginal carcinoma, and smoking [3]. Importantly, epidemiological studies have indicated that more than 8 out of 10 anal cancers are linked to HPV infection [4]. Since HPV infection incidence has also been on the rise [5], this could provide a potential explanation for the rising incidence of HPV-related cancers, such as cancers of the anus, cervix, vulva, vagina and oropharynx. It is worth noting that HPV-positive cancers appear to be more sensitive to radiation than HPV-negative cancers, although the mechanism underlying this difference is not clearly understood [6].

1.2.2 Types of anal cancer

Anal cancers can be broadly classified into two categories: cancers of the anal margin and cancers of the anal canal [7]. Figure 1-1 illustrates the anatomy of the anal region and shows the location of the anal canal and the anal margin.

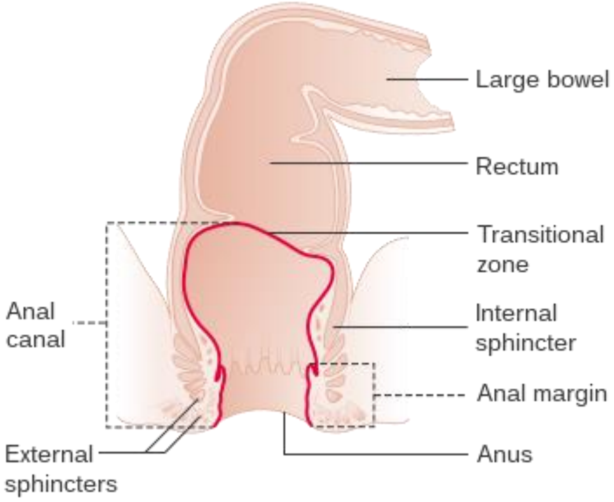


Figure 1-1. Cross-section of the rectum and anal canal, showing the position of the of the anal canal and the anal margin. CRUK [8]. Reproduced under the Creative Commons Attribution-Share Alike 4.0 International license [https://creativecommons.org/licenses/by-sa/4.0/].

85% of all anal cancer arises in between the anal canal and the anal verge. Anal margin cancers are relatively uncommon, comprising only about 15% of all anal cancers. Histologically, squamous cell carcinomas (SCC) are the most common lesions, comprising about three quarters of the total number of anal canal cancers [7].

Other pathology found at the anal verge or canal include adenocarcinoma (treated as rectal adenocarcinoma), and less commonly, cloacogenic carcinoma, basal cell carcinoma, Bowen's disease, and Paget's disease [9]. These patients are treated differently to patients found to have SCC. The work carried out and discussed in this thesis focuses specifically on SCC anal cancer, as it is the most common form of anal cancer and will be referred to as anal cancer from this point onwards.

1.2.3 Diagnosis and pre-treatment investigations

Anal cancer typically presents in the form of a perianal mass and the most frequently encountered non-specific symptoms include anal bleeding, anal or perianal pain and weight loss [10]. In cases where a patient presents with these symptoms and anal cancer is suspected, a patient will be referred to a colorectal surgeon. The diagnosing clinician then performs a digital rectal examination, as well as a physical examination of the inguinal lymph nodes to identify any abnormal growths [11]. If anal cancer is suspected, the patient will undergo an examination under anaesthetic and biopsy. The definitive diagnosis of anal cancer depends on biopsy-proven histology [12].

After the initial anal cancer diagnosis, various imaging modalities are utilised in order to collect more information on the extent of the disease. These include computed tomography (CT), positron emission tomography (PET)/CT and magnetic resonance imaging (MRI) scans. In most centres, multiple modalities are employed since each technique presents its own advantages. Contrast CT scans are mainly used to assess the presence of involved lymph nodes and distant metastases, PET/CT scans are used to identify the extent of local disease, involved lymph nodes and distant metastases, and MRI scans are highly effective in assessing locoregional disease and lymph node metastases [12].

The biopsy results as well as the various scan images are subsequently reviewed and discussed in a Colorectal Multidisciplinary Team (MDT) meeting [13]. This team consists of experts in different disciplines, such as clinical oncologists, medical oncologists, surgeons, radiologists, histopathologists, and clinical nurse specialists [14], who meet in order to review the patient’s diagnosis, assess the extent of the disease and select the most appropriate treatment plan according to the available evidence.

1.2.3.1 Anal cancer staging, grouping, and grading

During the MDT meeting, the patient’s cancer is staged according to the tumour-node-metastasis (TNM) staging system developed by the American Joint Committee on Cancer (AJCC) [15]. The TNM staging system is designed to classify the patient’s cancer into risk categories and considers the size of the tumour (T stage), spread to nearby lymph nodes (N stage) and to other organs (M stage). Over the past few decades, the AJCC Cancer Staging Manual has been updated multiple times to incorporate the latest evidence from research and clinical trials. Specifically for anal cancer, there have been substantial changes in nodal staging between the previous version (AJCC TNM v7) [16] and the most recent version (AJCC TNM v8) [17], which is being used since 2016, as presented in Table 1-1. The categorisation for T stage and M stage did not change.

Table 1-1. AJCC staging for anal cancer. For N stage, the classification is shown for both AJCC TNM version 7 and version 8 to emphasise the changes that took place.

T stage		**N stage (v7)**		**N stage (v8)**		**M stage**	
Tumour size		*Nodal involvement*		*Nodal involvement*		*Metastasis*	
Tx	Primary tumour not examined or cannot be assessed	Nx	Nodal involvement not examined or cannot be evaluated	Nx	Nodal involvement not examined or cannot be evaluated	Mx	Metastasis not examined or cannot be evaluated
T0	No evidence of primary tumour	N0	No lymph node involvement	N0	No lymph node involvement	M0	No metastasis to distant organs
Tis	Early-stage cancer that hasn’t spread to other tissue						
T1	<2cm	N1	Perirectal lymph node involvement	N1a	Inguinal, perirectal or internal iliac lymph node involvement	M1	Metastasis to distant organs
T2	>2cm and <5cm	N2	Internal iliac and/or inguinal node involvement	N1b	External iliac lymph node involvement		
T3	>5cm	N3	Perirectal and inguinal and/or internal iliac and/or inguinal node involvement	N1c	Inguinal, perirectal or internal iliac, and external iliac node involvement		
T4	Any size, cancer has invaded other nearby organs						

Once the cancer's TNM staging is determined, its overall stage group [17] can subsequently be assigned by combining the T, N and M stages together. The overall eight stage group categories are summarised in Table 1-2.

The tumour can also be further characterised by its histological grade [15], which refers to how the cells look under a microscope. A lower grade (G1) means that the cancer cells are well differentiated and resemble normal cells. G2 and G3 are assigned to cancers with cells that are moderately or poorly differentiated, respectively, and therefore look rather different from normal cells. This grading system helps clinicians predict how fast the cancer will grow, which may in turn influence the plans for management.

Table 1-2. AJCC v8 group staging for anal cancer [15].

Stage group	TNM stages included	Description
0	Tis – N0 – M0	Early-stage cancer that hasn't spread to other tissue. No nodal involvement and no metastasis to other organs.
I	T1 – N0 – M0	The tumour is smaller than 2cm across. No nodal involvement and no metastasis to other organs.
IIA	T2 – N0 – M0	The tumour is larger than 2cm and smaller than 5cm across. No nodal involvement and no metastasis to other organs.
IIB	T3 – N0 – M0	The tumour is larger than 5cm across. No nodal involvement and no metastasis to other organs.
IIIA	T1 – N1 – M0 or T2 – N1 – M0	The tumour is smaller than 5cm across. Nodal involvement but no metastasis to other organs.
IIIB	T4 – N0 – M0	The tumour is any size and is growing into other nearby organs. No nodal involvement and no metastasis to other organs.
IIIC	T3 – N1 – M0 or T4 – N1 – M0	The tumour is larger than 5cm across or it is any size and is growing into other nearby organs. Nodal involvement but no metastasis to other organs.
IV	Tany – Nany – M1	The tumour is any size and is or isn't growing into other nearby organs. Nodal involvement or no nodal involvement. The cancer has metastasised to distant organs.

1.2.4 Management of anal cancer

1.2.4.1 History of clinical trials in anal cancer

Treatment for anal cancer using radiotherapy combined with chemotherapy was first described by Nigro et al. in 1974 [18,19] and then supported by three landmark trials [20–22], which defined external beam radiotherapy and concurrent chemotherapy with 5-fluorouracil (5-FU) and mitomycin C (MMC) as the standard of care. Prior to this, the standard treatment involved abdominoperineal resection (APR) surgery. This procedure involved the complete removal of the sigmoid colon, rectum and the anus [23], resulting in permanent colostomy and the need for a stoma bag, which significantly impacted the patient's quality of life. This treatment regimen also yielded low survival rates as well as high locoregional and distant relapse rates.

Combined modality chemoradiotherapy led to a complete response in 84% of the patients, did not involve a permanent colostomy and thus preserved sphincter function in a majority of patients. Therefore, the improvement in quality of life conferred by this approach led to the widespread acceptance of chemoradiotherapy as the gold standard of care, with surgery left as a salvage treatment (except for local excision of some T1N0 anal margin tumours). Over the last four decades, a number of multicentre trials have been conducted in the UK and abroad, in order to evaluate and optimise the treatment of anal cancer with chemoradiation, as summarised in Table 1-3.

These include the ACT 1 [20], RTOG 8704 [21], EORTC [22], RTOG 9811 [24], EXTRA [25], ACCORD 03 [26], ACT 2 [27], ACCORD 16 [28], RTOG 0529 [29] trials.

Table 1-3. Summary of past anal cancer clinical trials. 5-FU: 5-fluorouracil; MMC: Mitomycin.

Trial name	Phase	Number of patients and randomisation	What was investigated	Conclusions
ACT 1 [20]	3	585 Randomised	Is combined modality therapy (radiotherapy and chemotherapy) better than radiotherapy alone in patients with anal cancer?	Combined modality treatment leads to lower risk of local failure and death from anal cancer. Radiotherapy and 5-FU should be the standard treatment for most patients.
RTOG 8704 [21]	3	310 Randomised	Is adding MMC to the standard regimen important? What is the role of salvage chemoradiation in patients with residual tumours after chemoradiation?	MMC leads to greater toxicity, but its addition to the standard chemoradiation regimen is justified. In patients with residual tumours after chemoradiation, salvage chemoradiation should be attempted before salvage surgery.

EORTC [22]	3	110 Randomised	Does concomitant radiotherapy and chemotherapy improve local disease control and reduce the need for colostomy?	Locoregional control rates were significantly improved when radiotherapy and chemotherapy were used concomitantly. The need for colostomy was reduced in patients with locally advanced disease. Late side effects did not increase significantly.
RTOG 9811 [24]	3	598 Randomised	Does changing the standard treatment (concurrent radiotherapy and chemotherapy with MMC and 5-FU) to induction chemotherapy with cisplatin and 5-FU, followed by concurrent radiotherapy and chemotherapy with cisplatin and 5-FU improve disease-free survival?	The combination of induction chemotherapy with cisplatin and 5-FU, followed by cisplatin, 5-FU and radiation did not significantly improve disease-free survival compared to MMC, 5-FU and radiation.
EXTRA [25]	2	21 Non-randomised	Is the combination of capecitabine, mitomycin and radiotherapy feasible, efficient and well-tolerated in anal cancer patients?	The combination of capecitabine, mitomycin and radiotherapy leads to minimal toxicity and acceptable compliance.
ACCORD 03 [26]	3	283 Randomised	Does an increase in radiotherapy boost dose or two cycles of induction chemotherapy lead to improved colostomy-free survival?	Neither an increase in radiotherapy boost dose nor two cycles of induction chemotherapy leads to improved colostomy-free survival.
ACT 2 [27]	3	940 Randomised	Does replacing mitomycin with cisplatin improve response to chemoradiotherapy? Does maintenance chemotherapy after chemoradiation improve survival?	Replacing mitomycin with cisplatin does not significantly improve response to chemoradiotherapy. Maintenance chemotherapy after chemoradiation does not improve survival. The standard treatment (5-fluorouracil and radiotherapy) should remain unchanged.
ACCORD 16 [28]	2	16 Non-randomised	Does adding cetuximab to standard chemoradiotherapy improve the objective response rate in patients with locally advanced disease?	The addition of cetuximab to standard chemoradiotherapy leads to significant and unacceptable toxicity.
RTOG 0529 [29]	2	52 Non-randomised	Does using dose-painted IMRT reduce grade 2+ toxicity?	The use of dose-painted IMRT is linked with reduced grade 2+ hematologic and grade 3+ dermatologic and gastrointestinal toxicity.

1.2.4.2 Radiotherapy planning and delivery – Treatment modalities

Before radiotherapy is delivered, careful treatment planning needs to be carried out by a multi-disciplinary team that includes clinical oncologists, dosimetrists, and physicists. In this process, the treatment target volumes are delineated on a CT scan in order to direct the radiotherapy treatment fields to the appropriate treatment targets [30]. In the past, radiotherapy planning was based on outlining structures in two dimensions. The first phase of radiotherapy treatment was non-conformal, where only two treatment fields (anterior and posterior) were outlined to cover the disease and any at-risk nodes. The second phase consisted of outlining a planned volume that covered the primary tumour and nodal disease, as illustrated in Figure 1-2.

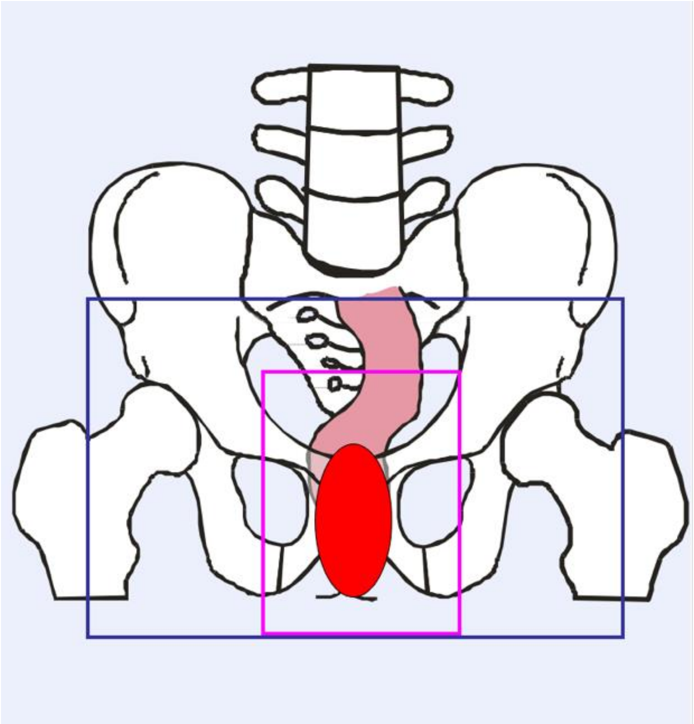


Figure 1-2. 2D Treatment plan with outlines of the anterior/posterior treatment fields, the primary tumour and nodal disease.

However, the introduction of three-dimensional conformal radiotherapy (3D-CRT) in the 1990s led to the adaptation of three-dimensional treatment target volumes and plans on axial CT images, which in turn led to increased treatment accuracy and therefore improved delivery.

Modern treatment plans include delineations of the gross tumour volume (GTV), the clinical target volume (CTV), the planning target volume (PTV) and the organs at risk (Figure 1-3) [31]. The GTV covers the primary tumour (GTV-T) and any involved nodes (GTV-N), whereas the CTV covers the entire elective volume, which comprises of the nodal volumes to which the cancer might have potentially spread. Therefore, the CTV is generally much larger than the GTV. The PTV is the overall treatment volume that includes the GTV and CTV, but is expanded further to account for all the uncertainties conferred by treatment planning and delivery [32]. The organs at risk (OAR) are also delineated. The OAR are critical healthy anatomical structures that may receive a significant radiation dose and are generally outside the PTV but may overlap. They rarely overlap with the CTV. In anal cancer, the OAR include the small bowel, large bowel, rectum, bladder, external genitalia, bone marrow, and the right and left femoral heads. These regions are spared and receive the lowest dose of radiation possible.

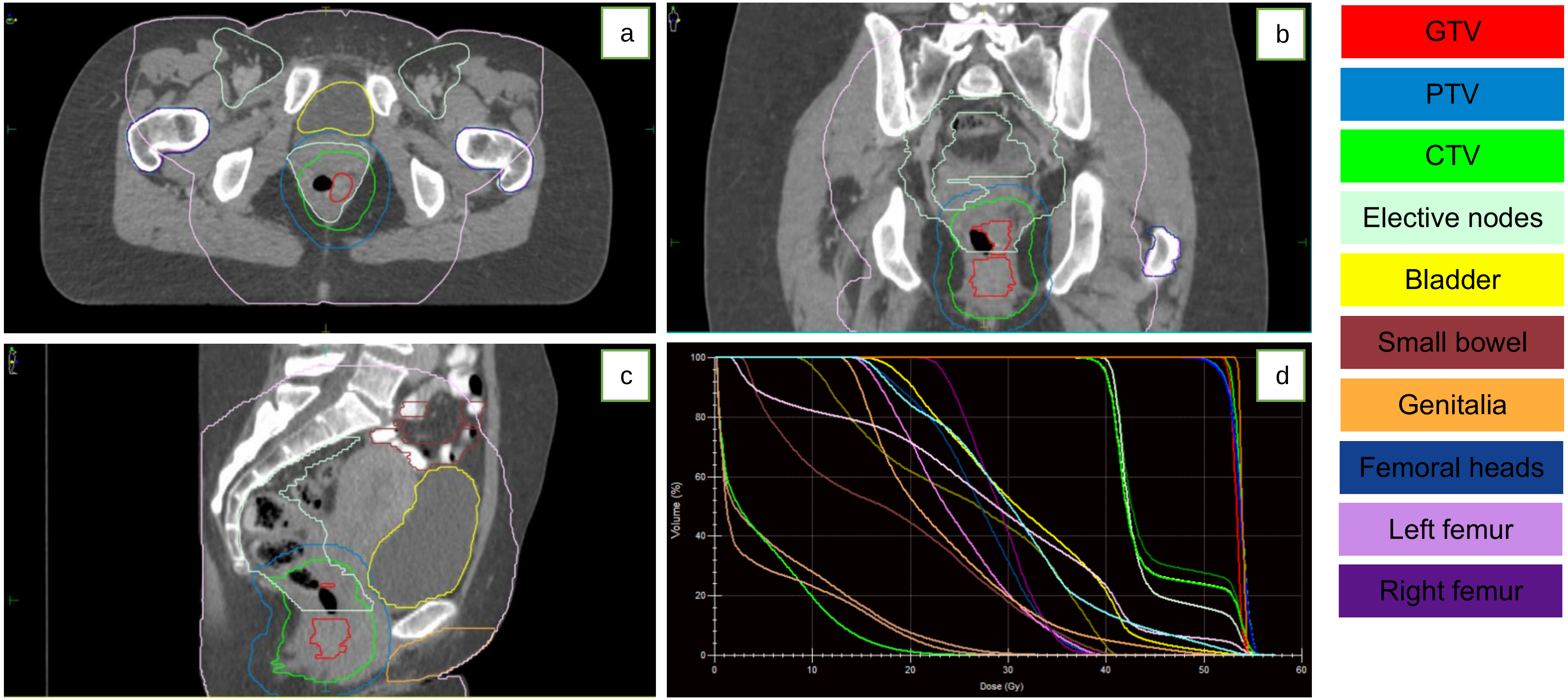


Figure 1-3. Conformal 3D treatment plan, showing the delineated volumes in multiple views: (a) transverse, (b) coronal, (c) sagittal. The dose-volume histogram is also shown in (d). GTV: Gross tumour volume; PTV: planning target volume; CTV: clinical target volume.

More recently, the use of MRI/CT and PET/CT images for target volume delineation has been explored and compared [33]. It has been demonstrated that there is a good agreement in GTV delineation between these modalities. Therefore, a combination of imaging modalities can be utilised for target volume delineation to improve accuracy.

In anal cancer, radiotherapy is commonly delivered as external-beam radiotherapy (EBRT) with linear accelerators and aims to adequately target and treat the tumour volume and associated lymph nodes, while sparing surrounding normal tissue and especially the OAR [34]. 3D-CRT used 3D imaging as well as multileaf collimators (MLCs) in the linear accelerator head to shape the beams so that they conform to the target's shape, therefore helping to shield normal tissue [35].

Further advances in the 2000s led to the introduction of intensity-modulated radiation therapy (IMRT), where MLCs move across the field within a beam, allowing even more control over the targeting of tumour volumes and sparing of normal tissue. IMRT also allows the delivery of a simultaneous integrated boost, in which varying doses can be given to different target volumes during the same treatment fraction [36]. The dose can therefore be delivered all at once instead of sequentially, leading to shorter treatment times. The effectiveness of this technique has been confirmed by the RTOG0529 [29] phase 2 trial (Table 1-3), which also indicated that it leads to lower levels of hematologic, dermatologic, and gastrointestinal toxicity. Thus, IMRT was established as the standard technique for radiotherapy delivery to anal cancer patients in the UK [32].

An enhanced type of IMRT, volumetric-modulated arc therapy (VMAT) is now used in the majority of treating hospitals. During VMAT, the radiation beams are delivered while the treatment gantry is moving around the patient instead of being static. Therefore, the tumour can be treated faster and more efficiently. The use of VMAT has also been linked to further improvements in patient outcomes and reductions in acute toxicity [37,38]. As a result, VMAT is currently the recommended technique for the treatment of patients with anal cancer and is currently the most commonly used type of IMRT in UK hospitals [12].

An alternative way to deliver radiation, other than EBRT, is via brachytherapy. Brachytherapy involves placing a sealed radioactive source directly into or next to the tumour, and it is an effective method of delivering high doses of radiation to the tumour whilst avoiding normal tissue [39]. Despite this technique not being commonly utilised for the treatment of anal cancer in the UK, a number of European centres favour its use.

High-dose brachytherapy is usually used in combination with EBRT to deliver a boost to the primary tumour [40]. Evidence has shown that due to the high dose delivered to the tumour, the reduced dose to healthy tissue and the reduction in overall treatment time conferred by brachytherapy, it results in excellent response to treatment, with improved local control and a reduction in treatment toxicity [40–42]. However, prospective RCTs need to be conducted in the future, in order to further confirm the role and the optimal dose of brachytherapy for the management of anal cancer. For the purpose of this project, no patients treated with brachytherapy were studied.

1.2.4.3 Standard radiotherapy in the UK

The current standard therapy for anal cancer in the UK involves concomitant radiotherapy, delivered via IMRT/VMAT, and chemotherapy. The primary aim of this treatment is to cure the disease, achieve locoregional control and preserve the function of the anus, whilst maintaining the best possible quality of life for the patient. The current radiotherapy schema is stratified according to tumour stage, as summarised in Table 1-4 [32].

In 2016, a prospective audit of anal cancer practice across the UK was carried out [43]. This study confirmed that the majority of UK centres delivered radiotherapy to patients with anal cancer with IMRT, which was linked to higher rates of radiotherapy completion, lower toxicity, and fewer treatment interruptions compared to two-phase conformal radiotherapy.

Table 1-4. Radiotherapy regimens for anal cancer patients in the UK, categorised by disease stage, according to the national guidance for IMRT in anal cancer [32]. The symbol # denotes radiotherapy fractions.

Tumour stage	Gross anal disease dose	Gross nodal disease dose	Elective nodes dose
T1/T2 N0 (and T2N1 at clinician's discretion)	50.4Gy in 28# (1.8Gy per #) in 5.5 weeks	50.4Gy in 28# (1.8Gy per #) in 5.5 weeks	40Gy in 28# (1.43Gy per #) in 5.5 weeks
T3/4N0 or Tany N2/3 (and T2N1 at clinician's discretion)	53.2Gy in 28# (1.9Gy per #) in 5.5 weeks	50.4Gy in 28# (1.8Gy per #) in 5.5 weeks.	40Gy in 28# (1.43Gy per #) in 5.5 weeks

1.2.4.4 Standard chemotherapy in the UK

Standard concomitant chemotherapy for localised anal cancer in the UK consists of mitomycin-C (12 mg/m^2 bolus day 1, capped at 20 mg) and 5-FU (1000 mg/m^2 in 1L normal saline over 24 hours, days 1-4 and days 29-32, capped at 2 m^2) or capecitabine (825 mg/m^2 twice daily on days of radiotherapy) [32].

The combination of 5-FU and cisplatin may also be prescribed in some centres, particularly to patients with advanced disease, or to patients who cannot receive mitomycin-C. Elderly and/or frail patients who cannot tolerate two chemotherapy drugs may receive 5-FU alone [44].

Additional alternative agents are also available and may sometimes be prescribed to patients with advanced anal cancer, or to patients that have already received standard chemotherapy with mitomycin-C and 5-FU. These alternative agent combinations include Carboplatin with paclitaxel (Taxol), Oxaliplatin with leucovorin and 5-FU, Docetaxel (Taxotere) with cisplatin and 5-FU, and finally cisplatin with Leucovorin and 5-FU [44].

1.2.4.5 Ongoing and future developments in anal cancer radiotherapy

The RTOG 9811 [24] and ACT 2 [27] trials reported that patients presenting with locally advanced tumours and nodal involvement have poorer outcomes after treatment. Additionally, different patterns of relapse were observed between early and late anal cancers after chemoradiotherapy. Therefore, the relationship between disease stage, radiotherapy dose and response to treatment needs to be investigated further. A small number of tumour control probability (TCP) models have been developed using literature-based data. These models were trained using large cohorts to explore the relationship between radiotherapy dose and treatment response [45,46]. The findings from these studies provide the rationale for individualised radiotherapy dosing in anal cancer patients, suggesting that a lower radiotherapy dose should be delivered to small tumours (dose de-escalation), and a higher radiotherapy dose should be delivered to large tumours (dose escalation).

The above hypothesis is being investigated in the ongoing PLATO trial [47]. PLATO consists of three separate trials (ACT3, ACT4, ACT5) that investigate the role of dose de-escalation and escalation according to how advanced the anal cancer is at diagnosis [48]. ACT3 is a non-randomised phase 2 trial in patients with T1N0 anal margin tumours, which evaluates the strategy of local excision where patients with microscopic margins >1mm have local excision only followed by observation, and those with ≤1mm margins receive additional lower-dose CRT using 41.4Gy in 23 fractions. ACT4 is a randomised phase 2 trial that aims to investigate whether dose de-escalation leads to equivalent or higher locoregional control rates whilst reducing side effects for patients with intermediate-risk anal cancer (T1-2N0). ACT5 is a randomised and integrated pilot - phase 2 - phase 3 trial that addresses high risk disease (T3-4N0 or TanyNode positive). It aims to explore whether dose escalation (using two alternative schedules – 58.8Gy in 28 fractions or 61.6Gy in 28 fractions) reduces local recurrence rates in patients with locally advanced anal cancer, without significantly increasing treatment side effects.

Apart from dose individualisation, other ongoing developments in anal cancer radiotherapy include the implementation of proton therapy, immunotherapy, and adaptive radiotherapy. Proton therapy is a type of radiation therapy that delivers protons instead of X-rays [49]. Currently, two ongoing trials (ClinicalTrials.gov IDs: NCT03018418 and NCT05055635) are exploring the role of proton therapy in the treatment of anal cancer. NCT03018418 is a phase 1 feasibility trial to evaluate if the dose to normal tissue is lower with the use of proton compared to photon beams in the primary treatment of anal cancer. NCT05055635 is a non-randomised phase 2 trial evaluating the role of proton beam radiotherapy in recurrent anal cancer.

Immunotherapy refers to the treatment of cancer by delivering drugs which activate or suppress the patient's immune system. Immunotherapy drugs commonly consist of targeted antibodies, tumour-infecting viruses, checkpoint inhibitors, or cytokines. Immunotherapy aims to trigger the immune system, in order to help it identify and destroy cancer cells more effectively [50]. The CORINTH (NCT04046133) phase 1b trial explores the safety and tolerability of the immunotherapy drug pembrolizumab, an immune checkpoint inhibitor, in patients with advanced (stage 3 and 4) anal cancer. The aim of the trial is to determine whether pembrolizumab can be added safely to standard CRT. The RADIANCE trial (NCT04230759) is a randomised phase 2 trial which aims to examine whether adding the immunotherapy drug durvalumab to the standard chemoradiotherapy regimen in patients with advanced anal cancer improves disease-free survival. Another phase 3 trial (NCT03233711) aims to determine the efficacy of giving the monoclonal antibody nivolumab to patients with high-risk disease after standard chemoradiotherapy and establish whether it improves disease-free survival for this group of patients.

Lastly, adaptive radiotherapy involves re-planning a patient after the initiation of radiotherapy, either at pre-specified time intervals, between the delivery of a certain number of fractions, or daily prior to treatment delivery, in order to account for the changes in tumour and normal tissue anatomy that occur during treatment [51,52]. The currently recruiting ROAR phase 2 trial (NCT05438836) will explore whether daily online adaptive radiotherapy significantly reduces treatment-related gastrointestinal toxicity. In this trial, a new treatment plan will be created each time a patient receives radiotherapy. Each new plan will account for the changes in tumour and normal tissue anatomy that occur during the course of the radiotherapy treatment [51].

1.2.5 Patient outcomes after radiotherapy

1.2.5.1 Follow-up procedure

Initial clinical response to treatment is evaluated at 6 weeks following completion of chemoradiotherapy through a digital rectal examination. Patients' will then have an MRI scan and/or a PET/CT scan at 3 months from the end of treatment to evaluate imaging response to treatment. Complete response is achieved when there are no residual tumours or ulcers. According to the ESMO-ESSO-ESTRO clinical practice guidelines, patients who have completely responded to treatment are then followed-up every 3 to 6 months for the first two years after treatment, and subsequently every 6 to 12 months for the next three years [12]. These follow-up appointments consist of discussion around symptoms of recurrence and side effects, a digital rectal examination, and palpation of the inguinal lymph nodes. In order to evaluate the presence or absence of locoregional and metastatic disease, a further MRI scan at 6 months from the end of treatment is standard based on the results from ACT2 trial. Patients will also undergo CT restaging scans at 12, 24 and 36 months after the end of their treatment (although individual centres practice may vary). In cases where residual or recurrent tumours are detected, additional imaging (e.g. further MRI or PET/CT scans) and/or examination under anaesthetic +/- biopsy will be organised if required, following discussion at MDT.

1.2.5.2 Locoregional failure

Locoregional failure is commonly defined as persistent disease or recurrence, either at the primary tumour site, or at the surrounding pelvic region and inguinal nodes [53,54]. In a study by Shakir et al. [53], which analysed the records of 385 anal cancer patients treated with conformal radiotherapy techniques in 5 UK centres, 86.7% of patients achieved a complete clinical response. Moreover, a three-year overall survival of 85.6% and a three-year disease-free survival of 75.6% was reported.

The study reported a disease recurrence rate of 19.2%, and notably the majority of relapses (83.4%) occurred at the site of the primary disease. This highlights the challenge of achieving locoregional tumour control in a subset of patients, and that most patients will fail locoregionally before getting metastatic disease. Currently, patients that have persistent disease after treatment or relapse locoregionally receive salvage surgery. This leads to acceptable overall survival and disease-free survival rates, but also high rates of surgical complications [55,56] that negatively impact the patient's quality of life. Therefore, efforts focusing on effective treatment of locoregional disease should be maximised.

1.2.5.3 Distant metastasis

Distant metastasis rates after chemoradiotherapy for anal cancer have been reported to be approximately 15% [12]. In most cases, the cancer metastasises to the liver, lungs, para-aortic nodes, or skin. Further treatment is planned according to the site and distribution of metastasis, and usually involves systemic treatment (e.g. chemotherapy) [57]. Patients with distant metastases have a poor prognosis, as only approximately 30% of patients survive for more than 5 years [58].

1.2.5.4 Treatment-related toxicity

Despite the relatively high survival rates achieved by modern treatment [37,59,60], chemoradiotherapy leads to numerous early and late side effects that may impact the quality of life of patients, even years after the end of treatment.

Common early side effects include radiation dermatitis, gastrointestinal toxicity (e.g. diarrhoea, bowel frequency and urgency), and urinary tract toxicity (e.g. dysuria) [61,62]. Additionally, patients receiving concurrent chemotherapy have a risk of various forms of haematological toxicity, such as anaemia, leukopenia and thrombocytopenia [12,61]. Therefore, full blood counts are taken from the patients weekly to track and manage these side effects. Various approaches are employed to reduce treatment-related side effects and improve tolerance to treatment, including the use of antibiotics and anti-emetics, as well as the provision of psychological support and advice regarding nutrition.

Additionally, chemoradiotherapy may result in fertility loss and/or early menopause, which is discussed with the patient before treatment. Female patients can choose to undergo cryopreservation of embryos or oocytes, and male patients may consider sperm banking.

One of the most common late side effects after chemoradiotherapy for anal cancer is faecal incontinence. It has been confirmed in various patient cohorts that more than a third of anal cancer survivors suffer from some degree of faecal incontinence [63,64]. Other late side effects include radiation proctitis, ulceration, dysuria, atrophy of the vaginal mucosa in female patients and impotence in male patients [62,64].

1.2.5.5 Patient-reported outcomes and health-related quality of life

Patient-reported outcomes (PROs) refer to reports received directly from patients in regards to their health-related quality of life [65]. PROs can provide valuable information on patients' experience during and after their treatment, as well as what side effects they are experiencing and how severe they are. They can be used alongside primary outcome measures assessed by the treating clinician, in order to acquire a more complete view of how the cancer and the treatment impact the patient's quality of life [66].

Despite the growth in PRO research in anal cancer, there is still room for improvement in the methodology used to collect data from patients, which will allow for the widespread adoption of PROs in clinical practice [67]. The EORTC quality of life group has addressed this issue through the development and validation of a questionnaire that is specific to anal cancer [68]. This questionnaire consists of 27 questions and can be used to collect PRO data from patients at different time-points during their treatment and follow-up. Efforts in collecting and analysing PROs from patients with anal cancer should be maximised, as they are fundamental in understanding the disease from a patient's perspective, which may in turn aid in the improvement of treatment for future patients.

1.2.5.6 Factors impacting patient outcomes and personalised medicine

Even though research and clinical trials carried out during past few decades have yielded substantial insights on anal cancer and its underlying biology, the translation of this information into novel therapies that can be implemented in the clinic remains a considerable challenge. One of the most important aspects that needs to be addressed is the heterogeneity in outcomes observed between patients.

Identifying which clinical factors, imaging factors and molecular biomarkers are prognostic and potentially predictive will help us better understand how this heterogeneity in outcomes arises. This insight can then be used to design innovative treatments for groups of patients with specific characteristics, leading to a more stratified or individualised approach to cancer treatment. Ongoing prospective trials [47] will address this in time, but valuable knowledge can in the meantime be extracted from data on routine treatment in local, national or international patient cohorts.

1.3 Data-driven approaches to improve patient outcomes

1.3.1 Limitations of randomised controlled trials

Most new discoveries and interventions in cancer are initially evaluated in randomised controlled trials (RCTs) before clinical adaptation, in order to demonstrate their safety and efficacy [69]. However, the majority of RCTs suffer from several limitations. Firstly, in many RCTs, there is a significant time interval between the trial conceptualisation and initiation phases [70]. Patient recruitment rates can also be slow, especially for rare cancers such as anal cancer, where the number of new patients diagnosed each year is usually small, even in large regional centres [71]. Therefore, there is often a long time interval between the conceptualising a trial and obtaining results. Other RCT limitations include limited follow-up periods, high costs, and high non-completion rates. Importantly, only a small percentage (5-15%) of the total number of patients participate in RCTs, which highlights that their results and conclusions might not be fully generalisable to the whole patient population [69].

Several studies have investigated this issue; they have demonstrated that women, children, patients with other common medical conditions, and elderly patients are commonly excluded, and therefore under-represented in RCTs [72–74]. Between the years of 2016 to 2018, more than a third of new cancers diagnoses in the UK were in people older than 75 years [75]. Despite this, almost 75% of older patients with colorectal cancer, including anal cancer, were deemed ineligible for inclusion in clinical trials [73,74]. This is problematic since the exclusion of this population of patients leads to significant gaps in knowledge on the benefits and risks conferred by the novel cancer treatments and strategies being tested in RCTs. As a result, access to new therapies for these patients can be severely restricted [74]. This is an even bigger challenge for rare cancers, where overall patient populations are small, making it even harder to test new treatment strategies, such as dose de-escalation [76], in elderly or frail patients.

One possible solution for this challenge is the collection and analysis of data from routine clinical practice. Even though routine data are generally of lower quality compared to data collected from RCTs, they can be more representative of the entire patient population and may therefore aid uncover novel insights on the disease in question. These may be more generalisable and could be used to improve treatment for future patients. The NHS England report from 2016 which set out a strategy for achieving world-class cancer outcomes further highlighted the importance of routine data in supporting improvements in cancer care and outcomes [77,78].

1.3.1.1 Routine data generation and collection

The cancer diagnosis, treatment and follow-up pathway previously described yields large amounts of routine data for each patient with anal cancer. This includes data that can be broadly classified in eight main categories: personal data, baseline clinical data, diagnostic data, radiotherapy planning and delivery data, non-radiotherapy related treatment data, follow-up data, outcome data, and supplementary data. A breakdown of each category is summarised in Table 1-5. Routinely collecting and storing these data for future use in research can provide real-world evidence to drive advances in anal cancer radiotherapy planning and delivery.

Table 1-5. Routine data collected for each patient with anal cancer after diagnosis, as well as during treatment and follow-up.

Category	Examples of data generated and collected
Personal data	Name, surname, sex and gender, date of birth, ethnicity, socioeconomic background.
Baseline clinical data	Date of diagnosis, place of diagnosis, site of tumour, tumour morphology, TNM stage, tumour size, number of nodes involved, health-related family history, previous cancer diagnosis, co-morbidities, performance status, site-specific information (such as HIV and HPV status for anal cancer).
Diagnostic data	Biopsy, diagnostic CT, MR and PET image.

Radiotherapy treatment planning and delivery data	Time between diagnosis and start of treatment, radiotherapy technique used for treatment (eg. IMRT/VMAT), radiotherapy dose and fractionation schedule (prescribed and delivered), full radiotherapy treatment data stored as DICOM files (including treatment plan, structure set, dose distribution and on-treatment imaging).
Non-radiotherapy related treatment data	Chemotherapy-specific data, surgery-specific data, immunotherapy-specific data.
Follow-up data	Biopsies, follow-up CT, MR and PET images.
Outcome data	Survival, local and regional failure, distant metastases, toxicity, patient-reported outcomes.
Supplementary data	Hospital admission during treatment, clinical trial information (considered/approached for inclusion, participation), surveys on patient experience.

1.3.1.2 Routine data in Leeds Cancer Centre

Currently, routine data from patients diagnosed and treated for anal cancer in LCC are stored in several distinct databases or clinical systems and can only be accessed by the patient's direct clinical care team, as well as by authorised individuals for the purposes of audit and research. Typically, patient data are stored in electronic health record systems such as the Patient Pathway Manager (PPM), in radiotherapy-specific software systems such as MOSAIQ and Monaco, and in picture archiving and communication systems (PACS). A more detailed description of these systems is provided in Chapter 2.

1.3.1.3 Leeds anal cancer database

As discussed in the previous section, routine data in LCC are stored in various systems and databases that are not fully linked. Therefore, neither of these include all relevant data for each individual patient. Consequently, clinicians or researchers looking to get a complete view of a patient's diagnosis, treatment, and follow-up journey need to access and collect data from multiple sources.

Once these data are collected, they need to be carefully pre-processed and cleaned, in order to validate their quality. Most real-world data suffer from several limitations that need to be addressed before they are analysed and used for research purposes [79]. Firstly, the data need to be carefully reviewed to ensure the number of errors or inaccuracies is minimised, especially when the data are automatically extracted from clinical systems. The presence of missing values is another important aspect that negatively affects the quality of a dataset. Dealing with missing values can pose a significant challenge. Even though the topic of missing value imputation has been studied extensively, selecting the correct approach is complex process [80,81] and may necessitate the expertise of data scientists and statisticians to adequately address. Lastly, different databases and systems may include overlapping information and patient cohorts. When extracting data from these, it is not uncommon to identify discrepancies in the data for the same patient. In these cases, additional evaluation of the data sources needs to take place, which may cause further delays. Assessing a dataset's quality is particularly complicated for non-clinical staff, who do not have first-hand experience in how data are generated and stored. For example, interpreting and extracting free-text data from clinicians' notes by researchers is not only arduous, but may also lead to more errors, impinging on the overall quality of the dataset.

As part of this PhD project, the Leeds anal cancer data warehouse was developed, with the aim of curating anal cancer data from various systems into a single data warehouse, in order to address the challenges discussed above. Chapter 2 of this thesis discusses in detail the work conducted to set up the data warehouse as well as the future plans to improve it.

1.3.1.4 National and international cancer databases

Institutional cancer databases can help us uncover new insights that can contribute to the improvement of treatment for future patients. However, these can be small and only include data from limited numbers of patients. In the past few decades, multiple national and international cancer databases have been created, aiming to link data from multiple cancer treatment centres. Due to the large amounts of data included and the variation in the patient cohorts, they can be analysed to identify patterns that would otherwise not be detectable by solely analysing data from single-institutional databases [82]. Through this approach, novel evidence-based interventions may be discovered, which may ultimately lead to improvements in patient survival and quality of life [83].

Despite their benefits, national and international cancer databases also suffer from several limitations. Notably, the validity of these databases has been scrutinised [84], as it is incredibly difficult to validate the accuracy of the data. This is because data originate from numerous sources, which may use different data entry and data coding approaches. Other limitations include the absence of various clinically important data items, duplicate reports from different sources and lastly large amounts of missing data [85]. All these need to be taken into consideration during the pre-processing and analysis phases when using data extracted from such databases.

1.3.2 Prognostic research overview

Prognostic research focuses on the estimation of the probability of a disease-related outcome after the end of treatment, given the specific characteristics of a patient at baseline (at diagnosis) [86,87]. Prognostic models have been proposed in cancer research for use in clinical treatment for more than 20 years [88] and have a wide range of potential applications, including the prediction of recurrence risk [89,90] and survival after the end of treatment [91–93].

The data generated from RCTs and from routine clinical practice can be used to develop prognostic models, which can yield insights into which factors impact patient outcomes after chemoradiotherapy for anal cancer. These models may ultimately help us stratify patients into risk groups, in order to develop more stratified or personalised treatment strategies. Currently, established prognostic factors are being used to stratify treatment for anal cancer in the UK, as shown in Table 1-4 [32]. Patients diagnosed with a small tumour (T1 or T2) and no involved nodes receive a lower radiotherapy dose than patients diagnosed with larger tumours (T3), tumours invading nearby organs (T4) or with involved nodes. This approach aims to minimise side effects in patients with lower risk disease, whilst still maintaining favourable oncological outcomes.

Prognostic models can also be used as clinical decision support tools, assisting clinicians in making informed decisions about patient management following a diagnosis [94]. Additionally, radiotherapy planning can be optimised through the development of normal tissue complication probability (NTCP) prediction models [95]. These models estimate the probability of dose-induced complications in normal tissue adjacent to the tumour and can therefore be used to compare the efficiency of different treatment strategies [96]. The majority of prognostic models fit into the data science framework of machine learning [97].

1.3.2.1 Prognostic factor research

According to the PROgnosis RESearch Strategy (PROGRESS) initiative [86], one of the main themes of prognostic research is to identify key factors that are associated with a certain prognosis [98].

Prognostic factors are defined as patient or disease-related characteristics that are linked with a certain outcome after therapy, such as death, locoregional failure, or distant metastasis. On an individual level, the identification of prognostic factors is vital, as they can be used to predict the chance of recovery from the disease, or the risk of disease relapse. This information can feed into the clinical and treatment-related decision-making prior to treatment. On a collective level, established prognostic factors can contribute to the design of new RCTs. For instance, anal cancer staging is currently being used in the PLATO trial to guide radiotherapy dose escalation and de-escalation strategies [47]. Prognostic factors can also be used as building blocks to develop robust prognostic models, as discussed in the next section.

However, current prognostic factor research methodology generally exhibits several weaknesses. Firstly, a large number of prognostic factor studies are poorly designed and do not employ appropriate statistical analysis techniques [99]. The reporting of prognostic studies is also often poor [100], leading to inadequate replication of their results by other studies [101]. In order for a prognostic factor to be truly informative, a factor's prognostic capacity should be generalisable across multiple studies examining different but similar patient cohorts. Therefore, more well-designed prognostic factor studies need to be conducted. These should ideally analyse data from large cohorts with the appropriate statistical techniques and should report the methodology employed as well as their results in a transparent manner.

1.3.2.2 Prognostic modelling research

Prognostic modelling research builds upon prognostic factor research, by using combinations of identified prognostic factors to develop models that can predict the probability of specific clinical outcomes in individual patients [102]. Prognostic model research consists of three phases: developing and internally validating the model, externally validating the model, and lastly establishing its clinical impact.

Firstly, robust model development highly depends on the quantity and quality of data used as well as the extent of missing data [103]. In general, models developed using large datasets consisting of high-quality data are more robust and can yield more accurate predictions. In addition, to ensure model robustness, the outcome measure and definition should be clearly defined, and the appropriate model form chosen. Once a model is developed, it needs to be validated. Internal validation is carried out using a dataset consisting of a similar population to the population used to train the model and ensures that the model is reproducible and does not suffer from overfitting. Overfitting refers to the model very closely fitting the set of data it was trained on (effectively fitted on "noise" in the data) but failing to adequately perform on a separate dataset with a similar patient population. Internal validation is commonly carried out by employing resampling techniques, such as bootstrapping and cross-validation to correct for over-optimism of the model's performance [104].

In external validation, the model's performance is evaluated on a completely different dataset in order to establish the generalisability of the model [105]. The model's performance can be determined in various ways, including measuring its discrimination, calibration, and overall performance [106,107]. Discrimination metrics determine a model's ability to accurately predict that an event will take place among patients who have that event ("case"), compared to patients that do not ("control"). One of the most commonly used metrics of model discrimination is Harrell's C-statistic, which can be visualised by constructing a Receiver Operating Characteristics (ROC) curve [108]. This statistic denotes the estimated probability that for a pair of "case" and "control" patients, the model assigns a higher risk to the "case" patient. Calibration metrics evaluate the model's ability to correctly predict the absolute risk of an event. In other words, a model that produces predictions that align with the observed values is thought to be well-calibrated [107]. The calibration of the model can be visualised by plotting a calibration or validation graph, where the x-axis represents the model's predictions, and the y-axis represents the observed outcomes. For binary outcomes, the y-axis only includes 0 and 1 values. The intercept and the slope of the calibration graph can then be assessed to deduce whether the model is well calibrated. The intercept can indicate whether model predictions are systematically too high or too low, whereas the calibration slope should be 1. Calibration slopes smaller than 1 may signify model overfitting, or that shrinkage of the model coefficients is needed [106]. Lastly, the R^2 Brier score [109] is a metric that can be used to measure the overall performance of the model, encompassing both the discrimination and the calibration aspects.

Once a model has been established, its development and validation strategy as well as its results need to be reported in a transparent and reproducible manner. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement is a 22-item checklist, which includes all aspects that should be addressed when reporting the development and validation of outcome prediction models [110,111].

Despite the large number of prognostic models reported in the literature, only few have been adopted for use in clinical practice [112–114]. This is mainly due to poor reporting of model development, inadequate external validation of results, and the lack of clinical impact studies, which assess the benefits and costs of using a model in clinical practice. These need to be addressed, in order to increase the impact of prognostic model research. This highlights the need for collaboration between different research groups, especially in the context of rare cancers. Through collaboration, more robust models can be developed by linking multiple datasets and analysing bigger patient cohorts. Moreover, one or a few datasets within the collaboration can be reserved for external validation purposes. Lastly, the impact and effectiveness of the model can then be assessed in multiple treatment centres.

1.3.2.3 Survival analysis

Survival analysis can be employed to predict the time duration until a specific outcome of interest occurs after cancer treatment, such as complete response to treatment, locoregional recurrence, distant metastasis, or death [115]. This analysis explicitly handles censored data and variation in follow-up between patients. Moreover, through survival analysis, the effect that various factors have on these events of interest can be determined. The majority of survival analyses employ a combination of statistical techniques, including the Kaplan-Meier estimator, log-rank tests, and Cox proportional hazards regression [116].

The Kaplan-Meier method makes use of two functions; the survival function and the hazard function, in order to estimate the survival curve [115,116]. The survival function S(*t*) estimates the probability of surviving to the time *t*, whereas the hazard function h(*t*) estimates the probability of dying at time *t*, given that the individual has survived up to that time. The survival curve is the plot of the Kaplan-Meier survival probability (S(*t*)) against time (*t)* and can be estimated using the observed survival times. The Kaplan-

Meier method does not assume an underlying probability distribution. Through this approach, the median survival time can be calculated. In order to compare the survival curves from two groups, a log-rank test can be carried out [115,116]. This is a non-parametric statistical hypothesis test that assesses whether the probability of an event occurring at any time point is the same for the two groups. However, this test cannot account for other explanatory factors.

The Cox proportional hazards regression model, first developed by David Cox more than 50 years ago [117], is analogous to a multiple regression model. Cox regression modelling can be conducted to explore the effect of multiple prognostic factors upon the time a specified event takes to happen, and to calculate the hazard ratios for each factor. The Cox proportional hazards regression approach assumes 'proportional hazards', i.e. that the effects of specific factors on the hazard function are independent of time. This approach assumes that the hazard function can be split into a time-dependent part, which does not depend on any of the covariates, and a time-independent part, which contains the co-variate effects. [116]. The model can be written as:

Equation 1-1

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\beta^T\boldsymbol{Z}) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p)$$

where $\lambda_0(t)$ is the baseline hazard function (baseline hazard when all the explanatory factors are 0); **Z** = {Z_1, Z_2, …, Z_p} is a p dimensional vector of explanatory factors; and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_p\}$ is a *p* dimensional vector of model coefficients. The factor coefficients are estimated from the data [118], using partial likelihood based methods. The Breslow's partial likelihood function is commonly used to estimate $\boldsymbol{\beta}$ using individual level patient data, in cases where tied event times are present [119], and is expressed as:

Equation 1-2

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^T \sum_{l\in\mathscr{D}_i} \mathbf{z}^l\right)}{\left[\sum_{l\in\mathscr{R}_i} \exp\left(\boldsymbol{\beta}^T \mathbf{z}^l\right)\right]^{d_i}},$$

where D is the total number of distinct event times; $\mathcal{D}i$ is the index set of subjects with observed events (e.g. death); $\mathcal{R}i$ is the index set of subjects at risk for the event, at the i-th distinct event time with i = 1, ..., D; di=|$\mathcal{D}i$| is the count of tied survival times at event time i. Finally, $\mathbf{z}^l$ = {$z^l{}_1$, $z^l{}_2$, …, $z^l{}_p$} is the realization of the p dimensional explanatory variable $\mathbf{Z}$ for a subject indicated by the superscript l.

The coefficients are then iteratively optimised until the partial likelihood is maximised, using the Newton-Raphson algorithm [120], which is expressed as:

Equation 1-3

$$\boldsymbol{\beta}^{\tau} = \boldsymbol{\beta}^{\tau-1} - \left[l''\left(\boldsymbol{\beta}^{\tau-1}\right)\right]^{-1} l'(\boldsymbol{\beta}^{\tau-1}).$$

where the parameters $\boldsymbol{\beta}^{\tau}$ are updated until convergence, in order to maximize the likelihood function at the τ-th iteration. Finally, the baseline hazard function in Breslow's approach can be defined as:

Equation 1-4

$$\lambda_0(t_i) = \frac{1}{\sum_{l \in \mathscr{R}_i} \exp(\widehat{\boldsymbol{\beta}}^{T} \mathbf{z}^l)},$$

where t_i is i-th distinct event time, and $\widehat{\boldsymbol{\beta}}'$ is the optimised estimate from *Equation 1-3* that maximizes the likelihood function.

The hazard ratio is calculated by exponentiating the factor coefficient (HR$_x$ = exp($\boldsymbol{\beta}_x$)) and is defined as the ratio of the risk of an event occurring at a particular point in time in one group (e.g. female patients) compared to another group (e.g. male patients). For instance, a hazard ratio of 2 with the female group being the baseline would signify that males are twice as likely to die at a particular point in time compared to females. The hazard ratio is time independent; therefore, the risk of males dying would be twice the risk of females dying at any point in time.

1.3.2.4 Stratified and personalised medicine

The ultimate goal of prognostic research is to identify prognostic factors which can be used to develop robust prognostic models, that can subsequently contribute to the advancement of novel stratified or personalised medicine approaches. Currently, the treatment of some types of cancer follows a one-size-fits-all approach, where the same treatment or the same dose is given to all patients with a specific cancer [121], with the exception of patients with metastases, which typically receive a different treatment to patients with localised cancer. However, every patient and every tumour are different, and these differences need to be considered in order to deliver the best possible care to the patient. In stratified medicine, each subgroup of patients with similar biological or risk characteristics is prescribed a targeted treatment [122]. Personalised medicine approaches proceed one step further; each individual patient receives a treatment that is targeted specifically to their distinct characteristics. Through stratified and personalised medicine, clinicians and patients together can select the most appropriate treatment that will confer the most clinical benefit and the least harm. This decision can be informed by using established and clinically impactful prognostic models. The characteristics of a newly diagnosed patient with cancer can be used as input for the model, which then returns the predicted risk group that the patient belongs to, or the predicted probability that the patient will have a certain event after their initial treatment. This information can be vital in making the correct decisions regarding a patient's care.

1.3.3 Prognostic research in anal cancer

Multiple prognostic factor studies have focused on investigating the effect of different factors on several disease-related outcomes in anal cancer, such as overall survival, disease-free survival, progression-free survival, and metastasis-free survival. The majority of these studies are retrospective, included small cohorts of less than 100 patients [123], or employed non-conformal radiotherapy techniques [124]. The results from studies that analysed cohorts treated with crude, non-conformal radiotherapy techniques may not be generalisable to current clinical practice, as non-conformal techniques were likely to deliver a radiotherapy dose that differed significantly from the prescribed dose [125]. Furthermore, due to the rarity of anal cancer, none of the few prospective RCTs that have already been completed have used conformal radiotherapy (3D-CRT / IMRT / VMAT) for the treatment of the participating patients and reported on factors that were deemed to be prognostic.

In a period of 20 years (between 2000 and 2020), only 19 studies analysed large cohorts of more than 100 patients treated with conformal radiotherapy and reported on survival and disease-related outcomes. The results and prognostic factors identified by these studies are discussed in detail in the systematic review of the literature that was carried out as part of this thesis Chapter 3. Overall, T stage, N stage, sex, pre-treatment biopsy HPV load, as well as the presence of baseline leukocytosis, neutrophilia and anaemia were found to be the most commonly identified prognostic factors for the outcomes explored. Recently published prognostic studies that fulfil the above criteria also confirm the prognostic role of T stage [126,127], sex [128], baseline neutrophilia [129] and anaemia [127] for a variety of anal cancer outcomes. Additional prognostic factors were also identified, including age [126], AJCC stage [126], baseline eosinophilia [127], baseline lymphopenia [129], and the hemo-eosinophils inflammation (HEI) index [128].

The number of robust prognostic studies in anal cancer remains small to this day. More prognostic studies need to be conducted in order to support the design of future RCTs and the growth of stratified and personalised medicine. These studies need to analyse sufficiently large datasets in order to develop statistically powerful models and identify the most relevant prognostic factors, especially factors with relatively limited effect size or with low prevalence. For rare cancers such as anal cancer, it might be a considerable challenge to collect enough patient data at single treatment centres. One possible solution for this issue it to establish multi-centre collaborations, in order to link and analyse datasets from multiple centres.

1.3.4 Multicentre collaboration in anal cancer

1.3.4.1 Sample size considerations and need for bigger datasets

New prognostic models must be developed using appropriately sized cohorts relative to the model complexity, to ensure that model predictions are accurate when applied to new patients [130]. Therefore, the minimum sample size required for the dataset used for model development should be calculated prior to the analysis. In the past, the 10 events-per-variable rule of thumb was considered as standard approach for calculation of the required sample size [131]. When applying this approach, the number of events of interest in the patient population is divided by 10 to dictate the number of parameters that can be included in the model. Conversely, if a number of prespecified parameters must be included in the model, the required minimum number of events in the dataset can be calculated. Over the past decade, the sample size calculation paradigm has been shifting away from this rule, since it was claimed to be too simplistic for use in clinical prognostic modelling [132–134]. More robust techniques have now been developed [135], which take into account the overall event rate, how long the cohort was followed up, the expected performance of the new model and the maximum degree of model overfitting allowed.

Small development datasets can lead to numerous issues relating to the performance of the resulting model. Prognostic models developed using small datasets are often over-optimistic [136]. Such models appear to perform well when tested on a similar patient cohort, but due to overfitting, their performance drops significantly when making predictions on a different cohort. Poor model calibration as a result of inadequate sample size has been reported in multiple studies [137]. Additionally, the power to identify relevant prognostic factors in studies with small datasets, especially factors with relatively limited effect size or with low prevalence, may be limited. Any factors identified and their effect estimates may suffer from small sample bias. Furthermore, externally validating a prognostic model using a small dataset can only provide information on how the model performs on that specific population or setting [138]. In order for the model to be considered generalisable to the whole patient population and thus clinically impactful, it needs to be externally validated using a large enough dataset that consists of a varied patient cohort, treatment periods and settings.

1.3.4.2 Multicentre prognostic studies in anal cancer

One of the most effective ways of accumulating enough data to carry out robust prognostic factor and modelling research for a rare cancer is by linking smaller institutional datasets together. Even though the number of published studies with large cohorts exploring factors affecting anal cancer outcomes after conformal radiotherapy is relatively small (n=23), more than half of these (n=13) were multicentre studies. Importantly, all 13 multicentre studies were published within the last 10 years, emphasising the direction that research in this area is heading towards.

The value of national and international collaboration has been demonstrated by numerous other large studies that analysed data from national and international anal cancer databases. As mentioned previously, a prospective national database consisting of 242 cases treated in 40 UK centres has been published [43]. Outcomes in the patient cohort treated with IMRT were reported, and were subsequently compared to the outcomes reported by the ACT 2 trial [27], which employed older, non-conformal radiotherapy techniques. The results indicated that IMRT confers superior outcomes, which helped lead to its widespread adoption throughout the UK. In France, 60 radiotherapy centres collaboratively developed the ANABASE multicentric cohort, which includes data from more than 1000 patients treated between 2015 and 2020 [139]. The final analysis, reported in abstract form only, demonstrated that treatment with IMRT (814 patients) leads to good outcomes for the majority of patients [140]. A complimentary analysis using the same cohort aimed to compare outcomes between HIV positive and HIV negative patients [139]. The study reported significantly poorer overall survival rates for HIV positive patients, but no significant differences in treatment toxicity. A similar study was conducted in Italy, which involved the analysis of a national cohort of 987 patients [141]. The study's results were in accordance with the UK and French studies regarding the efficacy of IMRT treatment. In addition, this study also confirmed the prognostic impact of lymph node involvement and histological grade on anal cancer outcomes.

Although the anal cancer scientific community calls for further collaboration, not only nationally, but also across international borders [142], there are still significant barriers that need to be addressed and overcome in order to form new collaborations.

1.3.4.3 Local barriers to using patient data for research

To begin with, in order to form a new collaboration, patient data need to be collected locally. This can be a challenge in itself, since gaining the necessary approvals required to access patient data is a lengthy process that may take months [143]. The researchers seeking to collect these data must be able to prove to data providers and governance bodies that their research will preserve patient privacy and that the results will be beneficial to patients and the public [144]. Appropriate governance is of utmost importance for patient confidentiality protection, however, approval processes for patient data collection require further streamlining [145]. Even when the necessary approvals to access and collect data are granted, physical access to the data can be inefficient [144]. In some cases, data can only be accessed in secure environments, which may involve travel outside of the researcher's usual environment. Furthermore, these secure environments may be limited in terms of computing capacity, software, and access hours.

1.3.4.4 Barriers to multicentre collaboration and possible solutions

The aforementioned barriers to using patient data for research can become even more prominent when the data need to be linked with data from external organisations and analysed in a multicentre setting [146].

An important barrier to multicentre collaboration is the lack of standardised procedures [143]. For instance, in many cases, the computer infrastructure used varies between different centres and across countries. The data generated throughout the patient treatment pathway may therefore be stored in different formats or using different scoring systems, making it very challenging to link datasets originating from multiple centres.

The quality of the data can also vary from centre to centre, rendering quality assurance processes incredibly complicated.

Importantly, ethical considerations and data protection regulations often limit sharing of data between centres and thus make data sharing across institutions and countries very challenging. Approval processes are usually time-consuming and bear unnecessarily high costs [143].

Distributed learning, a novel data analysis technique, offers a promising solution to some of the barriers discussed and may aid the foundation of future multicentre collaborations.

1.4 Distributed learning

1.4.1 What is distributed learning?

Distributed learning is an approach that can be applied to collaboratively develop robust models using local datasets originating from multiple centres, without having to exchange any sensitive patient data between centres [147]. Only non-identifiable, processed, and aggregated information in the form of mathematical parameters, such as model coefficients, is shared between centres in order to train and validate a distributed model. Consequently, distributed learning strategies preserve patient data privacy and minimise Information Governance issues related to sharing data with external organisations [148]. Notably, it has been proven by multiple studies that a number of distributed models are mathematically equivalent to models developed through the traditional, central learning approach, in which all data is shared between centres and collated into a single dataset [149,150]. This evidences that these distributed models and their centralised counterparts yield exactly the same results and exhibit the same performance. As a result, a wide range of research questions can be investigated within a multicentre setting through distributed learning. The differences between central learning and distributed learning model development are illustrated in Figure 1-4.

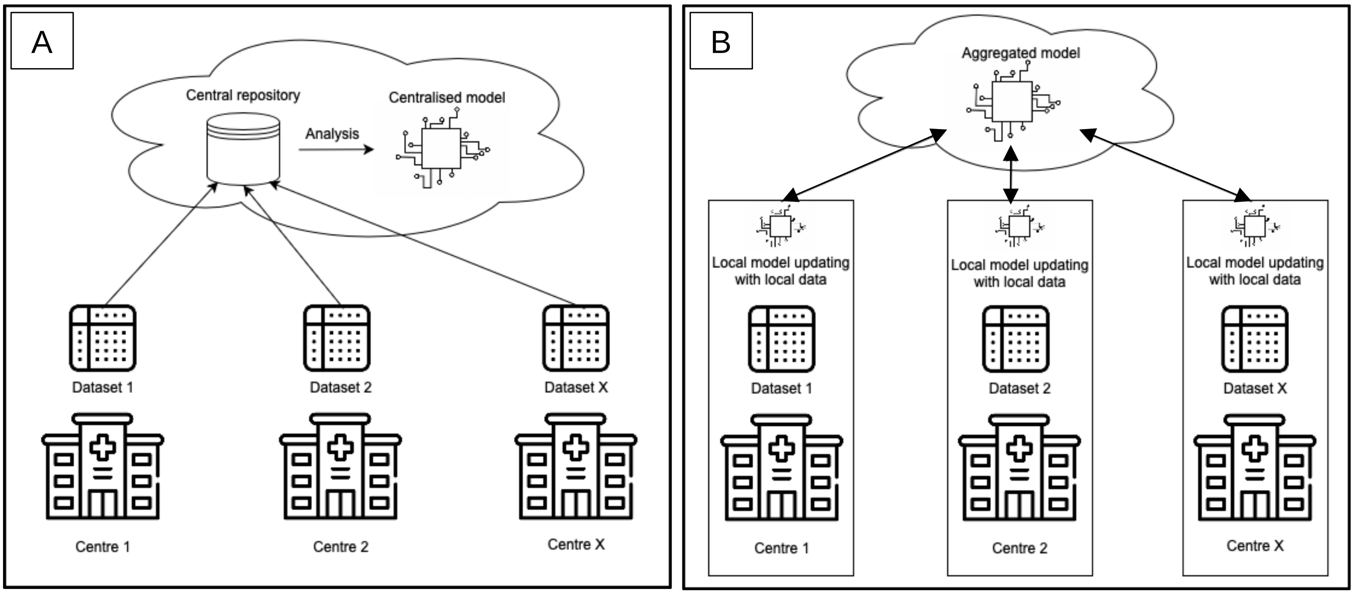


Figure 1-4. Centralised (A) vs distributed learning (B) approach to model development. In centralised model development (A), the datasets from each participating centre are combined to form a central repository, which is then analysed to create the centralised model. In distributed model development (B), the data from each participating centre never leave the originating centre. Instead, the model is sent to each centre, where it is updated locally. The updated local models are then sent back to a central location, where they are aggregated. This is an iterative approach, where the process is repeated until the aggregated model converges according to pre-specified convergence criteria, generating the final distributed model.

1.4.1.1 Distributed learning methods

There are three distinct distributed learning methods that can be differentiated according to various computational principles [147]. They all have in common the fact that a model is trained within a network that consists of multiple nodes. A node is set up at each individual participating centre and is linked to the local dataset. The three distributed learning methods, ensembling, split learning and federated learning, differ in three major ways, as demonstrated in Figure 1-5: (1) how the model parameters are transmitted within the network, (2) the way the nodes interact with each other within the network and (3) what type of data is exchanged between them.

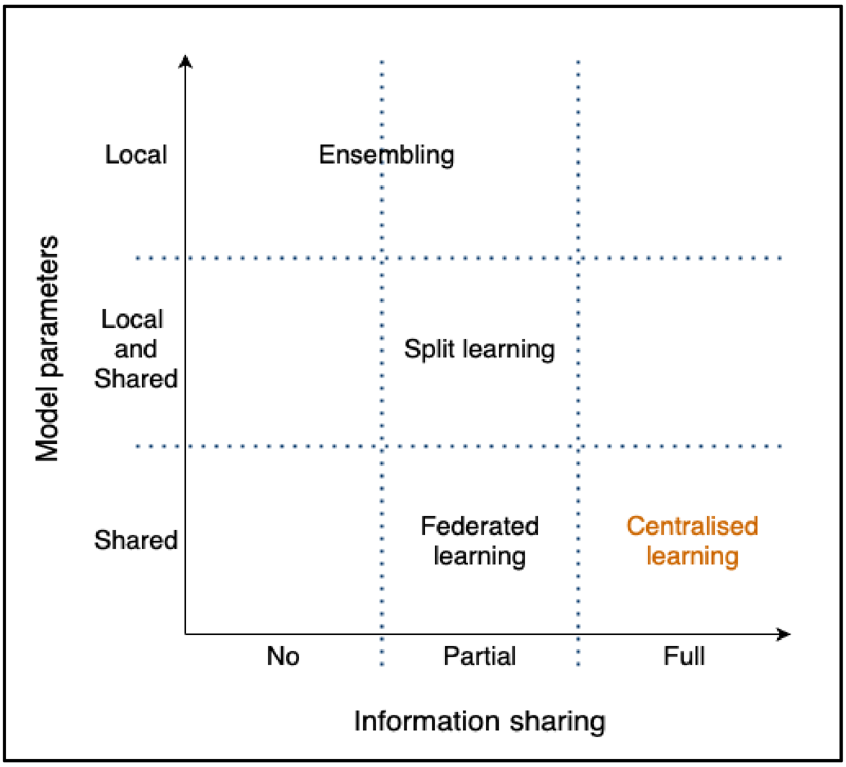


Figure 1-5. Diagram highlighting the differences between centralised learning and three types of distributed learning (ensembling, split learning and federated learning), according to the information and model parameters shared between nodes.

To begin with, the ensembling approach involves constructing multiple smaller or simpler models [151]. Through this approach, each node (or centre) in a multicentre collaboration trains its own local model using only its local dataset. When all models are trained, they can then be combined to yield the final results. There are various ways that the models can be combined. The simplest way is to obtain the output from all models and average it. A more sophisticated way is to combine all models together into a single meta-model, which uses the output from each simpler model as its input. Since each simpler model is trained independently, no individual level data needs to be shared between nodes. The only information being exchanged between nodes are aggregated statistics calculated during the model training phase at each node, or the resulting trained local models. One major limitation of this approach is that if the simpler models are trained using small datasets with inadequate number of samples, the resulting meta-model may not be robust.

The second distributed learning approach is split learning. Through this approach, only one model is constructed, but it is split into multiple sections and each section is trained by an individual node [152]. This method usually involves having a central node that receives the outputs from the other training nodes and aggregates them to make the final prediction. This is an iterative approach, where the central node sends back the error gradient to each of the training nodes, which then use it to update their outputs. The procedure is then repeated until the error from each of the training nodes is adequately small. Split learning only requires the exchange of extracted features from the training nodes and error gradients from the central node, and therefore preserves individual level data privacy.

The final distributed learning method is federated learning [153]. This approach involves a network of nodes collaboratively constructing a single model and a central server that handles the communication between the individual nodes. During the training phase, the global model is located at the central server, which sends it to each of the participating nodes. Subsequently, each node uses its local dataset to compute an update to the model and sends the updated model back to the central server, where updates from all nodes are aggregated and a new global model is computed. This is also an iterative approach that aims to minimise the prediction error. When the new global model is computed, the central server also calculates the error gradient. The model and error gradient are then sent back to all nodes, which use this information to further update the model. This process is repeated until the global model converges according to pre-specified convergence criteria, such as small error margins or robust model performance in terms of discrimination or predictive capability [149]. Only model updates, such as model coefficients and parameters, as well as error gradients are shared between the nodes and the central server. The nodes do not directly interact with each other, and there is no exchange of individual level data at any point during the model training phase.

This thesis focuses solely on federated learning methods. For simplicity, from this point onwards, the term distributed learning refers specifically to federated learning.

1.4.1.2 Privacy of distributed learning approaches

One of the main benefits conferred by distributed learning is privacy preservation, which can be separated into three aspects: data privacy, model privacy and model output privacy [148]. It is vital all three aspects of privacy are considered and preserved when employing distributed learning.

Preserving data privacy involves applying data de-identification procedures, and is essential when dealing with patient data [154]. Through these procedures, all available information that can be used to identify a certain individual, including names, dates of birth and addresses, is deleted. Any identifiable information that cannot be deleted because it needs to be analysed has to be converted into useable non-identifiable information. For instance, the date of death or date of cancer recurrence may need to be used for prognostic model development. However, this identifiable information first needs to be de-identified by converting the dates to number of days since baseline (e.g. date of diagnosis, or date of treatment start). Another way of preserving data privacy is to ensure that access to the data is restricted to authorised individuals only, and that no patient data leave the originating centre.

Ensuring model privacy and model output privacy in distributed learning can be challenging, but various cryptographic solutions and differentially private mechanisms have been developed to address this [155,156]. These aim to prevent leakage of individual level patient data during model training via distributed learning. Data leakage refers to the unauthorised passage of data from inside the originating organisation to a destination outside its secured network. Additionally, multiparty protocols are employed, ensuring that all computations and all communication between nodes and the central server are secure [149]. As this is an emerging and growing field, numerous recent studies have focused on developing new distributed learning approaches that aim to ensure model privacy [147,156–159].

In particular, the study by Brink et al. [159] has demonstrated that during the optimisation of a distributed Cox regression model, there is a risk of individual level patient data leakage. This can happen in cases where event times are unique (no ties) and there is no censoring. In such cases, individual level data from the patient that survives the longest need to be shared between centres during the distributed Cox regression model optimisation phase. As this is an iterative process, data from multiple patients can be reconstructed outside the originating centre. In cases where event times are not unique and there is censoring, the reconstruction of individual level patient data would still be possible but would be much more challenging. The authors have proposed an alternative approach that prevents the issue of data leakage. In this approach, the partial likelihood is calculated locally at each participating centre, and as a result, it is no longer necessary to share individual level data from the patient that survives the longest between centres. The limitation of this approach is that it depends on stratified

Cox regression using a local baseline hazard function that is unique to each centre. This means that this approach does not provide a global model which can be applied to an independent patient cohort afterwards. Moreover, a study by Huth et al. [160] has established that it is possible to reverse engineer individual-level data in some distributed learning systems. This is a significant limitation of a number of existing distributed learning frameworks. To address this limitation, possible defence strategies are being explored. The novel insights uncovered will enhance the privacy and security of distributed learning systems in the future.

1.4.1.3 Data considerations

In order to render distributed learning analysis feasible and reproducible, the data storage and pre-processing aspects need to be taken into consideration during the initial phases of the research [148]. Since centres may employ different workflows in terms of how the patient data are generated and stored in their local systems, it is vital that a detailed data model and data dictionary are agreed upon between all participating centres in a multicentre collaboration. The Findable, Accessible, Interoperable, Reusable (FAIR) Guiding Principles [161] have recently been developed with the aim of improving data management and accelerating the extraction of valuable knowledge from datasets. In order to make data originating from clinical systems FAIR, several tools need to be deployed [162] at each participating centre. Firstly, the clinical systems in which the relevant data can be sourced from need to be identified. Since multiple clinical system sources may be available at specific centres, it is important that the relevant data within a single centre are combined into a single dataset. Therefore, software needs to be developed that effectively extracts the data, combines, and transforms them to the required format, and subsequently loads it into a local data repository. For multicentre analyses, the next step is to standardise the data in all local repositories to ensure that they match. This can be achieved by using medical ontologies, such as the National Cancer Institute Thesaurus [163] and the Radiation Oncology Ontology [164]. These provide cross-mapped and controlled terminologies that match the data between the participating centres. As a result, clinical data from multiple disconnected local repositories are linked and transformed into FAIR data, supporting the detection of novel relationships [164].

1.4.2 Distributed learning applications in healthcare and oncology

The field of distributed learning in healthcare and oncology has been steadily growing over the last two decades, as indicated by the number of new publications in this field (Figure 1-6). A spike in publications during the COVID-19 pandemic (2020 and 2021) has been observed, further highlighting the utility of distributed learning as a way of overcoming data sharing barriers in research.

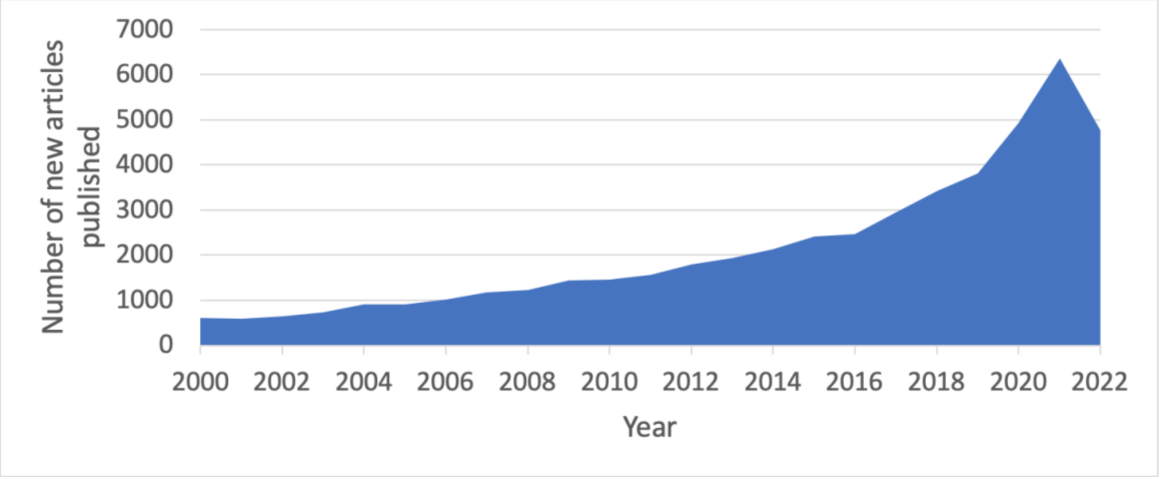


Figure 1-6. Number of new "distributed learning" or "federated learning" research articles published each year between 2000 and 2022. (Source: PubMed search [165])

A systematic review published in 2020 by Zerka et al. [148] identified 127 published research articles applying a distributed learning approach in the context of healthcare. The review reports a wide range of distributed learning applications in medicine. As an example, electronic health records were analysed using a distributed learning approach to study the relationship between the use of medication in pregnant women and foetal loss [166]. Another study, which also used data from electronic health records, aimed to predict hospitalisations due to heart disease [167]. Both of these studies demonstrated that applying distributed learning preserves patient data privacy whilst minimising bias and maintaining high statistical efficiency of the resulting models. Moreover, the use of distributed learning has been proven to be beneficial in the field of medical imaging [168] for a variety of applications, such as automated support for clinical diagnoses of retinopathy [169] and various highly heterogeneous psychiatric disorders [170]. Genome-wide association studies (GWAS) may also benefit from applying a distributed learning methodology to analyse larger number of samples through multicentre collaborations [171]. As a result, these studies may uncover key insights on the relationship between genetic variants and certain diseases.

Distributed learning approaches also have the potential to support the stratification and personalisation of cancer therapy through robust multicentre prognostic research [172]. The training and validation of prognostic models using a distributed learning framework has proven to be feasible in multiple instances. Jochems et al. developed a distributed Bayesian network model for non-small cell lung cancer to predict survival at two years, using data from two centres, across two countries [173]. Upon validation of this distributed model on an independent dataset from a third centre, it was concluded that its performance is comparable to the performance of a centralised model. Another study from the same research group constructed a distributed Bayesian network to predict dyspnoea after radiotherapy for lung cancer [174]. The model was trained and validated using data from five centres across three countries and also exhibited similar performance to its centralised counterpart. Both of these studies demonstrate the feasibility of distributed learning for the development of prognostic models in oncology research. More specifically, radiation oncology research could be advanced by the adoption of the distributed learning methodology. It has been established that radiomics data from CT scans could be incorporated for the training of distributed prognostic models [175,176]. The EuroCAT IT infrastructure has been developed by Deist et al. in order to facilitate the adoption of distributed learning methodologies for radiation oncology by more centres [149]. Currently, there are various available open-source and commercial platforms, including DistriM [177], the Varian Learning Portal [178], Clara [179] and GRIN [180] that can be used to support the execution of distributed learning projects.

The Vantage6 (priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange) platform [181] has been used to carry out the distributed learning analysis described in this thesis (Chapters 4 and 6). The Vantage6 architecture consists of multiple components, including a researcher, a server and one or multiple nodes. In summary, the researcher can pose a question that may be answered using data that are available at different centres. Each centre has its own separate node set up, which can only access the local data. The researcher sends the question to the server as a task by calling a function. This can be done using a range of programming languages, including R or Python. The server then processes the task and handles administrative functions, such as the authorisation and authentication of the nodes. The task is subsequently delivered to each of the participating nodes as a Docker image, where it is executed. When the task is complete and once a solution has been reached, the results are transmitted back to the researcher via the server. Through this approach, the data never leave the centre they originate from.

With the recent increase in number of centres adopting a distributed learning approach, more robust and generalisable distributed prognostic models will be developed in the near future. The ultimate aim of distributed learning in oncology would be to develop models that can be used as decision support tools that are fully integrated within the clinic, allowing information to be shared between centres across the globe in a standardised and dynamic fashion, therefore enabling truly personalised cancer therapy [175].

1.4.3 Distributed Cox proportional hazards algorithm

As previously discussed, traditional centralised Cox modelling (Section 1.3.2.3) involves collecting all available patient data in a single repository. However, Lu et al. [150] has adapted the Cox regression algorithm for use in a distributed learning setting. The study authors have also demonstrated that the distributed Cox regression model generates the same model outputs as a traditional centralised Cox regression model trained with the same data. In practice, a distributed and a centralised model produced near-identical model coefficients, with differences in the range of 10^{-15} to 10^{-12}. This paper also proved that the distributed and centralised models are mathematically equivalent under the Breslow likelihood assumption.

To construct a distributed model using data from *M* participating centres, the first and second order derivatives $l'(\beta)$ and $l''(\beta)$ of the partial likelihood function (Equation 1-2) need to be calculated by the distributed Cox regression algorithm:

Equation 1-5

$$l'_r(\boldsymbol{\beta}) = \sum_{k=1}^{M} \sum_{i=1}^{D} \sum_{l \in \mathscr{D}_i^k} z_r^l - \sum_{i=1}^{D} \left(\sum_{k=1}^{M} |\mathscr{D}_i^k| \right) \frac{\sum_{k=1}^{M} \sum_{l \in \mathscr{R}_i^k} z_r^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{k=1}^{M} \sum_{l \in \mathscr{R}_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}.$$

and

Equation 1-6

$$l''_{r,q}(\boldsymbol{\beta}) = -\sum_{i=1}^{D}\left(\sum_{k=1}^{M}|\mathcal{D}_i^k|\right)\left\{\frac{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k} z_r^l z_q^l \exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)}{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k}\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)} - \frac{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k} z_r^l \exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)}{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k}\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)}\,\frac{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k} z_q^l \exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)}{\sum_{k=1}^{M}\sum_{l\in\mathcal{R}_i^k}\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)}\right\},$$

where $\mathcal{D}^k{}_i$ and $\mathcal{R}^k{}_i$ are subsets of $\mathcal{D}_i$ and $\mathcal{R}_i$ signifying individuals from the k-th centre, and k = 1, 2, …, M. In Equation 1-6, the count d_i is replaced by $d_i = \sum^M{}_{k=1}|\mathcal{D}^k{}_i|$, so that it can be aggregated from multiple distributed centres. According to Equations 1-5 and 1-6, the derivatives of the log likelihood function are naturally decomposed through the calculation and the subsequent sharing of locally aggregated values from each centre. Through this decomposition, the sum of derivatives learned from the distributed centres is guaranteed to be exactly the same as the derivative calculated from the central repository that is analysed to develop the centralized Cox model. The following steps are executed to update the distributed Cox regression model using data from multiple distributed centres:

1. The first step involves the local initialisation for all centres. Based on the local data, each centre initialises index subsets $\mathcal{R}^k{}_i$ and $\mathcal{D}^k{}_i$. The aggregated statistic $\sum^D{}_{i=1}\sum_{l\in\mathcal{D}^k{}_i} z^l{}_r$ from each centre is then sent to the global server. Since this value remains unchanged during the entire learning process, sharing this value with the global server during the initialisation phase avoids additional communication overhead.
2. The next step involves global initialisation. The distinct event times from each centre are sent to the global server, in order to initialize the parameters D and $|\mathcal{D}^k{}_i|$. Subsequently, based on Equation 1-5, the global server aggregates the incoming statistics from all centres $z\hat{}_r=\sum^M{}_{k=1}\sum^D{}_{i=1}\sum_{l\in}\mathcal{D}^k{}_i z^l{}_r$. The server then initializes $\boldsymbol{\beta}^0$ and distributes it to each centre.
3. The third step involves a parallel update that is carried out in all centres. Each centre receives the updated $\boldsymbol{\beta}^\tau$ from the global server. Using this, the following aggregated statistics are calculated: $\sum_{l\in\mathcal{R}_i^k}\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)$, $\sum_{l\in\mathcal{R}_i^k} z_r^l\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)$ and $\sum_{l\in\mathcal{R}_i^k} z_r^l z_q^l\exp\left(\boldsymbol{\beta}^T\mathbf{z}^l\right)$. These statistics are then sent back to the global server.

4. Using the statistics received from each centre, the global server calculates the first and second derivatives of the likelihood function, according to Equations 1-5 and 1-6.
5. The statistic $\boldsymbol{\beta}^{\tau+}1$ is then updated using the Newton-Raphson algorithm (Equation 1-3), and the updated $\boldsymbol{\beta}^{\tau+1}$ is sent back to each site.
6. Steps 3 to 5 are repeated until the parameters converge.
7. In the final step, the converged model parameters to are sent to each centre.

This algorithm was adapted by a group of researchers at the Netherlands Comprehensive Cancer Organisation (IKNL) for use in the Vantage6 distributed learning infrastructure [182], and it has been used for the training of distributed outcome models in the atomCAT1 and atomCAT2 studies, which form part of this thesis (Chapters 4 and 6).

1.4.4 Distributed learning for anal cancer outcome modelling

A distributed learning approach may be ideally suited for outcome prediction modelling in rare cancers such as anal cancer. It could help in the acquisition of sufficient patient data from multiple different centres with the aim of developing robust generalisable models, while working around many of the barriers associated with physical data sharing. By using this novel technology to promote collaboration and link hospitals nationally and across the world, a diverse cohort of patients, which is representative of the overall patient population, may be analysed. Therefore, this ensures that we learn from every patient treated for anal cancer; not only from patients who are eligible to participate in RCTs.

1.5 Research overview

1.5.1 Aims and objectives

The primary aim of this research is to demonstrate the feasibility of conducting distributed learning across multiple institutions for anal cancer outcome modelling. The work carried out involves the evaluation of data availability, the creation of a local anal cancer patient data warehouse, the deployment of the distributed learning technical infrastructure, and finally development of distributed outcome models of increasing complexity, incorporating radiotherapy-specific data. The specific objectives of this research are to:

1. Establish a local data warehouse that consists of comprehensive data from patients treated for anal cancer in LCC.
2. Conduct a systematic review to identify prognostic factors for disease-related outcomes in anal cancer reported in the literature.
3. Carry out a proof-of-concept study in collaboration with two other European centres to demonstrate the feasibility of distributed learning in outcome modelling for rare cancers.
4. Develop a prospective study protocol and statistical analysis plan, detailing the plans for extension of the distributed learning outcome modelling work.
5. Establish a wider consortium that consists of anal cancer radiotherapy treatment centres across the world. Securely link international databases in individual centres to analyse and learn from complex individual-level patient data through distributed learning.

1.5.2 Chapter overview

Chapter 2: Development of a comprehensive institutional anal cancer data warehouse for real-world data analysis

The aim of this chapter is to describe the development of a comprehensive data warehouse consisting of data from patients treated for anal cancer at LCC since January 2013. The data warehouse is now available to clinicians and researchers on an ongoing basis. In this chapter, the data collection and quality evaluation procedures are documented, highlighting the proportion of data items that could be automatically extracted, and the amount of manual entry and manual review required to develop an institutional data warehouse. The work outlined in this chapter feeds into the subsequent chapters and forms a key development for the application of this research in clinical practice. The manuscript detailing this work is being prepared for publication.

Chapter 3: Prognostic factors for patients with anal cancer treated with conformal radiotherapy – a systematic review

In this chapter, the existing literature is evaluated by systematic review in order to identify established prognostic factors for a variety of disease-related outcomes in anal cancer, focusing on patients treated with curative intent using modern conformal radiotherapy techniques. The prognostic factors identified will be considered as potential predictor variables for the distributed learning models developed in subsequent chapters (Chapters 4 and 6). This work was published in the journal *BMC Cancer* [183].

Chapter 4: Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study

The aim of the work described in this chapter is to demonstrate the feasibility of distributed learning for outcome prediction modelling in a rare cancer. To achieve this, an overall survival model for anal cancer was developed through distributed learning. The model was collaboratively developed using data from patients treated at LCC (Leeds, UK), MAASTRO clinic (Maastricht, the Netherlands) and Oslo University Hospital (Oslo, Norway), without the exchange of any individual-level patient data between the three centres. This work supported the growth of the collaboration into a larger international consortium. This proof-of-concept study was published in the journal *Radiotherapy and Oncology* [184].

Chapter 5: Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study

The aim of this chapter is to outline the research proposal for the international multi-centre atomCAT2 project and the procedures to be followed during its course. The prospective study protocol includes details on the atomCAT2 study design, the patient population to be analysed, and the outcomes to be investigated. The protocol also pre-specifies all models to be developed as part of the analysis, including which prognostic factors are to be analysed in each model. Lastly, a prospective sample size calculation and a detailed statistical analysis plan are provided. The atomCAT2 study protocol was published in the journal *Diagnostic and Prognostic Research* [185].

Chapter 6: Prognostic models for anal cancer using distributed learning: the international multi-centre atomCAT2 study

In this chapter, the aim is to describe the development and validation of prediction models for anal cancer outcomes after chemoradiotherapy through distributed learning.

To achieve this, a consortium of 14 international cancer treatment centres based in the UK and Europe has been formed. A cohort of more than 1099 patients treated across 12 centres was analysed, in order to develop and validate models for overall survival, locoregional control and distant metastasis, as well as to identify key prognostic factors and their effect size. This has provided unique insights and may guide the design of future anal cancer clinical trials. The manuscript for the atomCAT2 study is currently being prepared for publication.

Chapter 7: Discussion

This chapter aims to bring together and synthesise the research described in the previous chapters as a whole. The main findings from the research conducted are discussed in relation to the literature. The limitations of this research are also discussed in more detail, and potential directions for future work are indicated.

1.6 References

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Pineros M, et al. Global Cancer Observatory: Cancer Today. France: International Agency for Research on Cancer. 2018. https://gco.iarc.fr/today (accessed September 29, 2020).

[2] Cancer Research UK. Anal cancer incidence by sex and UK country 2020. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/anal-cancer/incidence#heading-Zero (accessed September 29, 2020).

[3] Shridhar R, Shibata D, Chan E, Thomas CR. Anal cancer: current standards in care and recent changes in practice. CA Cancer J Clin 2015;65:139–62. https://doi.org/10.3322/caac.21259.

[4] Gami B, Kubba F, Ziprin P. Human Papilloma Virus and Squamous Cell Carcinoma of the Anus. Clin Med Insights Oncol 2014;8:CMO.S13241. https://doi.org/10.4137/CMO.S13241.

[5] Lin C, Franceschi S, Clifford GM. Human papillomavirus types from infection to cancer in the anus, according to sex and HIV status: a systematic review and meta-analysis. Lancet Infect Dis 2018;18:198–206. https://doi.org/10.1016/S1473-3099(17)30653-9.

[6] Liu C, Mann D, Sinha UK, Kokot NC. The molecular mechanisms of increased radiosensitivity of HPV-positive oropharyngeal squamous cell carcinoma (OPSCC): an extensive review. J Otolaryngol - Head Neck Surg J Oto-Rhino-Laryngol Chir Cervico-Faciale 2018;47:59. https://doi.org/10.1186/s40463-018-0302-y.

[7] Salati SA. Anal Cancer : A Review. Int J Health Sci 2012;6:206–30. https://doi.org/10.12816/0006000.

[8] Cancer Research UK. Diagram showing the anatomy of the anus CRUK 2014. https://commons.wikimedia.org/wiki/File:Diagram_showing_the_anatomy_of_the_anus_CRUK_282.svg (accessed November 8, 2022).

[9] Wietfeldt E, Thiele J. Malignancies of the Anal Margin and Perianal Skin. Clin Colon Rectal Surg 2009;22:127–35. https://doi.org/10.1055/s-0029-1223845.

[10] Sauter M, Keilholz G, Kranzbühler H, Lombriser N, Prakash M, Vavricka SR, et al. Presenting symptoms predict local staging of anal cancer: a retrospective analysis of 86 patients. BMC Gastroenterol 2016;16:46. https://doi.org/10.1186/s12876-016-0461-0.

[11] Siegel R, Werner RN, Koswig S, Gaskins M, Rödel C, Aigner F, et al. Clinical Practice Guideline: Anal Cancer—Diagnosis, Treatment and Follow-up. Dtsch Arzteblatt Int 2021;118:217–24. https://doi.org/10.3238/arztebl.m2021.0027.

[12] Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. Radiother Oncol J Eur Soc Ther Radiol Oncol 2014;111:330–9. https://doi.org/10.1016/j.radonc.2014.04.013.

[13] Dinneen A. The Colorectal Service Multidisciplinary Team (MDT) - Cancer Services Information for Patients. East North Herts NHS Trust 2015. https://clinical-pathways.org.uk/sites/default/files/leaflet/colorectal-cancer-mdt-updated-112016.pdf (accessed November 8, 2022).

[14] University College London Hospitals NHS Foundation Trust. Colorectal and anal cancer multidisciplinary team (MDT) - Information for patients, relatives and carers. Univ Coll Lond Hosp NHS Found Trust 2021. https://www.uclh.nhs.uk/patients-and-visitors/patient-information-pages/colorectal-and-anal-cancer-multidisciplinary-team-mdt (accessed November 8, 2022).

[15] Amin MB, American Joint Committee on Cancer, American Cancer Society, editors. AJCC cancer staging manual. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP ; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR-Technical editor ; Laura R. Meyer, CAPM-Managing editor. Chicago IL: American Joint Committee on Cancer, Springer; 2017.

[16] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. Ann Surg Oncol 2010;17:1471–4. https://doi.org/10.1245/s10434-010-0985-4.

[17] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging: The Eighth Edition AJCC Cancer Staging Manual. CA Cancer J Clin 2017;67:93–9. https://doi.org/10.3322/caac.21388.

[18] Nigro ND, Vaitkevicius VK, Considine B. Combined therapy for cancer of the anal canal: A preliminary report. Dis Colon Rectum 1974;17:354–6. https://doi.org/10.1007/BF02586980.

[19] Nigro ND, Vaitkevicius VK, Buroker T, Bradley GT, Considine B. Combined therapy for cancer of the anal canal: Dis Colon Rectum 1981;24:73–5. https://doi.org/10.1007/BF02604287.

[20] UKCCCR Anal Cancer Trial Working Party. Epidermoid anal cancer: results from the UKCCCR randomised trial of radiotherapy alone versus radiotherapy, 5-fluorouracil, and mitomycin. The Lancet 1996;348:1049–54. https://doi.org/10.1016/S0140-6736(96)03409-5.

[21] Flam M, John M, Pajak TF, Petrelli N, Myerson R, Doggett S, et al. Role of mitomycin in combination with fluorouracil and radiotherapy, and of salvage chemoradiation in the definitive nonsurgical treatment of epidermoid carcinoma of the anal canal: results of a phase III randomized intergroup study. J Clin Oncol 1996;14:2527–39. https://doi.org/10.1200/JCO.1996.14.9.2527.

[22] Bartelink H, Roelofsen F, Eschwege F, Rougier P, Bosset JF, Gonzalez DG, et al. Concomitant radiotherapy and chemotherapy is superior to radiotherapy alone in the treatment of locally advanced anal cancer: results of a phase III randomized trial of the European Organization for Research and Treatment of Cancer Radiotherapy and Gastrointestinal Cooperative Groups. J Clin Oncol 1997;15:2040–9. https://doi.org/10.1200/JCO.1997.15.5.2040.

[23] Perry W, Connaughton J. Abdominoperineal Resection: How Is It Done and What Are the Results? Clin Colon Rectal Surg 2007;20:213–20. https://doi.org/10.1055/s-2007-984865.

[24] Ajani JA, Winter KA, Gunderson LL, Pedersen J, Benson AB, Thomas C, et al. Intergroup RTOG 98–11: A phase III randomized study of 5-fluorouracil (5-FU), mitomycin, and radiotherapy versus 5-fluorouracil, cisplatin and radiotherapy in carcinoma of the anal canal. J Clin Oncol 2006;24:4009–4009. https://doi.org/10.1200/jco.2006.24.18_suppl.4009.

[25] Glynne-Jones R, Meadows H, Wan S, Gollins S, Leslie M, Levine E, et al. EXTRA—A Multicenter Phase II Study of Chemoradiation Using a 5 Day per Week Oral Regimen of Capecitabine and Intravenous Mitomycin C in Anal Cancer. Int J Radiat Oncol 2008;72:119–26. https://doi.org/10.1016/j.ijrobp.2007.12.012.

[26] Peiffert D, Tournier-Rangeard L, Gérard J-P, Lemanski C, François E, Giovannini M, et al. Induction Chemotherapy and Dose Intensification of the Radiation Boost in Locally Advanced Anal Canal Carcinoma: Final Analysis of the Randomized UNICANCER ACCORD 03 Trial. J Clin Oncol 2012;30:1941–8. https://doi.org/10.1200/JCO.2011.35.4837.

[27] James RD, Glynne-Jones R, Meadows HM, Cunningham D, Myint AS, Saunders MP, et al. Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II): a randomised, phase 3, open-label, 2×2 factorial trial. Lancet Oncol 2013;14:516–24. https://doi.org/10.1016/S1470-2045(13)70086-X.

[28] Deutsch E, Lemanski C, Pignon JP, Levy A, Delarochefordiere A, Martel-Lafay I, et al. Unexpected toxicity of cetuximab combined with conventional chemoradiotherapy in patients with locally advanced anal cancer: results of the UNICANCER ACCORD 16 phase II trial. Ann Oncol Off J Eur Soc Med Oncol 2013;24:2834–8. https://doi.org/10.1093/annonc/mdt368.

[29] Kachnic LA, Winter K, Myerson RJ, Goodyear MD, Willins J, Esthappan J, et al. RTOG 0529: A Phase 2 Evaluation of Dose-Painted Intensity Modulated Radiation Therapy in Combination With 5-Fluorouracil and Mitomycin-C for the Reduction of Acute Morbidity in Carcinoma of the Anal Canal. Int J Radiat Oncol 2013;86:27–33. https://doi.org/10.1016/j.ijrobp.2012.09.023.

[30] Rao S, Guren MG, Khan K, Brown G, Renehan AG, Steigen SE, et al. Anal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up☆. Ann Oncol 2021;32:1087–100. https://doi.org/10.1016/j.annonc.2021.06.015.

[31] Burnet NG. Defining the tumour and target volumes for radiotherapy. Cancer Imaging 2004;4:153–61. https://doi.org/10.1102/1470-7330.2004.0054.

[32] Muirhead R, Adams RA, Gilbert DC, Harrison M, Glynne-Jones R, Sebag-Montefiore D, et al. National guidance for IMRT in anal cancer 2016. http://analimrtguidance.co.uk/national-anal-imrt-guidance-v3.pdf (accessed January 31, 2022).

[33] Rusten E, Rekstad BL, Undseth C, Al-Haidari G, Hanekamp B, Hernes E, et al. Target volume delineation of anal cancer based on magnetic resonance imaging or positron emission tomography. Radiat Oncol 2017;12:147. https://doi.org/10.1186/s13014-017-0883-z.

[34] Glynne-Jones R, Tan D, Hughes R, Hoskin P. Squamous-cell carcinoma of the anus: progress in radiotherapy treatment. Nat Rev Clin Oncol 2016;13:447–59. https://doi.org/10.1038/nrclinonc.2015.218.

[35] Murray L, Lilley J. Radiotherapy: technical aspects. Radiotherapy 2020;48:p79-83. https://doi.org/10.1016/j.mpmed.2019.11.003.

[36] Martin D, Balermpas P, Winkelmann R, Rödel F, Rödel C, Fokas E. Anal squamous cell carcinoma - State of the art management and future perspectives. Cancer Treat Rev 2018;65:11–21. https://doi.org/10.1016/j.ctrv.2018.02.001.

[37] Franco P, Arcadipane F, Ragona R, Mistrangelo M, Cassoni P, Munoz F, et al. Volumetric modulated arc therapy (VMAT) in the combined modality treatment of anal cancer patients. Br J Radiol 2016;89:20150832. https://doi.org/10.1259/bjr.20150832.

[38] Possiel J, Ammon HE, Guhlich M, Conradi L-C, Ghadimi M, Wolff HA, et al. Volumetric Modulated Arc Therapy Improves Outcomes in Definitive Radiochemotherapy for Anal Cancer Whilst Reducing Acute Toxicities and Increasing Treatment Compliance. Cancers 2021;13:2533. https://doi.org/10.3390/cancers13112533.

[39] Chargari C, Deutsch E, Blanchard P, Gouy S, Martelli H, Guérin F, et al. Brachytherapy: An overview for clinicians. CA Cancer J Clin 2019;69:386–401. https://doi.org/10.3322/caac.21578.

[40] Frakulli R, Buwenge M, Cammelli S, Macchia G, Farina E, Arcelli A, et al. Brachytherapy boost after chemoradiation in anal cancer: a systematic review. J Contemp Brachytherapy 2018;10:246–53. https://doi.org/10.5114/jcb.2018.76884.

[41] Falk AT, Claren A, Benezery K, François E, Gautier M, Gerard J-P, et al. Interstitial high-dose rate brachytherapy as boost for anal canal cancer. Radiat Oncol 2014;9:240. https://doi.org/10.1186/s13014-014-0240-4.

[42] Ali ZS, Solomon E, Mann P, Wong S, Chan KKW, Taggar AS. High dose rate brachytherapy in the management of anal cancer: A review. Radiother Oncol 2022;171:43–52. https://doi.org/10.1016/j.radonc.2022.03.019.

[43] Muirhead R, Drinkwater K, O'Cathail SM, Adams R, Glynne-Jones R, Harrison M, et al. Initial Results from the Royal College of Radiologists' UK National Audit of Anal Cancer Radiotherapy 2015. Clin Oncol 2017;29:188–97. https://doi.org/10.1016/j.clon.2016.10.005.

[44] The American Cancer Society medical and editorial content team. Chemotherapy for Anal Cancer 2017. https://www.cancer.org/cancer/anal-cancer/treating/chemotherapy.html (accessed November 8, 2022).

[45] Muirhead R, Partridge M, Hawkins MA. A tumor control probability model for anal squamous cell carcinoma. Radiother Oncol 2015;116:192–6. https://doi.org/10.1016/j.radonc.2015.07.014.

[46] Johnsson A, Leon O, Gunnlaugsson A, Nilsson P, Höglund P. Determinants for local tumour control probability after radiotherapy of anal cancer. Radiother Oncol 2018;128:380–6. https://doi.org/10.1016/j.radonc.2018.06.007.

[47] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[48] Sebag-Montefiore D, Adams R, Bell S, Berkman L, Gilbert DC, Glynne-Jones R, et al. The Development of an Umbrella Trial (PLATO) to Address Radiation Therapy Dose Questions in the Locoregional Management of Squamous Cell Carcinoma of the Anus. Int J Radiat Oncol 2016;96:E164–5. https://doi.org/10.1016/j.ijrobp.2016.06.1006.

[49] Tian X, Liu K, Hou Y, Cheng J, Zhang J. The evolution of proton beam therapy: Current and future status (Review). Mol Clin Oncol 2017. https://doi.org/10.3892/mco.2017.1499.

[50] Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat Rev Immunol 2020;20:651–68. https://doi.org/10.1038/s41577-020-0306-5.

[51] Brock KK. Adaptive Radiotherapy: Moving Into the Future. Semin Radiat Oncol 2019;29:181–4. https://doi.org/10.1016/j.semradonc.2019.02.011.

[52] Morgan HE, Sher DJ. Adaptive radiotherapy for head and neck cancer. Cancers Head Neck 2020;5:1. https://doi.org/10.1186/s41199-019-0046-z.

[53] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[54] Nilsson MP, Nilsson ED, Johnsson A, Leon O, Gunnlaugsson A, Scherman J. Patterns of recurrence in anal cancer: a detailed analysis. Radiat Oncol 2020;15:125. https://doi.org/10.1186/s13014-020-01567-7.

[55] Guerra GR, Kong JC, Bernardi M-P, Ramsay RG, Phillips WA, Warrier SK, et al. Salvage Surgery for Locoregional Failure in Anal Squamous Cell Carcinoma. Dis Colon Rectum 2018;61:179–86. https://doi.org/10.1097/DCR.0000000000001010.

[56] Bogach J, Fenech D, Chu W, Ashamalla S, Ung Y, Taggar AS, et al. Salvage surgery for locally recurrent anal cancer after intensity modulated radiation therapy with concurrent chemotherapy. Cancer Treat Res Commun 2021;26:100287. https://doi.org/10.1016/j.ctarc.2020.100287.

[57] Sclafani F, Adams RA, Eng C, Benson AB, Glynne-Jones R, Sebag-Montefiore D, et al. InterAACT: An international multicenter open label randomized phase II advanced anal cancer trial comparing cisplatin (CDDP) plus 5-fluorouracil (5-FU) versus carboplatin (CBDCA) plus weekly paclitaxel (PTX) in patients with inoperable locally recurrent (ILR) or metastatic disease. J Clin Oncol 2015;33:TPS792–TPS792. https://doi.org/10.1200/jco.2015.33.3_suppl.tps792.

[58] Rao S, Sclafani F, Eng C, Adams RA, Guren MG, Sebag-Montefiore D, et al. International Rare Cancers Initiative Multicenter Randomized Phase II Trial of Cisplatin

and Fluorouracil Versus Carboplatin and Paclitaxel in Advanced Anal Cancer: InterAAct. J Clin Oncol 2020;38:2510–8. https://doi.org/10.1200/JCO.19.03266.

[59] Yordanov K, Cima S, Richetti A, Pesce G, Martucci F, Azinwi NC, et al. Concurrent chemoradiation with volumetric modulated Arc therapy of patients treated for anal cancer—acute toxicity and treatment outcome. J Gastrointest Oncol 2017;8:361–7. https://doi.org/10.21037/jgo.2017.03.09.

[60] Jhaveri J, Rayfield L, Liu Y, Chowdhary M, Tian S, Cassidy RJ, et al. Impact of intensity modulated radiation therapy on survival in anal cancer. J Gastrointest Oncol 2018;9:618–30. https://doi.org/10.21037/jgo.2018.05.07.

[61] Pepek JM, Willett CG, Wu QJ, Yoo S, Clough RW, Czito BG. Intensity-modulated radiation therapy for anal malignancies: a preliminary toxicity and disease outcomes analysis. Int J Radiat Oncol Biol Phys 2010;78:1413–9. https://doi.org/10.1016/j.ijrobp.2009.09.046.

[62] Koerber SA, Slynko A, Haefner MF, Krug D, Schoneweg C, Kessel K, et al. Efficacy and toxicity of chemoradiation in patients with anal cancer--a retrospective analysis. Radiat Oncol Lond Engl 2014;9:113. https://doi.org/10.1186/1748-717X-9-113.

[63] Bentzen AG, Guren MG, Vonen B, Wanderås EH, Frykholm G, Wilsgaard T, et al. Faecal incontinence after chemoradiotherapy in anal cancer survivors: Long-term results of a national cohort. Radiother Oncol 2013;108:55–60. https://doi.org/10.1016/j.radonc.2013.05.037.

[64] Pan YB, Maeda Y, Wilson A, Glynne-Jones R, Vaizey CJ. Late gastrointestinal toxicity after radiotherapy for anal cancer: a systematic literature review. Acta Oncol 2018;57:1427–37. https://doi.org/10.1080/0284186X.2018.1503713.

[65] Philpot LM, Barnes SA, Brown RM, Austin JA, James CS, Stanford RH, et al. Barriers and Benefits to the Use of Patient-Reported Outcome Measures in Routine Clinical Care: A Qualitative Study. Am J Med Qual 2018;33:359–64. https://doi.org/10.1177/1062860617745986.

[66] Weldring T, Smith SMS. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). Health Serv Insights 2013;6:HSI.S11093. https://doi.org/10.4137/HSI.S11093.

[67] Gilbert A, Francischetto EO, Blazeby J, Holch P, Davidson S, Sebag-Montefiore D, et al. Choice of a patient-reported outcome measure for patients with anal cancer for use in cancer clinical trials and routine clinical practice: a mixed methods approach. Lancet Lond Engl 2015;385 Suppl 1:S38. https://doi.org/10.1016/S0140-6736(15)60353-1.

[68] Sodergren SC, Johnson CD, Gilbert A, Darlington A-S, Cocks K, Guren MG, et al. International validation of the EORTC QLQ-ANL27, a field study to test the anal cancer-specific health-related quality of life questionnaire. Int J Radiat Oncol Biol Phys 2022;S0360301622035076. https://doi.org/10.1016/j.ijrobp.2022.11.002.

[69] Burbach JPM, Kurk SA, Coebergh van den Braak RRJ, Dik VK, May AM, Meijer GA, et al. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. Acta Oncol 2016;55:1273–80. https://doi.org/10.1080/0284186X.2016.1189094.

[70] Young RC. Cancer Clinical Trials — A Chronic but Curable Crisis. N Engl J Med 2010;363:306–9. https://doi.org/10.1056/NEJMp1005843.

[71] Bennette CS, Ramsey SD, McDermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting Low Accrual in the National Cancer Institute's Cooperative Group Clinical Trials. J Natl Cancer Inst 2016;108:djv324. https://doi.org/10.1093/jnci/djv324.

[72] Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review. JAMA 2007;297:1233. https://doi.org/10.1001/jama.297.11.1233.

[73] Bellera C, Praud D, Petit-Monéger A, McKelvie-Sebileau P, Soubeyran P, Mathoulin-Pélissier S. Barriers to inclusion of older adults in randomised controlled clinical trials on Non-Hodgkin's lymphoma: A systematic review. Cancer Treat Rev 2013;39:812–7. https://doi.org/10.1016/j.ctrv.2013.01.007.

[74] Canouï-Poitrine F, Lièvre A, Dayde F, Lopez-Trabada-Ataz D, Baumgaertner I, Dubreuil O, et al. Inclusion of Older Patients with Cancer in Clinical Trials: The SAGE Prospective Multicenter Cohort Survey. The Oncologist 2019;24:e1351–9. https://doi.org/10.1634/theoncologist.2019-0166.

[75] Cancer Research UK. Cancer incidence by age 2021. https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#ref- (accessed November 8, 2022).

[76] Charnley N, Choudhury A, Chesser P, Cooper RA, Sebag-Montefiore D. Effective treatment of anal cancer in the elderly with low-dose chemoradiotherapy. Br J Cancer 2005;92:1221–5. https://doi.org/10.1038/sj.bjc.6602486.

[77] NHS England. Achieving world-class cancer outcomes: a strategy for England 2015 – 2020 2017. https://www.england.nhs.uk/publication/achieving-world-class-cancer-outcomes-a-strategy-for-england-2015-2020/ (accessed November 8, 2022).

[78] Spencer K, Morris E. Collection of routine cancer data from private health-care providers. Lancet Oncol 2019;20:1202–4. https://doi.org/10.1016/S1470-2045(19)30545-5.

[79] Sajjadnia Z, Khayami R, Moosavi MR. Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services. Cancer Inform 2020;19:1176935120917955. https://doi.org/10.1177/1176935120917955.

[80] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91. https://doi.org/10.1016/j.jclinepi.2006.01.014.

[81] Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2020;53:1487–509. https://doi.org/10.1007/s10462-019-09709-4.

[82] The potential and limitations of data from population-based state cancer registries. Am J Public Health 2000;90:695–8. https://doi.org/10.2105/AJPH.90.5.695.

[83] Parkin DM. The role of cancer registries in cancer control. Int J Clin Oncol 2008;13:102–11. https://doi.org/10.1007/s10147-008-0762-6.

[84] Lyu HG, Haider AH, Landman AB, Raut CP. The opportunities and shortcomings of using big data and national databases for sarcoma research. Cancer 2019;125:2926–34. https://doi.org/10.1002/cncr.32118.

[85] Yang DX, Khera R, Miccio JA, Jairam V, Chang E, Yu JB, et al. Prevalence of Missing Data in the National Cancer Database and Association With Overall Survival. JAMA Netw Open 2021;4:e211793. https://doi.org/10.1001/jamanetworkopen.2021.1793.

[86] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ 2013;346:e5595–e5595. https://doi.org/10.1136/bmj.e5595.

[87] Kent P, Cancelliere C, Boyle E, Cassidy JD, Kongsted A. A conceptual framework for prognostic research. BMC Med Res Methodol 2020;20:172. https://doi.org/10.1186/s12874-020-01050-7.

[88] Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. J Med Syst 1991;15:11–9. https://doi.org/10.1007/BF00993877.

[89] Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. J Breast Cancer 2012;15:230. https://doi.org/10.4048/jbc.2012.15.2.230.

[90] Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl 2014;24:1311–6. https://doi.org/10.1007/s00521-013-1359-1.

[91] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113–27. https://doi.org/10.1016/j.artmed.2004.07.002.

[92] Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22:e184–90. https://doi.org/10.1093/bioinformatics/btl230.

[93] Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 2014;48:1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006.

[94] Abu-Hanna A, Lucas PJF. Prognostic Models in Medicine: AI and Statistical Approaches. Methods Inf Med 2001;40:1–5. https://doi.org/10.1055/s-0038-1634456.

[95] Marks LB, Yorke ED, Jackson A, Ten Haken RK, Constine LS, Eisbruch A, et al. Use of Normal Tissue Complication Probability Models in the Clinic. Int J Radiat Oncol 2010;76:S10–9. https://doi.org/10.1016/j.ijrobp.2009.07.1754.

[96] Troicki FT, Troicki FT, Troicki FT, Perez CA, Thorstad WL, Fisher BJ, et al. Normal Tissue Complication Probability (NTCP). In: Brady LW, Yaeger TE, editors. Encycl. Radiat. Oncol., Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, p. 560–560. https://doi.org/10.1007/978-3-540-85516-3_341.

[97] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[98] Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Med 2013;10:e1001380. https://doi.org/10.1371/journal.pmed.1001380.

[99] Simon R, Altman D. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85. https://doi.org/10.1038/bjc.1994.192.

[100] Kyzas PA, Loizou KT, Ioannidis JPA. Selective Reporting Biases in Cancer Prognostic Factor Studies. JNCI J Natl Cancer Inst 2005;97:1043–55. https://doi.org/10.1093/jnci/dji184.

[101] Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer 2007;43:2559–79. https://doi.org/10.1016/j.ejca.2007.08.030.

[102] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med 2013;10:e1001381. https://doi.org/10.1371/journal.pmed.1001381.

[103] Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: Springer; 2015.

[104] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer; 2009.

[105] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40. https://doi.org/10.1186/1471-2288-14-40.

[106] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010;21:128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[107] Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA 2017;318:1377. https://doi.org/10.1001/jama.2017.12126.

[108] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17. https://doi.org/10.1002/sim.4154.

[109] Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Weather Rev 1950;78:1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[110] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55. https://doi.org/10.7326/M14-0697.

[111] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015;162:W1. https://doi.org/10.7326/M14-0698.

[112] Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer 1982;45:361–6. https://doi.org/10.1038/bjc.1982.62.

[113] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation 1998;97:1837–47. https://doi.org/10.1161/01.CIR.97.18.1837.

[114] Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of Clinical Classification Schemes for Predicting Stroke: Results From the National Registry of Atrial Fibrillation. JAMA 2001;285:2864. https://doi.org/10.1001/jama.285.22.2864.

[115] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 2003;89:232–8. https://doi.org/10.1038/sj.bjc.6601118.

[116] Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care 2004;8:389. https://doi.org/10.1186/cc2955.

[117] Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol 1972;34:187–220.

[118] Fox J, Weisberg S. An R companion to applied regression. Third edition. Los Angeles: SAGE; 2019.

[119] Breslow NE. Analysis of Survival Data under the Proportional Hazards Model. Int Stat Rev Rev Int Stat 1975;43:45. https://doi.org/10.2307/1402659.

[120] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Hoboken, N.J: J. Wiley; 2002.

[121] Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. P T Peer-Rev J Formul Manag 2010;35:560–76.

[122] Hingorani AD, Windt DA v. d., Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ 2013;346:e5793–e5793. https://doi.org/10.1136/bmj.e5793.

[123] Jones MP, Hruby G, Metser U, Sridharan S, Capp A, Kumar M, et al. FDG-PET parameters predict for recurrence in anal cancer - results from a prospective, multicentre clinical trial. Radiat Oncol Lond Engl 2019;14:140. https://doi.org/10.1186/s13014-019-1342-9.

[124] Tomaszewski JM, Link E, Leong T, Heriot A, Vazquez M, Chander S, et al. Twenty-five-year experience with radical chemoradiation for anal cancer. Int J Radiat Oncol Biol Phys 2012;83:552–8. https://doi.org/10.1016/j.ijrobp.2011.07.007.

[125] Aggarwal A, Gayadeen S, Robinson D, Hoskin PJ, Mawdsley S, Harrison M, et al. Clinical target volumes in anal cancer: Calculating what dose was likely to have been delivered in the UK ACT II trial protocol. Radiother Oncol 2012;103:341–6. https://doi.org/10.1016/j.radonc.2012.03.007.

[126] Lu Y, Wang X, Li P, Zhang T, Zhou J, Ren Y, et al. Clinical characteristics and prognosis of anal squamous cell carcinoma: a retrospective audit of 144 patients from 11 cancer hospitals in southern China. BMC Cancer 2020;20:679. https://doi.org/10.1186/s12885-020-07170-z.

[127] Rimini M, Franco P, Bertolini F, Berardino DB, giulia ZM, Stefano V, et al. The Prognostic Role of Baseline Eosinophils in HPV-Related Cancers: a Multi-institutional Analysis of Anal SCC and OPC Patients Treated with Radical CT-RT. J Gastrointest Cancer 2022. https://doi.org/10.1007/s12029-022-00850-y.

[128] Rimini M, Franco P, De Bari B, Zampino MG, Vagge S, Frassinetti GL, et al. The Prognostic Value of the New Combined Hemo-Eosinophil Inflammation Index (HEI Index): A Multicenter Analysis of Anal Cancer Patients Treated with Concurrent Chemo-Radiation. Cancers 2021;13:671. https://doi.org/10.3390/cancers13040671.

[129] Kim E, Kim TH, Jung W, Kim K, Chang AR, Park HJ, et al. Prognostic impact of neutrophilia and lymphopenia on survival in anal cancer treated with definitive concurrent chemoradiotherapy: a retrospective multicenter study. Int J Clin Oncol 2022;27:553–62. https://doi.org/10.1007/s10147-021-02094-5.

[130] Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. https://doi.org/10.1136/bmj.m441.

[131] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2017;26:796–808. https://doi.org/10.1177/0962280214558972.

[132] Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol 2011;64:993–1000. https://doi.org/10.1016/j.jclinepi.2010.11.012.

[133] van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016;16:163. https://doi.org/10.1186/s12874-016-0267-3.

[134] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res 2019;28:2455–74. https://doi.org/10.1177/0962280218784726.

[135] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96. https://doi.org/10.1002/sim.7992.

[136] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study: Sample size considerations for validating a prognostic model. Stat Med 2016;35:214–26. https://doi.org/10.1002/sim.6787.

[137] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. https://doi.org/10.1016/j.jclinepi.2015.12.005.

[138] Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016:i3140. https://doi.org/10.1136/bmj.i3140.

[139] Evin C, Quéro L, Le Malicot K, Blanchet-Deverly S, François E, Buchalet C, et al. MO-0226 Clinical outcomes of HIV-positive patients with anal cancer in the ANABASE multicentric cohort. Radiother Oncol 2022;170:S185–6. https://doi.org/10.1016/S0167-8140(22)02328-3.

[140] Vendrely V, Lemanski C, Pommier P, Le Malicot K, Francois E, Rivin Del Campo E, et al. OC-0270 Final results of the French national cohort ANABASE: treatment and outcome in anal cancer. Radiother Oncol 2022;170:S226–7. https://doi.org/10.1016/S0167-8140(22)02528-2.

[141] Caravatta L, Mantello G, Valvo F, Franco P, Gasparini L, Rosa C, et al. Radiotherapy with Intensity-Modulated (IMRT) Techniques in the Treatment of Anal Carcinoma (RAINSTORM): A Multicenter Study on Behalf of AIRO (Italian Association of Radiotherapy and Clinical Oncology) Gastrointestinal Study Group. Cancers 2021;13:1902. https://doi.org/10.3390/cancers13081902.

[142] IMACC Faculty. The First International Multidisciplinary Anal Cancer Conference 2021. https://events.au.dk/imacc2021/conference (accessed November 7, 2022).

[143] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9. https://doi.org/10.1016/j.radonc.2014.10.001.

[144] Cavallaro F, Lugg-Widger F, Cannings-John R, Harron K. Reducing barriers to data access for research in the public interest—lessons from covid-19. BMJ Opin 2022. https://blogs.bmj.com/bmj/2020/07/06/reducing-barriers-to-data-access-for-research-in-the-public-interest-lessons-from-covid-19/ (accessed November 8, 2022).

[145] Ford E, Boyd A, Bowles JKF, Havard A, Aldridge RW, Curcin V, et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. Learn Health Syst 2019;3. https://doi.org/10.1002/lrh2.10191.

[146] Mourby MJ, Doidge J, Jones KH, Aidinlis S, Smith H, Bell J, et al. Health Data Linkage for UK Public Interest Research: Key Obstacles and Solutions. Int J Popul Data Sci 2019;4:1093. https://doi.org/10.23889/ijpds.v4i1.1093.

[147] Kirienko M, Sollini M, Ninatti G, Loiacono D, Giacomello E, Gozzi N, et al. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. Eur J Nucl Med Mol Imaging 2021;48:3791–804. https://doi.org/10.1007/s00259-021-05339-7.

[148] Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform 2020:184–200. https://doi.org/10.1200/CCI.19.00047.

[149] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[150] Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc JAMIA 2015;22:1212–9. https://doi.org/10.1093/jamia/ocv083.

[151] Verbraeken J, Wolting M, Katzy J, Kloppenburg J, Verbelen T, Rellermeyer JS. A Survey on Distributed Machine Learning. ACM Comput Surv 2021;53:1–33. https://doi.org/10.1145/3377454.

[152] Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: Distributed deep learning without sharing raw patient data 2018. https://doi.org/10.48550/ARXIV.1812.00564.

[153] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated Learning: Strategies for Improving Communication Efficiency. ArXiv161005492 Cs 2017.

[154] Kayaalp M. Modes of De-identification. AMIA Annu Symp Proc AMIA Symp 2017;2017:1044–50.

[155] Pinkas B. Cryptographic techniques for privacy-preserving data mining. ACM SIGKDD Explor Newsl 2002;4:12–9. https://doi.org/10.1145/772862.772865.

[156] Froelicher D, Troncoso-Pastoriza JR, Pyrgelis A, Sav S, Sousa JS, Bossuat J-P, et al. Scalable Privacy-Preserving Distributed Learning 2020. https://doi.org/10.48550/ARXIV.2005.09532.

[157] Jayaraman B, Wang L, Evans D, Gu Q. Distributed Learning without Distress: Privacy-Preserving Empirical Risk Minimization. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Adv. Neural Inf. Process. Syst., vol. 31, Curran Associates, Inc.; 2018.

[158] Chamikara MAP, Bertok P, Khalil I, Liu D, Camtepe S. Privacy preserving distributed machine learning with federated learning. Comput Commun 2021;171:112–25. https://doi.org/10.1016/j.comcom.2021.02.014.

[159] Brink C, Hansen CR, Field M, Price G, Thwaites D, Sarup N, et al. Distributed learning optimisation of Cox models can leak patient data: Risks and solutions 2022.

[160] Huth M, Gusinow R, Contento L, Tacconelli E, Hasenauer J. Accessibility of covariance information creates vulnerability in Federated Learning frameworks. Bioinformatics; 2022. https://doi.org/10.1101/2022.10.09.511497.

[161] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[162] Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiother Oncol 2020;144:189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[163] Sioutos N, Coronado S de, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40:30–43. https://doi.org/10.1016/j.jbi.2006.02.013.

[164] Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys 2018;45:e854–62. https://doi.org/10.1002/mp.12879.

[165] National Library of Medicine. PubMed.gov search: “distributed learning” or “federated learning.” PubMedGov n.d. https://pubmed.ncbi.nlm.nih.gov/?term=%28distributed+learning%29+OR+%28federated+learning%29&sort= (accessed November 9, 2022).

[166] Duan R, Boland MR, Liu Z, Liu Y, Chang HH, Xu H, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. J Am Med Inform Assoc 2020;27:376–85. https://doi.org/10.1093/jamia/ocz199.

[167] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis ICh, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inf 2018;112:59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[168] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2:305–11. https://doi.org/10.1038/s42256-020-0186-1.

[169] Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018;25:945–54. https://doi.org/10.1093/jamia/ocy017.

[170] Dluhoš P, Schwarz D, Cahn W, van Haren N, Kahn R, Španiel F, et al. Multi-center machine learning in imaging psychiatry: A meta-model approach. NeuroImage 2017;155:10–24. https://doi.org/10.1016/j.neuroimage.2017.03.027.

[171] Constable SD, Tang Y, Wang S, Jiang X, Chapin S. Privacy-preserving GWAS analysis on federated genomic datasets. BMC Med Inform Decis Mak 2015;15:S2. https://doi.org/10.1186/1472-6947-15-S5-S2.

[172] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, et al. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109:159–64. https://doi.org/10.1016/j.radonc.2013.07.007.

[173] Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol 2017;99:344–52. https://doi.org/10.1016/j.ijrobp.2017.04.021.

[174] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 2016;121:459–67. https://doi.org/10.1016/j.radonc.2016.10.002.

[175] Lambin P, Zindler J, Vanneste BGL, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. Adv Drug Deliv Rev 2017;109:131–53. https://doi.org/10.1016/j.addr.2016.01.006.

[176] Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Sci Data 2019;6:218. https://doi.org/10.1038/s41597-019-0241-0.

[177] Radiomics. Radiomics Insight-based decision making 2022. https://radiomics.bio (accessed November 9, 2022).

[178] Czeizler E, Wiessler W, Koester T, Hakala M, Basiri S, Jordan P, et al. Using federated data sources and Varian Learning Portal framework to train a neural network model for automatic organ segmentation. Phys Med 2020;72:39–45. https://doi.org/10.1016/j.ejmp.2020.03.011.

[179] Powell K. NVIDIA Clara Platform to Usher in Next Wave of Medical Instruments. Clara Dram Boost Capab Leg Instrum Setting Future AI Med Devices 2018. https://blogs.nvidia.com/blog/2018/09/12/nvidia-clara-platform/ (accessed November 9, 2022).

[180] Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH, et al. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. Genet Med 2020;22:371–80. https://doi.org/10.1038/s41436-019-0646-3.

[181] Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. AMIA Annu Symp Proc AMIA Symp 2020;2020:870–7.

[182] IKNL. Distributed Cox regression algorithm 2019.

[183] Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, et al. Prognostic factors for patients with anal cancer treated with conformal radiotherapy—a systematic review. BMC Cancer 2022;22:607. https://doi.org/10.1186/s12885-022-09729-4.

[184] Theophanous S, Choudhury A, Lønne P-I, Samuel R, Guren MG, Berbee M, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed

learning – A proof-of-concept study. Radiother Oncol 2021;159:183–9. https://doi.org/10.1016/j.radonc.2021.03.013.

[185] Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Dennis K, et al. Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study. Diagn Progn Res 2022;6:14. https://doi.org/10.1186/s41512-022-00128-8.

Chapter 2 - Development of a comprehensive institutional anal cancer data warehouse for real-world data analysis

2.1 Abstract

2.1.1 Introduction

Analysis of routine patient data generated from clinical practice can support the continuous development and improvement of clinical practice. However, these data are commonly stored in multiple systems/databases without seamless interoperability. Information Governance and data regulatory policies are a potential barrier to accessing data for research purposes. This highlights the need for cancer data warehouses which integrate heterogeneous data from multiple sources, with well-specified data dictionaries and clear processes for data quality assurance and cleaning. In this study, we aimed to develop a comprehensive data warehouse of patients treated at Leeds Cancer Centre, using anal cancer as an exemplar.

2.1.2 Materials and methods

The data warehouse development process encompassed the identification of relevant data items and their respective data sources. A combination of automatic and manual data extraction was employed, with considerable quality assessment of initial data extraction as part of the data warehouse development. A process for data de-identification was implemented to facilitate access to clinicians and researchers on an ongoing basis for research and audit purposes. Further maintenance and data updating was designed to be as automated as possible.

2.1.3 Results

Retrospective data from 568 patients treated with radiotherapy for anal cancer in LCC between January 2013 and September 2022 were collected. The data warehouse consists of 194 anal cancer-related data items, which were defined in a comprehensive data dictionary. Automatically extracted data items were evaluated to be of high quality, although the quality of manually collected data items requires further improvement.

2.1.4 Conclusion

Authorised local researchers and clinicians can now access the available high quality, de-identified data, which are ready for analysis in future research studies. Ultimately, the insights obtained by analysing these data may help advance future clinical management of patients diagnosed with anal cancer.

2.2 Introduction

The analysis of routine patient data generated from clinical practice throughout the cancer diagnosis, treatment, and follow-up pathway can provide cancer treatment centres with a number of benefits [1]. Firstly, analysing these data can yield a better understanding of changes in patient referral and treatment patterns. Additionally, routinely collected patient data can be used to learn directly from cohorts with characteristics that reflect the makeup of local patients. Therefore, such analyses provide real-world data on treatments and outcomes which do not suffer from the issues that are prevalent in randomised controlled trials (RCTs) [2]. Even though RCTs remain the gold standard for initially demonstrating the efficacy of novel interventions, they suffer from several weaknesses. RCTs are carried out under strict conditions, aiming to minimise bias and ensure that the outcome being measured is only affected by the intervention under investigation. However, in most cases, this does not reflect real-world healthcare settings, where numerous external factors influence patient outcomes during and after treatment [3]. Moreover, the patient population that participates in RCTs may not be fully representative of the overall patient population [4]. Evidence in RCTs is typically derived from patients that are much younger and fitter than the average patient [5,6]. The analysis of routinely-collected patient data can therefore be used to complement RCT evidence [7], particularly in cases where rare populations or under-represented groups are being considered.

However, routine clinical data are commonly stored in multiple different systems/databases without seamless interoperability. Additionally, due to the highly sensitive nature of these data, Information Governance (IG) and data regulatory policies are a potential barrier to accessing data for research purposes. While essential from a patient privacy perspective, such policies can introduce inefficiencies for researchers aiming to collect and analyse patient data.

Raw routine data are seldomly generated with audit and research as the main purpose. Consequently, routine data may need considerable pre-processing to ensure high quality prior to the analysis phase [8]. This is often done on a ‘per project’ basis, with limited consideration of future (re-)use of data; i.e. with limited thoughts to documentation of data providence, coding, interoperability and visibility. Many clinical researchers will recognise dealing with old project spreadsheets, with perplexing column names and coding systems, and no updates once a project is complete and an abstract or paper submitted. These studies may have partly overlapping cohorts and information, signifying duplication of effort and inefficient use of resources.

The above challenges call for the development of cancer data warehouses which integrate heterogeneous data from multiple sources, with well-specified data dictionaries and clear processes for data quality assurance and cleaning [9]. To ensure safe use of data for research, such data warehouses also need appropriate governance processes to ensure control of data access, and technical infrastructure to support these. This paper describes the methodology employed to develop a cancer data warehouse fulfilling the above requisites, within the context of a radiotherapy data warehouse in Leeds Cancer Centre (LCC). We use anal cancer as an exemplar.

Anal cancer is a rare cancer with an increasing incidence rate. it is treated with a combination of radiotherapy and chemotherapy [10], with three-year overall survival of 85.6% and a three-year disease-free survival of 75.6% [11,12]. Using real-world data to understand patient outcomes on a local, national, and international level is particularly relevant in cases of rare cancer, where access to high quality data limits our ability to develop models with sufficient data to establish new models of care.

The aim of this study was to develop a comprehensive data warehouse of patients with anal cancer treated at LCC which is updated in a process that is as automated as possible. By implementing this approach, we aimed to establish a methodology that can be reproduced across disease sites, allowing for easy and efficient utilisation of real-world data. Here, we document the information governance arrangements, as well as the data collection and quality evaluation procedures, highlighting the proportion of data items that could be automatically extracted, and the amount of manual entry and manual review required to develop an institutional data warehouse.

2.3 Materials & Methods

2.3.1 Ethics and governance for data collection and data warehouse development

This project sits within the framework of a HRA reviewed and REC approved research database (called 'LeedsCAT'); REC reference 19/YH/0300, IRAS project ID 255585. LeedsCAT aims to learn from every radiotherapy patient treated at LCC, by repurposing existing oncology databases for evaluation, audit, and research purposes. The LeedsCAT Governance board consists of clinicians, scientific leads, radiographers, scientific computing leads, an Information and Technology Security Officer, a Research and Innovation representative, a Patient and Public Involvement (PPI) representative and a project manager. All LCC proposed projects involving retrospective patient data analysis are considered by the Governance board. All patient data is rigorously de-identified at the earliest opportunity and access to any patient data including de-identified data is strictly controlled with only named individuals having access. The data storage is split up into four tiers where data access is allowed to clinical staff, LCC research staff, LCC non-research staff and distributed learning, respectively. For this specific work on anal cancer, approval was obtained from the LeedsCAT Governance board (See Appendix D for the LeedsCAT letter of approval).

2.3.2 Identification of relevant data items and data sources

The initial phase of the data warehouse development involved identifying and defining relevant data items. This process was carried out by a team of researchers, medical physicists, and clinical oncologists to ensure all data items that might be useful in future research are included in the data warehouse. In order to reach consensus, the current literature on anal cancer was explored to identify data items that are commonly analysed [13]. Additional data items were added, including items that were analysed in previous anal cancer studies conducted at LCC, as well as items deemed relevant according to the clinical expertise of the involved clinical oncologists, and according to the results of the core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC) consensus [14].

Upon identifying all relevant data items, the most appropriate data sources were pinpointed. There were four main data sources: Patient Pathway Manager (PPM), Monaco (Elekta AB), MOSAIQ (Elekta AB), and ChemoCare. PPM is LCCs electronic health record [15]. It includes personal and baseline clinical data, histopathology reports, and general clinical notes, as well as summary radiotherapy, chemotherapy, and surgery-related data. Clinical notes often include important free-text information on the patient's diagnosis, multidisciplinary team meeting discussions, and outcome data. Monaco [16] is a treatment planning system developed by Elekta, which is used to generate radiotherapy plans, and can be used to carry out tasks such as treatment volume delineation and plan optimisation. MOSAIQ [17], which is a radiation oncology information system also developed by Elekta, is used to manage radiotherapy booking and scheduling. Moreover, MOSAIQ aids with the communication and control of the linear accelerators that are used to deliver radiation to patients. Specifically, it links the patient's radiotherapy treatment plan to the linear accelerator. MOSAIQ can record and verify, as well as store radiotherapy-specific notes. Lastly, ChemoCare [18] is the clinical system used at LTHT to schedule chemotherapy treatments and manage chemotherapy prescriptions. Through ChemoCare, administered dosages and patient appointments are tracked and recorded, in order to facilitate communication between oncologists, pharmacists, and nurses.

A plan was devised detailing how to collect data for each item. Data collection for each item depended on which database or clinical system the data item could be sourced from, and in which form it was originally stored in. Data items were classified into three main categories: (1) data items to be primarily automatically extracted from their source, (2) data items to be primarily manually collected from their source, and (3) data items to be automatically calculated from other data items. A selected list of data items that were automatically extracted from their source were also manually reviewed by a researcher or a clinician to ensure accuracy and high quality (see Data Quality Evaluation section).

2.3.3 Data dictionary and structure of the data warehouse

The next phase entailed the generation of a data dictionary (Appendix E). For each data item, the data dictionary specified the data type, the data length, whether null values were allowed or not, as well as a short description defining the data item. Additionally, for numerical data items, the units of measurement were specified (SI units were used where possible). The coding system devised for all categorical data items was also denoted where relevant. The data dictionary indicated where each data item was sourced and how the data could be collected (automatically extracted, manually collected, automatically calculated, or via a combination of automatic extraction and manual review). Finally, the data quality score (discussed below) of each data item was included, along with a justification.

The data warehouse structure consisted of multiple inter-linked tables. The first table included all identifiable patient data, such as name, surname, and all dates, including date of birth, date of diagnosis (defined as the date that the radiotherapy treatment referral form was added to MOSAIQ), and date of start of radiotherapy. A unique, project-specific identification number (ID) was assigned to each patient, which was then used in all other tables. All other tables included de-identified patient data only, as indicated in Table 2-1. All dates were converted to number of days after the start of radiotherapy. Lastly, a link table was created with the aim of linking together all tables that used different ID numbers. A visual representation of the data warehouse architecture is provided in Figure 2-1.

Table 2-1. Structure of the Leeds anal cancer data warehouse. MEPP signifies tables that can include multiple entries per patient.

Table	Data included	Specific examples of data items
A	All identifiable patient data	Patient name, surname, date of birth, date of death (if relevant), date of diagnosis, date of first radiotherapy fraction.
B	Demographics, pre-existing comorbidities, diagnostic data, follow-up data	Age, baseline TNM staging, baseline performance status, tumour histology, data on treatment response, outcome data (survival, locoregional control, freedom from distant metastasis).
C	Radiotherapy data	Prescribed radiotherapy dose to primary tumour, involved nodes and elective nodes, overall treatment duration, delivered radiotherapy doses.
D	Chemotherapy data (MEPP)	Chemotherapy regimen, duration of chemotherapy, number of chemotherapy cycles, when did the patient receive chemotherapy.
E	Surgery data (MEPP)	Type of surgery, intent, when did the patient receive surgery.
F	Hospital admissions (MEPP)	Reason for admission, when was the patient admitted and discharged.

G	Clinical trial data	Which trial does/did the patient participate in, when was the patient recruited, outcome
H	Link table	Linking all other tables together by matching identification numbers.

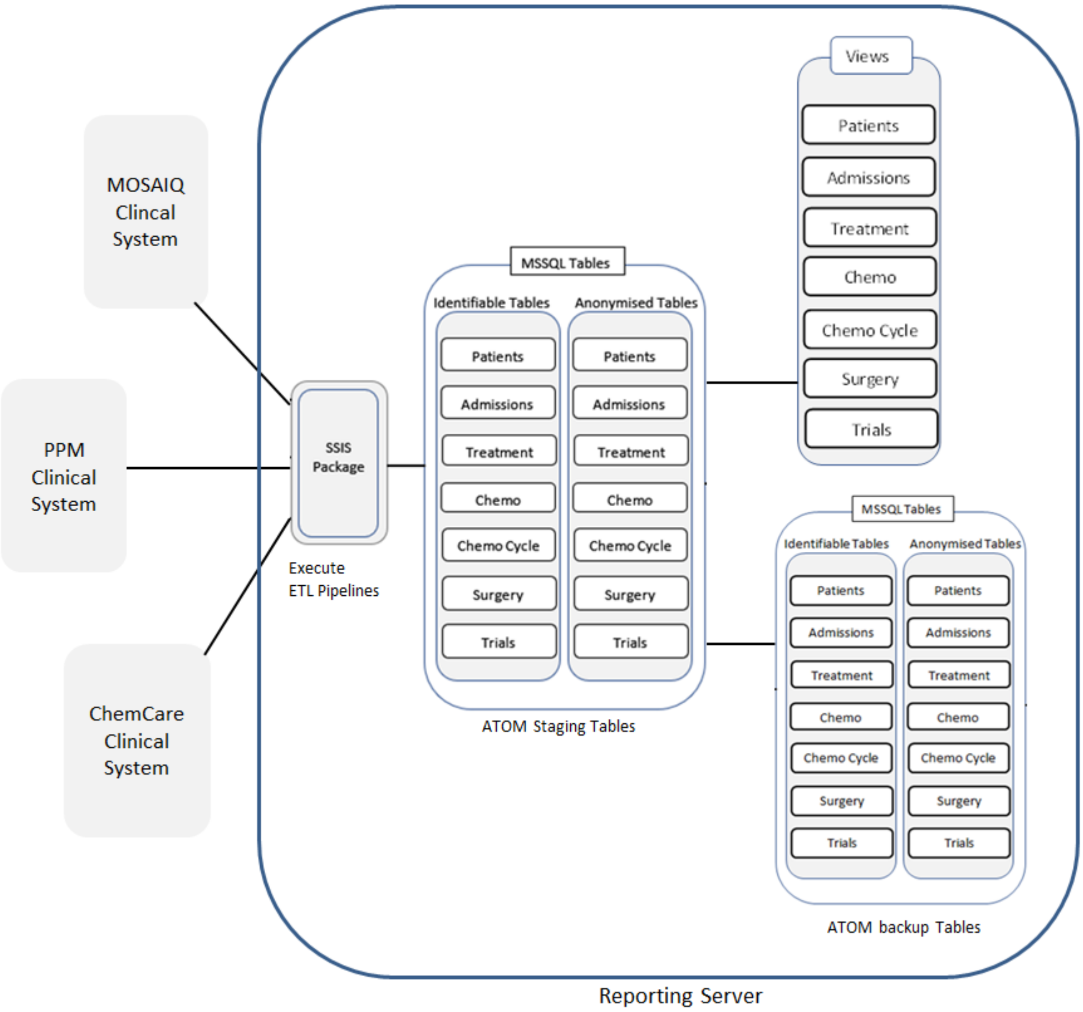


Figure 2-1. Diagram depicting the Leeds anal cancer data warehouse architecture.

2.3.4 Identification of patients treated for anal cancer

The main inclusion criterion for the data warehouse was pelvic radiotherapy for anal cancer. No exclusion took place according to whether patients received chemotherapy or not. Similarly, patients that were treated with palliative intent or who had surgery prior to radiotherapy were not excluded. The following strategy was used to identify patients treated in LCC for inclusion: firstly, the radiotherapy planning code (a code used at LCC, which is generated at the point of creating a radiotherapy plan) was used to search MOSAIQ for patients treated with radiotherapy at LCC between 2013 and 2020. The results of this search were filtered by the radiotherapy treatment site. Only patients with radiotherapy treatment sites that included the keywords “anus” or “anal” were selected. A similar search was conducted using the International Classification of Diseases 10th revision (ICD10) diagnosis code in MOSAIQ. The same radiotherapy treatment site filters were applied to the results of this search as well. The two lists of identified patients were merged and duplicate IDs were removed. Patients that were identified in only one of the two searches were manually reviewed. Patient records were accessed in PPM by a researcher to ensure these patients were indeed diagnosed and treated for anal cancer. The final list only encompassed patients that received pelvic radiotherapy for anal cancer. A small number of patients that only received surgery and no radiotherapy were not identified and were therefore not included in the data warehouse.

2.3.5 Data collection procedure

The data collection procedure was divided into three phases: automatic data extraction, manual data collection and automatic calculation from already collected data.

Data items that could be sourced from MOSAIQ or ChemoCare were automatically extracted using a series of SQL queries. The SQL queries were extensively tested and validated to ensure integrity and consistency in results, and were executed using Microsoft SQL Server Integration Services (SSIS). SSIS is an Extract Transform Load (ETL) tool which is widely implemented in the healthcare sector for pre-processing, aggregation, and integration of data from multiple databases and clinical systems. SSIS supports the seamless long-term maintenance and updating of the data warehouse. Separate ETL pipelines were developed for the tables including identifiable and de-identified data, which utilised the SQL queries to collect data from MOSAIQ and ChemoCare. A number of pre-processing steps were applied within the ETL pipeline to clean and condition the data for ingestion into the data warehouse. As illustrated in Figure 2-2, these pipelines were then integrated together within SSIS to ensure that all the tables in the data warehouse are updated with minimal effort.

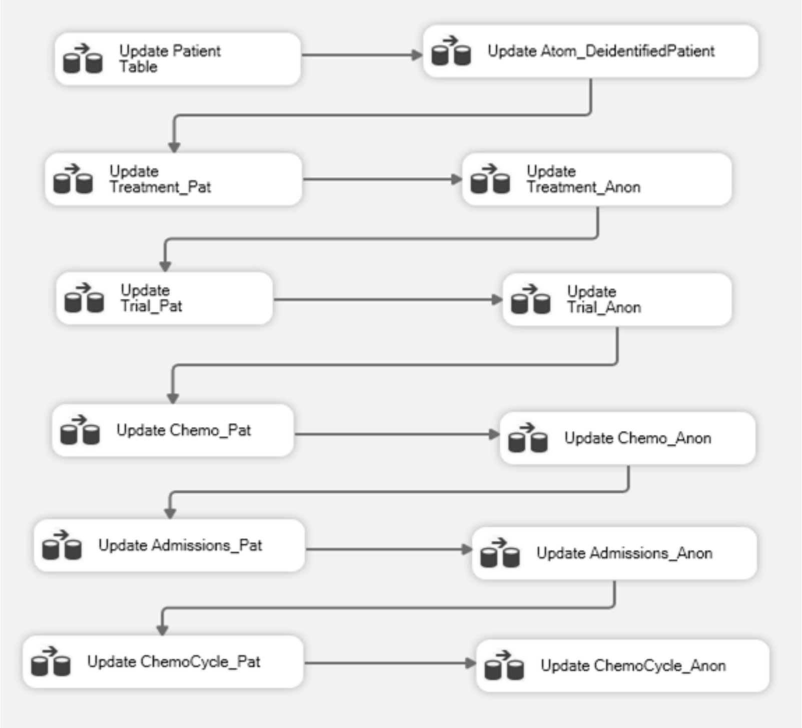


Figure 2-2. Schematic depicting the separate ETL pipelines developed for each table.

Data items that could be sourced from PPM had to be manually collected due to access restrictions to the system's backend (the backend communicates with the system and renders content to push to the frontend, which is accessible by end users), which were put in place by LCC's IT team. Therefore, bulk data extraction from PPM was not possible. Additionally, data that could be sourced from the radiotherapy plans stored in Monaco had to be manually collected. A decision was made to limit manual data collection to a subset of the patients treated with IMRT/VMAT (as these would be included in the atomCAT studies, see Chapters 4 and 6). In PPM, patient records were individually accessed by a researcher (ST) and the relevant data were recorded in a spreadsheet. A number of data items were collected from clinic letters, multidisciplinary team meeting letters or clinician's notes. These items were only available in free-text format (baseline TNM staging, locoregional control and freedom from distant metastasis data). A subset of the manually collected data were reviewed by a clinical oncologist (AG) to confirm high quality, and to resolve any ambiguities. Radiotherapy treatment plans in Monaco were accessed by ST in order to collect primary tumour gross tumour volume (GTV) data for patients treated with VMAT. A medical physicist (AA) helped resolve any ambiguities. Lastly, all manually collected data were imported into the data warehouse and merged with the automatically extracted data.

2.3.6 Missing data

Missing data in the data warehouse were stratified into three categories: (1) “Not available”, referring to data with an identified source which we attempted to obtain but the data could not be found in the originating database or clinical system; (2) “Not relevant”, referring to data that were not relevant for specific patients (e.g. date of death for patients that were alive); and (3) “Not assessed”, referring to information which may exist for an individual patient but was not obtained one way or another; e.g. data whose source could not be identified (e.g. nuclear medicine data and HPV status), or data with an identified source that we did not attempt to obtain (e.g. comorbidity and treatment toxicity data). The data dictionary specifies how each type of missing data was coded for relevant data items.

2.3.7 Data quality evaluation

Data quality within the warehouse was evaluated using two approaches. The quality of data items was evaluated and recorded in the data dictionary. A quality score between 0 and 10 was assigned to data items, which depended on the following criteria: (1) which database or clinical system the data item was sourced from, (2) how the data were originally recorded in the source database/clinical system (if known), (3) whether the data item was automatically extracted or manually collected, (4) if the data were automatically extracted, whether manual review was required, (5) if manual review was required, the rate of agreement between the automatically extracted and manually collected data, and (6) if there was a high rate of disagreement, the underlying reason. A justification for the quality score assigned to each data item was also provided in the data dictionary. The quality of data items that were added to the database for internal use (e.g. ID numbers), and of data items with missing data stratified as “Not assessed” (see previous section) for the majority of patients was not assessed.

The quality of nine data items that were automatically extracted was evaluated quantitatively, to assess the robustness of the automatic data extraction process. This list of data items encompassed a range of data types (dates, categorical and numerical data items). For a subset of patients (n=246), data for a number of these items were also manually collected from a different database or clinical system for further manual review. For instance, the prescribed radiotherapy dose to the primary tumour was initially extracted from MOSAIQ automatically but was also manually collected from clinical notes in PPM. The manually collected data formed the initial data quality evaluation dataset. In order to render this evaluation procedure more robust, the data quality evaluation dataset was expanded using data from four existing research anal cancer datasets. These datasets consisted of anal cancer data that were previously manually collected by clinicians (AG, AS, CJ, RS) in LCC for research purposes. The datasets encompassed different but overlapping cohorts of patients (n=133, 90, 289, 76) and data items for patients treated for anal cancer between July 2008 and May 2018. The final data quality evaluation dataset consisted of data from 476 patients.

The data quality evaluation dataset was compared against the corresponding automatically extracted data. The date of diagnosis varied considerably across three of the existing research datasets. Therefore, the date of diagnosis comparison between automatically extracted and manually collected data was carried out separately for each dataset. The data collected for each patient via the two methods were compared to estimate the rate of agreement and the rate of discrepancy between the two data collection methods. Any discrepancies in data between the two data collection methods were explored further to identify the source and reasons of discrepancy, and the data stored in the data warehouse were updated where necessary. Lastly, additional columns were added to the data warehouse to signify whether the patient had been included in specific research cohorts.

2.3.8 Updating the data warehouse with new patients

A plan was devised with the aim of updating the data warehouse with newly diagnosed patients. The initial database used for development included patients treated up until January 2020. Thus, as a test of viability, a cohort of patients treated for anal cancer from January 2020 to September 2022 was added to the data warehouse. The strategy for identification of patients treated for anal cancer specified above was repeated using the following date limits: 1st January 2020 to 31st September 2022. The identified patient NHS IDs were used for the automatic data extraction from MOSAIQ and ChemoCare and for the manual collection of data from PPM and Monaco, following the methodology outlined above. The ETL pipelines used to update the de-identified patient data and chemotherapy data tables are provided as exemplars (Figures 2-3 and 2-4, respectively).

The automatically extracted data were directly added into the data warehouse. The manually collected data were recorded in a separate spreadsheet, which was then merged with the data warehouse.

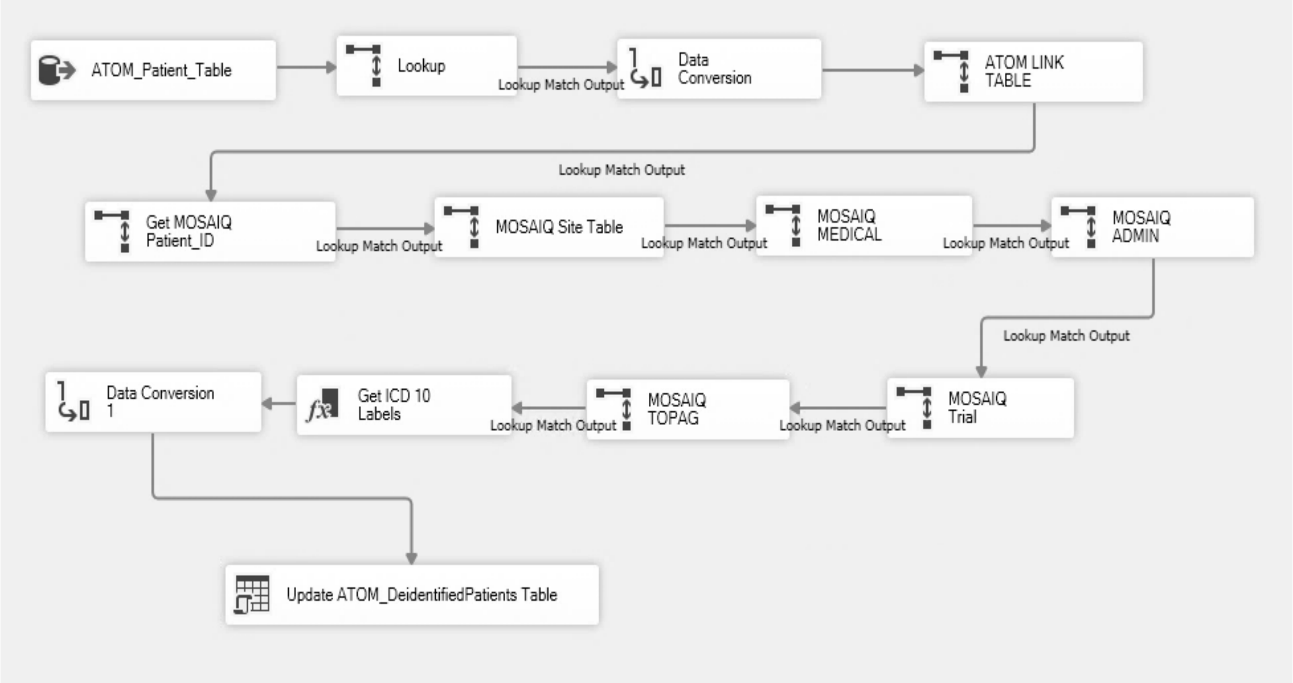
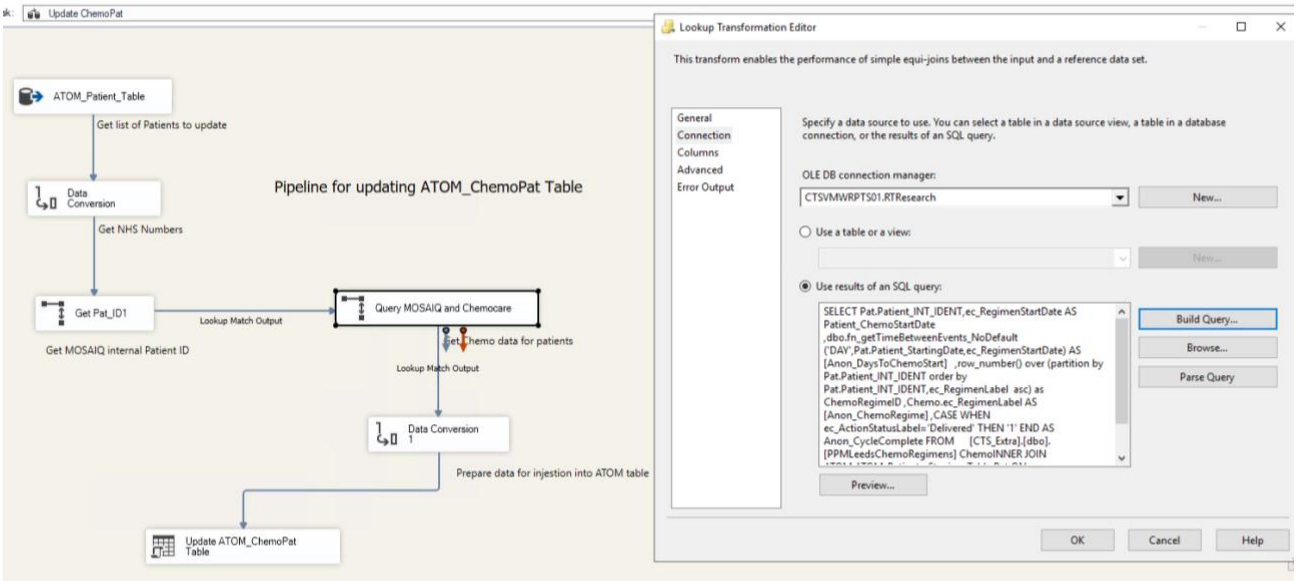


Figure 2-3. ETL pipeline used to update the de-identified patient data table.


Update ChemoPat
ATOM_Patient_Table
Get list of Patients to update
Data Conversion
Get NHS Numbers
Pipeline for updating ATOM_ChemoPat Table
Get Pat_ID1
Lookup Match Output
Query MOSAIQ and Chemocare
Get MOSAIQ internal Patient ID
Get Chemo data for patients
Lookup Match Output
Data Conversion 1
Prepare data for injestion into ATOM table
Update ATOM_ChemoPat Table
Lookup Transformation Editor
General
Connection
Columns
Advanced
Error Output
New...
Build Query...
Browse...
Parse Query
Preview...
OK
Cancel
Help

This transform enables the performance of simple equi-joins between the input and a reference data set.

Specify a data source to use. You can select a table in a data source view, a table in a database connection, or the results of an SQL query.

OLE DB connection manager: CTSVMWRPTS01.RTResearch

○ Use a table or a view:

● Use results of an SQL query:

```
SELECT Pat.Patient_INT_IDENT,ec_RegimenStartDate AS
Patient_ChemoStartDate
,dbo.fn_getTimeBetweenEvents_NoDefault
('DAY',Pat.Patient_StartingDate,ec_RegimenStartDate) AS
[Anon_DaysToChemoStart]  ,row_number() over (partition by
Pat.Patient_INT_IDENT order by
Pat.Patient_INT_IDENT,ec_RegimenLabel  asc) as
ChemoRegimeID ,Chemo.ec_RegimenLabel AS
[Anon_ChemoRegime] ,CASE WHEN
ec_ActionStatusLabel='Delivered' THEN '1' END AS
Anon_CycleComplete FROM     [CTS_Extra].[dbo].
[PPMLeedsChemoRegimens] ChemoINNER JOIN
```

Figure 2-4. ETL pipeline used to update the chemotherapy table.

2.3.9 Technical infrastructure and data access

The data stored in the data warehouse can be accessed via SQL Views, which provides an intuitive interface to end users whilst ensuring data security. This approach also enables end users to access data in a number of formats and interfaces. Data can be loaded directly into other applications, such as Microsoft Excel, SPSS, Power BI, as well as integrated development environments for R and Python, for further analysis.

The data warehouse is currently hosted on a dedicated clinical oncology server which is based on the Microsoft SQL Server platform. The server is fully secured behind a firewall with strict access control in place. Access to the SQL views and the underlying tables is managed using Schema and Object level permissions, ensuring that only users with the correct level of authorisation are able to access the data.

2.4 Results

The flow of data across the various phases of the study is summarised in Figure 2-5.

2.4.1 Data integrity and completeness

The patient search yielded a list of 579 patients. Upon manually reviewing this list, a small number of patients were identified as false positive hits. Specifically, 2% (n=11) of the patients identified by the search were patients diagnosed with and treated for rectal cancer instead of anal cancer. These patients were removed from the data warehouse, yielding the final list of 568 patients treated with radiotherapy for anal cancer in LCC between January 2013 and September 2022. A total of 80% of these patients were treated with IMRT/VMAT (n=463). Figure 2-6 illustrates the number of patients treated per year. A median of 58 patients were treated for anal cancer in LCC each year. As Figure 2-6 indicates, the largest number of patients were treated in 2016 (n=69).

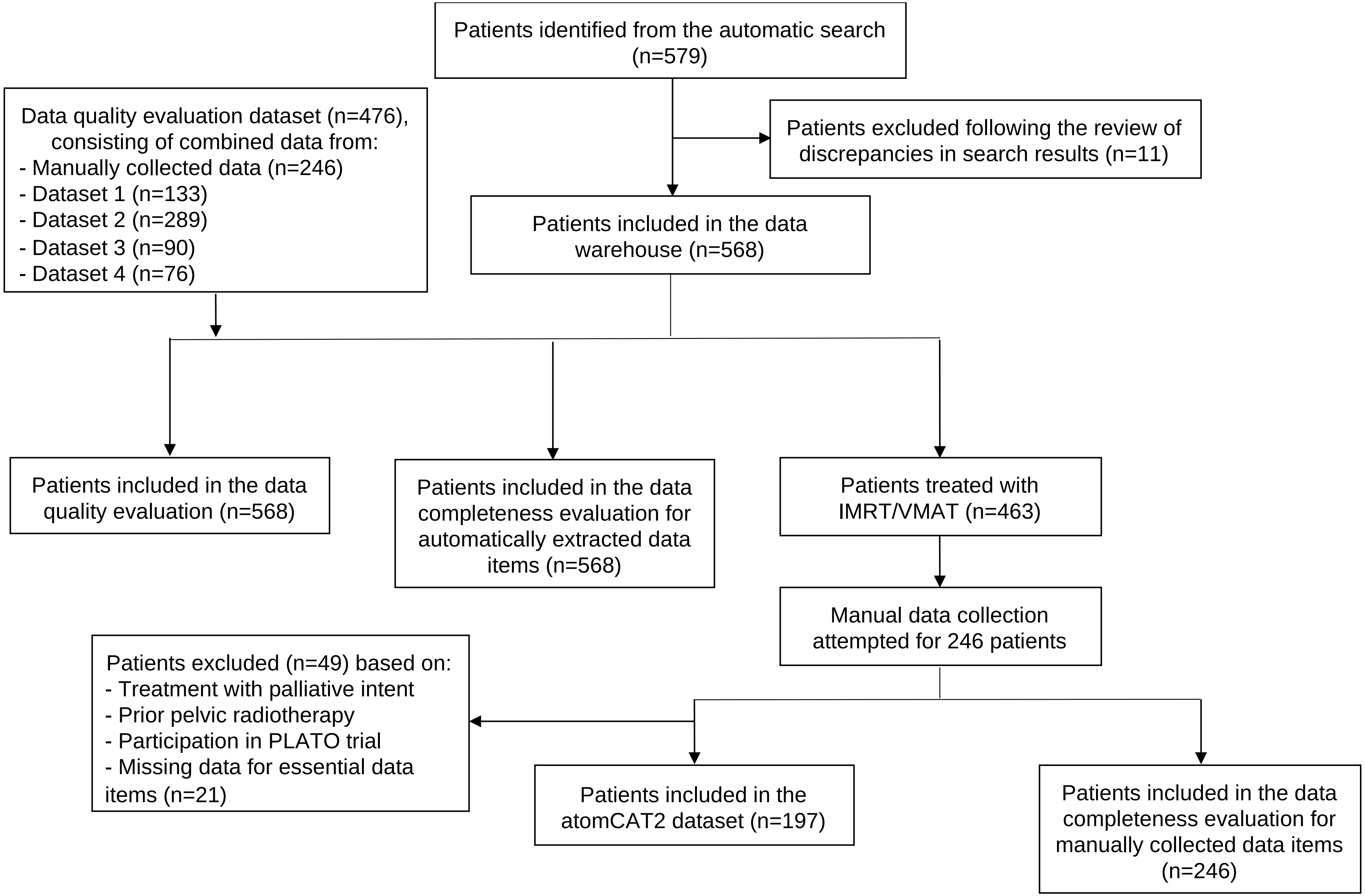

Patients identified from the automatic search (n=579)
Patients excluded following the review of discrepancies in search results (n=11)
Data quality evaluation dataset (n=476), consisting of combined data from:
- Manually collected data (n=246)
- Dataset 1 (n=133)
- Dataset 2 (n=289)
- Dataset 3 (n=90)
- Dataset 4 (n=76)
Patients included in the data warehouse (n=568)
Patients included in the data quality evaluation (n=568)
Patients included in the data completeness evaluation for automatically extracted data items (n=568)
Patients treated with IMRT/VMAT (n=463)
Manual data collection attempted for 246 patients
Patients excluded (n=49) based on:
- Treatment with palliative intent
- Prior pelvic radiotherapy
- Participation in PLATO trial
- Missing data for essential data items (n=21)
Patients included in the atomCAT2 dataset (n=197)
Patients included in the data completeness evaluation for manually collected data items (n=246)

Figure 2-5. Summary of the flow of data across all phases of the study.

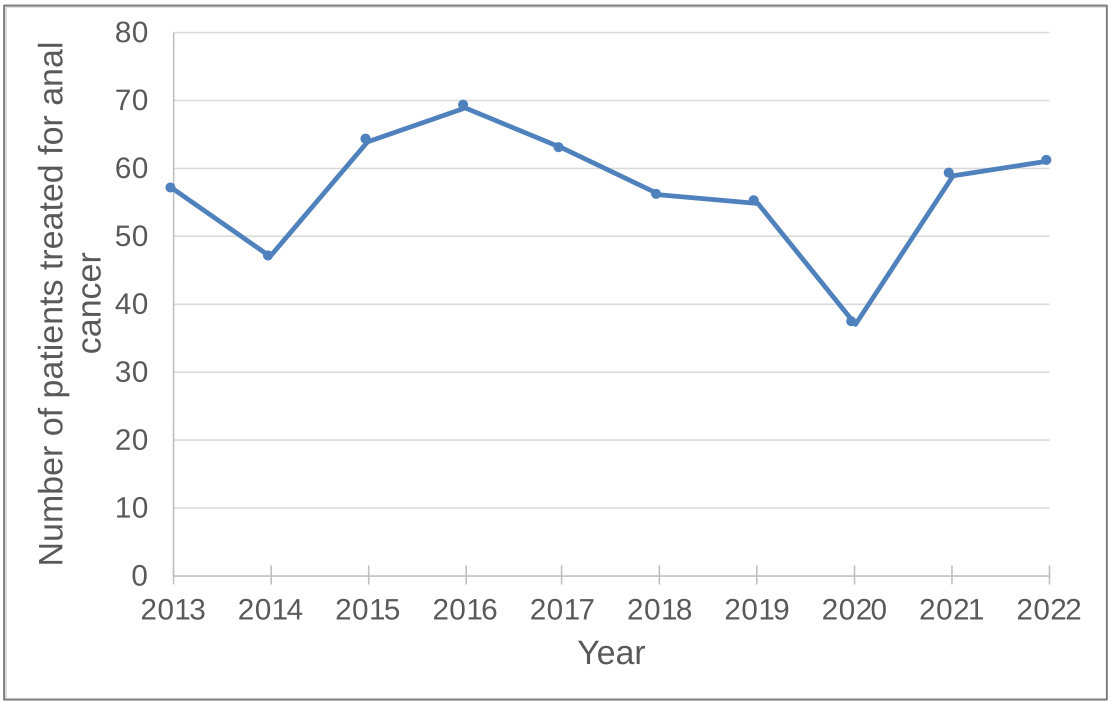


Figure 2-6. Number of patients with anal cancer treated with radiotherapy each year (2013-2022) at Leeds Cancer Centre.

The data warehouse consists of 194 anal cancer-related data items stored in eight tables, excluding data items related to patient IDs and data items added for internal use. From these, a total of 47 data items were automatically extracted, 31 were manually collected, 18 were calculated from other data items (14 from automatically extracted and 4 from manually collected data items), and 98 were not assessed. Data items that were not assessed consist of data items related to comorbidity (n=53) and treatment toxicity (n=27), as well as nuclear medicine-related data items (n=6) and data items whose source could not be identified (n=12). These data items have not been fully extracted yet; therefore, they were not included in subsequent evaluations of data completeness and quality.

The automatically extracted data items exhibit high levels of data completeness, as indicated in Figure 2-7. In summary, from the 47 automatically extracted data items, 43 have complete data for more than 80% of patients, three (diagnosis site, histology, total number of hospital admission days) have complete data for 61-80% of patients, and one data item (histology grade) has complete data for only 25% of patients. No automatically extracted data items have missing data for more than 80% of patients.

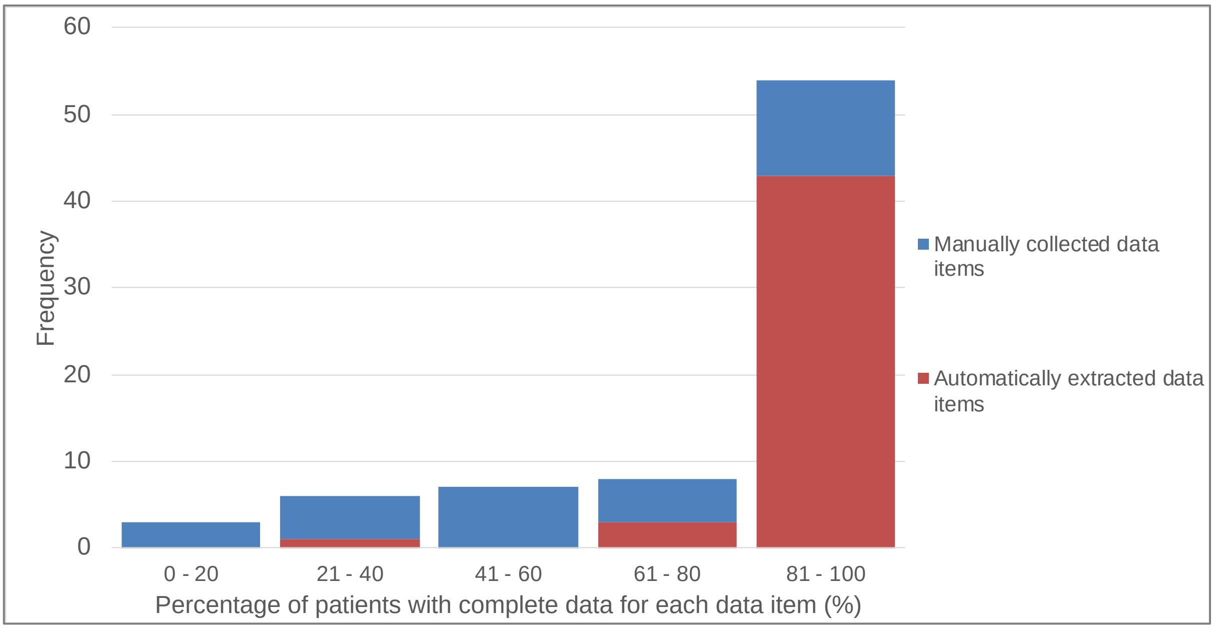


Figure 2-7. Histogram summarising the results of the data completeness evaluation for the automatically extracted and manually collected data items.

The data completeness pattern of the manually collected data items slightly differs (Figure 2-7). A small number of these data items have missing data for more than 80% of patients (n=3, 10%). The majority of manually collected data items have complete data for 21%-80% of patients (n=17, 55%), and 11 data items (35%) have complete data for 81%-100% of patients. The manually collected data items with complete data for more than 81% of patients are the following: originating hospital (the hospital that originally referred the patient to LCC), locoregional recurrence status, locoregional recurrence date, locoregional recurrence site, distant metastasis status (diagnosed at follow-up), distant metastasis date, distant metastasis site (diagnosed at follow-up), T stage, N stage, M stage, and primary tumour GTV.

2.4.2 Data quality

The data quality of the 96 data items that were automatically extracted, manually collected, and automatically calculated from other data items was evaluated. The data quality scores throughout all the evaluated data items range from 1 to 10, with a mean data quality score of 6.9. The mean quality score is 8.3 for automatically extracted data items, 4.6 for manually collected data items, and 7.4 for automatically calculated data items. The quality score distribution of all 96 data items is presented in Figure 2-8.

Only two data items are deemed to be of low quality with a score of 2 or less, both of which are manually collected data items (whether the tumour was located in the anal canal or in the anal margin, and whether the patient had an altered chemotherapy schedule, e.g. missed doses). The lowest scoring automatically collected data items include surgery date, surgery type and surgery intent, all of which have a quality score of 4. The low quality scores can be attributed to the lack of access to data about surgical treatment outside of Leeds Teaching Hospitals NHS Trust. The majority of data items have a high quality score of 7 or more (n=52, 54%), most of which are automatically extracted data items (n=39). Only two manually collected data items have a quality score of 7 or higher; primary tumour size (GTV) and originating hospital.

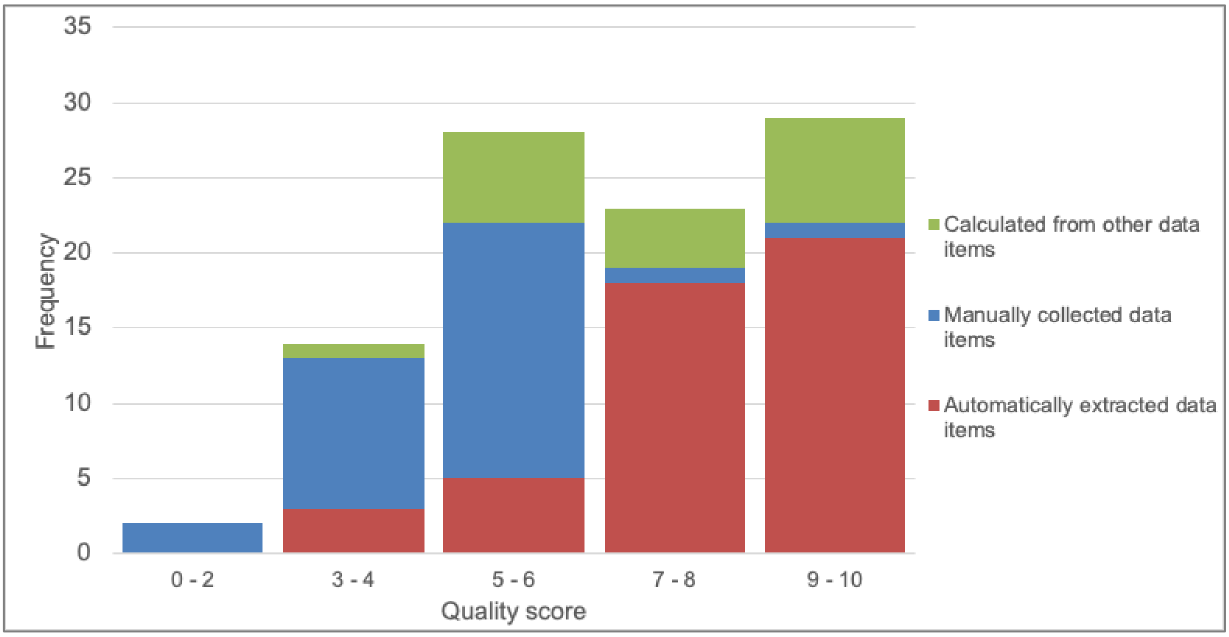


Figure 2-8. Distribution of data item quality scores, stratified by the type of data collection/extraction method.

The results from the additional evaluation of automatically extracted data are summarised in Table 2-2.

Table 2-2. Quality evaluation of automatically extracted data through comparison with manually collected data. Manually collected data were available from the combined data quality evaluation dataset (476 patients). The date of diagnosis varied across the research datasets that formed the data quality evaluation dataset and were therefore evaluated separately. Datasets 1, 2, and 3 included data from 133, 289, and 90 patients, respectively. DQED: Data quality evaluation dataset; N/A: Not applicable; Gy: Gray.

Data item	Manually collected data available in	Available data by both automatic extraction and manual collection	Agreement between automatically extracted and manually collected data	Differences between automatic and manual extraction	
Date of death	DQED	77	100%	No differences	
Histology	DQED	381	100%	No differences	
Histological grade	DQED	144	100%	No differences	
Prescribed radiotherapy fractions	DQED	330	92%	1 fraction	7
				2 fractions	3
				3 to 5 fractions	12
				More than 5 fractions	5
Prescribed dose to primary tumour	DQED	282	84%	Equal to or less than 1Gy	18
				Between 1.01Gy and 2Gy	4
				Between 2.01Gy and 5Gy	11
				More than 5Gy	19
Delivered radiotherapy fractions	DQED	246	95%	1 fraction	3
				2 fractions	1
				3 to 5 fractions	7
				More than 5 fractions	2
Delivered dose to primary tumour	DQED	246	1%	Equal to or less than 1Gy	32
				Between 1.01Gy and 2Gy	185
				Between 2.01Gy and 5Gy	7
				More than 5Gy	19
Chemotherapy regimen	DQED	119	100%	No differences	
Date of diagnosis	Dataset 1	101	0%	Equal to or less than 7 days	4
				Between 8 and 14 days	3
				Between 15 and 28 days	53
				More than 28 days	41
	Dataset 2	150	10%	Equal to or less than 7 days	19
				Between 8 and 14 days	31
				Between 15 and 28 days	46
				More than 28 days	39
	Dataset 3	88	1%	Equal to or less than 7 days	11
				Between 8 and 14 days	34
				Between 15 and 28 days	26
				More than 28 days	16

Overall, the automatically extracted data and the manually collected data in the data quality evaluation dataset are in perfect agreement for all three categorical data items (histology, histological grade, and chemotherapy regimen). The same applies for the date of death. For the number of prescribed fractions and the number of delivered fractions, there is a high rate of agreement between the two data collection methods (92% and 95%, respectively). For the date of diagnosis, low rates of agreement are observed between the automatically extracted data and the manually collected data in Dataset 1 (0%), Dataset 2 (10%) and Dataset 3 (0%). For the disagreements observed between automatically extracted and manually collected data through the comparison with all three datasets (n=323), diagnosis dates differ by less than 8 days in 11% of cases, between 8 and 28 days in 59% of cases, and by more than 28 days in 30% of cases. High rates of disagreement between each of the three research datasets are also prevalent, highlighting the issue of non-uniform definitions of date of diagnosis used across datasets. This will be discussed further later in this chapter. Low rates of agreement between the two data collection methods are also observed for the prescribed and delivered dose to primary tumour. Where the differences in prescribed and delivered dose between the automatic extraction and manual collection were large (more than 2Gy), the patient records in MOSAIQ were individually accessed by a researcher (ST) to determine whether the automatic extraction yielded incorrect data. The automatic extraction was incorrect in 15 cases for the prescribed dose and only 5 cases for the delivered dose. These errors were found in patients who had a replan, irrespective of the radiotherapy technique used. In these cases, the replan was not captured by the automatic extraction, and therefore the replanned prescribed and delivered doses were missing altogether. Only the doses from the original plan were registered. The data warehouse was updated following the manual review.

2.4.3 Data extraction for use in the atomCAT2 study

Data from the data warehouse were extracted for use in the atomCAT2 international multicentre study, which aims to develop prediction models for anal cancer outcomes through distributed learning [19]. The data warehouse was firstly filtered to identify patients treated with VMAT for primary anal cancer, for which manual data collection was attempted (n=246). Patients treated with palliative intent, patients who had prior pelvic radiotherapy or who were participants in the PLATO trial [20] were excluded (n=28). A further 21 patients were excluded due to missing data for essential data items

(See Chapters 5 and 6). A final list of 197 patients with complete data for essential data items was identified and formed the atomCAT2 study cohort from LCC. For these patients, a variety of baseline patient data were extracted from the data warehouse, including age at the start of radiotherapy, sex, baseline TNM staging, histology, primary tumour GTV, prescribed primary tumour dose, radiotherapy technique used for treatment, and chemotherapy regimen. Moreover, overall survival, locoregional control and freedom from distant metastasis data were also extracted. The resulting dataset was analysed in combination with anal cancer data from 11 other centres, in order to identify key prognostic factors for the outcomes that explored (Chapter 6). The models that were developed will allow for the prediction of outcomes in individual patients, which may inform current clinical practice and subsequently aid the stratification or personalisation of anal cancer treatment.

2.5 Discussion

We achieved our aim of developing a comprehensive data warehouse of anal cancer patients treated at Leeds Cancer Centre. This framework was developed with the potential to be updated and modified to expand the type of data included. The data warehouse development process initially encompassed the identification of 194 relevant data items and their respective data sources in collaboration with clinicians who treat anal cancer. A combination of automatic and manual data extraction was employed, with considerable quality assessment of initial data extraction as part of the data warehouse development. A process for data de-identification was implemented to facilitate access to clinicians and researchers on an ongoing basis for research and audit purposes. Further maintenance and data updating was designed to be as automated as possible.

Preserving patient data privacy was of utmost importance for this project. By taking advantage of the governance processes established in LCC through LeedsCAT, we ensured that only authorised individuals can access the data available in the data warehouse. Central to LeedsCAT is a philosophy of least needed data access, which was also adopted for this project. Compliance with Caldicott principles, IG and research governance is also embedded in the LeedsCAT approach. LeedsCAT expedites radiotherapy research by creating and maintaining reusable and sustainable project resources. By following the LeedsCAT principles, we aimed to support research activity at LCC that not only results in the improvement of clinical practice, but also promotes the improvement of clinical data quality at the source, thus improving the data quality for both future clinical practice and future research.

Numerous challenges arose throughout the data warehouse development and updating process. To begin with, agreeing on the relevant data items and subsequently documenting them effectively was a complex iterative task that required input from a multidisciplinary team. Notably, research projects that utilise retrospective patient data often overlook the importance of fully defining relevant data items that are analysed [21,22]. This is one aspect in which RCTs excel at and that we aimed to address within this project. Many data items included in the warehouse, including TNM staging, are dynamic and may be updated multiple times when additional diagnostic procedures (e.g. additional imaging) are undertaken. For such data items, a robust definition had to be pre-specified in the data dictionary to ensure that data collection across patients is consistent. For the purpose of the data warehouse, the baseline TNM staging was defined as the final staging at the point of radiotherapy initiation. For all patients, baseline TNM staging data were manually collected from the last clinic letter before the beginning of radiotherapy. Upon generating a comprehensive data dictionary, the identification of data sources posed a significant challenge for a range of data items. For instance, the more accurate baseline TNM staging and outcome data were found in free-text form within clinic letters (PPM). While TNM staging is documented in PPM, the coded staging is often inaccurate, as is it is not systematically updated once the final diagnostic imaging is reviewed. For response assessment outcome data, this information may only be found in free-text imaging reports or summarised within clinic letters, and therefore cannot be readily extracted in an automatic fashion. This renders the continuous updating of the data warehouse for existing and new patients difficult, unless additional structures are introduced to ensure consistent recording in a single clinical system.

Despite the comprehensive data dictionary and data warehouse developed, there are several limitations. Firstly, the patient identification procedure leads to a small percentage of false positive hits. These most likely result from miscoding of the radiotherapy treatment site during manual data entry. In order to detect these false positive hits, a manual review of the patient list generated by the search may need to be undertaken. This can be carried out during the manual data collection, since accessing patient records can confirm whether the patient was treated for anal cancer or for a different cancer. However, there is no way of identifying false negative hits, as in, patients that were treated for anal cancer but were misclassified as being treated for a different type of cancer. Estimating how many patients with anal cancer were not classified as such would be challenging.

In terms of the data collection process, this would ideally be carried out in a fully automated fashion for all data items. However, this is currently not possible. In order to render the automated extraction of data from patient record systems (e.g. PPM) possible, access to the system's backend would be required and natural language processing techniques would need to be implemented to extract the data [23]. Another barrier to fully automating the data collection process is that, at present, the primary tumour GTV can only be manually collected from the radiotherapy treatment plans in Monaco. To address this, scripting could be employed to handle the automatic data extraction of these data from DICOM [24] treatment plan stores.

The current version of the data warehouse does not include data for a large number of data items, such as comorbidity and treatment toxicity data items. These data are currently not consistently recorded but could potentially be manually collected from clinic letters in PPM. However, there are often more than 10 clinic letters available for each patient, and all these would need to be reviewed comprehensively to extract all relevant data. To automate the data collection process for these data items, complex natural language processing approaches would need to be employed [25]. Lastly, the HPV status of the majority of patients included in the data warehouse is missing, despite this being an important factor that is commonly analysed in anal cancer research [26–28]. This is due to the status not being routinely assessed in LCC at present.

The data quality evaluation process has also identified several aspects that could be improved. Firstly, the date of diagnosis appears to vary considerably between the automatically extracted data and the manually collected data. There are also substantial discrepancies in diagnosis dates between the research datasets that were used to carry out the data quality evaluation. This highlights that the definition of this data item is inconsistent throughout the whole patient pathway and that currently there is no single standard way of defining it. Additional variation can be introduced when these data are manually collected by different individuals, depending on the timing of the data collection and where the data are sourced from. Specifically, data for the diagnosis date that were automatically extracted from MOSAIQ were added to the data warehouse. This date of diagnosis reflects the date that the radiotherapy treatment referral form was added to the system and therefore it is not defined as the clinical diagnosis date. Alternative definitions for the diagnosis date include the date of pathological diagnosis, the date of MDT review, and the date of the first clinic letter, among others. As a result, the automatically extracted data that were included in the data warehouse are not the ground truth. Despite this, by automatically extracting date of diagnosis data from a single source, we ensured that the data item is fully standardised and consistent across all patients. The delivered radiotherapy dose also differs between the automatically extracted and manually collected data, with the majority of differences being less than 2Gy. These small differences can be attributed to rounding errors during manual data collection or may be linked to MOSAIQ recording the dose in the dose specification point for each beam as the per-fraction dose delivered, which differs from the mean dose to the entire volume. This was a historical issue relating to how information about doses was transferred from Monaco to MOSAIQ, which has been resolved since 2020. Large differences between the automatically extracted and manually collected data in both prescribed and delivered doses (more than 2Gy) were observed in cases where patients had a replan or a two-phase plan that was not recorded correctly, and therefore not captured by the automatic extraction. All large discrepancies were manually reviewed to ensure that the final data added to the warehouse were correct.

Plans have been devised with the aim of improving numerous aspects of the data warehouse in the future. Firstly, we are hoping to develop a pipeline that automatically updates the data warehouse at regular intervals. This would involve updating the existing patients with more up-to-date data, as well as incorporating data for newly diagnosed patients. In order to achieve this, the data collection methodology needs to be as automated as possible. Therefore, access to PPM backend needs to be obtained in order to be able to automatically extract data. In order to extract data items that are only available in free-text form, we aim to test and implement natural language processing algorithms. Lastly, we aim to develop an automated system that handles the automatic extraction of GTV values from radiotherapy treatment plans. The next step would be to establish a workflow to prospectively collect a core set of data for all new patients diagnosed with anal cancer. This would involve collaborating with the Leeds clinical oncology team to identify which data items are essential, and to generate an appropriate and user-friendly data input approach. Future work on the data warehouse could also involve expanding the data collection to other cancer sites and incorporating more complex data which are not currently available, such as nuclear medicine data, patient reported outcome data, as well as radiotherapy planning and imaging data. The data warehouse could also be linked to other existing anal cancer research at LCC, including linkage to tissue and blood immune signatures and predictive biomarkers. This local data warehouse could also be potentially linked with national or international anal cancer data warehouses and registries to carry out even more robust analyses. Lastly, an intuitive user interface could be developed, including a customisable dashboard. This would allow researchers and clinicians accessing the data warehouse to quickly get a high-level overview of the data they need.

In conclusion, routine data from patients treated for anal cancer in Leeds Cancer Centre were collected from multiple sources, using automatic data extraction and manual data collection approaches, and were collated into a comprehensive data warehouse. Researchers and clinicians at LCC can now access the available high quality, de-identified data, which are ready for analysis in future research studies (Chapters 4 and 6) and in local quality improvement work. Ultimately, the insights obtained by analysing these data may help advance clinical management of patients diagnosed with anal cancer in the future.

2.6 References

[1] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. Can Med Assoc J 2016;188:E158–64. https://doi.org/10.1503/cmaj.150653.

[2] Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16:495. https://doi.org/10.1186/s13063-015-1023-4.

[3] Rothwell PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?” The Lancet 2005;365:82–93. https://doi.org/10.1016/S0140-6736(04)17670-8.

[4] Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. Npj Digit Med 2020;3:67. https://doi.org/10.1038/s41746-020-0277-8.

[5] He J, Morales DR, Guthrie B. Exclusion rates in randomized controlled trials of treatments for physical conditions: a systematic review. Trials 2020;21:228. https://doi.org/10.1186/s13063-020-4139-0.

[6] Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review. JAMA 2007;297:1233. https://doi.org/10.1001/jama.297.11.1233.

[7] Arlett P, Kjær J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. Clin Pharmacol Ther 2022;111:21–3. https://doi.org/10.1002/cpt.2479.

[8] Ferrão J, Oliveira M, Janela F, Martins H. Preprocessing structured clinical data for predictive modeling and decision support: A roadmap to tackle the challenges. Appl Clin Inform 2016;07:1135–53. https://doi.org/10.4338/ACI-2016-03-SOA-0035.

[9] Goldacre B, Morley J. Better, Broader, Safer: Using Health Data for Research and Analysis. A review commissioned by the Secretary of State for Health and Social Care. Department of Health and Social Care.; 2022.

[10] Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. Int J Epidemiol 2016:dyw276. https://doi.org/10.1093/ije/dyw276.

[11] Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol 2014;40:1165–76. https://doi.org/10.1016/j.ejso.2014.07.030.

[12] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[13] Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, et al. Prognostic factors for patients with anal cancer treated with conformal radiotherapy—a systematic review. BMC Cancer 2022;22:607. https://doi.org/10.1186/s12885-022-09729-4.

[14] Fish R, Sanders C, Adams R, Brewer J, Brookes ST, DeNardo J, et al. A core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC): a patient and health-care professional consensus. Lancet Gastroenterol Hepatol 2018;3:865–73. https://doi.org/10.1016/S2468-1253(18)30264-4.

[15] The Leeds Teaching Hospitals NHS Trust. Patient Pathway Manager n.d. https://www.leedsth.nhs.uk/ppm/ (accessed October 25, 2022).

[16] Elekta. Monaco - High-precision treatment planning for radiation therapy n.d. https://www.elekta.com/products/radiation-therapy/monaco/ (accessed October 25, 2022).

[17] Elekta. MOSAIQ for Medical Oncology n.d. https://www.elekta.com/products/oncology-informatics/mosaiq-plaza/medical-oncology/ (accessed October 25, 2022).

[18] ChemoCare - Innovative electronic chemotherapy prescribing and patient management n.d. https://www.cis-healthcare.com/chemocare/ (accessed October 25, 2022).

[19] Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Dennis K, et al. Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study. Diagn Progn Res 2022;6:14. https://doi.org/10.1186/s41512-022-00128-8.

[20] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[21] Shenvi EC, Meeker D, Boxwala AA. Understanding data requirements of retrospective studies. Int J Med Inf 2015;84:76–84. https://doi.org/10.1016/j.ijmedinf.2014.10.004.

[22] Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol 2021;157:1362. https://doi.org/10.1001/jamadermatol.2021.3129.

[23] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 2022. https://doi.org/10.1007/s11042-022-13428-4.

[24] Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. J Am Med Inform Assoc 1997;4:199–212. https://doi.org/10.1136/jamia.1997.0040199.

[25] Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inform Assoc 2011;18:539–539. https://doi.org/10.1136/amiajnl-2011-000501.

[26] Daling JR, Madeleine MM, Johnson LG, Schwartz SM, Shera KA, Wurscher MA, et al. Human papillomavirus, smoking, and sexual practices in the etiology of anal cancer. Cancer 2004;101:270–80. https://doi.org/10.1002/cncr.20365.

[27] Gami B, Kubba F, Ziprin P. Human Papilloma Virus and Squamous Cell Carcinoma of the Anus. Clin Med Insights Oncol 2014;8:CMO.S13241. https://doi.org/10.4137/CMO.S13241.

[28] Balermpas P, Martin D, Wieland U, Rave-Fränk M, Strebhardt K, Rödel C, et al. Human papilloma virus load and PD-1/PD-L1, CD8+ and FOXP3 in anal cancer patients treated with chemoradiotherapy: Rationale for immunotherapy. OncoImmunology 2017;6:e1288331. https://doi.org/10.1080/2162402X.2017.1288331.

Chapter 3 - Prognostic factors for patients with anal cancer treated with conformal radiotherapy - a systematic review

3.1 Abstract

3.1.1 Aims

Anal cancer is primarily treated using concurrent chemoradiotherapy (CRT), with conformal techniques such as intensity modulated radiotherapy (IMRT) and volumetric arc therapy (VMAT) now being the standard techniques utilised across the world. Despite this, there is still very limited consensus on prognostic factors for outcome following conformal CRT. This systematic review aims to evaluate the existing literature to identify prognostic factors for a variety of oncological outcomes in anal cancer, focusing on patients treated with curative intent using contemporary conformal radiotherapy techniques.

3.1.2 Materials and methods

A literature search was conducted using Medline and Embase to identify studies reporting on prognostic factors for survival and cancer-related outcomes after conformal CRT for anal cancer. The prognostic factors which were identified as significant in univariable and multivariable analysis, along with their respective factor effects (where available) were extracted. Only factors reported as prognostic in more than one study were included in the final results.

3.1.3 Results

The results from 19 studies were analysed. In both univariable and multivariable analysis, N stage, T stage, and sex were found to be the most prevalent and reliable clinical prognostic factors for the majority of outcomes explored. Only a few biomarkers have been identified as prognostic by more than one study – pre-treatment biopsy HPV load, as well as the presence of leukocytosis, neutrophilia and anaemia at baseline measurement. The results also highlight the lack of studies with large cohorts exploring the prognostic significance of imaging factors.

3.1.4 Conclusion

Establishing a set of prognostic and potentially predictive factors for anal cancer outcomes can guide the risk stratification of patients, aiding the design of future clinical trials. Such trials will in turn provide us with greater insight into how to effectively treat this disease using a more personalised approach.

3.2 Background

First reported in 1974 by Nigro et al. [1] and established by two phase III trials [2, 3], concurrent chemoradiotherapy (CRT) is the current standard of care for localised anal squamous cell carcinoma (ASCC). The introduction of three-dimensional conformal radiotherapy (3D-CRT), intensity modulated radiotherapy (IMRT) and latterly volumetric arc therapy (VMAT) [4] has allowed for substantial reduction in dose to pelvic organs at risk (OAR) and associated toxicity, with far fewer unplanned treatment breaks as a result. The current UK standard for anal cancer comprises of IMRT/VMAT and concurrent chemotherapy with 5-fluorouracil (5-FU) or capecitabine and mitomycin C (MMC), with surgery reserved as salvage treatment [5].

Anal cancer is a rare cancer, and only a handful of late phase clinical trials have been conducted over the last four decades [2, 3, 6–9]. Other than the single arm phase II RTOG 0529 [10] trial, these trials were conducted prior to widespread adoption of conformal radiotherapy techniques, such as 3D-CRT or IMRT/VMAT. Similarly, much of the published literature on prognostic factors in anal cancer consists of retrospective series, often small cohorts [11, 12] or cohorts of patients treated with older techniques [13, 14]. No systematic review of studies identifying prognostic factors after treatment with conformal radiotherapy has previously been conducted.

Despite advances in radiotherapy planning and delivery, locoregional control remains challenging, and patients usually fail locoregionally before getting metastatic disease. A UK multi-centre retrospective review by Shakir et al. [15] analysed 385 anal cancer patients treated with contemporary radiotherapy techniques, and demonstrated a 85.6% three-year overall survival. Initial complete clinical response rates were high at 86.7%, but over time 24.4% of patients relapsed, with the majority of relapses (83.4%) being local.

Establishing risk factors for oncological outcomes, in particular locoregional control following conformal chemoradiotherapy, could help optimise future treatment strategies and aid in the design and analysis of new clinical trials [16]. A consensus on prognostic factors could inform research by determining specific patient risk groups and the development of personalised treatment approaches, tailored to individual patient characteristics [17], and/or the introduction of novel agent combinations. This systematic review evaluates the literature to identify prognostic factors for a variety of disease-related outcomes in anal cancer, focusing on patients treated with curative intent using conformal radiotherapy techniques and contemporary treatment schedules.

3.3 Methods

A systematic review was undertaken according to PRISMA 2020 [18]. A comprehensive literature search was conducted using the Medline and Embase databases, to identify studies reporting on prognostic factors for survival and cancer-related outcomes after conformal chemoradiotherapy for anal cancer. The search terms included ‘radiotherapy’ AND ‘anal cancer’ AND ‘prognostic factor’, as well as related terms (see Supplementary material A - Section 3.8.1, for the full search strategies). Only studies published after 1st January 2000 and up to and including 30th June 2020 were considered. An initial scoping search showed that no studies conducted prior to 2000 had a majority of patients treated using conformal techniques.

Studies were included if they: (1) comprised of at least 70% of patients treated with solely conformal radiotherapy techniques (3D-defined targets on CT, beams conformed to targets e.g. using multileaf collimators, 3D dose calculation and dose distribution optimisation), (2) reported survival or disease-related outcomes and (3) examined prognostic factors for outcomes using univariable (UVA) or multivariable (MVA) analysis. Studies were excluded if (1) patients were treated with 2D radiotherapy techniques and/or fields based solely on bony landmarks, if (2) cohorts included less than 100 patients or (3) were derived from population-level databases, or if (4) treatment with palliative intent. The cut-off of 100 patients was chosen to ensure that the prognostic factors identified are generalisable and to decrease the likelihood of identifying spurious prognostic factors from studies that suffer from small sample size bias. All (5) meta-analysis studies, reviews, animal model studies, conference abstracts/letters and studies without English translation were also excluded.

Two independent reviewers (ST and RS) screened and reviewed all relevant articles. A third independent reviewer (AA) assisted in reconciling differences in cases of disagreement. One reviewer (ST) extracted and analysed data from all relevant articles, including: study location, publication year, study design, source of participants, participant selection criteria, number of patients included, treatment period, radiotherapy technique administered, radiotherapy schedule, chemotherapy regimen, follow-up procedure, core clinical/patient characteristics, outcomes reported/definitions, statistical analysis used, prognostic variables tested, prognostic variables identified as significant and corresponding effect estimates. An independent reviewer (RS) repeated the data extraction from a subset (20%) of all relevant articles to ensure that the data extraction process was reproducible. The methodological quality of all relevant articles was assessed independently by two reviewers (ST, RS) using the National Institutes of Health (NIH) Quality Assessment Tool for Case Series Studies [19]. Any disagreements were reviewed independently by a third reviewer (AA) to achieve consensus.

Reported outcomes and outcome definitions were extracted from each study and stratified into nine categories for further analysis. Disease activity and survival outcomes were firstly grouped according to the CORMAC review [20], which was used as the initial reporting framework for outcome stratification. Additional categories were inductively derived after the data extraction process.

For each study, factors analysed for their prognostic impact were extracted, whether they were shown to have a significant relationship with outcome, and the statistical method used for analysis. The factors were grouped into three broader categories: clinical factors, biomarkers and imaging factors. The total number of times a factor was tested in UVA for each of the nine outcomes was counted across all studies. Where factors tested were not reported explicitly, it was assumed that all reported patient characteristics were tested. Prognostic factors which were identified as significant in each study, along with their respective factor effect in the form of hazard ratios (HRs) were extracted (where available), and the proportion of times each factor was identified as prognostic for each outcome was calculated. Since the majority of studies did not report which factors were tested in MVA for each distinct outcome, the total number of times each factor was tested could not be counted. Therefore, only the prognostic factors and their respective factor effects were extracted. Only factors reported as prognostic in more than one study were included in the final results.

3.4 Results

3.4.1 Literature search

1567 studies published between 1st January 2000 and 30th June 2020 were identified, 404 of which were duplicates. Titles and abstracts of 1163 unique studies were screened. 1021 were excluded and the final 142 studies assessed for eligibility, of which 123 were excluded after reviewing the full text. 48 studies employed non-conformal radiotherapy techniques in more than 30% of patients. Other main factors for exclusion were sample size less than 100 (n=29) and incomplete reporting on the radiotherapy technique (n=21). Ultimately, 19 studies [15, 21–38] were included in this literature review (Figure 3-1).

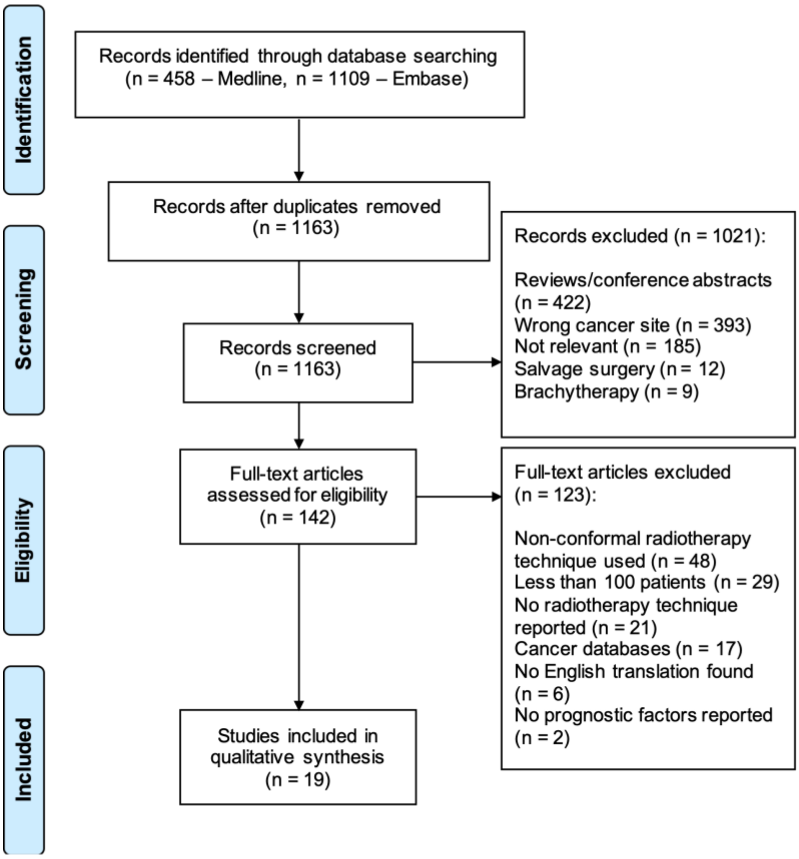

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 458 – Medline, n = 1109 – Embase)
Records after duplicates removed (n = 1163)
Records screened (n = 1163)
Records excluded (n = 1021):
Reviews/conference abstracts (n = 422)
Wrong cancer site (n = 393)
Not relevant (n = 185)
Salvage surgery (n = 12)
Brachytherapy (n = 9)
Full-text articles assessed for eligibility (n = 142)
Full-text articles excluded (n = 123):
Non-conformal radiotherapy technique used (n = 48)
Less than 100 patients (n = 29)
No radiotherapy technique reported (n = 21)
Cancer databases (n = 17)
No English translation found (n = 6)
No prognostic factors reported (n = 2)
Studies included in qualitative synthesis (n = 19)

Figure 3-1. PRISMA flow diagram depicting the number of studies that were identified, included and excluded, and the reasons for exclusion.

3.4.2 Study characteristics

Included studies were retrospective case series (n=19), either single institutional (n=10) or multi-institutional (n=9). Patients were treated between 1989-2018 with median follow-up range of 14.9-70.0 months. The most common radiotherapy techniques employed were a combination of 3D-CRT and IMRT/VMAT (n=9), followed by IMRT only (n=6). Dose ranged from 45Gy/25 fractions to 63Gy/35 fractions and chemotherapy regimens were mainly MMC and 5-FU based, with three studies including the option of cisplatin. Statistical techniques for UVA were log-rank tests (n=12) and univariable Cox regression (n=9), with four studies using both. Multivariable Cox regression was applied for MVA in all but one study, which used logistic regression instead. Regarding quality, 16 were deemed good and three deemed fair (Supplementary material B, Section 3.8.2). A short follow-up (of less than 36 months, as used for the primary endpoint in the PLATO trial [17]) was a common issue in eight studies. Due to the lack of universal reporting of effect sizes for prognostic factors, it was not possible to carry out a meta-analysis on the data. Table 3-1 presents the main characteristics for all included studies (Supplementary material C, Section 3.8.3, presents a more detailed version including information on cancer subtype and location in the included cohorts, TNM staging version used and all predictors tested).

3.4.3 Outcomes

Outcome definitions varied considerably. Supplementary material D, Section 3.8.4 presents the definitions extracted from each study and how they were categorised. Nine outcome categories were used: three disease activity (freedom-from-disease, locoregional failure (LRF) and distant failure) as well as six survival categories (overall survival (OS), disease-free survival (DFS), colostomy-free survival (CFS), cancer-specific survival, local failure-free survival and metastasis-free survival (MFS)). Disease-free survival and progression-free survival were grouped together, as definitions overlapped in most papers. Local and regional failures were grouped with locoregional failures, due to the small number of studies reporting only on the latter. Freedom-from-disease, a category which was not included in CORMAC, was devised in order to include definitions of time-to-recurrence, time-to-failure (not specified as local, regional or distant) and disease-free survival where death was not considered an event. Commonly investigated outcomes were OS (n=17), LRF (n=11) and DFS (n=11). Supplementary material E, Section 3.8.5 lists all outcomes reported, along with all factors tested.

Table 3-1. Overview of study characteristics, including treatment techniques and regimens. ***** *used to differentiate between two studies by the same author published in the same year. MC: multi-centre. SC: single-centre. EU: Europe. NA: North America. IN: International. NR: not reported. Gy: Gray. MMC: mitomycin C. Cap: capecitabine. 5-FU: 5-fluorouracil. Cisp: cisplatin. UV: univariable. BV: bivariable. MV: multivariable. Cox: Cox regression. Log-rank: log-rank statistical test.*

#	Study	Location	Number of patients	Years of treatment	Radiotherapy technique	Radiotherapy regimen	Chemotherapy regimen	Median follow-up (months)	Type of statistical analysis used	Quality
1	Shakir et al. (2020) [15]	MC, EU	385	2013-2018	IMRT	50.4Gy/28 fractions for T1/2N0, 53.2Gy/28 fractions for T1/2N+ or T3/4Nany	MMC and Cap or 5-FU	24.0	UV Cox, MV Cox	Good
2	Martin et al. (2020) [21]	SC, EU	223	1996-2017	3D-CRT (58%) IMRT (42%)	50–50.4Gy in 1.8–2Gy/fraction, boost of 5.4–9 Gy	5-FU and MMC or Cisp	46.0	UV Cox, MV Cox	Good
3	de Bellefon et al. (2020) [22]	SC, EU	193	2005-2017	IMRT	45Gy in 1.8Gy/fraction, boost of 14.4–20Gy (1.8–2 Gy/fraction)	5-FU and MMC	70.0	UV Cox, MV Cox	Good
4	Brown et al. (2019) [23]	SC, EU	189	2008-2016	2D/ 3D-CRT (79%) VMAT (21%)	49.6Gy in 1.8Gy/fraction	5-FU and MMC	35.1	MV logistic	Good
5	Rouard et al. (2019) [24]	MC, EU	165	2006-2016	IMRT	45–50Gy in 1.8 or 2Gy/fraction, boost of 15–20 Gy	5-FU and MMC	33.8	BV Cox, MV Cox	Good
6	Franco et al. (2018) [25]	MC, EU	161	NR	IMRT	50–50.4Gy in 1.8–2Gy/fraction	5-FU and MMC	27.0	Log-rank, UV Cox, MV Cox	Good
7	Call et al. (2016) [26]	MC, NA	152	NR	IMRT	51.25Gy/28 fractions	5-FU and MMC (75% of patients)	26.8	Log-rank, MV Cox	Fair
8	Balermpas et al. (2017) [27]	MC, EU	150	NR	3D-CRT IMRT	53.4Gy in 1.8–2Gy/fraction	5-FU and MMC	40.0	Log-rank, MV Cox	Good
9	Rodel et al. (2018) [28]	MC, EU	140	NR	3D-CRT IMRT	53.4Gy (range 46.8–64.8Gy)	5-FU and MMC	40.0	Log-rank, MV Cox	Good

10	Schernberg et al. (2017) [29]	MC, EU	133	2000-2015	IMRT (77%) 3D-CRT (23%)	49.5Gy/30 fractions (centre 1), 45Gy/25 fractions (centre 2)	Cisp and 5-FU or Cap / MMC and 5-FU or Cap	37.4	Log-rank, MV Cox	Good
11	Martin et al. (2019) [30]	SC, EU	126	2004-2016	IMRT (65%) 3D-CRT (35%)	59.4Gy in 1.8 or 2Gy/fraction	5-FU and MMC	NR	Log-rank, UV Cox, MV Cox	Good
12	Oehler-Janne et al. (2008) [31]	MC, IN	121	1997-2006	3D-CRT	52Gy-60Gy depending on centre	5-FU and MMC or Cisp	36.0	Log-rank, UV Cox, MV Cox	Good
13	Susko et al. (2020) [32]	SC, NA	111	2005-2018	3D-CRT IMRT	55.8Gy/30 fractions	5-FU and MMC	28.0	Log-rank, UV Cox, MV Cox	Good
14	Cardenas et al. (2017) [33]	SC, NA	110	2003-2013	IMRT (75%) 2D-CRT (25%)	50.4Gy/28 fractions for T2N0, 54Gy/30 fractions for T3/4Nany	5-FU and MMC	28.6	UV Cox, MV Cox	Fair
15	Bitterman et al. (2015) [34]	SC, NA	109	2004-2013	IMRT (60%) 3D-CRT (40%)	45Gy+ in 1.8Gy/fraction	5-FU and MMC	14.9	Log-rank, MV Cox	Good
16	Fraunholz et al. (2013) [35]	MC, EU	103	1989-2011	3D-CRT	50.4Gy in 1.8 or 2Gy/fraction	5-FU and MMC	44.0	Log-rank, MV Cox	Good
17	Schernberg et al. (2017)* [36]	SC, EU	103	2006-2016	IMRT (53%) 3D-CRT (47%)	45Gy/25 fractions of 1.8Gy or 44Gy/22 fractions of 2Gy	5-FU and MMC or Cap	38.7	Log-rank, MV Cox	Good
18	Hosni et al. (2018) [37]	SC, NA	101	2008-2013	IMRT	45Gy/25 fractions for T1N0, 54Gy/30 fractions for T1/2N+ or 63Gy/35 fractions for T3/4Nany	5-FU and MMC	56.5	UV Cox, MV Cox	Fair
19	Oblak et al. (2016) [38]	SC, EU	100	2003-2013	3D-CRT IMRT	45Gy/25 fractions	5-FU and MMC or Cap	52.0	Log-rank, MV Cox	Good

3.4.4 Clinical prognostic factors

Table 3-2 presents clinical factors identified as prognostic for each outcome in more than one study, categorised by UVA and MVA. For prognostic factors identified in MVA, the range of factor effects (HRs) across studies is also reported. Eight unique prognostic factors were established by more than one study in UVA and seven in MVA (See Supplementary material F, Section 3.8.6 for full results).

*Table 3-2. Clinical factors identified as prognostic for worse outcomes by more than one study. These clinical factors were identified through univariable and multivariable analysis, and were stratified by outcome. A number of studies reported on "gender", however this was analysed in conjunction with "sex" throughout the study, since "sex" is used when reporting on biological factors instead of gender identity, or psychosocial or cultural factors. HR: Hazard Ratio. N/A: Not available. *Factor effects (HRs) were provided by only one study for this prognostic factor, therefore the effect range could not be determined.*

Univariable analysis				
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Total times tested**	**Studies which identified factor as prognostic**
Overall survival (n=17)	Higher N stage	10	16	[15],[21],[22],[25],[26],[27],[28],[35],[36],[38]
	Higher T stage	9	16	[15],[21],[22],[27],[28],[35],[36],[37], [38]
	Male sex	7	12	[15],[21],[25],[27],[28],[29],[37]
	Worse performance status	3	4	[15],[29],[38]
	Older age	3	4	[24],[27],[37]
	Incomplete/interrupted RT or breaks	2	2	[15],[24]
	Longer CRT duration	2	5	[36],[38]
Locoregional failure (n=11)	Higher N stage	7	11	[15],[21],[26],[27],[28],[30],[38]
	Higher T stage	7	11	[15],[21],[26],[27],[28],[32],[38]
	Male sex	5	9	[15],[21],[27],[28],[29]
	Worse performance status	4	4	[15],[24],[29],[38]
	Longer CRT duration	2	2	[32],[38]
Disease-free survival (n=11)	Male sex	5	8	[21],[27],[29],[30],[37]
	Higher N stage	4	9	[21],[22],[27],[30]
	Higher T stage	4	10	[21],[22],[28],[37]

Metastasis-free survival (n=5)	Higher T stage	5	5	[21],[22],[30],[35],[36]
	Higher N stage	4	4	[21],[30],[35],[36]
	Male sex	2	4	[21],[30]
Freedom from disease (n=4)	Higher N stage	4	4	[15],[28],[31],[38]
	Male sex	2	3	[15],[28]
	Higher T stage	2	3	[15],[38]
Colostomy-free survival (n=4)	Higher T stage	3	4	[22],[26],[37]
Cancer-specific survival (n=3)	Higher T stage	2	3	[35],[38]
	Higher N stage	2	3	[35],[38]
Multivariable analysis				
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Factor effect range (HR)**	**Studies which identified factor as prognostic**
Overall survival (n=17)	Male sex	7	1.92 – 4.80	[15],[21],[25],[27],[28],[29],[37]
	Higher T stage	3	2-88 – 4.98	[22],[34],[37]
	Older age	3	1.05 – 2.43	[24],[37],[37]
	Higher N stage	3	1.88 – 5.80	[25],[26],[36]
	Higher AJCC stage	2	2.23 – 2.82	[22],[38]
Locoregional failure (n=11)	Male sex	4	2.08 – 3.40	[15],[21],[27],[29]
	Higher N stage	3	2.23 – 3.58	[15],[21],[30]
	Incomplete/interrupted RT or breaks	2	2.47 – 4.96	[15],[22]
	Worse performance status	2	3.82 – 5.50	[24],[29]
Disease-free survival (n=11)	Male sex	4	2.13 – 3.60	[21],[27],[29],[37]
	Higher T stage	3	2.57 – 7.02	[22],[23],[37]
	Higher N stage	2	N/A*	[21],[23]
Metastasis-free survival (n=5)	Male sex	2	3.87 – 4.08	[21],[23]
	Higher T stage	2	2.61 – 3.54	[21],[22]
	Higher N stage	2	2.41 – 4.49	[21],[30]
Freedom from disease (n=4)	Male sex	2	2.16 – 2.16	[15],[28]
Colostomy-free survival (n=4)	Higher T stage	3	3.65 – 4.10	[22],[26],[37]

In UVA, T stage, N stage and sex were the most commonly tested factors for all seven outcomes for which prognostic factors were identified (Table 3-1). T stage was prognostic for all outcomes; in 56% of the studies that tested it for OS, in 64% for LRF, in 40% for DFS, in 100% for MFS, in 67% for freedom-from-disease, in 75% for CFS and in 67% for cancer-specific survival. Similarly, N stage was prognostic for six of seven outcomes. It was prognostic in 63% of the studies testing for OS, in 64% for LRF, in 44% for DFS, in 100% for MFS, in 100% for freedom-from-disease and in 67% for cancer-specific survival. The third most identified prognostic factor in UVA was sex. It was prognostic for five of the seven outcomes, in 58% of the studies that tested it for OS, in 56% for LRF, in 63% for DFS, in 50% for MFS and in 67% for freedom-from-disease. Performance status was also identified as prognostic in 75% of the studies that tested it for OS, and in 100% of studies that tested it for LRF.

In MVA, sex retained its prognostic significance, appearing as the predominant prognostic factor for six of the seven outcomes, altogether identified in nine studies [15, 21, 22, 25, 27–29, 35, 37]. Other commonly identified prognostic factors included higher T stage (OS, DFS, MFS and CFS; identified in seven studies [21–23, 26, 28, 34, 37]) and higher N stage (OS, LRF, DFS, MFS; identified in seven studies [15, 21, 23, 25, 26, 30, 36]). The rest of the factors were identified as prognostic for a single outcome only; age and AJCC stage for OS, as well as incomplete/interrupted radiotherapy and performance status for LRF.

3.4.5 Biomarkers and imaging prognostic factors

A smaller number of studies (n=8) examined the prognostic significance of biomarkers [25, 27–30, 35, 36, 38]. Only four unique biomarkers were deemed prognostic overall by more than one study in both UVA and MVA (Table 3-3 and Supplementary material G, Section 3.8.7).

In UVA, HPV16 load from pre-treatment biopsies was found to be prognostic for OS (2/3 – 67% of studies [27, 28]) and for LRF (2/3 – 67% of studies [27, 28]), whereas the presence of baseline neutrophilia (circulating blood neutrophil count of more than 7500/mm3 in one study and more than 7G/L in the second study) was found to be prognostic for OS (2/2 – 100% of studies [29, 36]) and DFS (2/2 – 100% of studies [29, 36]). Additionally, baseline anaemia (haemoglobin count <13g/dL) was deemed prognostic for OS only (2/2 – 100% of studies [29, 36]) and the presence of baseline leukocytosis markers (white blood cell count >10000/mm3 in one study and more than 10G/L in the second study) for DFS only.

Table 3-3. Biomarkers identified as prognostic for worse outcomes by more than one study. These biomarkers were identified through univariable and multivariable analysis and were stratified by outcome. HPV: human papillomavirus. HR: Hazard ratio.

Univariable analysis				
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Total times tested**	**Studies which identified factor as prognostic**
Overall survival (n=17)	Lower HPV16 load	2	3	[27],[28]
	Neutrophilia	2	2	[29],[36]
	Anaemia	2	2	[29],[36]
Locoregional failure (n=11)	Lower HPV16 load	2	3	[27],[28]
Disease-free survival (n=11)	Leukocytosis	2	2	[29],[36]
	Neutrophilia	2	2	[29],[36]
Multivariable analysis				
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Factor effect range (HR)**	**Studies which identified factor as prognostic**
Overall survival (n=17)	Leukocytosis	2	4.60 – 19.90	[29],[36]
	Neutrophilia	2	4.40 – 22.70	[29],[36]
Locoregional failure (n=11)	Lower HPV16 load	2	3.57 – 4.51	[27],[28]
Disease-free survival (n=11)	Leukocytosis	2	6.90 – 7.10	[29],[36]
	Neutrophilia	2	5.00 – 7.60	[29],[36]
	Anaemia	2	2.50 – 5.30	[29],[36]

In MVA, baseline neutrophilia retained its prognostic significance for both OS (two studies [29, 36]) and DFS (two studies [29, 36]), whereas HPV16 load retained its prognostic significance for LRF (two studies [27, 28]) only. Baseline leukocytosis was found to be prognostic for DFS (two studies [29, 36]) and for OS (two studies [29, 36]). Lastly, baseline anaemia was identified as prognostic for DFS (two studies [29, 36]) only.

Only two studies [23, 33] investigated imaging-related prognostic factors. In UVA, one study [33] identified post-treatment PET-CT SUVmax (positron emission tomography and computed tomography maximum standardized uptake value) and change in SUVmax (pre- vs. post-treatment) to be prognostic for OS. The pre-treatment and post-treatment SUVmax values were both found to be prognostic for local failure-free survival. In MVA, the post-treatment SUVmax and the change in SUVmax retained prognostic significance for OS. In the second study [23], a selection of radiomics markers were identified as prognostic for DFS (Supplementary material H, Section 3.8.8). For local failure-free survival, only the high post-treatment SUVmax was deemed prognostic in MVA (Supplementary material H, Section 3.8.8).

3.5 Discussion

This systematic review summarises the findings from studies examining prognostic factors for anal cancer outcomes following CRT with contemporary conformal radiotherapy techniques. By limiting our findings to studies with cohorts treated with conformal radiotherapy techniques, we aimed to ensure that the prognostic factors identified are the most informative to current practice and are representative of the more prevalent HPV-driven biology and the higher survival rates which have been observed in the past few years. N stage, T stage, and sex were established as the most prevalent and reliable clinical prognostic factors for the majority of outcomes explored, in both UVA and MVA. Few biomarkers have been identified as prognostic by more than one study: pre-treatment biopsy HPV load, as well as the presence of leukocytosis, neutrophilia and anaemia at baseline measurement. The review also highlighted the lack of studies with large cohorts exploring the prognostic significance of imaging factors.

Due to the rarity of anal cancer, only few randomised prospective clinical trials have been conducted to date; none of which have employed conformal radiotherapy techniques and reported on prognostic factors. Reports from randomised trials using non-conformal radiotherapy techniques support the prognostic role of N stage, T stage and sex [3, 39]. Male sex and a higher N stage were found to be strong prognostic indicators for worse OS [3, 40, 41], for higher risk of local failure [3, 42] and LRF [41]. The prognostic role of T stage was less apparent, since higher T stage was only found to be prognostic for worse OS [40] and local failure [42]. Our results suggest that a higher T stage is prognostic for higher risk of LRF in UVA, but not in MVA. Although the aforementioned trials used highly standardised approaches and studied a relatively large number of patients, crude radiotherapy techniques were employed, therefore the prescribed and received radiotherapy doses are likely to differ significantly [43].

In terms of tumour biomarkers, HPV status is the strongest previously-established prognostic indicator in anal cancer [44, 45]. A previous study [46] also established the prognostic significance of p16INK4A in anal cancer, a biomarker commonly used as a surrogate for HPV involvement. In line with these findings, our results confirm the prognostic role of pre-treatment biopsy HPV load in anal cancer. Treatment modification based on HPV status is currently being tested in a head and neck cancer clinical trial, where treatment is stratified based on the HPV status of the cancer [47]. Apart from HPV load, no other tumour biomarkers were identified as prognostic in this review. In terms of haematological biomarkers, long-term outcome data from the ACT1 randomised controlled trial reported that a higher baseline white blood cell count is prognostic for worse OS [41], supporting our results (Table 3-3). Baseline anaemia, another haematological biomarker identified as prognostic in our review, may carry important clinical implications. Although not predictive of OS in the ACT1 data, it was independently predictive of anal cancer death. In cervical cancer, another HPV-driven cancer, blood transfusions are given if haemoglobin levels are below 10g/dl prior to CRT and this may be an area of future clinical consideration in anal cancer treatment.

Due to the lack of studies exploring imaging factors, it is difficult to put our review findings into perspective. Future radiomics research in this setting should focus on multicentre cohorts; but we also noted the lack of secondary or explorative radiomics research from prospective trials. Further research in this area may for instance help identify tumour volumes of greater radiotherapy resistance for boosting.

Three other reviews have previously investigated prognostic factors for anal cancer. One systematic review focused solely on biomarkers and did not include any information on general, pathological or treatment-related prognostic factors [48]. A second systematic review examined the prognostic factors for the specific subset of HIV-positive anal cancer patients undergoing highly active antiretroviral therapy (HAART) [49]. The third review [50] explored clinical, treatment-related as well as molecular prognostic factors, but was a narrative rather than a systematic review. None focused specifically on identifying prognostic factors for outcomes after conformal radiotherapy.

The current work has several limitations. As anal cancer is rare, reports exploring this topic are often single-centre studies with small cohorts, meaning that the power to identify relevant prognostic factors, especially factors with relatively limited effect size or with low prevalence, may be limited. Any factors identified and their effect estimates may suffer from small sample bias [51]. We opted for a sample size of 100 patients as the cut-off point, following an initial screen of available studies, in order to ensure that a reasonable number of studies could be included in the final analysis and the factors identified were generalisable. Through the initial screen, only 43 studies which had cohorts of more than 20 patients were identified. If studies with 20-100 patients had been included, seven additional studies exploring biomarkers and 12 additional studies exploring imaging factors would have been considered, and a larger number of factors would potentially be identified as prognostic. Only few of the studies included in this review distinguished between cancers of the anal canal and perianal cancers (Supplementary material C, Section 3.8.3). Therefore, it was not possible to identify prognostic factors for a specific tumour location or subtype. Additionally, the TNM staging version used varied from the 6th edition to the 8th edition across studies (Supplementary material C, Section 3.8.3) and some studies did not report the version used at all. As a result, in this review all tumour and nodal staging information was analysed together, without accounting for the version used.

There was large variation in treatment regimens, factors tested and outcome definitions between studies. This renders the identification of prognostic factors for anal cancer challenging and highlights the need for uniform outcome definitions, not only in clinical trials and research, but also in routine clinical practice [52]. The studies themselves suffer from several limitations as well, especially in the statistical methodology. The majority of studies applied a univariable screening technique to select factors for MVA. Generally, univariable screening should be avoided for such analyses, as it invalidates the effect and significance estimates in MVA [53, 54], and more robust approaches should be used instead [53, 55]. Moreover, a considerable number of studies did not report on factor effects acquired from UVA or MVA, therefore we could not summarise factor effects across studies. Since a meta-analysis could not be conducted, only a summary of factor effects is reported in this review. Lastly, the proportion of times each factor was identified as prognostic, which is a better indicator of the reliability of the prognostic significance of a factor, could not be calculated from MVA results, due to a lack of detail about the total number of times each factor was tested for each outcome.

Overall, this study confirms the prognostic value of only few well-established clinical factors and biomarkers relevant to contemporary clinical practice. No novel prognostic factors have been identified. This emphasises the lack of studies with large cohorts treated with conformal radiotherapy that report on prognostic factors, especially studies exploring biomarkers and imaging factors. In spite of the remarkable advances in anal cancer treatment efficacy and the reduction of toxicity through conformal CRT, our understanding of the biomarker and imaging factors that predict the outcomes of this disease is still very limited. To tackle the challenge of prognostic factor identification, larger multi-institutional studies and prospective clinical trials would need to be conducted, not only on a national scale, but also on an international scale using approaches that link data across borders [56].

3.6 Conclusions

This systematic review confirms the following prognostic factors for outcomes following anal cancer treatment with conformal CRT: T stage, N stage, sex, pre-treatment biopsy HPV load, as well as the presence of baseline leukocytosis, neutrophilia and anaemia. The prognostic information presented can be used as a starting point for variable selection in future prognostic modelling studies. Additionally, by establishing a set of prognostic and potentially predictive factors for anal cancer outcomes, we may be able to stratify patients into risk groups in order to design more personalised clinical trials in the future. Radiotherapy dose modification based on risk by T and N stage is being evaluated in the currently recruiting PLATO clinical trial [17], with translational research into prognostic biomarkers and imaging embedded within the trial design. This will in turn provide us with greater insight into how to effectively treat this disease using a more personalised approach.

3.7 References

[1] Nigro ND, Vaitkevicius VK, Considine B. Combined therapy for cancer of the anal canal: A preliminary report. Dis Colon Rectum 1974;17:354–6. https://doi.org/10.1007/BF02586980.

[2] Flam M, John M, Pajak TF, Petrelli N, Myerson R, Doggett S, et al. Role of mitomycin in combination with fluorouracil and radiotherapy, and of salvage chemoradiation in the definitive nonsurgical treatment of epidermoid carcinoma of the anal canal: results

of a phase III randomized intergroup study. J Clin Oncol 1996;14:2527–39. https://doi.org/10.1200/JCO.1996.14.9.2527.

[3] Bartelink H, Roelofsen F, Eschwege F, Rougier P, Bosset JF, Gonzalez DG, et al. Concomitant radiotherapy and chemotherapy is superior to radiotherapy alone in the treatment of locally advanced anal cancer: results of a phase III randomized trial of the European Organization for Research and Treatment of Cancer Radiotherapy and Gastrointestinal Cooperative Groups. J Clin Oncol 1997;15:2040–9. https://doi.org/10.1200/JCO.1997.15.5.2040.

[4] Murray LJ, Lilley J. Radiotherapy: technical aspects. Medicine (Baltimore) 2020;48:79–83. https://doi.org/10.1016/j.mpmed.2019.11.003.

[5] Muirhead R, Adams RA, Gilbert DC, Harrison M, Glynne-Jones R, Sebag-Montefiore D, et al. National guidance for IMRT in anal cancer 2016. http://analimrtguidance.co.uk/national-anal-imrt-guidance-v3.pdf (accessed January 31, 2022).

[6] UKCCCR Anal Cancer Trial Working Party. Epidermoid anal cancer: results from the UKCCCR randomised trial of radiotherapy alone versus radiotherapy, 5-fluorouracil, and mitomycin. The Lancet 1996;348:1049–54. https://doi.org/10.1016/S0140-6736(96)03409-5.

[7] Gunderson LL, Winter KA, Ajani JA, Pedersen JE, Moughan J, Benson AB 3rd, et al. Long-term update of US GI intergroup RTOG 98-11 phase III trial for anal carcinoma: survival, relapse, and colostomy failure with concurrent chemoradiation involving fluorouracil/mitomycin versus fluorouracil/cisplatin. J Clin Oncol 2012;30:4344–51. https://doi.org/10.1200/JCO.2012.43.8085.

[8] Peiffert D, Tournier-Rangeard L, Gérard J-P, Lemanski C, François E, Giovannini M, et al. Induction Chemotherapy and Dose Intensification of the Radiation Boost in Locally Advanced Anal Canal Carcinoma: Final Analysis of the Randomized UNICANCER ACCORD 03 Trial. J Clin Oncol 2012;30:1941–8. https://doi.org/10.1200/JCO.2011.35.4837.

[9] James RD, Glynne-Jones R, Meadows HM, Cunningham D, Myint AS, Saunders MP, et al. Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II): a randomised, phase 3, open-label, 2×2 factorial trial. Lancet Oncol 2013;14:516–24. https://doi.org/10.1016/S1470-2045(13)70086-X.

[10] Kachnic LA, Winter K, Myerson RJ, Goodyear MD, Willins J, Esthappan J, et al. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. Int J Radiat Oncol Biol Phys 2013;86:27–33. https://doi.org/10.1016/j.ijrobp.2012.09.023.

[11] Jones MP, Hruby G, Metser U, Sridharan S, Capp A, Kumar M, et al. FDG-PET parameters predict for recurrence in anal cancer - results from a prospective, multicentre clinical trial. Radiat Oncol Lond Engl 2019;14:140. https://doi.org/10.1186/s13014-019-1342-9.

[12] Wang J, Zhang H, Chuong M, Latifi K, Tan S, Choi W, et al. Prediction of Anal Cancer Recurrence After Chemoradiotherapy Using Quantitative Image Features Extracted From Serial 18F-FDG PET/CT. Front Oncol 2019;9:934. https://doi.org/10.3389/fonc.2019.00934.

[13] Das P, Bhatia S, Eng C, Ajani JA, Skibber JM, Rodriguez-Bigas MA, et al. Predictors and Patterns of Recurrence After Definitive Chemoradiation for Anal Cancer. Int J Radiat Oncol 2007;68:794–800. https://doi.org/10.1016/j.ijrobp.2006.12.052.

[14] Tomaszewski JM, Link E, Leong T, Heriot A, Vazquez M, Chander S, et al. Twenty-five-year experience with radical chemoradiation for anal cancer. Int J Radiat Oncol Biol Phys 2012;83:552–8. https://doi.org/10.1016/j.ijrobp.2011.07.007.

[15] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[16] Halabi S, Owzar K. The Importance of Identifying and Validating Prognostic Factors in Oncology. Semin Oncol 2010;37:e9–18. https://doi.org/10.1053/j.seminoncol.2010.04.001.

[17] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[18] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021:n71. https://doi.org/10.1136/bmj.n71.

[19] National Institutes of Health. Study quality assessment tools 2014. https://www. nhlbi. nih. gov/health-topics/study-quality-assessment-tools (accessed June 11, 2021).

[20] Fish R, Sanders C, Ryan N, der Veer SV, Renehan AG, Williamson PR. Systematic review of outcome measures following chemoradiotherapy for the treatment of anal cancer (CORMAC). Colorectal Dis Off J Assoc Coloproctology G B Irel 2018;20:371–82. https://doi.org/10.1111/codi.14103.

[21] Martin D, Rödel F, von der Grün J, Rödel C, Fokas E. Acute organ toxicity correlates with better clinical outcome after chemoradiotherapy in patients with anal carcinoma. Radiother Oncol 2020;149:168–73. https://doi.org/10.1016/j.radonc.2020.05.016.

[22] de Meric de Bellefon M, Lemanski C, Castan F, Samalin E, Mazard T, Lenglet A, et al. Long-term follow-up experience in anal canal cancer treated with Intensity-Modulated Radiation Therapy: Clinical outcomes, patterns of relapse and predictors of failure. Radiother Oncol 2020;144:141–7. https://doi.org/10.1016/j.radonc.2019.11.016.

[23] Brown PJ, Zhong J, Frood R, Currie S, Gilbert A, Appelt AL, et al. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging 2019;46:2790–9. https://doi.org/10.1007/s00259-019-04495-1.

[24] Rouard N, Peiffert D, Rio E, Mahé M-A, Delpon G, Marchesi V, et al. Intensity-modulated radiation therapy of anal squamous cell carcinoma: Relationship between delineation quality and regional recurrence. Radiother Oncol J Eur Soc Ther Radiol Oncol 2019;131:93–100. https://doi.org/10.1016/j.radonc.2018.10.021.

[25] Franco P, Montagnani F, Arcadipane F, Casadei C, Andrikou K, Martini S, et al. The prognostic role of hemoglobin levels in patients undergoing concurrent chemo-radiation for anal cancer. Radiat Oncol Lond Engl 2018;13:83. https://doi.org/10.1186/s13014-018-1035-9.

[26] Call JA, Prendergast BM, Jensen LG, Ord CB, Goodman KA, Jacob R, et al. Intensity-modulated Radiation Therapy for Anal Cancer: Results From a Multi-Institutional

Retrospective Cohort Study. Am J Clin Oncol 2016;39:8–12. https://doi.org/10.1097/COC.0000000000000009.

[27] Balermpas P, Martin D, Wieland U, Rave-Fränk M, Strebhardt K, Rödel C, et al. Human papilloma virus load and PD-1/PD-L1, CD8+ and FOXP3 in anal cancer patients treated with chemoradiotherapy: Rationale for immunotherapy. OncoImmunology 2017;6:e1288331. https://doi.org/10.1080/2162402X.2017.1288331.

[28] Rödel F, Steinhäuser K, Kreis N-N, Friemel A, Martin D, Wieland U, et al. Prognostic impact of RITA expression in patients with anal squamous cell carcinoma treated with chemoradiotherapy. Radiother Oncol 2018;126:214–21. https://doi.org/10.1016/j.radonc.2017.10.028.

[29] Schernberg A, Huguet F, Moureau-Zabotto L, Chargari C, Rivin Del Campo E, Schlienger M, et al. External validation of leukocytosis and neutrophilia as a prognostic marker in anal carcinoma treated with definitive chemoradiation. Radiother Oncol 2017;124:110–7. https://doi.org/10.1016/j.radonc.2017.06.009.

[30] Martin D, Rödel F, Balermpas P, Winkelmann R, Fokas E, Rödel C. C-Reactive Protein-to-Albumin Ratio as Prognostic Marker for Anal Squamous Cell Carcinoma Treated With Chemoradiotherapy. Front Oncol 2019;9:1200. https://doi.org/10.3389/fonc.2019.01200.

[31] Oehler-Jänne C, Huguet F, Provencher S, Seifert B, Negretti L, Riener M-O, et al. HIV-specific differences in outcome of squamous cell carcinoma of the anal canal: a multicentric cohort study of HIV-positive patients receiving highly active antiretroviral therapy. J Clin Oncol 2008;26:2550–7. https://doi.org/10.1200/JCO.2007.15.2348.

[32] Susko M, Wang C-CJ, Lazar AA, Kim S, Laffan A, Feng M, et al. Factors Impacting Differential Outcomes in the Definitive Radiation Treatment of Anal Cancer Between HIV-Positive and HIV-Negative Patients. The Oncologist 2020. https://doi.org/10.1634/theoncologist.2019-0824.

[33] Cardenas ML, Spencer CR, Markovina S, DeWees TA, Mazur TR, Weiner AA, et al. Quantitative FDG-PET/CT predicts local recurrence and survival for squamous cell carcinoma of the anus. Adv Radiat Oncol 2017;2:281–7. https://doi.org/10.1016/j.adro.2017.04.007.

[34] Bitterman DS, Grew D, Gu P, Cohen RF, Sanfilippo NJ, Leichman CG, et al. Comparison of anal cancer outcomes in public and private hospital patients treated at a single radiation oncology center. J Gastrointest Oncol 2015;6:524–33. https://doi.org/10.3978/j.issn.2078-6891.2015.061.

[35] Fraunholz I, Rödel F, Kohler D, Diallo-Georgiopoulou M, Distel L, Falk S, et al. Epidermal growth factor receptor expression as prognostic marker in patients with anal carcinoma treated with concurrent chemoradiation therapy. Int J Radiat Oncol Biol Phys 2013;86:901–7. https://doi.org/10.1016/j.ijrobp.2013.03.039.

[36] Schernberg A, Escande A, Rivin Del Campo E, Ducreux M, Nguyen F, Goere D, et al. Leukocytosis and neutrophilia predicts outcome in anal cancer. Radiother Oncol 2017;122:137–45. https://doi.org/10.1016/j.radonc.2016.12.009.

[37] Hosni A, Han K, Le LW, Ringash J, Brierley J, Wong R, et al. The ongoing challenge of large anal cancers: prospective long term outcomes of intensity-modulated radiation therapy with concurrent chemotherapy. Oncotarget 2018;9:20439–50. https://doi.org/10.18632/oncotarget.24926.

[38] Oblak I, Cesnjevar M, Anzic M, Hadzic JB, Ermenc AS, Anderluh F, et al. The impact of anaemia on treatment outcome in patients with squamous cell carcinoma of anal canal and anal margin. Radiol Oncol 2016;50:113–20. https://doi.org/10.1515/raon-2015-0015.

[39] Ajani JA, Winter KA, Gunderson LL, Pedersen J, Benson AB, Thomas CR, et al. Prognostic factors derived from a prospective database dictate clinical biology of anal cancer: The intergroup trial (RTOG 98-11). Cancer 2010;116:4007–13. https://doi.org/10.1002/cncr.25188.

[40] Bilimoria KY, Bentrem DJ, Rock CE, Stewart AK, Ko CY, Halverson A. Outcomes and Prognostic Factors for Squamous-Cell Carcinoma of the Anal Canal: Analysis of Patients From the National Cancer Data Base. Dis Colon Rectum 2009;52:624–31. https://doi.org/10.1007/DCR.0b013e31819eb7f0.

[41] Glynne-Jones R, Sebag-Montefiore D, Adams R, Gollins S, Harrison M, Meadows HM, et al. Prognostic factors for recurrence and survival in anal cancer: generating hypotheses from the mature outcomes of the first United Kingdom Coordinating Committee on Cancer Research Anal Cancer Trial (ACT I). Cancer 2013;119:748–55. https://doi.org/10.1002/cncr.27825.

[42] Johnsson A, Leon O, Gunnlaugsson A, Nilsson P, Höglund P. Determinants for local tumour control probability after radiotherapy of anal cancer. Radiother Oncol 2018;128:380–6. https://doi.org/10.1016/j.radonc.2018.06.007.

[43] Aggarwal A, Gayadeen S, Robinson D, Hoskin PJ, Mawdsley S, Harrison M, et al. Clinical target volumes in anal cancer: Calculating what dose was likely to have been delivered in the UK ACT II trial protocol. Radiother Oncol 2012;103:341–6. https://doi.org/10.1016/j.radonc.2012.03.007.

[44] Gilbert DC, Wakeham K, Langley RE, Vale CL. Increased risk of second cancers at sites associated with HPV after a prior HPV-associated malignancy, a systematic review and meta-analysis. Br J Cancer 2019;120:256–68. https://doi.org/10.1038/s41416-018-0273-9.

[45] Serup-Hansen E, Linnemann D, Skovrider-Ruminski W, Høgdall E, Geertsen PF, Havsteen H. Human Papillomavirus Genotyping and p16 Expression As Prognostic Factors for Patients With American Joint Committee on Cancer Stages I to III Carcinoma of the Anal Canal. J Clin Oncol 2014;32:1812–7. https://doi.org/10.1200/JCO.2013.52.3464.

[46] Gilbert DC, Serup-Hansen E, Linnemann D, Høgdall E, Bailey C, Summers J, et al. Tumour-infiltrating lymphocyte scores effectively stratify outcomes over and above p16 post chemo-radiotherapy in anal cancer. Br J Cancer 2016;114:134–7. https://doi.org/10.1038/bjc.2015.448.

[47] Owadally W, Hurt C, Timmins H, Parsons E, Townsend S, Patterson J, et al. PATHOS: a phase II/III trial of risk-stratified, reduced intensity adjuvant treatment in patients undergoing transoral surgery for Human papillomavirus (HPV) positive oropharyngeal cancer. BMC Cancer 2015;15:602. https://doi.org/10.1186/s12885-015-1598-x.

[48] Lampejo T, Kavanagh D, Clark J, Goldin R, Osborn M, Ziprin P, et al. Prognostic biomarkers in squamous cell carcinoma of the anus: a systematic review. Br J Cancer 2010;103:1858–69. https://doi.org/10.1038/sj.bjc.6605984.

[49] Correa RJM, Louie AV, Virine B, Dinniwell R, Kaiser A, Mishra MV. A Systematic Review of Clinical Outcomes Following Chemoadiation Therapy for Anal Cancer in

HIV-Positive Patients on Highly Active Antiretroviral Therapy. Int J Radiat Oncol 2017;99:E142. https://doi.org/10.1016/j.ijrobp.2017.06.940.

[50] Das P, Crane CH, Eng C, Ajani JA. Prognostic factors for squamous cell cancer of the anal canal. Gastrointest Cancer Res GCR 2008;2:10–4.

[51] Hackshaw A. Small studies: strengths and limitations. Eur Respir J 2008;32:1141–3. https://doi.org/10.1183/09031936.00136408.

[52] Fish R, Sanders C, Adams R, Brewer J, Brookes ST, DeNardo J, et al. A core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC): a patient and health-care professional consensus. Lancet Gastroenterol Hepatol 2018;3:865–73. https://doi.org/10.1016/S2468-1253(18)30264-4.

[53] Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J 2018;60:431–49. https://doi.org/10.1002/bimj.201700067.

[54] Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996;49:907–16. https://doi.org/10.1016/0895-4356(96)00025-X.

[55] Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inf 2018;116:10–7. https://doi.org/10.1016/j.ijmedinf.2018.05.006.

[56] Theophanous S, Choudhury A, Lønne P-I, Samuel R, Guren MG, Berbee M, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study. Radiother Oncol 2021;159:183–9. https://doi.org/10.1016/j.radonc.2021.03.013.

3.8 Supplementary material

3.8.1 Supplementary material A*. Full search strategies employed in Embase and Medline to identify relevant papers between January 1st 2000 and June 30th 2020.*

Table 3-A1. Complete search strategy used in Embase.

Database: Embase Classic+Embase <1947 to 2020 June 30>	
1	exp radiotherapy/ (599698)
2	Radiation Oncology/ (3434)
3	(radiotherap* or radiotreat* or roentgentherap* or radiosurg*).tw. (285655)
4	((radiat* or radio* or irradiat* or roentgen or x-ray or xray) adj4 (therap* or treat* or repair* or oncolog* or surg*)).tw. (396128)
5	(RT or RTx or XRT).tw. (308595)
6	exp chemoradiotherapy/ (52074)
7	(chemoradiotherap* or radiochemotherap* or chemoradiation*).tw. (53407)
8	(CRT or CRTx or CCRT or NCRT or RCTx or RT-CT or chemoRT).tw. (38794)
9	or/1-8 [radiotherapy or chemeradiotherapy] (1107122)
10	exp Anus cancer/ (8197)
11	((anus or anal) adj5 (cancer* or neoplas* or carcinoma* or tumo?r*)).tw,kw. (10640)
12	or/10-11 [anal cancer] (13164)
13	(predict* and (outcome* or risk* or model*)).tw. (1210603)
14	(validate or rule*).tw. (376396)
15	predict*.ti. (470931)
16	((history or variable* or criteria or scor* or characteristic* or finding* or factor*) and (predict* or model* or decision* or identify or prognose)).tw. (3153437)
17	(prognostic and (history or variable* or criteria or scor* or characteristic* or finding* or factor* or model*)).tw. (343683)
18	ROC Curve/ (57455)
19	(stratification or discrimination or discriminate or c-statistic or c statistic or area under the curve or AUC or calibration or indices or algorithm or multivariable or (model and outcome) or classif*).tw. (2013800)
20	((model* or clinical).tw. or logistics models/) and decision.tw. (183621)
21	or/13-20 [predictive factor or outcomes] (5491959)
22	9 and 12 and 21 [radiotherapy or chemoradiotherapy and anal cancer and predictive factors for outcomes] (1219)
23	limit 22 to yr="2000 -Current" (1134)
24	remove duplicates from 23 (1109)

Table 3-A2. Complete search strategy used in Medline.

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily <1946 to June 30, 2020>	
1	exp Radiotherapy/ (184934)
2	Radiation Oncology/ (4114)
3	(radiotherap* or radiotreat* or roentgentherap* or radiosurg*).tw. (179332)
4	((radiat* or radio* or irradiat* or roentgen or x-ray or xray) adj4 (therap* or treat* or repair* or oncolog* or surg*)).tw. (244503)
5	(RT or RTx or XRT).tw. (201228)
6	exp Chemoradiotherapy/ (14534)
7	(chemoradiotherap* or radiochemotherap* or chemoradiation*).tw. (30229)
8	(CRT or CRTx or CCRT or NCRT or RCTx or RT-CT or chemoRT).tw. (18166)
9	or/1-8 [radiotherapy or chemeradiotherapy] (615114)
10	exp Anus Neoplasms/ (6335)
11	((anus or anal) adj5 (cancer* or neoplas* or carcinoma* or tumo?r*)).tw,kw. (6355)
12	or/10-11 [anal cancer] (9210)
13	(predict* and (outcome* or risk* or model*)).tw. (847317)
14	(validate or rule*).tw. (265142)
15	predict*.ti. (325883)
16	((history or variable* or criteria or scor* or characteristic* or finding* or factor*) and (predict* or model* or decision* or identify or prognose)).tw. (2253192)
17	(prognostic and (history or variable* or criteria or scor* or characteristic* or finding* or factor* or model*)).tw. (216359)
18	ROC Curve/ (57771)
19	(stratification or discrimination or discriminate or c-statistic or c statistic or area under the curve or AUC or calibration or indices or algorithm or multivariable or (model and outcome) or classif*).tw. (1421211)
20	((model* or clinical).tw. or logistics models/) and decision.tw. (123422)
21	or/13-20 [predictive factor or outcomes] (3986478)
22	9 and 12 and 21 [radiotherapy or chemoradiotherapy and anal cancer and predictive factors for outcomes] (522)
23	limit 22 to yr="2000-2020" (458)

3.8.2 **Supplementary material B**. *Complete results from the study quality appraisal by both reviewers (ST and RS), including the assessment criteria used. Y: Yes. N: No. NR: Not reported.*

Case series study appraisal criteria:

1. Was the study question or objective clearly stated?
2. Was the study population clearly and fully described, including a case definition?
3. Were the cases consecutive?
4. Were the subjects comparable? Reasonably homogeneous study population
5. Was the intervention clearly described?
6. Were the outcome measures clearly defined, valid, reliable, and implemented consistently across all study participants?
7. Was the length of follow-up adequate? 3 years according to PLATO
8. Were the statistical methods well-described?
9. Were the results well-described?

Table 3-B1. Study quality appraisal by ST.

Study/Criterion	1	2	3	4	5	6	7	8	9	Quality rating (Good/Fair/Poor)	Type of study
Shakir et al. (2020)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Martin et al. (2020)	Y	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
de Bellefon et al. (2020)	N	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series
Brown et al. (2019)	Y	Y	Y	Y	N	Y	N	Y	Y	Good	Case series
Rouard et al. (2019)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Franco et al. (2018)	Y	Y	NR	Y	Y	Y	N	Y	Y	Good	Case series
Call et al. (2016)	N	Y	NR	Y	Y	Y	N	Y	Y	Fair	Case series
Balermpas et al. (2017)	N	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Rodel et al. (2018)	N	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Schernberg et al. (2017)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series
Martin et al. (2019)	Y	Y	NR	Y	Y	Y	NR	Y	Y	Good	Case series
Oehler-Janne et al. (2008)	Y	Y	Y	N	Y	Y	Y	Y	Y	Good	Case series
Susko et al. (2020)	Y	Y	NR	Y	Y	Y	N	Y	Y	Good	Case series
Cardenas et al. (2017)	Y	Y	NR	Y	Y	N	N	Y	Y	Fair	Case series
Bitterman et al. (2015)	N	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Fraunholz et al. (2013)	Y	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Schernberg et al. (2017)*	Y	Y	Y	Y	Y	N	Y	Y	Y	Good	Case series
Hosni et al. (2018)	N	Y	NR	Y	Y	N	Y	Y	Y	Fair	Case series
Oblak et al. (2016)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series

Table 3-B2. Study quality appraisal by RS.

Study/Criterion	1	2	3	4	5	6	7	8	9	Quality rating (Good/Fair/Poor)	Type of study
Shakir et al. (2020)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Martin et al. (2020)	Y	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
de Bellefon et al. (2020)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series
Brown et al. (2019)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Rouard et al. (2019)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series

Franco et al. (2018)	Y	Y	NR	Y	Y	Y	N	Y	Y	Good	Case series
Call et al. (2016)	Y	Y	NR	Y	Y	Y	N	Y	Y	Good	Case series
Balermpas et al. (2017)	N	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Rodel et al. (2018)	N	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Schernberg et al. (2017)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series
Martin et al. (2019)	Y	Y	NR	Y	Y	Y	NR	Y	Y	Good	Case series
Oehler-Janne et al. (2008)	N	Y	Y	N	Y	Y	Y	Y	Y	Good	Case series
Susko et al. (2020)	Y	Y	Y	Y	Y	N	N	Y	Y	Good	Case series
Cardenas et al. (2017)	Y	Y	NR	Y	Y	N	N	Y	Y	Fair	Case series
Bitterman et al. (2015)	Y	Y	Y	Y	Y	Y	N	Y	Y	Good	Case series
Fraunholz et al. (2013)	Y	Y	NR	Y	Y	Y	Y	Y	Y	Good	Case series
Schernberg et al. (2017)*	Y	Y	Y	Y	Y	N	Y	Y	Y	Good	Case series
Hosni et al. (2018)	N	Y	NR	Y	Y	N	Y	Y	Y	Fair	Case series
Oblak et al. (2016)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Good	Case series

3.8.3 Supplementary material C. Complete overview of study characteristics, including the predictors tested in each study. NR: not reported. SCC: Squamous cell carcinoma. RT: radiotherapy. CRT: chemoradiotherapy. MMC: Mitomycin C.

#	Study	Year of publication	Location	Study design	Number of patients	Years of treatment	RT technique	Cancer subtype and location in cohort	TNM staging version used	Median follow-up (months)	Type of statistical analysis used	Predictors tested	Quality score
1	Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy with a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma [Shakir et al. (2020)]	2020	Multi-centre, Europe	Retrospective	385	2013-2018	IMRT	SCC of the anal canal and anal margin	7 & 8	24.0	Univariable Cox regression, multivariable Cox regression	Age, sex, performance status, T stage, N stage (TNM 7 and 8), RT completion, chemotherapy type	Good
2	Acute organ toxicity correlates with better clinical outcome after chemoradiotherapy in patients with anal carcinoma [Martin et al. (2020)]	2020	Single centre, Europe	Retrospective	223	1996-2017	3D-CRT (58%) and IMRT (42%)	SCC (location not specified)	7	46.0	Univariable Cox regression, multivariable Cox regression	T stage, N stage, age, sex, high grade acute organ toxicity (HGAOT)	Good
3	Long-term follow-up experience in anal canal cancer treated with Intensity-Modulated Radiation Therapy: Clinical outcomes, patterns of relapse and predictors of failure [de Bellefon et al. (2020)]	2020	Single centre, Europe	Retrospective	193	2005-2017	IMRT	SCC of the anal canal	7 & 8	70.0	Univariable Cox regression, multivariable Cox regression	Only significant factors reported - T stage, N stage, AJCC stage, sex, RT breaks, exclusive RT, lack of MMC, residual disease	Good
4	Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT [Brown et al. (2019)]	2019	Single centre, Europe	Retrospective	189	2008-2016	2D/ 3D-CRT (79%) and VMAT (21%)	SCC (location not specified)	NR	35.1	Multivariable logistic regression	Only significant factors reported - Multiple FDG-PET scan variables, sex, age, T stage, N stage	Good

5	Intensity-modulated radiation therapy of anal squamous cell carcinoma: Relationship between delineation quality and regional recurrence [Rouard et al. (2019)]	2019	Multi-centre, Europe	Retrospective	165	2006-2016	IMRT	SCC of the anal canal and anal margin	NR	33.8	Bivariable Cox regression, multivariable Cox regression	Sex, age, performance status, immunodepression, smoking status, tumour localisation, T stage, N stage, location of involved lymph nodes, tumour size, keratinisation, differentiation, HPV status, RT breaks, tumour boost technique, tumour total dose, chemotherapy type, multiple delineation variables	Good
6	The prognostic role of hemoglobin levels in patients undergoing concurrent chemo-radiation for anal cancer [Franco et al. (2018)]	2018	Multi-centre, Europe	Retrospective	161	NR	IMRT	SCC of the anal canal and anal margin	NR	27.0	Log-rank analysis, univariable Cox regression, multivariable Cox regression	Age, sex, T stage, N stage, response to treatment, overall treatment duration, RT total dose, boost, OTT, basal haemoglobin levels	Good
7	Intensity-modulated Radiation Therapy for Anal Cancer Results From a Multi-Institutional Retrospective Cohort Study [Call et al. (2016)]	2016	Multi-centre, North America	Retrospective	152	NR	IMRT	SCC (location not specified)	NR	26.8	Log-rank analysis, multivariable Cox regression	Dose, N stage, T stage, RT duration	Fair
8	Human papilloma virus load and PD-1/PD-L1, CD8+ and FOXP3 in anal cancer patients treated with chemoradiotherapy: Rationale for immunotherapy [Balermpas et al. (2017)]	2017	Multi-centre, Europe	Retrospective	150	NR	3D-CRT and IMRT	SCC (location not specified)	NR	40.0	Log-rank analysis, multivariable Cox regression	Age, gender, T stage, N stage, grade, HPV load, and CD8, PD1, PD-L1, FOXP3, pCASP8 expression	Good
9	Prognostic impact of RITA expression in patients with anal squamous cell carcinoma treated with chemoradiotherapy [Rodel et al. (2018)]	2018	Multi-centre, Europe	Retrospective	140	NR	3D-CRT and IMRT	SCC (location not specified)	NR	40.0	Log-rank analysis, multivariable Cox regression	Gender, T stage, N stage, HPV-16 DNA load, RITA expression	Good

10	External validation of leukocytosis and neutrophilia as a prognostic marker in anal carcinoma treated with definitive chemoradiation [Schernberg et al. (2017)]	2017	Multi-centre, Europe	Retrospective	133	2000-2015	IMRT (77%) and 3D-CRT (23%)	SCC (location not specified)	7	37.4	Log-rank analysis, multivariable Cox regression	Age, sex, T stage, N stage, performance status, leukocytosis, neutrophilia, anaemia, lymphopenia, monocytosis, thrombocytosis	Good
11	C-Reactive Protein-to-Albumin Ratio as Prognostic Marker for Anal Squamous Cell Carcinoma Treated with Chemoradiotherapy [Martin et al. (2019)]	2019	Single centre, Europe	Retrospective	126	2004-2016	IMRT (65%) and 3D-CRT (35%)	SCC (location not specified)	7	NA	Log-rank analysis, univariable Cox regression, multivariable Cox regression	Age, sex, HIV status, T stage, N stage, grade, C reactive protein to Albumin ratio, RT modality, RT total dose	Good
12	HIV-specific differences in outcome of squamous cell carcinoma of the anal canal: a multicentric cohort study of HIV-positive patients receiving highly active antiretroviral therapy [Oehler-Janne et al. (2008)]	2008	Multi-centre, International	Retrospective	121	1997-2006	3D-CRT	SCC of the anal canal	NR	36.0	Log-rank analysis, univariable Cox regression, multivariable Cox regression	Not explicitly reported - Age, sex, WHO performance status, histologic subtype, tumour size, N stage, M stage, CDC stage, CD4 count, viral load, HAART type	Good
13	Factors Impacting Differential Outcomes in the Definitive Radiation Treatment of Anal Cancer Between HIV-Positive and HIV-Negative Patients [Susko et al. (2020)]	2020	Single centre, North America	Retrospective	111	2005-2018	3D-CRT and IMRT	SCC (location not specified)	NR	28.0	Log-rank analysis, univariable Cox regression, multivariable Cox regression	Age, sex, T stage, N stage, HIV status, time from diagnosis to treatment, treatment duration	Good
14	Quantitative FDG-PET/CT predicts local recurrence and survival for squamous cell carcinoma of the anus [Cardenas et al. (2017)]	2017	Single centre, North America	Retrospective	110	2003-2013	IMRT (75%) and 2D-CRT (25%)	SCC (location not specified)	NR	28.6	Univariable Cox regression, multivariable Cox regression	Multiple FDG-PET scan variables, RT modality, chemotherapy, T stage, N stage, HIV status	Fair
15	Comparison of anal cancer outcomes in public and private hospital patients treated at a single radiation oncology center [Bitterman et al. (2015)]	2015	Single centre, North America	Retrospective	109	2004-2013	IMRT (60%) and 3D-CRT (40%)	SCC (location not specified), cloacogenic (n=2) and adeno (n=2) carcinomas	NR	14.9	Log-rank analysis, multivariable Cox regression	Referral from public hospital, HIV status, T stage, RT technique, RT duration, RT delay	Good

16	Epidermal Growth Factor Receptor Expression As Prognostic Marker in Patients With Anal Carcinoma Treated With Concurrent Chemoradiation Therapy [Fraunholz et al. (2013)]	2013	Multi-centre, Europe	Retrospective	103	1989-2011	3D-CRT	SCC, basaloid, cloacogenic (location not specified)	7	44.0	Log-rank analysis, multivariable Cox regression	Age, sex, HIV status, T stage, N stage, grade, EGFR expression	Good
17	Leukocytosis and neutrophilia predicts outcome in anal cancer [Schernberg et al. (2017)*]	2017	Single centre, Europe	Retrospective	103	2006-2016	IMRT (53%) and 3D-CRT (47%)	SCC (location not specified)	6	38.7	Log-rank analysis, multivariable Cox regression	Leukocytosis, neutrophilia, anaemia, T stage, N stage, CRT duration	Good
18	The ongoing challenge of large anal cancers: prospective long term outcomes of intensity-modulated radiation therapy with concurrent chemotherapy [Hosni et al. (2018)]	2018	Single centre, North America	Retrospective	101	2008-2013	IMRT	SCC of the anal canal, SCC of the anal canal with perianal extension	7	56.5	Univariable Cox regression, multivariable Cox regression	T stage, N stage, sex, age, grade, maximum tumour size, RT interruption	Fair
19	The impact of anaemia on treatment outcome in patients with squamous cell carcinoma of anal canal and anal margin [Oblak et al. (2016)]	2016	Single centre, Europe	Retrospective	100	2003-2013	3D-CRT and IMRT	SCC of the anal canal and anal margin	7	52.0	Log-rank analysis, multivariable Cox regression	Pre-treatment Hb, on-treatment Hb, end-of-treatment Hb, performance status, T stage, N stage, stage, histology, tumour site, blood transfusion, overall radiation time, operation	Good

3.8.4 Supplementary material D. Outcome definitions given in each study, stratified into nine categories. The final stratification yielded three disease activity outcome categories and six survival outcome categories.

#	Outcomes (number of studies reporting outcome)	Outcomes included	Study	Definition
1	**Overall survival (n=17)**	Overall survival	Martin et al. (2020) [21]	Survival times were calculated from start of CRT to the date of respective events or last follow-up. Assessed with death of any cause as the respective event.
		Overall survival	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: death from any cause.
		Overall survival	Rouard et al. (2019) [24]	The time between the first day of RT and the death (all causes). Surviving patients were censored at the date of last follow-up or five years after D1.
		Overall survival	Franco et al. (2018) [25]	Calculated from the date of diagnosis to that of death from any cause or lost at observation.
		Overall survival	Call et al. (2016) [26]	Not defined.
		Overall survival	Balermpas et al. (2017) [27]	Calculated from the beginning of CRT to death for any reasons or to cancer-related death, or the day of the last follow-up.
		Overall survival	Rodel et al. (2018) [28]	Defined from the beginning of CRT to the day of death from any reasons.
		Overall survival	Schernberg et al. (2017) [29]	The time between the diagnosis and the time of death.
		Overall survival	Martin et al. (2019) [30]	Calculated from start of CRT to the date of event or last follow-up. Assessed with death of any cause as the respective event.
		Overall survival	Oehler-Janne et al. (2008) [31]	Calculated from the beginning of RT to the day of death or the date of last follow-up.
		Overall survival	Susko et al. (2020) [32]	The time from last radiation treatment to date of death or last follow-up.
		Overall survival	Cardenas et al. (2017) [33]	Not defined.
		Overall survival	Bitterman et al. (2015) [34]	The time from initiation of CRT to death due to any cause or most recent follow-up.
		Overall survival	Fraunholz et al. (2013) [35]	The time from start of CRT until death resulting from any cause, or the date of last follow-up visit.
		Overall survival	Schernberg et al. (2017)* [36]	Not defined
		Overall survival	Hosni et al. (2018) [37]	Not defined
		Overall survival	Oblak et al. (2016) [38]	The time interval from the beginning of the treatment to the death due to any cause.
2	**Locoregional failure (n=11)**	Locoregional recurrence	Shakir et al. (2020) [15]	All failures at site of primary tumor, within the pelvis or inguinal nodes, with or without distant failure, including both patients who failed to achieve CR at 6 months and those occurring more than 6 months after completion of CRT after initial CR.
		Local failure	Shakir et al. (2020) [15]	Persistence or recurrence at the site of initial primary tumor. The site of failure was determined based on physical examination, imaging, and pathology.
		Regional failure	Shakir et al. (2020) [15]	Persistence or recurrence elsewhere in the pelvis or inguinal nodes at any point. The site of failure was determined based on physical examination, imaging, and pathology.
		Local relapse-free survival	Martin et al. (2020) [21]	Survival times were calculated from start of CRT to the date of respective events or last follow-up. Calculated using non-complete response at first restaging or locoregional recurrence after initial complete response as event.

		Locoregional failure	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: residual disease, local and/or regional recurrences.
		Locoregional recurrence	Rouard et al. (2019) [24]	The time between the first day of RT and the date of first local or regional recurrence.
		Local recurrence	Rouard et al. (2019) [24]	The time between the first day of RT and the date of local recurrence.
		Regional recurrence	Rouard et al. (2019) [24]	The time between the first day of RT and the date of regional recurrence.
		Local control	Call et al. (2016) [26]	Defined as the time to local relapse. *No definition for locoregional failure given. Local and regional failure definitions stated separately only.
		Regional control	Call et al. (2016) [26]	Defined as the time to regional relapse. *No definition for locoregional failure given. Local and regional failure definitions stated separately only.
		Cumulative incidence of locoregional failure	Balermpas et al. (2017) [27]	Calculated from the beginning of CRT to non-complete response at restaging or locoregional tumor detection after initial complete response.
		Cumulative incidence of locoregional failure	Rodel et al. (2018) [28]	The time to non-complete response at restaging or locoregional tumour detection after initial complete response. All time-to-event end points were measured from the start of CRT.
		Locoregional control	Schernberg et al. (2017) [29]	The time between the diagnosis and the time of loco-regional recurrence.
		Locoregional control rate	Martin et al. (2019) [30]	Calculated from start of CRT to the date of event or last follow-up. Calculated using non-complete response at first restaging or locoregional recurrence after initial complete response as event.
		Freedom from local recurrence	Susko et al. (2020) [32]	The time from last radiation treatment to locally recurrent disease or last follow-up.
		Locoregional control	Oblak et al. (2016) [38]	The time interval from the beginning of the treatment to the appearance of local and/or regional progression.
3	**Disease-free survival (n=11)**	Disease-free survival	Martin et al. (2020) [21]	Survival times were calculated from start of CRT to the date of respective events or last follow-up. Calculated using the date of diagnosis of locoregional failure, distant metastases, or death of any cause.
		Disease-free survival	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: death from any cause or recurrence.
		Progression-free survival	Brown et al. (2019) [23]	Comprises of locoregional failure (LRF), new distant metastatic disease and death, based on which occurred first.
		Disease-free survival	Rouard et al. (2019) [24]	The time between the first day of RT and the date of local, regional or metastatic recurrence or death, whichever occurred first.
		Progression-free survival	Franco et al. (2018) [25]	The time interval between diagnosis and disease recurrence and/or progression at any site, death or lost at follow-up.
		Disease-free survival	Balermpas et al. (2017) [27]	Measured from the beginning of CRT to the day of locoregional failure or distant recurrence, or death from any cause.
		Progression-free survival	Schernberg et al. (2017) [29]	The time between the diagnosis and the time of recurrence or death.
		Disease-free survival	Martin et al. (2019) [30]	Calculated from start of CRT to the date of event or last follow-up. Calculated using the date of diagnosis of locoregional failure, distant metastases, or death of any cause.
		Disease-free survival	Bitterman et al. (2015) [34]	The time from initiation of CRT to the occurrence of local, regional, or distant recurrence, death, or most recent follow-up.
		Progression-free survival	Schernberg et al. (2017)* [36]	Not defined
		Disease-free survival	Hosni et al. (2018) [37]	Not defined
4	**Distant failure (n=5)**	Distant relapse	Shakir et al. (2020) [15]	Development of disease outside of the pelvis or inguinal nodes independent of locoregional status at any point. Failure within the common iliac nodes was considered distant failure.
		Distant control	Call et al. (2016) [26]	Defined as the time to distant relapse.

		Cumulative incidence of distance metastases	Rodel et al. (2018) [28]	Any occurrence of distant metastasis during CRT, at re-staging, or during follow-up. All time-to-event end points were measured from the start of CRT.
		Freedom from distant metastasis	Susko et al. (2020) [32]	The time from last radiation treatment to distant recurrence of disease or last follow-up.
		Distant metastases control	Schernberg et al. (2017)* [36]	The time between the diagnosis and the time of distant metastasis.
5	**Metastasis-free survival (n=5)**	Distant metastasis-free survival	Martin et al. (2020) [21]	Survival times were calculated from start of CRT to the date of respective events or last follow-up. Calculated using the date of diagnosis of distant metastases or death of any cause as event.
		Metastasis-free survival	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: death or distant relapse.
		Distant metastasis-free survival	Martin et al. (2019) [30]	Calculated from start of CRT to the date of event or last follow-up. Calculated using the date of diagnosis of distant metastases or death of any cause as event.
		Distant metastases-free survival	Fraunholz et al. (2013) [35]	The time from the start of CRT to the diagnosis of distant metastases or to death, or the date of last follow-up visit.
		Distant failure-free survival	Schernberg et al. (2017)* [36]	Not defined
6	**Freedom from disease (n=4)**	Disease-free survival	Shakir et al. (2020) [15]	Event defined as either a failure to achieve CR at 6 months or subsequent relapse (local, regional, or distant).
		Time to failure	Shakir et al. (2020) [15]	Interval from start of CRT to date of detection of recurrence. Last follow-up was considered the last clinic visit or date of death.
		Disease-free survival	Rodel et al. (2018) [28]	Defined from the beginning of CRT to the day of locoregional failure or distant recurrence.
		Time to recurrence	Oehler-Janne et al. (2008) [31]	Calculated from the beginning of RT to the day of recurrence or the date of last follow-up.
		Disease-free survival	Oblak et al. (2016) [38]	The time interval from the beginning of the treatment to the appearance of local and/or regional progression and/or appearance of distant metastases.
7	**Colostomy-free survival (n=4)**	Colostomy-free survival	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: death or definitive colostomy. A colostomy performed before radiotherapy was considered as a failure on the first day of treatment as long as it was not reversed later on.
		Colostomy-free survival	Call et al. (2016) [26]	Defined as the time to the date of a colostomy procedure.
		Colostomy-free survival	Bitterman et al. (2015) [34]	Measured from initiation of CRT to diverting colostomy or salvage abdominoperineal resection (APR), death, or most recent follow-up without surgery.
		Colostomy-free survival	Hosni et al. (2018) [37]	Not defined
8	**Cancer-specific survival (n=3)**	Cancer-specific survival	de Bellefon et al. (2020) [22]	Calculated starting from the first day of radiotherapy and defined as follows: death from SCCAC.
		Cancer-specific survival	Fraunholz et al. (2013) [35]	The time from start of CRT until death resulting from the cancer, or the date of last follow-up visit.
		Disease-specific survival	Oblak et al. (2016) [38]	The time interval from the beginning of the treatment to the death because of cancer.
9	**Local failure-free survival (n=2)**	Local recurrence-free survival	Cardenas et al. (2017) [33]	Not defined.
		Local failure-free survival	Fraunholz et al. (2013) [35]	The time from start of CRT to the first local tumor detection after CRT (ie. noncomplete response or local tumor recurrence after complete response) or to death (if the latter event occurred before a local failure was diagnosed), or the date of last follow-up visit.

3.8.5 Supplementary material E*. All outcomes reported in each study, along with all factors tested in both univariable and multivariable analysis.*

#	Study	Outcomes	Factors identified as prognostic using univariable analysis	Factors identified as prognostic using multivariable analysis
1	Shakir et al. (2020) [15]	Locoregional recurrence, distant relapse, persistent disease, disease-free survival, overall survival	Sex, performance status, T stage, N stage, RT completion, chemotherapy type	Sex, N stage, RT completion, performance status
2	Martin et al. (2020) [21]	Local relapse free survival, distant metastasis-free survival, disease-free survival, overall survival	T stage, N stage, gender, high grade acute organ toxicity	T stage, N stage, gender, high grade acute organ toxicity
3	de Bellefon et al. (2020) [22]	Locoregional failure, overall survival, colostomy-free survival, disease-free survival, metastasis-free survival	T stage, AJCC stage, N stage, exclusive RT, lack of MMC, RT breaks	T stage, N stage, AJCC stage, sex, RT breaks, exclusive RT, lack of MMC, residual disease
4	Brown et al. (2018) [23]	Progression-free survival	*N/A - No univariable analysis performed.*	T stage, N stage, planned total RT dose, planned total RT fractions, Minimum CT value, GLCM entropy log10- PET, GLCM entropy log2- PET, NGLDM busyness- PET, total SMTV, total TLG
5	Rouard et al. (2019) [24]	Overall survival, locoregional recurrence, local recurrence, regional recurrence	Age, immunodepression, definitive RT break, anal tumour boost technique, anal tumour total dose, performance status, active smoking, differentiation, lack of MMC, N stage, external iliac involvement at diagnosis, inguinal involvement at diagnosis, keratinisation, PLNA with NC delineation, involved LN not boosted, internal iliac delineation	Age, immunodepression, performance status, active smoking, external iliac involvement at diagnosis, PLNA with NC delineation
6	Franco et al. (2018) [25]	Progression-free survival, overall survival	Sex, N stage, basal haemoglobin levels, response to treatment	Sex, N stage, basal haemoglobin levels, response to treatment
7	Call et al. (2016) [26]	Overall survival, local control, regional control, distant control, colostomy-free survival	N stage, T stage	N stage, T stage, RT duration
8	Balermpas et al. (2017) [27]	Cumulative incidence of locoregional failure, disease-free survival, overall survival	Age, sex, T stage, N stage, HPV16 load, p16, CD8, PD-1, PD-L1, FOXP3, pCASP-8	Age, sex, HPV16 load, p16, CD8, PD-1, PD-L1, FOXP3, pCASP-8
9	Rodel et al. (2018) [28]	Cumulative incidence of locoregional failure, cumulative incidence of distance metastases, disease-free survival, overall survival	Gender, T stage, N stage, HPV16 load, RITA expression	Gender, T stage, N stage, HPV16 load, RITA expression
10	Schernberg et al. (2017) [29]	Overall survival, progression-free survival, locoregional control, distant metastases control	Leukocytosis, neutrophilia, anaemia, sex, performance status	Leukocytosis, neutrophilia, anaemia, sex, performance status
11	Martin et al. (2019) [30]	Locoregional control rate, disease-free survival, distant metastasis-free survival, overall survival	C reactive Protein to Albumin Ratio (CAR), gender, N stage	C reactive Protein to Albumin Ratio (CAR), N stage

12	Oehler-Janne et al. (2008) [31]	Overall survival, time to recurrence	*N/A - No univariable analysis performed.*	N stage, severe acute skin toxicity
13	Susko et al. (2020) [32]	Freedom from local recurrence, freedom from distant metastasis, overall survival	T stage, time from diagnosis to RT initiation, RT duration	T stage, time from diagnosis to RT initiation, RT duration
14	Cardenas et al. (2017) [33]	Local recurrence–free survival, overall survival	Pretreatment SUVmax, posttreatment SUVmax, ΔSUVmax, 5-FU/MMC chemotherapy, use of IMRT	Posttreatment SUVmax, ΔSUVmax, 5-FU/MMC chemotherapy, use of IMRT
15	Bitterman et al. (2015) [34]	Overall survival, disease-free survival, colostomy-free survival	*N/A - No univariable analysis performed.*	T stage, use of IMRT
16	Fraunholz et al. (2013) [35]	Local failure-free survival, distant metastases-free survival, cancer-specific survival, overall survival	Sex, T stage, N stage, grade, EGFR expression	Sex, N stage, grade
17	Schernberg et al. (2017)* [36]	Overall survival, progression-free survival, locoregional failure-free survival, distant failure-free survival	Leukocytosis, neutrophilia, anaemia, T stage, N stage, CRT duration	Leukocytosis, neutrophilia, anaemia, N stage, CRT duration
18	Hosni et al. (2018) [37]	Colostomy-free survival, disease-free survival, colostomy-free survival	*N/A - Univariable analysis performed but no significant prognostic factors identified*	T stage, sex, age, anal canal cancer with perianal extension
19	Oblak et al. (2016) [38]	Locoregional control, disease-free survival, disease-specific survival, overall survival	Pretreatment Hb level, mean on-treatment Hb level, end-of-treatment Hb level, performance status, T stage, N stage, overall disease stage, histologic tumour type, tumour site, blood transfusion, overall ratiation time, operation	Pre-treatment Hb level, overall disease stage

3.8.6 Supplementary material F*. Clinical factors identified as prognostic for worse outcomes through univariable and multivariable analysis, stratified by outcome. Where available, factor effects and parameterisation used for analysis are also included.*

Univariable analysis						
Outcome (number of studies reporting outcome)	**Risk factor**	**Times identified as prognostic**	**Total times tested**	**Factor effect (HR, 95% CI)**	**Note**	**Study**
Overall survival (n=17)	Higher N stage	10	16	3.40 (1.59-7.27)	Multiple categories (N0,N1,N2,N3)	Shakir et al. (2020) [15]
				N/A	N0 vs N+	Martin et al. (2020) [21]
				N/A	N0 vs N+	de Bellefon et al. (2020) [22]
				2.11 (1.31-2.90)	N0 vs N+	Franco et al. (2018) [25]
				N/A	Multiple categories (N0,N1,N2,N3)	Call et al. (2016) [26]
				N/A	N0 vs N+	Balermpas et al. (2017) [27]
				N/A	N0 vs N+	Rodel et al. (2018) [28]
				N/A	N0 vs N+	Fraunholz et al. (2013) [35]
				N/A	N0 vs N+	Schernberg et al. (2017)* [36]
				N/A	N0 vs N+	Oblak et al. (2016) [38]
	Higher T stage	9	16	4.15 (1.21-14.25)	Multiple categories (T1,T2,T3,T4)	Shakir et al. (2020) [15]
				N/A	T1-2 vs T3-4	Martin et al. (2020) [21]
				N/A	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
				N/A	T1-2 vs T3-4	Balermpas et al. (2017) [27]
				N/A	T1-2 vs T3-4	Rodel et al. (2018) [28]
				N/A	T1-2 vs T3-4	Fraunholz et al. (2013) [35]
				N/A	T1-2 vs T3-4	Schernberg et al. (2017)* [36]
				3.59 (1.30-9.88)	T1-2 vs T3-4	Hosni et al. (2018) [37]
				N/A	T1-3 vs T4	Oblak et al. (2016) [38]
	Male sex	7	12	2.93 (1.64-5.24)	Female/Male	Shakir et al. (2020) [15]
				N/A	Female/Male	Martin et al. (2020) [21]
				2.23 (1.42-3.05)	Female/Male	Franco et al. (2018) [25]
				N/A	Female/Male	Balermpas et al. (2017) [27]
				N/A	Female/Male	Rodel et al. (2018) [28]
				N/A	Female/Male	Schernberg et al. (2017) [29]
				3.38 (1.09-10.50)	Female/Male	Hosni et al. (2018) [37]
	Worse performance status	3	4	11.61 (2.56-52.75)	Multiple categories (PS0,PS1,PS2,PS3)	Shakir et al. (2020) [15]
				N/A	0 vs 1/2	Schernberg et al. (2017) [29]
				N/A	0 vs 1-3	Oblak et al. (2016) [38]
	Older age	3	4	2.15 (1.16-3.98)	<65 vs ≥65	Rouard et al. (2019) [24]
				N/A	≤59 vs >59	Balermpas et al. (2017) [27]
				1.05 (1.00-1.09)	Continuous	Hosni et al. (2018) [37]
	Incomplete/interrupted RT or breaks	2	2	6.21 (2.98-12.95)	No/Yes	Shakir et al. (2020) [15]
				3.25 (1.15-9.13)	No/Yes	Rouard et al. (2019) [24]
	Longer CRT duration	2	5	N/A	No/Yes	Schernberg et al. (2017)* [36]

				N/A	≤ 1.08 months vs > 1.08 months	Oblak et al. (2016) [38]
	Immunodepression	1	1	3.70 (1.30-10.51)	Yes/No	Rouard et al. (2019) [24]
	External RT	1	2	2.38 (1.05-5.55)	Brachytherapy vs External RT	Rouard et al. (2019) [24]
	Lower anal tumour total dose	1	3	2.04 (1.04-4.00)	≥64Gy vs <64Gy	Rouard et al. (2019) [24]
	No response to treatment	1	1	6.26 (2.73-14.40)	Yes/No	Franco et al. (2018) [25]
	Diagnosis to RT initiation	1	1	1.02 (1.00-1.04)	Continuous	Susko et al. (2020) [32]
	Lack of 5-FU/MMC chemotherapy	1	1	12.5	No/Yes	Cardenas et al. (2017) [33]
	Higher tumour grade	1	5	N/A	G1-2 vs G3	Fraunholz et al. (2013) [35]
	Anal canal cancer with perianal extension	1	1	3.04 (1.10-8.38)	No/Yes	Hosni et al. (2018) [37]
	Larger maximum primary tumor size	1	3	1.16 (1.02-1.32)	Continuous	Hosni et al. (2018) [37]
	Higher AJCC stage	1	1	N/A	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]
	Histologic tumour type	1	2	N/A	Basaloid vs squamous	Oblak et al. (2016) [38]
	Blood transfusion	1	1	N/A	No/Yes	Oblak et al. (2016) [38]
	Operation	1	1	N/A	No/Yes	Oblak et al. (2016) [38]
Locoregional failure (n=11)	Higher N stage	7	11	3.05 (1.63-5.73)	Multiple categories (N0,N1,N2,N3)	Shakir et al. (2020) [15]
				N/A	N0 vs N+	Martin et al. (2020) [21]
				N/A	Multiple categories (N0,N1,N2,N3)	Call et al. (2016) [26]
				N/A	N0 vs N+	Balermpas et al. (2017) [27]
				N/A	N0 vs N+	Rodel et al. (2018) [28]
				N/A	N0 vs N+	Martin et al. (2019) [30]
				N/A	N0 vs N+	Oblak et al. (2016) [38]
	Higher T stage	7	11	5.17 (1.55-17.28)	Multiple categories (T1,T2,T3,T4)	Shakir et al. (2020) [15]
				N/A	T1-2 vs T3-4	Martin et al. (2020) [21]
				N/A	T1-2 vs T3-4	Call et al. (2016) [26]
				N/A	T1-2 vs T3-4	Balermpas et al. (2017) [27]
				N/A	T1-2 vs T3-4	Rodel et al. (2018) [28]
				4.43 (1.93-10.16)	Multiple categories (T1,T2,T3,T4)	Susko et al. (2020) [32]
				N/A	T1-3 vs T4	Oblak et al. (2016) [38]
	Male sex	5	9	1.78 (1.09-2.91)	Female/Male	Shakir et al. (2020) [15]
				N/A	Female/Male	Martin et al. (2020) [21]
				N/A	Female/Male	Balermpas et al. (2017) [27]
				N/A	Female/Male	Rodel et al. (2018) [28]
				N/A	Female/Male	Schernberg et al. (2017) [29]
	Worse performance status	4	4	1.90 (1.14-3.19)	Multiple categories (PS0,PS1,PS2,PS3)	Shakir et al. (2020) [15]
				3.01 (1.05-8.66)	No/Yes	Rouard et al. (2019) [24]
				N/A	0 or 1 vs ≥2	Schernberg et al. (2017) [29]
				N/A	0 vs 1-3	Oblak et al. (2016) [38]
	Longer RT duration	2	2	1.05 (1.02-1.08)	Continuous	Susko et al. (2020) [32]
				N/A	≤ 1.08 months vs > 1.08 months	Oblak et al. (2016) [38]
	Incomplete/interrupted RT	1	2	5.29 (2.83-9.90)	No/Yes	Shakir et al. (2020) [15]
	Active smoking	1	1	2.22 (1.07-4.61)	No/Yes	Rouard et al. (2019) [24]
	Differentiation	1	1	4.31 (1.25-14.89)	Poorly/moderately/well differentiated	Rouard et al. (2019) [24]
	Lack of MMC chemotherapy	1	1	2.56 (1.16-5.88)	Yes/No	Rouard et al. (2019) [24]

	Immunodepression	1	1	3.54 (1.06-11.83)	No/Yes	Rouard et al. (2019) [24]
	External RT	1	1	3.57 (1.05-12.5)	Brachytherapy vs External RT	Rouard et al. (2019) [24]
	Higher number of involved LN at diagnosis	1	1	3.69 (1.24-11.04)	<2 vs ≥2	Rouard et al. (2019) [24]
	External iliac involvement at diagnosis	1	1	4.65 (1.55-13.93)	No/Yes	Rouard et al. (2019) [24]
	Inguinal involvement at diagnosis	1	1	3.16 (1.10-9.11)	No/Yes	Rouard et al. (2019) [24]
	No keratinisation	1	1	3.13 (1.03-9.10)	Yes/No	Rouard et al. (2019) [24]
	Higher PLNA with NC delineation	1	1	5.77 (1.29-25.78)	<10 vs ≥10	Rouard et al. (2019) [24]
	Higher number of involved LN not boosted	1	1	3.30 (1.03-10.52)	0 or 1 vs ≥2	Rouard et al. (2019) [24]
	NC Internal iliac delineation	1	1	4.20(1.17-15.08)	Conforming vs NC	Rouard et al. (2019) [24]
	Longer time to RT initiation from diagnosis	1	1	1.05 (1.02-1.08)	Continuous	Susko et al. (2020) [32]
	Higher AJCC stage	1	1	N/A	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]
	Squamous histologic tumour type	1	1	N/A	Basaloid vs squamous	Oblak et al. (2016) [38]
	Operation	1	1	N/A	No/Yes	Oblak et al. (2016) [38]
Disease-free survival (n=11)	Male sex	5	8	N/A	Female/Male	Martin et al. (2020) [21]
				N/A	Female/Male	Balermpas et al. (2017) [27]
				N/A	Female/Male	Schernberg et al. (2017) [29]
				N/A	Female/Male	Martin et al. (2019) [30]
				2.33 (1.00-5.46)	Female/Male	Hosni et al. (2018) [37]
	Higher N stage	4	9	N/A	N0 vs N+	Martin et al. (2020) [21]
				N/A	N0 vs N+	de Bellefon et al. (2020) [22]
				N/A	N0 vs N+	Balermpas et al. (2017) [27]
				N/A	N0 vs N+	Martin et al. (2019) [30]
	Higher T stage	4	10	N/A	T1-2 vs T3-4	Martin et al. (2020) [21]
				N/A	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
				N/A	T1-2 vs T3-4	Rodel et al. (2018) [28]
				6.25 (2.70-17.40)	T1-2 vs T3-4	Hosni et al. (2018) [37]
	Worse performance status	1	2	N/A	0 vs 1/2	Schernberg et al. (2017) [29]
	High grade acute organ toxicity	1	1	N/A	No/Yes	Martin et al. (2020) [21]
	Anal canal cancer with perianal extension	1	1	2.92 (1.26-6.75)	No/Yes	Hosni et al. (2018) [37]
	Larger maximum primary tumor size	1	2	1.23 (1.12-1.34)	Continuous	Hosni et al. (2018) [37]
Distant failure (n=5)	Male sex	1	3	N/A	Female/Male	Rodel et al. (2018) [28]
	Higher T stage	1	5	N/A	T1-2 vs T3-4	Rodel et al. (2018) [28]
	Higher N stage	1	5	N/A	N0 vs N+	Rodel et al. (2018) [28]
	Worse performance status	1	1	N/A	0 vs 1/2	Schernberg et al. (2017) [29]
Metastasis-free survival (n=5)	Higher T stage	5	5	N/A	T1-2 vs T3-4	Martin et al. (2020) [21]
				N/A	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
				N/A	T1-2 vs T3-4	Martin et al. (2019) [30]
				N/A	T1-2 vs T3-4	Fraunholz et al. (2013) [35]
				N/A	T1-2 vs T3-4	Schernberg et al. (2017)* [36]
	Higher N stage	4	5	N/A	N0 vs N+	Martin et al. (2020) [21]
				N/A	N0 vs N+	Martin et al. (2019) [30]

				N/A	N0 vs N+	Fraunholz et al. (2013) [35]
				N/A	N0 vs N+	Schernberg et al. (2017)* [36]
	Male sex	2	4	N/A	Female/Male	Martin et al. (2020) [21]
				N/A	Female/Male	Martin et al. (2019) [30]
	Higher tumour grade	1	3	N/A	In HIV- patients only	Fraunholz et al. (2013) [35]
	Longer CRT duration	1	1	N/A	<50 days vs >50 days	Schernberg et al. (2017)* [36]
Freedom from disease (n=4)	Higher N stage	4	4	4.02 (2.25-7.17)	Multiple categories (N0,N1,N2,N3)	Shakir et al. (2020) [15]
				N/A	N0 vs N+	Rodel et al. (2018) [28]
				N/A	N0 vs N+	Oblak et al. (2016) [38]
				N/A	In HIV- patients	Oehler-Janne et al. (2008) [31]
	Male sex	2	3	1.85 (1.18-2.92)	Female/Male	Shakir et al. (2020) [15]
				N/A	Female/Male	Rodel et al. (2018) [28]
	Higher T stage	2	3	4.48 (1.56-12.87)	Multiple categories (T1,T2,T3,T4)	Shakir et al. (2020) [15]
				N/A	T1-3 vs T4	Oblak et al. (2016) [38]
	Worse performance status	1	3	1.94 (1.21-3.12)	Multiple categories (PS0,PS1,PS2,PS3)	Shakir et al. (2020) [15]
	Incomplete/interrupted RT	1	1	4.98 (2.74-9.05)	No/Yes	Shakir et al. (2020) [15]
	Severe acute skin toxicity	1	1	N/A	No/Yes In HIV- patients	Oehler-Janne et al. (2008) [31]
	Higher AJCC stage	1	1	N/A	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]
	Squamous histologic tumour type	1	2	N/A	Basaloid vs squamous	Oblak et al. (2016) [38]
	Longer overall radiation time	1	1	N/A	≤ 1.08 months vs > 1.08 months	Oblak et al. (2016) [38]
	Operation	1	1	N/A	No/Yes	Oblak et al. (2016) [38]
Colostomy-free survival (n=4)	Higher T stage	3	4	N/A	T3-4 vs T1-2	de Bellefon et al. (2020) [22]
				N/A	T3-4 vs T1-2	Call et al. (2016) [26]
				3.83 (1.68-8.77)	T3-4 vs T1-2	Hosni et al. (2018) [37]
	Anal canal cancer with perianal extension	1	1	3.47 (1.56-7.74)	No/Yes	Hosni et al. (2018) [37]
	Larger maximum primary tumour size	1	1	1.18 (1.08-1.29)	Continuous	Hosni et al. (2018) [37]
Cancer-specific survival (n=3)	Higher T stage	2	3	N/A	T1-2 vs T3-4	Fraunholz et al. (2013) [35]
				N/A	T1-3 vs T4	Oblak et al. (2016) [38]
	Higher N stage	2	3	N/A	N0 vs N+	Fraunholz et al. (2013) [35]
				N/A	N0 vs N+	Oblak et al. (2016) [38]
	Higher tumour grade	1	1	N/A	G1-2 vs G3	Fraunholz et al. (2013) [35]
	Higher AJCC stage	1	3	N/A	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]
	Longer overall radiation time	1	1	N/A	≤ 1.08 months vs > 1.08 months	Oblak et al. (2016) [38]
	Operation	1	1	N/A	No/Yes	Oblak et al. (2016) [38]
Local failure-free survival (n=2)	Lack of 5-FU/MMC chemotherapy	1	1	4.76	No/Yes	Cardenas et al. (2017) [33]
	Lack of IMRT radiotherapy	1	1	5.56	No/Yes	Cardenas et al. (2017) [33]
	Male sex	1	1	N/A	Female/Male	Fraunholz et al. (2013) [35]
	Higher T stage	1	2	N/A	T1-2 vs T3-4	Fraunholz et al. (2013) [35]
	Higher N stage	1	2	N/A	N0 vs N+	Fraunholz et al. (2013) [35]
	Higher tumour grade	1	1	N/A	G1-2 vs G3	Fraunholz et al. (2013) [35]

Multivariable analysis					
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Factor effect (HR, 95% CI)**	**Note**	**Study**
Overall survival (n=17)	Male sex	7	4.00 (2.11-7.56)	Female/Male	Shakir et al. (2020) [15]
			1.92 (1.10-3.45)	Female/Male	Martin et al. (2020) [21]
			3.66 (1.56-8.60)	Female/Male	Franco et al. (2018) [25]
			3.13 (1.47-6.66)	Female/Male	Balermpas et al. (2017) [27]
			3.05 (1.42-6.55)	Female/Male	Rodel et al. (2018) [28]
			4.80 (1.60-14.50)	Female/Male	Schernberg et al. (2017) [29]
			4.50 (1.42-14.27)	Female/Male	Hosni et al. (2018) [37]
	Higher T stage	3	4.91 (2.25-10.72)	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
			2.88 (1.12-7.46)	T1-2 vs T3-4	Bitterman et al. (2015) [34]
			4.98 (1.69-14.72)	T1-2 vs T3-4	Hosni et al. (2018) [37]
	Older age	3	2.43 (1.29-4.60)	<65 vs ≥65	Rouard et al. (2019) [24]
			2.32 (1.13-4.73)	≤59 vs >59	Balermpas et al. (2017) [27]
			1.05 (1.00-1.09	Continuous	Hosni et al. (2018) [37]
	Higher N stage	3	2.25 (1.00-5.17)	N0 vs N+	Franco et al. (2018) [25]
			1.88 (1.16-3.10)	Multiple categories (N0,N1,N2,N3)	Call et al. (2016) [26]
			5.80	N0 vs N+	Schernberg et al. (2017)* [36]
	Higher AJCC stage	2	2.82 (1.22-6.53)	I/II/III vs IV	de Bellefon et al. (2020) [22]
			2.23 (1.17-4.26)	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]
	Worse performance status	1	10.71 (1.94-58.95)	Multiple categories (PS0,PS1,PS2,PS3)	Shakir et al. (2020) [15]
	Incomplete/interrupted RT	1	4.22 (1.78-10.00)	No/Yes	Shakir et al. (2020) [15]
	Exclusive RT	1	3.38 (1.29-10.72)	No/Yes	de Bellefon et al. (2020) [22]
	Lack of MMC	1	1.88 (0.92-3.85)	No/Yes	de Bellefon et al. (2020) [22]
	Immunodepression	1	5.05 (1.72-14.80)	No/Yes	Rouard et al. (2019) [24]
	No response to treatment	1	6.96 (2.96–16.50)	Yes/No	Franco et al. (2018) [25]
	Longer diagnosis to RT initiation	1	1.02 (1.00-1.05)	Continuous	Susko et al. (2020) [32]
	Lack of 5-FU/MMC chemotherapy	1	9.09	No/Yes	Cardenas et al. (2017) [33]
	Lack of IMRT radiotherapy	1	4.00 (1.30-12.5)	No/Yes	Bitterman et al. (2015) [34]
Locoregional failure (n=11)	Male sex	4	2.08 (1.24-3.48)	Female/Male	Shakir et al. (2020) [15]
			2.22 (1.16-4.38)	Female/Male	Martin et al. (2020) [21]
			2.56 (1.04-6.25)	Female/Male	Balermpas et al. (2017) [27]
			3.40 (1.30-9.40)	Female/Male	Schernberg et al. (2017) [29]
	Higher N stage	3	2.23 (1.13-4.39)	Multiple categories (N0,N1,N2,N3)	Shakir et al. (2020) [15]
			3.00 (1.55-5.81)	N0 vs N+	Martin et al. (2020) [21]
			3.58 (1.25-10.26)	N0 vs N+	Martin et al. (2019) [30]
	Incomplete/interrupted RT or breaks	2	4.96 (2.40-10.27)	No/Yes	Shakir et al. (2020) [15]
			2.47 (1.15-5.30)	No/Yes	de Bellefon et al. (2020) [22]

	Worse performance status	2	3.82 (1.31-11.09)	<2 vs ≥2	Rouard et al. (2019) [24]
			5.50 (2.20-14.00)	0 vs 1/2	Schernberg et al. (2017) [29]
	Exclusive RT	1	3.41 (1.21-9.57)	No/Yes	de Bellefon et al. (2020) [22]
	Lack of MMC	1	3.11 (1.28-7.56)	No/Yes	de Bellefon et al. (2020) [22]
	Active smoking	1	2.31 (1.11-4.82)	No/Yes	Rouard et al. (2019) [24]
	Immunodepression	1	7.25 (1.54-34.20)	No/Yes	Rouard et al. (2019) [24]
	External iliac involvement at diagnosis	1	7.89 (2.54-24.56)	No/Yes	Rouard et al. (2019) [24]
	Higher PLNA with NC delineation	1	9.09 (1.96-42.15)	<10 vs ≥10	Rouard et al. (2019) [24]
	Higher T stage	1	4.37 (1.83-10.47)	Multiple categories (T1,T2,T3,T4)	Susko et al. (2020) [32]
	Longer time to RT initiation from diagnosis	1	1.06 (1.03-1.010)	Continuous	Susko et al. (2020) [32]
Disease-free survival (n=11)	Male sex	4	2.13 (1.19-3.85)	Female/Male	Martin et al. (2020) [21]
			2.27 (2.38-4.35)	Female/Male	Balermpas et al. (2017) [27]
			3.60 (1.50-8.60)	Female/Male	Schernberg et al. (2017) [29]
			2.46 (1.04-5.73)	Female/Male	Hosni et al. (2018) [37]
	Higher T stage	3	2.57 (1.42-4.66)	Categorical	de Bellefon et al. (2020) [22]
			7.02 (2.76-17.83)	T1-2 vs T3-4	Hosni et al. (2018) [37]
			NA	Multiple categories (T1,T2,T3,T4), variable weighting reported (-0.011)	Brown et al. (2019) [23]
	Higher N stage	2	3.06 (1.70-5.49)	N0 vs N+	Martin et al. (2020) [21]
			NA	Multiple categories (N0,N1,N2,N3), variable weighting reported (-0.019)	Brown et al. (2019) [23]
	Lower planned total RT dose	1	NA	Continuous, variable weighting reported (0.007)	Brown et al. (2019) [23]
	Fewer planned total RT fractions	1	NA	Continuous, variable weighting reported (0.012)	Brown et al. (2019) [23]
	High grade acute organ toxicity	1	2.13 (1.20-3.70)	No/Yes	Martin et al. (2020) [21]
	Higher AJCC stage	1	2.23 (0.99-5.01)	I/II/III vs IV	de Bellefon et al. (2020) [22]
	Worse performance status	1	4.90 (2.10-11.50)	0 vs 1/2	Schernberg et al. (2017) [29]
	Longer CRT duration	1	33.33	<50 days vs >50 days	Schernberg et al. (2017)* [36]
Distant failure (n=5)	Male sex	1	3.83 (1.20-12.27)	Female/Male	Rodel et al. (2018) [28]
	Higher T stage	1	4.24 (1.43-12.57)	T1-2 vs T3-4	Rodel et al. (2018) [28]
Metastasis-free survival (n=5)	Male sex	2	4.08 (1.63-10.19)	Female/Male	Martin et al. (2020) [21]
			3.87 (1.08-13.84)	Female/Male	Fraunholz et al. (2013) [35]
	Higher T stage	2	3.54 (1.52-8.23)	T1-2 vs T3-4	Martin et al. (2020) [21]
			2.61 (1.45-4.70)	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
	Higher N stage	2	2.41 (1.0405.62)	N0 vs N+	Martin et al. (2020) [21]
			4.49 (1.20-16.80)	N0 vs N+	Martin et al. (2019) [30]
	Higher AJCC stage	1	3.05 (1.41-6.62)	I/II/III vs IV	de Bellefon et al. (2020) [22]
	Higher tumour grade	1	5.88 (1.72-20.00)	G1-2 vs G3	Fraunholz et al. (2013) [35]
Freedom from disease (n=4)	Male sex	2	2.16 (1.34-3.48)	Female/Male	Shakir et al. (2020) [15]
			2.16 (1.09-4.26)	Female/Male	Rodel et al. (2018) [28]
	Higher N stage	1	2.73 (1.43-5.21)	Multiple categories (N0,N1,N2,N3)	Shakir et al. (2020) [15]
	Incomplete/interrupted RT	1	4.50 (2.26-8.97)	No/Yes	Shakir et al. (2020) [15]
Colostomy-free survival (n=4)	Higher T stage	3	4.10 (2.23-7.52)	T1-2 vs T3-4	de Bellefon et al. (2020) [22]
			4.00 (1.03-17.09)	T1-2 vs T3-4	Call et al. (2016) [26]

			3.65 (1.59-8.37)	T1-2 vs T3-4	Hosni et al. (2018) [37]
	Male sex	1	1.90 (1.10-3.10)	Female/Male	de Bellefon et al. (2020) [22]
	Residual disease	1	7.78 (3.41-17.77)	No/Yes	de Bellefon et al. (2020) [22]
	Exclusive RT	1	3.03 (1.39-6.57)	No/Yes	de Bellefon et al. (2020) [22]
	Anal canal cancer with perianal extension	1	3.17 (1.42-7.09)	No/Yes	Hosni et al. (2018) [37]
Cancer-specific survival (n=3)	Male sex	1	4.13 (1.24-13.63)	Female/Male	Fraunholz et al. (2013) [35]
	Higher N stage	1	6.25 (1.51-25.00)	N0 vs N+	Fraunholz et al. (2013) [35]
	Higher AJCC stage	1	3.52 (1.38-9.03)	I/II vs IIIA/IIIB	Oblak et al. (2016) [38]

***3.8.7 Supplementary material G**. Biomarkers identified as prognostic for worse outcomes through univariable and multivariable analysis, stratified by outcome. Where available, factor effects and parameterisation used for analysis are also included.*

Univariable analysis						
Outcome (number of studies reporting outcome)	**Factor**	**Times identified as prognostic**	**Total times tested**	**Factor effect (HR, 95% CI)**	**Note**	**Study**
Overall survival (n=17)	Lower HPV16 load	2	3	N/A	>/≤ median	Balermpas et al. (2017) [27]
				N/A	>/≤ median	Rodel et al. (2018) [28]
	Neutrophilia	2	2	N/A	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
				N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Anaemia	2	2	N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017) [29]
				N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017)* [36]
	Lower basal heamoglobin levels	1	1	2.00 (1.20-3.33)	Continuous	Franco et al. (2018) [25]
	Lower CD8 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-1 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower RITA expression	1	1	N/A	>/≤ WS6	Rodel et al. (2018) [28]
	Leukocytosis	1	2	N/A	Present (leukocytes >10G/L) vs absent	Schernberg et al. (2017) [29]
	High C reactive protein to albumin ratio	1	1	N/A	≤/> 0.117	Martin et al. (2019) [30]
	Lower pre-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower mean on-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower end-of-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
Locoregional failure (n=11)	Lower HPV16 load	2	3	N/A	>/≤ median	Balermpas et al. (2017) [27]
				N/A	>/≤ median	Rodel et al. (2018) [28]
	Lower p16 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower CD8 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-1 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-L1 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Weaker FOXP3 phosporylation	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Weaker pCasp-8 phosporylation	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower RITA expression	1	1	N/A	>/≤ WS6	Rodel et al. (2018) [28]
	Leukocytosis	1	1	N/A	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
	Neutrophilia	1	1	N/A	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
	Anaemia	1	1	N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017) [29]
	High C reactive protein to albumin ratio	1	1	N/A	≤/> 0.117	Martin et al. (2019) [30]
	Lower pre-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower mean on-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower end-of-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Leukocytosis	2	2	N/A	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]

Disease-free survival (n=11)				N/A	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	2	2	N/A	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
				N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Lower CD8 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-1 expression	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Weaker FOXP3 phosporylation	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Weaker pCasp-8 phosporylation	1	1	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Lower HPV16 load	1	2	N/A	>/≤ median	Balermpas et al. (2017) [27]
	Anaemia	1	2	N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017) [29]
	High C reactive protein to albumin ratio	1	1	N/A	≤/> 0.117	Martin et al. (2019) [30]
Distant failure (n=5)	Lower HPV16 load	1	1	N/A	>/≤ median	Rodel et al. (2018) [28]
	Lower RITA expression	1	1	N/A	>/≤ WS6	Rodel et al. (2018) [28]
	Leukocytosis	1	1	N/A	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
	Neutrophilia	1	1	N/A	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
Metastasis-free survival (n=5)	High C reactive protein to albumin ratio	1	1	N/A	≤/> 0.117	Martin et al. (2019) [30]
	Leukocytosis	1	1	N/A	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	1	1	N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Anaemia	1	1	N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017)* [36]
Freedom from disease (n=4)	Lower HPV16 load	1	1	N/A	>/≤ median	Rodel et al. (2018) [28]
	Lower RITA expression	1	1	N/A	>/≤ WS6	Rodel et al. (2018) [28]
	Lower pre-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower end-of-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
Cancer-specific survival (n=3)	EGFR expression	1	1	N/A	Intermediate/Intense vs Absent/Weak	Fraunholz et al. (2013) [35]
	Lower pre-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
	Lower end-of-treatment heamoglobin levels	1	1	N/A	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
Local failure-free survival (n=2)	Leukocytosis	1	1	N/A	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	1	1	N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]

Multivariable analysis

Outcome (number of studies reporting outcome)	Factor	Times identified as prognostic	Total times tested	Factor effect (HR, 95% CI)	Note	Study
Overall survival (n=17)	Leukocytosis	2		4.60 (1.40-14.90)	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
				19.90	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	2		4.40 (1.30-14.80)	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
				22.70	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Lower basal heamoglobin levels	1		1.89 (1.15-3.03)	Continuous	Franco et al. (2018) [25]
	Lower HPV16 load	1		2.27 (1.05-5.00)	>/≤ median	Balermpas et al. (2017) [27]
	Lower RITA expression	1		3.19 (1.29-7.86)	>/≤ WS6	Rodel et al. (2018) [28]
	High C reactive protein to albumin ratio	1		4.47 (1.53-13.03)	≤/> 0.117	Martin et al. (2019) [30]
	Anaemia	1		5.40	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017)* [36]

	Lower pre-treatment heamoglobin levels	1	2.38 (1.08-5.26)	> 120 g/L vs ≤ 120 g/L	Oblak et al. (2016) [38]
Locoregional failure (n=11)	Lower HPV16 load	2	3.57 (1.29-10)	>/≤ median	Balermpas et al. (2017) [27]
			4.51 (1.15-13.46)	>/≤ median	Rodel et al. (2018) [28]
	Lower p16 expression	1	3.13 (1.30-7.14)	>/≤ median	Balermpas et al. (2017) [27]
	Lower CD8 expression	1	4.00 (1.20-14.29)	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-1 expression	1	3.45 (1.39-8.33)	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-L1 expression	1	3.70 (1.11-12.5)	>/≤ median	Balermpas et al. (2017) [27]
	Lower RITA expression	1	4.35 (1.45-13.02)	>/≤ WS6	Rodel et al. (2018) [28]
	Leukocytosis	1	4.50 (1.30-15.60)	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
	Neutrophilia	1	3.60 (1.20-11.60)	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
	Anaemia	1	4.10 (1.30-12.40)	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017) [29]
Disease-free survival (n=11)	Leukocytosis	2	7.10 (2.50-20.20)	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
			6.90	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	2	5.00 (1.70-14.50)	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
			7.60	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Anaemia	2	5.30 (1.90-14.70)	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017) [29]
			2.50	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017)* [36]
	Lower CD8 expression	1	2.38 (1.15-5.00)	>/≤ median	Balermpas et al. (2017) [27]
	Lower PD-1 expression	1	2.17 (1.16-4.00)	>/≤ median	Balermpas et al. (2017) [27]
	Weaker FOXP3 phosporylation	1	1.85 (1.00-3.45)	>/≤ median	Balermpas et al. (2017) [27]
	Weaker pCasp-8 phosporylation	1	2.04 (1.06-3.84)	>/≤ median	Balermpas et al. (2017) [27]
	Lower HPV16 load	1	2.50 (1.27-5.00)	>/≤ median	Balermpas et al. (2017) [27]
Distant failure (n=5)	Leukocytosis	1	4.00 (1.60-10.30)	Absent vs present (leukocytes >10G/L)	Schernberg et al. (2017) [29]
	Neutrophilia	1	3.30 (1.20-9.10)	Absent vs present (neutrophils >7G/L)	Schernberg et al. (2017) [29]
Metastasis-free survival (n=5)	Leukocytosis	1	N/A	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	1	N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]
	Anaemia	1	N/A	Absent vs present (hemoglobin count < 13.0 g/dL)	Schernberg et al. (2017)* [36]
Freedom from disease (n=4)	Lower HPV16 load	1	2.28 (1.08-4.79)	>/≤ median	Rodel et al. (2018) [28]
	Lower RITA expression	1	2.19 (1.07-4.47)	>/≤ WS6	Rodel et al. (2018) [28]
Local failure-free survival (n=2)	Leukocytosis	1	N/A	Absent vs present (leukocytes >10000/mm3)	Schernberg et al. (2017)* [36]
	Neutrophilia	1	N/A	Absent vs present (neutrophils >7500/mm3)	Schernberg et al. (2017)* [36]

*3.8.8 **Supplementary material H**. Imaging factors identified as prognostic for worse outcomes through univariable and multivariable analysis, stratified by outcome. Where available, factor effects are also included.*

Univariable analysis

Outcome (number of studies reporting outcome)	Factor	Times identified as prognostic	Total times tested	Factor effect (HR)	Study
Overall survival (n=17)	Higher posttreatment SUVmax	1	1	3.23	Cardenas et al. (2017) [33]
	Smaller ΔSUVmax	1	1	4.35	Cardenas et al. (2017) [33]
Local failure-free survival (n=2)	Lower pretreatment SUVmax	1	1	3.57	Cardenas et al. (2017) [33]
	Higher posttreatment SUVmax	1	1	4.35	Cardenas et al. (2017) [33]

Multivariable analysis

Outcome (number of studies reporting outcome)	Factor	Times identified as prognostic	Factor effect	Study
Overall survival (n=17)	Higher posttreatment SUVmax	1	2.77	Cardenas et al. (2017) [33]
	Smaller ΔSUVmax	1	3.33	Cardenas et al. (2017) [33]
Distant failure (n=5)	Minimum CT value	1	N/A	Brown et al. (2019) [23]
	GLCM entropy log10- PET	1	N/A	Brown et al. (2019) [23]
	GLCM entropy log2- PET	1	N/A	Brown et al. (2019) [23]
	NGLDM busyness- PET	1	N/A	Brown et al. (2019) [23]
	Total SMTV	1	N/A	Brown et al. (2019) [23]
	Total TLG	1	N/A	Brown et al. (2019) [23]
Local failure-free survival (n=1)	Higher posttreatment SUVmax	1	5.88	Cardenas et al. (2017) [33]

Chapter 4 - Predicting outcomes in anal cancer patients using multi-centre data and distributed learning - a proof-of-concept study

4.1 Abstract

4.1.1 Background and purpose

Predicting outcomes is challenging in rare cancers. Single-institutional datasets are often small and multi-institutional data sharing is complex. Distributed learning allows machine learning models to use data from multiple institutions without exchanging individual patient-level data. We demonstrate this technique in a proof-of-concept study of anal cancer patients treated with chemoradiotherapy across multiple European countries.

4.1.2 Materials and methods

atomCAT is a three-centre collaboration between Leeds Cancer Centre (UK), MAASTRO Clinic (The Netherlands) and Oslo University Hospital (Norway). We trained and validated a Cox proportional hazards regression model in a distributed fashion using data from 281 patients treated with radical, conformal chemoradiotherapy for anal cancer in three institutions. Our primary endpoint was overall survival. We selected disease stage, sex, age, primary tumour size, and planned radiotherapy dose (in EQD2) a priori as predictor variables.

4.1.3 Results

The Cox regression model trained across all three centres found worse overall survival for high risk disease stage (HR=2.02), male sex (HR=3.06), older age (HR=1.33 per 10 years), larger primary tumour volume (HR=1.05 per 10cm^3) and lower radiotherapy dose (HR=1.20 per 5 Gy). A mean concordance index of 0.72 was achieved during validation, with limited variation between centres (Leeds=0.72, MAASTRO=0.74, Oslo=0.70). The global model performed well for risk stratification for two out of three centres.

4.1.4 Conclusions

Using distributed learning, we accessed and analysed one of the largest available multi-institutional cohorts of anal cancer patients treated with modern radiotherapy techniques. This demonstrates the value of distributed learning in outcome modelling for rare cancers.

4.2 Introduction

Prediction models for cancer outcomes can support clinical decision making, and hold the promise for individualisation of cancer treatment and radiotherapy plan optimisation. Development of robust and validated models is often hampered by lack of access to data, however, especially across countries and institutions. This is particularly the case for rare cancers.

“Distributed learning” facilitates the development and validation of statistical models using data across multiple institutions without transferring individual patient data outside the originating institution. This is one of several novel methodologies developed to preserve patient data privacy [1,2], such as differential privacy and encryption [3]. Our distributed learning approach is an open-source solution (Vantage6) which prevents insider attacks by blocking any direct connection between data hosts [4,5]. Only locally aggregated statistics (model coefficients and fit errors) are exchanged between the data centres and the central server. Model development in the distributed learning framework is an iterative mathematical optimization problem where the coefficients of a single globally-convergent model will be determined by minimizing the total error [6]. The general methodology has been shown to be scalable up to vast numbers of patients [7].

The distributed learning approach may be ideally suited to rare diseases, where single-institutional datasets are limited in size and sharing data between institutions is restricted by data protection regulations and related ethical considerations [8]. One such example is anal cancer; a rare disease with an incidence rate around 2.1 per 100,000 person-years in Northern Europe and twice the incidence in women relative to men [9]. Currently, the standard treatment for localised disease involves concomitant radiotherapy and chemotherapy [10], which leads to a complete response in approximately 3 out of 4 patients. 5-year overall survival rates of 75% have previously been reported [11–13]. Further improvements in disease control and survival have proven challenging, and questions remain around optimal tumour dose [14–16]. Additionally, patients that undergo standard treatment commonly suffer from various early and late side effects, such as gastrointestinal symptoms that range from mild to severe [17]. This highlights the need for a personalised approach to anal cancer chemoradiotherapy. Such individualisation will be dependent on the development of outcome prediction models [18], which again require sufficient data for model training and validation. A distributed learning approach may help obtaining sufficient patient data from different institutions in order to develop robust and generalisable models, while circumventing many of the barriers associated with individual-level patient data sharing.

In this proof-of-concept study, we aimed to show the feasibility of our distributed learning approach for patients with anal cancer receiving radical chemoradiotherapy. A prediction model for overall survival (OS), employing established baseline clinical factors and radiotherapy dose as predictors, was applied on data across institutions in three European countries. OS was chosen as our outcome of interest as this is an important outcome measure in anal cancer research [19] and a robust endpoint across institutions.

We hypothesise that a global Cox proportional hazards model developed without exchange of any individual-level patient data is highly reproducible in a multi-centre setting, when evaluated through an “internal-external” validation cycle [20], despite the small sample sizes within each participating centre. Furthermore, we hypothesize that we can define risk groups across institutions.

4.3 Materials & Methods

The study protocol was developed collaboratively by the three participating institutions prior to study initiation: Leeds Cancer Centre (UK), MAASTRO Clinic (The Netherlands), Oslo University Hospital (Norway). Patients were treated with chemoradiotherapy with radical intent for anal squamous cell carcinoma (ASCC), with conformal radiotherapy (forward-planned 3D conformal (3D-CRT) or intensity-modulated radiation therapy/volumetric modulated arc therapy (IMRT/VMAT)). Baseline, treatment and outcome data were available. The main outcome of interest for this proof-of-concept study was overall survival (OS). Death from any cause was counted as an event, with patients censored at the time of local data collection. Survival interval was calculated from the date of the first fraction of radiotherapy, to either date of death or the last follow-up date if alive.

For candidate outcome predictors, the literature on anal cancer chemoradiotherapy was reviewed, and expert input sought from three consultant clinical oncologists specialising in anal cancer. Importantly, we considered only predictors available at start of treatment (thus not radiotherapy compliance or treatment gaps). The following predictor variables were chosen, based on published data, clinical experience, and data availability in participating institutions: disease stage - low risk (Stage I-II, T1N0 or T2N0 or T3N0) versus high risk (Stage III, T4N(any) or T(any)N+) according to TNM v8 [21]; sex; age; primary tumour size (gross tumour volume, GTV, on planning CT); and primary tumour prescribed dose (converted from physical dose to equivalent dose in 2 Gy per fraction, $EQD2_{\alpha/\beta=10Gy}$). For disease stage, there is ongoing debate as to whether T3N0 tumours should be regarded as low or high risk [16]. The model was thus also fitted with T3N0 tumours assigned to the high rather than the low risk group. Additionally, histology (basaloid SCC: yes/no) was identified as a potential predictor, but was not included in the final analysis due to a large proportion of missing SCC subtype data in one institution. A data code book was shared between all institutions, for standardised data collection and reporting.

4.3.1 Patient data collection

For Leeds Cancer Centre, a subset of patients treated for anal cancer between 2015 and 2018 with baseline and outcome data available were included. All patients were treated with VMAT and simultaneous integrated boost (SIB). Patients were identified through existing research databases, and additional data was sourced as necessary from clinical databases. Tumour volumes were extracted manually from radiotherapy plans. Survival data were based on patient electronic records, which are automatically linked to the NHS England death registry.

At MAASTRO Clinic, patients treated by radiotherapy for primary anal cancer with radical intent between 2008 and 2017 were retrieved from electronic treatment records. All radiotherapy was in the form of either 3D-CRT (n=26; prior to 2013) or VMAT (n=55; after 2013), with dose to the primary tumour escalated by either sequential boost or SIB. Tumour volumes were extracted manually from radiotherapy planning delineations.

Dates of death were obtained from the electronic patient records, which were automatically updated from a Dutch citizens registry.

For Oslo University Hospital, anal cancer patients enrolled in the prospective ANCARAD trial (ClinicalTrials registration NCT01937780) receiving treatment between 2013 and 2017 were included. All patients received chemoradiotherapy using 3D-CRT (40 patients), IMRT (11 patients) or VMAT (69 patients), with boosts delivered either sequentially (109 patients) or as SIB (11 patients). Baseline and outcome data were prospectively collected as part of the ANCARAD trial. Additional baseline data were retrieved as necessary from clinical databases. Tumour volumes were extracted from radiotherapy structure sets in the treatment planning system using an in-house script.

Details on the radiotherapy and concomitant chemotherapy schedules used at each centre are shown in Table 4-1.

Table 4-1. Radiotherapy and concomitant chemotherapy treatment schedules used at each of the three centres.

	Leeds	**MAASTRO**	**Oslo**
Radiotherapy regimen	Most patients were prescribed 50.4-53.2 Gy to the primary tumour, 50.4Gy to involved nodes and 40 Gy to elective nodal volumes in 28 fractions. 5 patients were treated with doses above 53.2 Gy.	All patients were prescribed 54-66 Gy to the primary tumour and 39-49.5 Gy to elective lymph nodes in 30-33 fractions.	All patients were prescribed 54-58 Gy to the primary tumour and pathological lymph nodes and 46 Gy to elective nodal volumes in 27-29 fractions.
Chemotherapy regimen	Mitomycin-C (12 mg/m2 bolus day 1, capped at 20 mg) and 5-FU (1000 mg/m2 in 1 L normal saline over 24 hours, days 1-4 and days 29-32, capped at 2 m2).	Mitomycin-C (10 mg/m2 bolus day 1) plus either capecitabine (2 x 825 mg/m2 per radiotherapy treatment day) or continuous 5-FU (750 mg/m2 days 1-5 and 29-33); 11 patients who were elderly/frail or had T1N0M0 disease were treated with 66 Gy radiotherapy only.	Mitomycin-C (10 mg/m2 bolus day 1, capped at 20 mg) and 5-FU (1000 mg/m2 in 1 L normal saline over 24 hours, days 1-4), according to national guidelines. Patients with T1-T2 and N0 tumours received a single cycle (5-FU: days 1-4, MMC: day 1); patients with T3-4 tumours or N+ received two cycles (additional cycle in the fifth treatment week; 5-FU days 29-32, MMC day 29).

4.3.2 Institutional data access & data protection approvals

Each institution acquired separate local approvals for accessing and collecting patient data for research. As no individual patient data were exchanged between institutions, no data sharing agreements or additional patient consent were needed. Local information governance and data protection review of the distributed learning infrastructure were obtained wherever appropriate. In Leeds, the study was approved by LeedsCAT; a radiotherapy-specific institutional research governance board. In MAASTRO, IRB approval was obtained to extract patient data from electronic records. In Oslo, Regional Ethics Committee approval was obtained for re-use of data from the ANCARAD trial (via an amendment), and the local data protection officer reviewed and approved the distributed learning infrastructure.

4.3.3 Distributed learning architecture

We used the Vantage6 v0.2.4 software to set up three components; (1) “nodes” where patient-level data is accessed and where local model coefficients are computed, (2) a trusted coordinating “server” that performs aggregation of coefficients, and (3) a “researcher” that provides the model to be trained. The purpose was to fit a distributed Cox model for overall survival for anal cancer (see Figure 4-1). For additional security, all patient data were pseudonymized and stripped of protected health information (e.g. dates of treatments, dates of birth/death, generic medical record numbers, etc).

Nodes were set up on common personal computers (either physical or virtual) running any one of well-supported operating systems (Windows/MacOS/Ubuntu) with an installation of Python (v3.6 or later), Docker Desktop community edition, and Vantage6 v0.2.4. The complete source code for the infrastructure implementation is available [https://github.com/IKNL/vantage6 - Version 0.2.4]. Network connectivity was fully compliant with local institutional policies, and only one secured network port through the institution firewall was enabled for Vantage6 traffic.

The Leeds node was set up as a Windows 10 Pro virtual machine (Intel(R) Xeon(R) Gold 5118 CPU, 16GB RAM), and only accessible by NHS Trust users granted the appropriate permissions. Patient data were extracted from a clinical database, de-identified, and forwarded to the virtual machine. The MAASTRO node was set up as a physical Surface Book 2 laptop (Intel(R) Core i7-8650 CPU, 16GB RAM) running

Windows 10 Pro, and pseudonymized patient data accessed via a mapped folder directing to an internal storage server. The Oslo node was set up on a Lenovo ThinkPad laptop (Intel(R) i7-4600M CPU, 16GB RAM), running Ubuntu Linux 18.04 as a virtual machine, which can easily be cloned when setting up nodes for new projects. The Oslo node was physically decoupled from the hospital network, and pseudonymized data was transferred to the machine via an encrypted external hard disk drive.

The central coordination server takes the role of trusted messaging “broker” for the collaboration network. Only key-authenticated messages were allowed to pass between researcher and server, and between node and server. The server administrator maintains a registry of collaborations, researchers, institutions and institution administrators, as well as unique encryption keys for each role. For this proof-of-concept run, the server was set up by MAASTRO as an Ubuntu Linux 18.04 virtual instance (30GB storage, 4 GB memory) on the Microsoft Azure cloud computing service based in Europe.

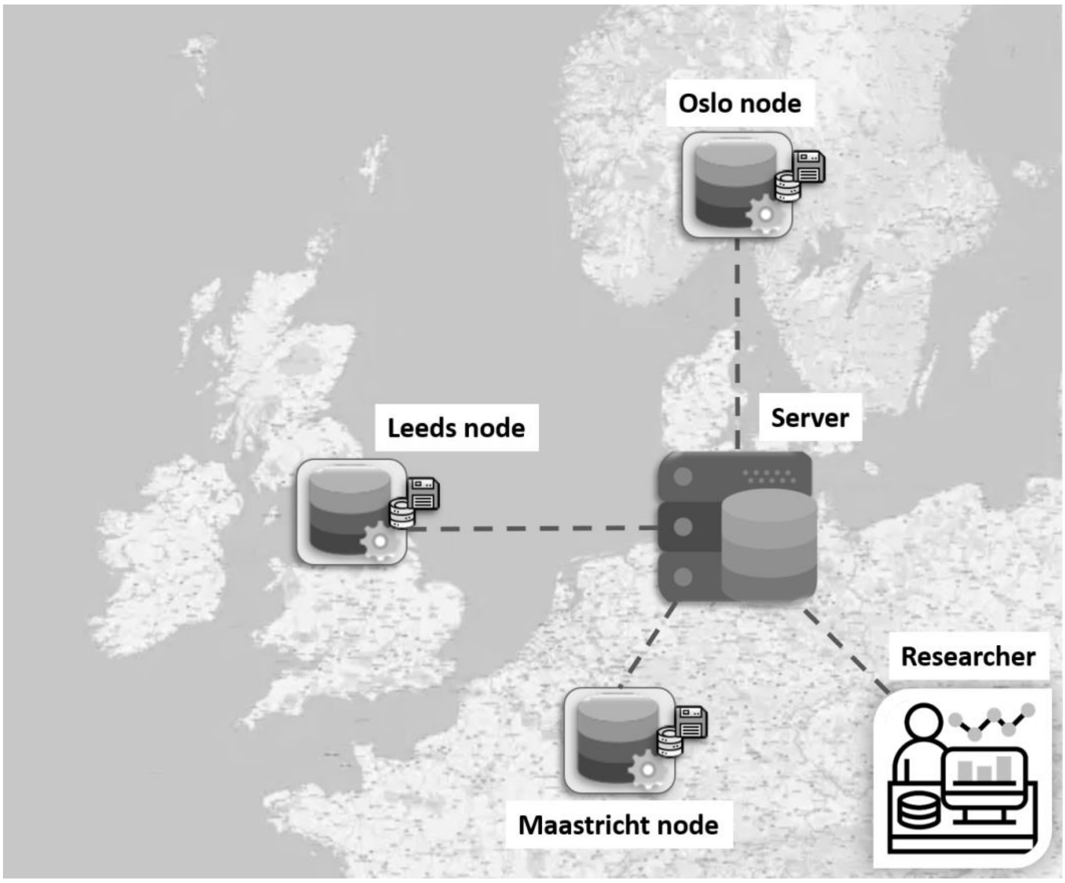


Figure 4-1. Distributed learning as a multinational collaboration to train a distributed Cox model for overall survival in anal cancer across three data nodes with a trusted coordination server in the Microsoft Azure cloud.

4.3.4 Descriptive data analysis

Summary statistics were exchanged between centres in order to explore cohort differences prior to modelling. Categorical variables were tested using a chi-squared test, and numerical variables were tested using a one-way ANOVA test. All tests were carried out using summary statistics (number of patients, mean and standard deviation values) rather than individual patient data. Estimated 3-year survival rates and potential follow-up times were calculated by each centre individually using the ‘survival’ package in R [22], employing the Kaplan-Meier estimator. Median follow-up time was based on the inverse Kaplan Meier estimator [23].

4.3.5 Distributed Cox algorithm

The Distributed Cox algorithm developed by Lu et al. [2] was adapted to the Vantage6 v0.2.4 infrastructure as R scripts (v.3.6.2). The source code has been made openly accessible on GitHub (https://github.com/AnanyaCN/d_coxph). Scripts for computing model coefficients, median risk score, and leave-one-centre-out model validation were packaged as application “containers” (via Docker) that were locally executed in each node.

4.3.6 Cox model development and validation

The primary analysis involved the development and validation of a Cox proportional hazards model across all centres. The performance of the model was initially assessed using Harrell’s concordance index (c-index) [24] on a per-centre basis. The global model’s performance was assessed on all data from all three institutions, which has been recommended by Steyerberg and Altman and TRIPOD [20,25] since small datasets should not be split during the model training phase. A more robust estimate for out-of-sample performance was obtained using a closed-loop “leave-one-centre-out” method [20], where new models were trained using data from two sites and then validated on the third site. This was repeated three times to cover the possible combinations, thus resulting in different c-indices which provide an estimate of the over-optimism of the global model. Additionally, the Schoenfeld residuals for each model variable were examined on a per centre level for the global model, and were tested for association with time, in order to examine whether the proportional hazard assumptions were fulfilled [26].

4.3.7 Visualisation of model performance

We evaluated the performance of the global model for risk stratification on a centre level. The individual patient risk score was defined as the overall risk for a patient relative to the baseline and was calculated as the exponent of the patient's linear predictor (LP) value (risk=e[LP]). A global median risk score from the global Cox regression model was estimated in an iterative procedure, with the median of the medians as a starting value. The global median risk score was used as cut-off for defining risk categories (high vs low risk), based on individual patient risk scores. Each centre subsequently produced a Kaplan-Meier data object independently in R, with their local survival curves stratified by risk categories, and then shared these objects. These only contained the coordinate points required to plot events and censored patients in a figure.

4.4 Results

A total of 281 patients were included in the analysis - 80 patients from Leeds, 81 patients from MAASTRO, and 120 patients from Oslo; see Table 4-2 for patient characteristics.

There were no significant differences in disease stage, age at the start of radiotherapy, or primary tumour GTV between the three cohorts. The Oslo cohort had a significantly higher proportion of female to male patients, as expected from the Norwegian anal cancer epidemiology [27]. EQD2 had the highest variance between cohorts, with a difference of 7.4Gy between the highest (MAASTRO) and lowest (Leeds) in mean dose. Moreover, all three cohorts had comparable outcomes and follow-up times. The 3-year survival estimates of Leeds were comparable to both other centres, while the 95% confidence intervals of MAASTRO and Oslo did not overlap.

Table 4-2. Overview of patient and treatment characteristics categorised by centre. P-values represent cohort comparisons using either chi-squared or one-way ANOVA tests. GTV: Gross tumour volume. EQD2: Equivalent dose in 2 Gy fractions (α/β=10Gy). IQR: Interquartile range. CI: Confidence interval.

	Leeds	MAASTRO	Oslo	p-value
Disease stage				
Low risk (T1-3N0)	28 (35%)	33 (41%)	58 (48%)	0.16
High risk (T4N(any) or T(any)N+)	52 (65%)	48 (59%)	62 (52%)	
Sex				
Female	53 (66%)	46 (57%)	88 (73%)	0.05
Male	27 (34%)	35 (43%)	32 (27%)	
Age at the start of radiotherapy (years)				
Mean (sd, range)	60 (12, 29-86)	61 (11, 28-84)	62 (10, 40-89)	0.44
Primary tumour GTV (cm^3)				
Mean (sd, range)	64.8 (58.7, 2.1-284.9)	57.5 (72.4, 0.8-433.0)	78.1 (69.4, 4.1-459.4)	0.09
Primary tumour dose (EQD2)				
Mean (sd, range)	52.8 (2.7, 49.1-62.6)	60.2 (2.7, 59.4-66.2)	56.3 (2.0, 54.0-58.1)	<0.0001
Potential follow-up time (months)				
Median (IQR)	46 (38-51)	42 (32-63)	49 (39-61)	N/A
Estimated 3-year survival				
Survival (std error, 95%CI)	83% (4%, 76-92%)	78% (5%, 70-88%)	93% (2%, 89-98%)	N/A
Outcome				
Alive	66 (83%)	63 (78%)	107 (89%)	N/A
Dead	14 (17%)	18 (22%)	13 (11%)	

The results of the global Cox regression model, trained on all three nodes, are summarised in Table 4-3 in the form of hazard ratio (HR) estimates.

Table 4-3. Results of the global distributed multivariate Cox regression analysis across all three centres. Age, primary tumour GTV and primary tumour dose were treated as continuous variables. The HRs represent a change of 10 years in age; 10cm^3 in primary tumour GTV; and 5 Gy in primary tumour dose (EQD2). GTV: Gross tumour volume. EQD2: Equivalent dose in 2 Gy fractions (α/β=10Gy). CI: Confidence interval

	Hazard ratio (95% CI)
High risk disease (compared to low risk disease)	2.02 (0.90-4.54)
Male sex (compared to female sex)	3.06 (1.54-6.11)
Age at the start of RT	1.33 (0.98-1.82)
Primary tumour GTV	1.05 (1.02-1.09)
Primary tumour dose (EQD2)	0.83 (0.48-1.43)

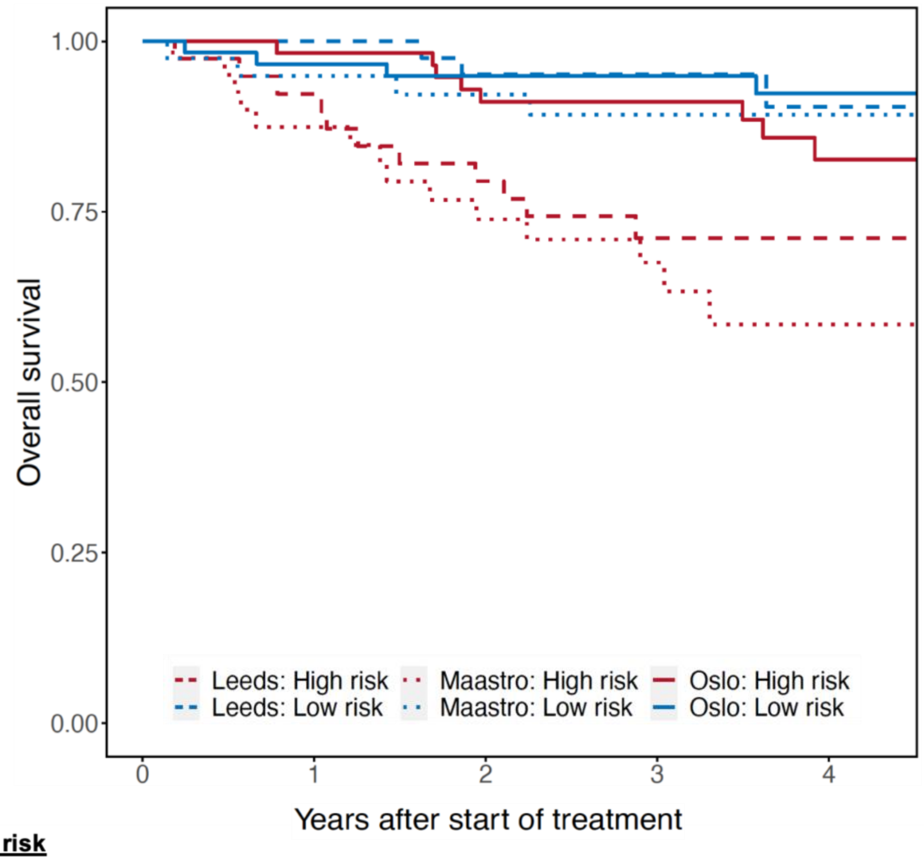
The results of the global model suggest that higher risk disease, older age at the start of radiotherapy, male sex, lower radiotherapy dose, and a greater volume primary tumour (GTV) are associated with worse overall survival. The global model's performance was assessed on each node, yielding a c-index of 0.72 for Leeds, 0.74 for MAASTRO, and 0.70 for Oslo. The c-indices from all three nodes are similar, suggesting that the model performs consistently well across centres.

In addition, the c-indices from the leave-one-centre-out validation runs (Table 4-4) suggest that the model performance remains stable when model training is carried out using data from only two centres and validated on a third, completely independent dataset. Moreover, the effects of factors are similar across centres, as all three runs produced similar hazard ratios for all variables. The only exception is prescription dose, where one model showed somewhat discordant effects. Notably, the effect of the primary tumour GTV is most consistent across the three validation runs. The overall results of the global model as well as the leave-one-centre-out validation runs were not considerably impacted when including T3N0 tumours in the high risk group (Supplementary material A, Section 4.7.1). The Schoenfeld test results convey that the proportional hazard assumptions were fulfilled for all variables in all three centres (Supplementary material B, Section 4.7.2).

Table 4-4. Results from the three leave-one-centre-out validation runs. Each column represents one run, consisting of model training (and associated hazard ratios, HR) on two nodes and validation on the third, independent node. Factor effects are presented in terms of hazard ratios with 95% confidence intervals; HR (95% CI). The HRs represent a change of 10 years in age; 10cm^3 in primary tumour GTV; and 5 Gy in primary tumour dose (EQD2). The resulting c-index from each validation run is also reported. GTV: Gross tumour volume. EQD2: Equivalent dose in 2 Gy fractions (α/β=10Gy).

Training nodes	MAASTRO Oslo	Leeds Oslo	Leeds MAASTRO
Validation node	Leeds	MAASTRO	Oslo
High risk disease (compared to low risk disease)	2.52 (0.93-6.78)	1.96 (0.68-5.67)	1.85 (0.71-4.86)
Male sex (compared to female sex)	3.59 (1.55-8.33)	3.83 (1.57-9.37)	2.12 (0.92-4.90)
Age at the start of RT	1.10 (0.74-1.64)	1.47 (0.99-2.17)	1.48 (1.05-2.10)
Primary tumour GTV	1.04 (1.00-1.08)	1.08 (1.03-1.13)	1.07 (1.03-1.11)
Primary tumour dose (EQD2)	0.97 (0.46-2.04)	0.35 (0.14-0.87)	0.97 (0.59-1.59)
Validation c-index	0.70	0.73	0.68

Risk scores were calculated using the global model. A global median risk score of 0.98 was used as the cut-off to define risk categories. Patients with individual risk scores lower than 0.98 were assigned in the low risk category, whereas patients with risk scores greater than 0.98 were assigned in the high risk category. The low risk category consisted of 141 patients (Leeds: 41, MAASTRO: 40, Oslo: 60); the high risk category included 140 patients (Leeds: 39, MAASTRO: 41, Oslo: 60). The Kaplan-Meier curves (Figure 4-2) convey that there is a good separation in overall survival between the low and high risk categories for two of the centres. For the third centre, the separation is small compared to the other centres.



Patients at risk					
Leeds: High risk	39	36	31	21	11
Leeds: Low risk	41	41	39	31	16
MAASTRO: High risk	41	33	26	17	8
MAASTRO: Low risk	40	37	33	25	18
Oslo: High risk	60	56	51	44	25
Oslo: Low risk	60	57	54	50	30

Figure 4-2. Kaplan-Meier overall survival curves for each centre's cohort, stratified into low and high risk categories. The curves were constructed using the global model, which was trained on data from all three centres. The HR of the high risk category relative to the low risk category is 4.39 [95% CI = 1.22-15.73] for Leeds, 4.02 [1.32-12.23] for MAASTRO and 1.73 [0.56-5.31] for Oslo.

4.5 Discussion

This proof-of-concept study demonstrates the feasibility of privacy preserving distributed learning for anal cancer. We trained and validated a Cox proportional hazards regression model [28] in a distributed fashion, using patient data from three European institutions, with clinical and treatment-related factors, and demonstrated robust model performance. Our approach is unique compared to previously published studies employing distributed learning, since we developed and applied a Cox proportional hazards regression model with a time-to-event outcome for a rare cancer. In contrast, other studies have explored binary outcomes using support-vector machines [6] or logistic regression [7]. In addition, the distributed learning architecture employed in our study is public, open-source and uses Docker containers for enhanced security.

Our analysis involved data for 281 patients treated with modern conformal radiotherapy techniques, including radiotherapy-specific data (GTV volume and prescription dose). This makes for one of the largest available cohorts of anal cancer patients treated with modern radiotherapy, and the only such study with robust multi-centre validation of outcome predictors. Shakir et al [29] reported outcome data from 385 patients treated with IMRT in five UK centres, with median follow-up of 24 months. de Meric de Bellefon et al [30] recently published long-term outcomes, including late toxicity data, for 193 patients treated with IMRT in a single French centre. No other studies have reported on cohorts of this size, and none with multi-national data. Our study could only be realised using the distributed learning methodology, which averted any need for data sharing agreements and data protection reviews.

We found, as expected, worse outcomes for patients with more advanced disease. This, and worse survival for males, mirrors previous results in the literature, including long-term data from RTOG 98-11 [31] and data from a large, prospective Nordic database [32]. Uniquely, by utilising data from 3D planned radiotherapy, we were able to include a volumetric measure of primary tumour size (GTV volume); with an increased risk observed for larger tumours even in multivariate analysis taking staging into account. Tumour size appears to be the most stable factor across all model runs. The relatively weak predictive power of radiation dose was expected as overall survival, and not tumour control, was used as endpoint. Still, the observed effect size was equivalent to that reported for local control in the study by Johnsson et al [14].

Our analysis was limited to data available in routine clinical records for two of the participating centres, and as such potential predictors for outcome were restricted. We selected up front the three clinical factors which we expected to have the largest impact on survival (stage, age, sex), in addition to two radiotherapy-related factors (GTV volume, dose). This process necessarily required some prioritisation, and other factors could equally well have been included such as HPV status, chemotherapy prescription, anatomical site (anal canal versus anal margin), and performance status. We did not examine non-linear effects of age, dose or GTV volume, nor interactions between factors; all of which might be of interest in a more definitive study. Other limitations include variation in staging and GTV definition between centres, as one would expect from a non-prospective multi-centre analysis.

Importantly, the current study was designed to test the feasibility of distributed learning in a rare cancer, with the prospect of accessing combined patient cohorts rivalling the largest reported in the literature. It was not designed as a quality improvement exercise, and as such did not attempt to compare outcomes between centres for specific tumour stages or other patient subgroups. Neither did we set out to produce a definitive model to guide treatment or to test novel predictors for outcome. In its current state, this model is not ready to be used for individual patient predictions. In addition to the inherent limitations related to the medium-size data set, a global baseline survival curve cannot be provided, which prevents individual patient survival risk estimates. This is a deficiency in the current implementation of Vantage6, which will be addressed in future versions. We examined the use of our global model for risk stratification on an individual centre level, and found good results for two centres. The inability of the model to properly stratify patients in the third centre (Figure 4-2) may possibly be caused by the high overall survival in that data subset. This emphasises that more centres, with more diverse data, will be needed to develop definitive models.

For optimisation and individualisation of anal cancer radiotherapy, models for locoregional tumour control and late toxicity are needed. For this, more complex radiotherapy data, such as dose volume histogram metrics for both tumour and normal tissue and detailed toxicity and recurrence data are required. Studies also suggest a role for imaging biomarkers for outcome prediction [33–35]. We plan to extend our distributed learning analysis to include both, in a larger network of centres.

We note further that distributed learning per se is not unique and is not perfect when the number of patients per centre is low. We used containerised applications, which provide an isolated execution space to the software and are easily shareable. Containerisation technologies also make it difficult for external parties to tamper with the software. This makes the model algorithms re-usable and agnostic to the specifics of each node installation. We have shown that this implementation works with a diverse collection of hardware and operating systems.

In conclusion, we have demonstrated the utility of privacy preserving distributed learning for analysing multi-national cohorts of patients with rare cancers. We aim to expand the network with more institutions, and also the complexity of our outcome prediction models.

4.6 References

[1] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 2016;121:459–67. https://doi.org/10.1016/j.radonc.2016.10.002.

[2] Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: A web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015:ocv083. https://doi.org/10.1093/jamia/ocv083.

[3] Dwork C. Differential Privacy: A Survey of Results. In: Agrawal M, Du D, Duan Z, Li A, editors. Theory Appl. Models Comput., vol. 4978, Berlin, Heidelberg: Springer Berlin Heidelberg; 2008, p. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1.

[4] Martin F, Sieswerda M, Soest JV, Moncada-Torres A, IntGRen, Codacy Badger. IKNL/vantage6: 1.0.0a1. Zenodo; 2020. https://doi.org/10.5281/ZENODO.3686944.

[5] Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Gelijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. AMIA Annu Symp Proc 2020.

[6] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[7] Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiother Oncol 2020;144:189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[8] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9. https://doi.org/10.1016/j.radonc.2014.10.001.

[9] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today 2018. https://gco.iarc.fr/today (accessed September 14, 2020).

[10] Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. Eur J Surg Oncol 2014;40:1165–76. https://doi.org/10.1016/j.ejso.2014.07.030.

[11] Ajani JA. Fluorouracil, Mitomycin, and Radiotherapy vs Fluorouracil, Cisplatin, and Radiotherapy for Carcinoma of the Anal Canal: A Randomized Controlled Trial. JAMA 2008;299:1914. https://doi.org/10.1001/jama.299.16.1914.

[12] James RD, Glynne-Jones R, Meadows HM, Cunningham D, Myint AS, Saunders MP, et al. Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II): a randomised, phase 3, open-label, 2×2 factorial trial. Lancet Oncol 2013;14:516–24. https://doi.org/10.1016/S1470-2045(13)70086-X.

[13] Peiffert D, Tournier-Rangeard L, Gérard J-P, Lemanski C, François E, Giovannini M, et al. Induction Chemotherapy and Dose Intensification of the Radiation Boost in Locally Advanced Anal Canal Carcinoma: Final Analysis of the Randomized UNICANCER ACCORD 03 Trial. J Clin Oncol 2012;30:1941–8. https://doi.org/10.1200/JCO.2011.35.4837.

[14] Johnsson A, Leon O, Gunnlaugsson A, Nilsson P, Höglund P. Determinants for local tumour control probability after radiotherapy of anal cancer. Radiother Oncol 2018;128:380–6. https://doi.org/10.1016/j.radonc.2018.06.007.

[15] Muirhead R, Partridge M, Hawkins MA. A tumor control probability model for anal squamous cell carcinoma. Radiother Oncol 2015;116:192–6. https://doi.org/10.1016/j.radonc.2015.07.014.

[16] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282 (accessed December 21, 2020).

[17] Gilbert A, Drinkwater K, McParland L, Adams R, Glynne-Jones R, Harrison M, et al. UK national cohort of anal cancer treated with intensity-modulated radiotherapy: One-year oncological and patient-reported outcomes. Eur J Cancer 2020;128:7–16. https://doi.org/10.1016/j.ejca.2019.12.022.

[18] El Naqa I, editor. A guide to outcome modeling in radiotherapy and oncology: listening to the data. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2018.

[19] Fish R, Sanders C, Adams R, Brewer J, Brookes ST, DeNardo J, et al. A core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC): a patient and health-care professional consensus. Lancet Gastroenterol Hepatol 2018;3:865–73. https://doi.org/10.1016/S2468-1253(18)30264-4.

[20] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[21] Amin MB, American Joint Committee on Cancer, American Cancer Society, editors. AJCC cancer staging manual. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR-Technical editor ; Laura R. Meyer, CAPM-Managing editor. Chicago IL: American Joint Committee on Cancer, Springer; 2017.

[22] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: 2014.

[23] Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343–6. https://doi.org/10.1016/0197-2456(96)00075-X.

[24] Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011. https://doi.org/10.1002/sim.4154.

[25] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis

Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015;162:W1. https://doi.org/10.7326/M14-0698.

[26] Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982;69:239–41. https://doi.org/10.1093/biomet/69.1.239.

[27] Guren MG, Aagnes B, Nygård M, Dahl O, Møller B. Rising Incidence and Improved Survival of Anal Squamous Cell Carcinoma in Norway, 1987-2016. Clin Colorectal Cancer 2019;18:e96–103. https://doi.org/10.1016/j.clcc.2018.10.001.

[28] Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspect Clin Res 2011;2:145. https://doi.org/10.4103/2229-3485.86872.

[29] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[30] de Meric de Bellefon M, Lemanski C, Castan F, Samalin E, Mazard T, Lenglet A, et al. Long-term follow-up experience in anal canal cancer treated with Intensity-Modulated Radiation Therapy: Clinical outcomes, patterns of relapse and predictors of failure. Radiother Oncol 2020;144:141–7. https://doi.org/10.1016/j.radonc.2019.11.016.

[31] Gunderson LL, Moughan J, Ajani JA, Pedersen JE, Winter KA, Benson AB 3rd, et al. Anal carcinoma: impact of TN category of disease on survival, disease relapse, and colostomy failure in US Gastrointestinal Intergroup RTOG 98-11 phase 3 trial. Int J Radiat Oncol Biol Phys 2013;87:638–45. https://doi.org/10.1016/j.ijrobp.2013.07.035.

[32] Leon O, Guren M, Hagberg O, Glimelius B, Dahl O, Havsteen H, et al. Anal carcinoma - Survival and recurrence in a large cohort of patients treated according to Nordic guidelines. Radiother Oncol J Eur Soc Ther Radiol Oncol 2014;113:352–8. https://doi.org/10.1016/j.radonc.2014.10.002.

[33] Brown PJ, Zhong J, Frood R, Currie S, Gilbert A, Appelt AL, et al. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging 2019;46:2790–9. https://doi.org/10.1007/s00259-019-04495-1.

[34] Jones M, Hruby G, Coolens C, Driscoll B, Stanwell P, Kumar M, et al. A prospective, multi-centre trial of multi-parametric MRI as a biomarker in anal carcinoma. Radiother Oncol 2020;144:7–12. https://doi.org/10.1016/j.radonc.2019.10.001.

[35] Rusten E, Rekstad BL, Undseth C, Klotz D, Hernes E, Guren MG, et al. Anal cancer chemoradiotherapy outcome prediction using 18F-fluorodeoxyglucose positron emission tomography and clinicopathological factors. Br J Radiol 2019;92:20181006. https://doi.org/10.1259/bjr.20181006.

4.7 Supplementary material

***4.7.1 Supplementary material A**. When including T3N0 tumours in the high risk group,* 2 patients changed category in the Leeds cohort, 3 in the MAASTRO cohort and 12 in the Oslo cohort. Table 4-A1 presents the results of the global model with this alternative grouping for disease stage. The validation c-indices derived from the global model are 0.74 for Leeds, 0.73 for MAASTRO, and 0.71 for Oslo. The results from the leave-one-centre-out validation runs using the alternative grouping for disease stage are shown in Table 4-A2.

Table 4-A1. Results of the global distributed multivariate Cox regression analysis across all three centres, carried out with the alternative grouping for disease stage, for which T3N0 tumours were assigned to the "high risk" group. Age, primary tumour GTV and primary tumour dose were treated as continuous variables. The HRs represent a change of 10 years in age; 10cm^3 in primary tumour GTV; and 5 Gy in primary tumour dose (EQD2). GTV: Gross tumour volume. EQD2: Equivalent dose in 2 Gy fractions (α/β=10Gy). CI: Confidence interval

	Hazard ratio (95% CI)
High risk disease (compared to low risk disease)	1.46 (0.62-3.45)
Male sex (compared to female sex)	2.40 (1.26-4.59)
Age at the start of RT	1.30 (0.95-1.78)
Primary tumour GTV	1.06 (1.02-1.09)
Primary tumour dose (EQD2)	0.87 (0.54-1.41)

Table 4-A2. Results from the three leave-one-centre-out validation runs, carried out with the alternative grouping for disease stage, for which T3N0 tumours were assigned to the "high risk" group. Each column represents one run, consisting of model training on two nodes and validation on the third, independent node. Factor effects are presented in terms of hazard ratios with 95% confidence intervals; HR (95% CI). The HRs represent a change of 10 years in age; 10cm^3 in primary tumour GTV; and 5 Gy in primary tumour dose (EQD2). The resulting c-index from each validation run is also reported. GTV: Gross tumour volume. EQD2: Equivalent dose in 2 Gy fractions (α/β=10Gy).

Training nodes	MAASTRO Oslo	Leeds Oslo	Leeds MAASTRO
Validation node	Leeds	MAASTRO	Oslo
High risk disease (compared to low risk disease)	1.58 (0.53-4.70)	1.89 (0.62-5.70)	1.85 (0.71-4.86)
Male sex (compared to female sex)	3.02 (1.32-6.87)	2.70 (1.19-6.16)	2.12 (0.92-4.90)
Age at the start of RT	1.06 (0.71-1.58)	1.46 (0.98-2.17)	1.48 (1.05-2.10)
Primary tumour GTV	1.05 (1.00-1.09)	1.07 (1.02-1.12)	1.07 (1.03-1.11)
Primary tumour dose (EQD2)	0.98 (0.47-2.06)	0.45 (0.22-0.94)	0.97 (0.59-1.59)
Validation c-index	0.69	0.73	0.67

4.7.2 Supplementary material B. *Table 4-B1. Schoenfeld test results for all variables included in our model, categorised by centre.*

Variable	Schoenfeld test p-value		
	Leeds	MAASTRO	Oslo
Disease stage	0.74	0.62	0.22
Sex	0.10	0.38	0.14
Age at the start of RT	0.11	0.98	0.99
Primary tumour GTV	0.22	0.20	0.38
Primary tumour dose (EQD2)	0.76	0.45	0.77

Chapter 5 - Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study

5.1 Abstract

5.1.1 Background

Anal cancer is a rare cancer with rising incidence. Despite the relatively good outcomes conferred by state-of-the-art chemoradiotherapy, further improving disease control and reducing toxicity has proven challenging. Developing and validating prognostic models using routinely collected data may provide new insights for treatment development and selection. However, due to the rarity of the cancer, it can be difficult to obtain sufficient data, especially from single centres, to develop and validate robust models. Moreover, multi-centre model development is hampered by ethical barriers and data protection regulations that often limit accessibility to patient data. Distributed (or federated) learning allows models to be developed using data from multiple centres without any individual-level patient data leaving the originating centre, therefore preserving patient data privacy. This work builds on the proof-of-concept three-centre atomCAT1 study and describes the protocol for the multi-centre atomCAT2 study, which aims to develop and validate robust prognostic models for three clinically important outcomes in anal cancer following chemoradiotherapy.

5.1.2 Methods

This is a retrospective multi-centre cohort study, investigating overall survival, locoregional control and freedom from distant metastasis after primary chemoradiotherapy for anal squamous cell carcinoma. Patient data will be extracted and organised at each participating radiotherapy centre (n=18). Candidate prognostic factors have been identified through literature review and expert opinion. Summary statistics will be calculated and exchanged between centres prior to modelling. The primary analysis will involve developing and validating Cox Proportional Hazards models across centres for each outcome through distributed learning. Outcomes at

specific timepoints of interest and factor effect estimates will be reported, allowing for outcome prediction for future patients.

5.1.3 Discussion

The atomCAT2 study will analyse one of the largest available cross-institutional cohorts of patients with anal cancer treated with chemoradiotherapy. The analysis aims to provide information on current international clinical practice outcomes and may aid the personalisation and design of future anal cancer clinical trials through contributing to a better understanding of patient risk stratification.

5.2 Background

Anal cancer is a rare disease encompassing only approximately 0.3% of all cancer cases across the world [1, 2], but with a gradually increasing incidence [3]. A combination of radiotherapy and chemotherapy has been established as the standard treatment for localised disease for the last three decades [4–6]. This treatment confers relatively good outcomes, with 75% overall survival rates reported at 5 years [7–10]. Despite this, it has proven challenging to determine the optimal therapeutic radiotherapy dose and to further improve disease control [11–13].

A study by Shakir et al. [14], which analysed data from 385 patients with anal cancer treated in five UK centres with conformal radiotherapy techniques, reported that the site of primary disease was the most common site of relapse (83.4% of cases). In addition, the majority of patients experienced locoregional failure prior to getting metastatic disease. This emphasises the need to establish an effective treatment for locoregional control with an optimal radiotherapy dose. Even though ongoing prospective clinical trials [13] are focusing on this issue, clinical data acquired through standard practice can also be analysed for the development and validation of prognostic models, to further inform clinical practice [15, 16].

Prognostic and predictive models have been proposed in cancer research for more than 20 years [17] and have a wide range of potential applications, including prediction of cancer susceptibility [18, 19], recurrence risk [20, 21] and survival [22–24]. In particular, prognostic models can be used as decision support tools in the clinic, assisting clinicians in making informed decisions about patient management following a diagnosis [25].

However, developing robust prognostic models for anal cancer is particularly challenging. Due to the rarity of the cancer, it can be difficult to obtain sufficient data for robust model training and validation. In addition, ethical barriers and data protection regulations often limit the ability to share data between centres and thus render multi-centre model development unfeasible [26]. A novel data analysis methodology called distributed learning (DL) [27] has paved the way towards model development between institutions and across international borders.

Distributed learning, also sometimes referred to as federated learning, is a privacy-preserving approach that facilitates the development of robust statistical models using data distributed over multiple sites [28]. The main premise of this approach is that no individual-level patient data leaves the originating centre; only non-identifiable aggregated statistics (model coefficients and fit errors) are exchanged between institutions and a central server. Consequently, adopting this methodology minimises privacy issues related to patient data sharing since it does not breach data privacy barriers. DL algorithms operate in an iterative process where the local dataset in each centre is used to calculate local model coefficients and fit errors. These are sent to the central server, where a single globally-convergent model is determined by minimizing the total error [29]. This methodology is applicable for the development of models with a relatively small number of patients [27], but has also been proven to be upscalable to more than 20,000 patients [30].

A DL approach may be ideally suited for prognostic modelling in rare cancers such as anal cancer. It could facilitate acquisition of sufficient patient data from multiple international centres with the aim of developing robust generalisable models, while working around many of the barriers associated with physical data sharing. The feasibility of this approach in anal cancer has previously been demonstrated in the atomCAT1 proof-of-concept study [31]. Using data from three international radiotherapy centres, a Cox proportional hazards model for overall survival was trained and validated in a distributed fashion. The study analysed one of the largest available cohorts of patients with anal cancer treated with conformal radiotherapy and carried out robust multi-centre validation of outcome predictors. However, the analysis was limited to a single outcome only (overall survival), whereas other clinically important outcomes such as locoregional control and freedom from distant metastasis were not taken into consideration.

The atomCAT2 (Anal cancer Treatment Outcome Modelling with Computer Aided Theragnostics) study aims to develop prediction models for anal cancer outcomes after chemoradiotherapy through distributed learning. To achieve this, a consortium of 18 international cancer treatment centres based in the UK, Europe, Australia and Canada has been formed. A cohort of more than 1,000 patients will be analysed to develop and validate models for overall survival, locoregional control and distant metastasis, as well as to identify key prognostic factors and their effect size. This will provide unique insights and may aid the personalisation of treatment according to each patient’s unique characteristics.

5.3 Methods

5.3.1 Study design and patient population

This is a retrospective multi-centre cohort study, investigating outcomes after primary (chemo)radiotherapy for anal squamous cell carcinoma (ASCC). The inclusion and exclusion criteria are summarised in Table 5-1. Patient data will be extracted and organised within the informatics infrastructure at 18 participating radiotherapy centres, where subjects have consented to treatment with chemoradiotherapy. Routine and standard of care data will be used, and no prospective data collection will be explicitly carried out for the purpose of this study. Using a pragmatic approach, centres will be encouraged to include data for all patients treated in their centre fulfilling the inclusion criteria (Table 5-1). However, pre-existing patient cohorts, representing a subset of available patient cases, will be accepted. Future expansion to more participating centres internationally is planned.

Patients have been treated according to each participating centre’s protocol, which may include radiotherapy only or varying chemoradiotherapy regimens. Centres will be asked to briefly outline their main treatment and follow-up protocols as part of study participation.

Table 5-1. Participant inclusion and exclusion criteria. 3D-CRT: Three-dimensional conformal radiation therapy. IMRT: Intensity-modulated radiation therapy. VMAT: Volumetric modulated arc therapy.

Inclusion criteria
• Radical intent external beam radiotherapy treatment for primary anal squamous cell carcinoma, with or without concomitant chemotherapy
• Radiotherapy delivered using modern radiotherapy techniques (3D-CRT, IMRT or VMAT)
Exclusion criteria
• Palliative treatment
• Prior pelvic radiotherapy
• Brachytherapy (either primary or as boost treatment)

5.3.2 Outcome definitions

Three outcomes will be explored: overall survival, locoregional control and freedom from distant metastasis. These were identified as key outcome research measures in anal cancer by the CORMAC initiative [32].

5.3.2.1 Overall survival

Overall survival will be calculated in days from the first fraction of radiotherapy to either event or censoring, whichever happens first. An event is defined as death from any cause at any point during follow-up. Patients will be censored at the last clinical follow-up date if alive.

5.3.2.2 Locoregional control

Time to locoregional control will be calculated in days from the first fraction of radiotherapy to either event or censoring, whichever happens first. An event is defined as any of the following as a first event: (1) Abdominoperineal resection to control locoregional disease at any point during follow-up. This will always take precedence in terms of date for locoregional recurrence. (2) Locoregional disease progression, during treatment or in follow-up (irrespective of whether complete or partial response have been initially achieved), not managed by surgery. This will preferably be confirmed with biopsy, in which case the date of biopsy will count, but will alternatively be based on imaging and clinical examination only (date of imaging will be used). (3) Lack of complete response (non-clearance of disease) at 26 weeks (6 months) from first fraction of radiotherapy, as defined by clinical examination, imaging and/or biopsy [33]. In case of uncertainty or where limited information is available, the date where treatment failure or locoregional recurrence is first noted in the patient records will be used.

Patients will be censored at death, at last clinical follow-up, if undergoing abdominoperineal resection for non-disease related reasons (e.g. due to treatment complications), or in case of distant metastases.

The site of failure (primary tumour versus pelvic / initially involved nodes) will be noted to allow for separate analysis of local and locoregional failure. Failures in pelvic lymph nodes (inguinal, perirectal, internal iliac or external iliac nodes) or in lymph nodes which were part of the original treatment volumes (which may be the case e.g. for common iliac or para-aortic lymph nodes) will be defined as locoregional failures.

5.3.2.3 Freedom from distant metastasis

Freedom from distant metastasis will be calculated in days from the start of radiotherapy to either event or censoring, whichever happens first. An event is defined as distant disease recurrence (previously untreated lymph node metastasis outside the pelvis, or other metastatic sites such as lung, liver, bone) as a first event. This may be confirmed with biopsy, in which case the date of biopsy will count as the date of recurrence, or alternatively based on imaging (date of imaging will be used). In case of any uncertainty or where limited information is available, the date where distant progression is first noted in the patient records will be used. Site(s) of failure will be noted. Patients will be censored at local recurrence, at death, or at last clinical follow-up.

5.3.3 Identification of relevant prognostic factors

Already-established prognostic factors for the outcomes in question have been identified through a systematic review of the literature [34]. Studies published after 2000 which reported on disease-related outcomes and examined prognostic factors in multivariable analysis for overall survival, locoregional control, and freedom from distant metastasis were reviewed. In these studies, at least 70% of patients were treated with conformal radiotherapy techniques (3D-defined targets on computed tomography (CT), beams conformed to targets e.g. using multi-leaf collimators (MLCs), 3D dose calculation and optimisation of dose distributions). This approach identified the initial list of relevant data to be collected; this was subsequently reviewed by three senior clinical oncologists with expertise in anal cancer treatment, and additional factors were added.

5.3.4 Data collection and completeness

Relevant patient data will be identified and extracted from existing research and clinical databases. Data extraction from databases will be carried out in an automated fashion where possible, with additional manual review if needed. Each participating centre will be responsible for ensuring good data quality by spot checking all extracted data to identify any outliers and to make sure the coding system used is correct, according to the data dictionary provided. Data items are classified as either “essential” or “optional”. For “essential” data items, centres will aim for at most 10% missing data for any given data item across their study cohort. If more than 10% of data is missing for an individual data item, imputation techniques will be implemented according to the framework set out below (see “Missing Data” section). For “optional” data items, missing data will be accepted. Each centre will contribute data from a minimum of 40 patients to ensure a representative sample and achieve a reasonable balance of patient heterogeneity, as well as limit reporting of subgroups with one or only a few patients. See Supplementary material A (Section 5.6.1) for full definition and coding of data items.

5.3.5 Missing data

For outcome data, Complete Case analysis will be used for each of the three outcomes. That is, if data is missing for a specific outcome for a patient, that patient will not contribute to the corresponding analysis. For potential prognostic factors, a mixed approach will be used: If more than 90% of patients per centre have complete data for all factors for a given analysis, then Complete Case analysis will be used as the primary analysis for that centre. If not, missing value imputation [35] will be used according to the framework set out below before any models are fitted, and Complete Case analysis will be performed as a robustness check. Missing data imputation has only been sparsely explored in the context of distributed learning and there is only limited precedence to guide best practise [36, 37]. Initially, we will implement the missing data imputation framework described below, but this may be adapted based on our ability to implement more robust techniques in a distributed setting.

Where data for potential prognostic factors is missing for a small number of patients in individual centres (>10% but ≤50%), local data imputation techniques will be employed. Missing data will be imputed using only the local dataset, and prior to any distributed model optimisation. The k-Nearest Neighbour (KNN) algorithm [38, 39] will be used to carry out the imputation [40]. Using this algorithm, each missing value will be replaced by a value that is as close as possible to the true value, obtained from related cases in the whole dataset. This technique aims to preserve the original structure of the dataset and avoids distorting the distribution of the imputed data item. To implement KNN, an appropriate value of k will be first determined through exploration of the data in each centre, using the square root of the sample size as a starting point [38, 41]. All available essential data items, as well as outcome data [42], will be included in the imputation model.

Where data for potential prognostic factors is systematically missing in specific centres (>50% data missing for any specific item), the general assumption will be that imputation based on the local centre data will be unreliable. In this case, consortium-wide regression will be implemented to impute the missing data items.

As an initial plan, a regression model will first be trained using data from all centres apart from the centre where the data item is missing. This model will then be run in the centre where the data is missing to impute the missing values. If two or more centres are missing the same data item, this approach will not be technically feasible due to limitations of the DL infrastructure that will be used. In this case, for continuous data items, the mean from each centre (apart from the centres with missing data) will be used to calculate the global “median of means” value for that data item. This value will be assigned to all patients in the centres where the data item is missing. For categorical data items, the frequency of each category across the global cohort for the data item that is missing will be calculated (excluding centres with the missing data item). Categories will then be assigned to each patient at random in centres where the data item is missing, ensuring the local frequency distribution is the same as the global frequency distribution.

In parallel to the main consortium analysis, an independent exploratory study will be carried out to evaluate the feasibility of various imputation techniques in the context of distributed learning. Once feasible and robust techniques have been identified, we will look at implementing these to improve on the missing data imputation framework described above.

5.3.6 Sample size

The atomCAT1 proof-of-concept study [31] demonstrated the ability to combine data from three centres for 281 patients. A Cox regression model for overall survival was fitted to the global dataset, taking five baseline factors into account. Its performance was evaluated using Harrell's concordance index (c-index). The internal-external validation approach returned a c-index of 0.70, which is considered as "good" model performance.

The "pmsampsize" [43] package in R was used to estimate the minimum sample size required for the atomCAT2 models. The parameters required to carry out this calculation include: R-squared (calculated from the c-index), number of candidate predictor parameters, shrinkage (level of reduction of the estimated predictor effect estimates to address overfitting), overall event rate in the population, mean follow-up time anticipated for individuals in the model development dataset and timepoint of interest for prediction.

Table 5-2 illustrates how many patients will be needed to fit a Cox proportional hazards model for overall survival in atomCAT2, aiming for similar performance to the model developed in the atomCAT1 study, with varying number of parameters. These estimates assume an event rate of 16% at 36 months, with a median follow-up of 46 months, a c-index of 0.70 (corresponding to R2CSapp of 0.0676) and a shrinkage value of 0.90. Sample size estimates in this setting are very robust to variations in event rate, and are thus also valid for models for locoregional control and freedom from distant metastases (with an event rate of 25% and 15% at 5 years, respectively [7]).

The number of prognostic factors which will be included in the final models will be based on the total number of patients available across the consortium. The analysis plan will be finalised when the total number of available patients are confirmed. Currently, we aim to include data from at least 1,000 patients, which would allow for eight parameters to be estimated per model. The number of prognostic factors included in the model could be the same or different to the number of parameters depending on the number of categories for the categorical factors and the parameterisation of the continuous factors.

Table 5-2. Estimated minimum sample size for a range of parameters, for the overall survival model (also valid for the locoregional control and freedom from distant metastasis models).

Parameters included in the model	Minimum sample size
5	641
6	769
7	897
8	1025
9	1153
10	1283
11	1409

5.3.7 Statistical analysis

5.3.7.1 Descriptive data analysis

Summary statistics will be shared with the central study team in order to explore cohort differences prior to modelling. Categorical variables will be summarised as proportions to the total number of patients per centre, expressed as percentages. Summary statistics will be calculated for numerical variables (mean, standard deviation, range, variance).

Summary statistics for the global cohort will be reported. Categorical variables will be summarised as proportions to the total number of patients in the global cohort, expressed as percentages. For numerical variables, random effect meta-analysis will be used (using the “meta” package in R), with inverse variance weighting for pooling, reporting the overall mean and 95% confidence intervals (calculated from overall standard deviations). The range will be reported as the lowest and highest values across the global cohort, calculated from the range from each centre.

Estimated 3-year survival / freedom from recurrence rates will be calculated by each centre individually, using the Kaplan-Meier estimator (using the “survival” package in R). The median potential follow-up time will be calculated based on the inverse Kaplan-Meier estimator for each outcome for each centre separately.

5.3.7.2 Model specification

The prognostic factors that will be included in each of the three primary models are specified in Table 5-3. The factors are listed in the order that they will be prioritised for analysis based directly on the findings from the systematic review and expert opinion. The factors are ordered according to the total number of times found to be prognostic in multivariable analysis. Additional factors were added by senior clinical oncologists. Factors found to be prognostic in univariable analysis but not in multivariable analysis may be included in the secondary models. The number of factors included in each model will depend on the final sample size for each outcome, as detailed above. For each factor, the primary parameterisation used (e.g. categorisation for categorical variables) is listed, with alternatives to be explored in secondary analyses. The parameterisation of all variables for the primary and secondary models was determined after a detailed discussion with clinical oncologists and represents the relationship they expect to see from clinical experience, as well as the expected data distribution. See Supplementary material B (Section 5.6.2) for secondary model specification.

Table 5-3. Specification of the primary models for overall survival, locoregional control and freedom from distant metastasis. N stage: nodal stage. T stage: tumour stage. GTV: Gross tumour volume. EQD2: Equivalent dose in 2Gy fractions (α/β=10Gy). SCC: Squamous cell carcinoma. 3D-CRT: Three-dimensional conformal radiation therapy. IMRT: Intensity-modulated radiation therapy. VMAT: Volumetric modulated arc therapy.

	Prognostic factors to be included in the primary models		
	Overall survival model	**Locoregional control model**	**Freedom from distant metastasis model**
1	N stage: N0 vs N+	Sex: Female vs Male	N stage: N0 vs N+
2	T stage: T1-2 vs T3-4	N stage: N0 vs N+	T stage: T1-2 vs T3-4
3	Sex: Female vs Male	T stage: T1-2 vs T3-4	Sex: Female vs Male
4	Age: Modelled as a continuous, linear factor	Age: Modelled as a continuous, linear factor	Age: Modelled as a continuous, linear factor
5	Primary tumour GTV (cm^3): Modelled as a continuous, log-transformed factor	Primary tumour GTV (cm^3): Modelled as a continuous, log-transformed factor	Primary tumour GTV (cm^3): Modelled as a continuous, log-transformed factor

6	Primary tumour dose (EQD2): Modelled as a continuous, linear factor	Primary tumour dose (EQD2): Modelled as a continuous, linear factor	Primary tumour dose (EQD2): Modelled as a continuous, linear factor
7	Histology: SCC vs Basaloid SCC	Histology: SCC vs Basaloid SCC	Histology: SCC vs Basaloid SCC
8	Chemotherapy regimen: [No chemotherapy] vs [Mitomycin C-based regimen] vs [Cisplatin-based regimen]	Chemotherapy regimen: [No chemotherapy] vs [Mitomycin C-based regimen] vs [Cisplatin-based regimen];	Chemotherapy regimen: [No chemotherapy] vs [Mitomycin C-based regimen] vs [Cisplatin-based regimen];
9	RT technique: [3D-CRT] vs [IMRT] vs [VMAT]	RT technique: 3D-CRT vs IMRT vs VMAT	RT technique: 3D-CRT vs IMRT vs VMAT

5.3.7.3 Cox model development and reporting

The primary analysis will involve the development and internal validation (Type 2b validation according to TRIPOD [44]) of Cox Proportional Hazards models using distributed learning [45] across all participating centres, separately for each outcome (overall survival, locoregional control, and freedom from distant metastasis). The primary model to be developed for each outcome is detailed above. Secondary analyses (Supplementary material B, Section 5.6.2) will be used to explore the robustness of the results to the choices made for the primary model. As an additional assessment of model robustness, another secondary analysis will be conducted. In this analysis, the specified models (Table 5-3) will be trained only on datasets comprising of more than 20 events, as a way of testing whether the number of events per centre affects the behaviour of the models.

The factor effects from each model will be reported in the form of Hazard Ratios (HRs) along with 95% confidence intervals (CIs). The 'baseline' outcome rate at specific timepoints of interest (e.g. 2 years, 3 years and 5 years) will be calculated. Combining the baseline outcome rate with the factor effect estimates (HRs) will allow for outcome prediction for a future patient, rendering the model useable for future prediction.

5.3.8 Evaluation and visualisation of model performance

Model performance will be initially assessed using Harrell's concordance index (c-index) [46] on a per-centre basis, with a weighted average c-index (and standard deviation) also reported. A more robust estimate for out-of-sample performance will be obtained using a closed-loop internal-external "leave-one-centre-out" cross-validation method [47], where the model will be optimised using data from all but one sites and then validated on the last site. This will be repeated to cover the possible combinations, resulting in different c-indices which provide an estimate of the over-optimism of the global model. The weighted average and interquartile range (IQR) of these c-index values will be reported. The factor effects from each of these validation models will be aggregated and the summary effects will be reported in the form of HR range for each factor across all models.

The calibration of the global model (performance check for the prediction aspect of the model) will be assessed by constructing calibration curves and quantifying the calibration slope [48, 49], on a local (per-centre) level. Calibration curves will use three groups per centre (low/medium/high risk, based on their predicted outcomes), and will compare average predicted outcome within each group with the observed outcome at 3 years, using the Kaplan-Meier estimator. This is the initial plan for evaluation of the model calibration, and the final plan may be altered depending on the size of each centre's cohort, as well as the number of events per centre.

The model development and validation procedure and results will be reported in accordance with the TRIPOD statement and checklist [50]. This protocol has also been checked against the relevant parts of the TRIPOD checklist for prediction model development and validation.

5.3.9 Distributed learning infrastructure

The infrastructure that will be used for this study is very similar to the infrastructure implemented in atomCAT1 [31]. The Distributed Cox algorithm developed by Lu et al. [45] was adapted to the Vantage6 v2.0 infrastructure as R scripts (v.3.6.2). The source code will be made openly accessible on GitHub. Scripts for computing model coefficients and leave-one-centre-out model validation will be packaged as application "containers" (via Docker) and will be locally executed in each centre. All other scripts that will be used for the data analysis will be uploaded in a GitLab repository, which will be made public at the end of the project.

5.3.10 Organisation and policies

The atomCAT2 study will be conducted as part of a wider atomCAT consortium. Details of consortium engagement and project management will be described in detail in a collaborative research agreement, which will be signed by all participating centres.

Medical Data Works BV (MDW, https://medicaldataworks.nl/) implements a privacy preserving distributed infrastructure that investigators in atomCAT2 will use. Therefore, an Infrastructure User Agreement will be signed as a contractual agreement between each centre and MDW. MDW will not be considered as a “processor” of clinical data according to the definition in the EU General Data Protection Regulation but is solely the provider of the information technology infrastructure and the central server. As the infrastructure provider, MDW will enforce the legal use of algorithms and data stations, and this agreement shall define the terms and conditions for the use of the infrastructure.

5.4 Discussion

This paper describes the protocol and statistical analysis plan for the international multi-centre atomCAT2 study. The study will aim to develop and validate robust prognostic models for three clinically important outcomes in anal cancer after treatment with conformal radiotherapy. Key prognostic factors for each outcome will also be identified and validated.

Only patients treated with conformal radiotherapy techniques (e.g. 3D-conformal, intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT)), will be included in the cohort for analysis as these techniques have been proven to reduce the dose delivered to normal tissues, minimising toxicity and reducing overall treatment duration and the need for treatment breaks [51–54]. Therefore, by limiting our cohort to patients treated with conformal radiotherapy, we ensure that the prognostic models developed will be informative to current clinical practice. These models will include a range of established prognostic factors, identified through a comprehensive review of the literature and confirmed by three experts from three different centres. A range of additional less-established prognostic factors will also be tested in secondary models, to quantify their effect size and assess their eligibility as clinically relevant predictors of outcome.

Most of the literature which reports on outcomes and prognostic factors in anal cancer after conformal radiotherapy are retrospective studies which include small cohorts from a single centre. The results from the prognostic models developed in these studies may therefore suffer from small sample size bias and might not be generalisable [55] across centres and countries. To our knowledge, only three previous studies have analysed more than 200 patients with conformal radiotherapy [14, 31, 56], only one of which was multi-national and conducted multi-centre validation of outcome predictors. The cohorts that will be included in atomCAT2 will not only be significantly larger in size, but also more heterogeneous, since treatment dose and delivery schedules vary between radiotherapy centres, especially across different countries.

The analysis will be limited to retrospective data that is readily available in clinical and radiotherapy planning databases in a large number of centres. Therefore, some factors that could potentially be prognostic, such as HPV status and baseline performance status, may not be included in the primary models as they are not routinely collected in all centres. Since atomCAT2 is a non-prospective multi-centre analysis, it is expected that some data will vary between centres, including tumour staging, GTV definitions and outcome definitions. Steps have been taken to take the variation into consideration and minimise it as much as possible, including providing pre-specified definitions for all three outcomes and asking centres to indicate the staging version and GTV definition used. Despite this, some variation is unavoidable, which may affect the results. Additionally, it is expected that some essential data will be missing in a number of centres.

The methods for handling missing data have been specified in the protocol, however, these are substantially limited to what can currently be implemented in the DL setting without having to share individual-level patient data between centres. The field of missing data imputation in the context of DL is still in its infancy and does not currently have established standards. So far, only few studies have been conducted with the aim of developing or evaluating imputation techniques that can be implemented in a DL setting [36, 37]. Our initial imputation plan for data missing for a small number of patients in individual centres proposes the implementation of the KNN algorithm, which is a single imputation approach. In this case, one unique value will be imputed for each patient with missing data, resulting in a single complete dataset [57, 58]. This will likely produce relatively unbiased estimates, especially if only a small proportion of the data is missing [57, 59]. However, it is worth noting that these approaches fail to take into account the uncertainty of the missing values [60], which often results in underestimation of the variability and standard errors that are too small [61]. If data for a single data item is missing in the majority of patients in an individual centre, we also propose single imputation (using the data from the remaining centres), but assigning the same value to all missing data from the centre in question (also referred to as single value imputation). We recognise that this approach may introduce significant bias, leading to a change in the distribution shape and a significant decrease in standard deviation of the data item being imputed [62]. Using more advanced approaches to impute missing data, such as multiple imputation by chained equations (MICE) [63], would be ideal but cannot be applied through the DL infrastructure at this point. Further methodological research is needed to incorporate robust data imputation techniques to a privacy-preserving setting in order to tackle the problem of missing data, which is particularly common in medical datasets.

Future research beyond atomCAT2 will include incorporation of imaging and radiomics data to the models to increase their complexity and the potential insights gained. A number of studies have reported various imaging-related prognostic factors in anal cancer [64–66], which might prove to be clinically relevant. Moreover, strong efforts from the research community are being put into increasing the utility of DL in medical research by adapting different statistical methods and models to the existing infrastructure. In the future, it may be possible to develop competing risk models [67] in a distributed fashion, allowing multiple outcomes to be analysed in combination. Additionally, other algorithms such as random survival forests [68], may be implemented in DL to carry out the analysis instead of Cox regression. Random survival forests allow for a larger number of factors to be considered and factor selection is embedded within the methodology, which may in turn improve learning performance. This will be particularly useful in cases where many factors need to be considered. Alternative approaches to DL could also be considered for prognostic model development without having to share individual-level patient data between centres. For example, a multivariate meta-analysis approach [69–71] could be adopted, where summary statistics and regression coefficients from different prognostic models can be combined into a new prediction model. However, there are various issues with this approach, which may have a negative impact on the performance of the resulting prediction model, such as inconsistent covariate adjustment across models and high levels of model heterogeneity [72]. One significant advantage of the distributed learning approach over a meta-analysis approach is that a distributed Cox regression model generates the same model outputs as a centralised Cox regression model trained with the same data [45]. It has also been proven that distributed and centralised Cox regression models are equivalent from a mathematical perspective. This might not be true in all cases where meta-analysis approaches to prognostic model development are employed.

In conclusion, the atomCAT2 models will be developed using one of the largest cohorts of patients with anal cancer treated with conformal radiotherapy techniques ever analysed and will be validated across centres and countries. The models will allow for the prediction of outcomes in individual patients, which will inform current clinical practice and may subsequently aid with the personalisation of anal cancer treatment. The results of the atomCAT2 study may guide patient risk stratification, which may in turn facilitate the design of future prospective clinical trials in anal cancer.

5.5 References

[1] Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. Int J Epidemiol 2016:dyw276. https://doi.org/10.1093/ije/dyw276.

[2] Salati SA. Anal Cancer : A Review. Int J Health Sci 2012;6:206–30. https://doi.org/10.12816/0006000.

[3] van der Zee RP, Richel O, de Vries HJC, Prins JM. The increasing incidence of anal cancer: can it be explained by trends in risk groups? Neth J Med 2013;71:401–11.

[4] Nigro ND, Vaitkevicius VK, Considine B. Combined therapy for cancer of the anal canal: A preliminary report. Dis Colon Rectum 1974;17:354–6. https://doi.org/10.1007/BF02586980.

[5] Rao S, Guren MG, Khan K, Brown G, Renehan AG, Steigen SE, et al. Anal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up☆. Ann Oncol 2021;32:1087–100. https://doi.org/10.1016/j.annonc.2021.06.015.

[6] Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol 2014;40:1165–76. https://doi.org/10.1016/j.ejso.2014.07.030.

[7] Ajani JA. Fluorouracil, Mitomycin, and Radiotherapy vs Fluorouracil, Cisplatin, and Radiotherapy for Carcinoma of the Anal Canal: A Randomized Controlled Trial. JAMA 2008;299:1914. https://doi.org/10.1001/jama.299.16.1914.

[8] James RD, Glynne-Jones R, Meadows HM, Cunningham D, Myint AS, Saunders MP, et al. Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II): a randomised, phase 3, open-label, 2×2 factorial trial. Lancet Oncol 2013;14:516–24. https://doi.org/10.1016/S1470-2045(13)70086-X.

[9] Peiffert D, Tournier-Rangeard L, Gérard J-P, Lemanski C, François E, Giovannini M, et al. Induction Chemotherapy and Dose Intensification of the Radiation Boost in Locally Advanced Anal Canal Carcinoma: Final Analysis of the Randomized UNICANCER ACCORD 03 Trial. J Clin Oncol 2012;30:1941–8. https://doi.org/10.1200/JCO.2011.35.4837.

[10] Sekhar H, Malcomson L, Kochhar R, Sperrin M, Alam N, Chakrbarty B, et al. Temporal improvements in loco-regional failure and survival in patients with anal cancer treated with chemo-radiotherapy: treatment cohort study (1990–2014). Br J Cancer 2020;122:749–58. https://doi.org/10.1038/s41416-019-0689-x.

[11] Johnsson A, Leon O, Gunnlaugsson A, Nilsson P, Höglund P. Determinants for local tumour control probability after radiotherapy of anal cancer. Radiother Oncol 2018;128:380–6. https://doi.org/10.1016/j.radonc.2018.06.007.

[12] Muirhead R, Partridge M, Hawkins MA. A tumor control probability model for anal squamous cell carcinoma. Radiother Oncol 2015;116:192–6. https://doi.org/10.1016/j.radonc.2015.07.014.

[13] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[14] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[15] Sturdza A, Pötter R, Fokdal LU, Haie-Meder C, Tan LT, Mazeron R, et al. Image guided brachytherapy in locally advanced cervical cancer: Improved pelvic control and survival in RetroEMBRACE, a multicenter cohort study. Radiother Oncol 2016;120:428–33. https://doi.org/10.1016/j.radonc.2016.03.011.

[16] Tanderup K, Fokdal LU, Sturdza A, Haie-Meder C, Mazeron R, van Limbergen E, et al. Effect of tumor dose, volume and overall treatment time on local control after radiochemotherapy including MRI guided brachytherapy of locally advanced cervical cancer. Radiother Oncol 2016;120:441–6. https://doi.org/10.1016/j.radonc.2016.05.014.

[17] Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. J Med Syst 1991;15:11–9. https://doi.org/10.1007/BF00993877.

[18] Waddell M, Page D, Shaughnessy J. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. Proc. 5th Int. Workshop Bioinforma. - BIOKDD 05, Chicago, Illinois: ACM Press; 2005, p. 21. https://doi.org/10.1145/1134030.1134035.

[19] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration. Cancer 2010;116:3310–21. https://doi.org/10.1002/cncr.25081.

[20] Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. J Breast Cancer 2012;15:230. https://doi.org/10.4048/jbc.2012.15.2.230.

[21] Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl 2014;24:1311–6. https://doi.org/10.1007/s00521-013-1359-1.

[22] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113–27. https://doi.org/10.1016/j.artmed.2004.07.002.

[23] Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22:e184–90. https://doi.org/10.1093/bioinformatics/btl230.

[24] Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 2014;48:1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006.

[25] Abu-Hanna A, Lucas PJF. Prognostic Models in Medicine: AI and Statistical Approaches. Methods Inf Med 2001;40:1–5. https://doi.org/10.1055/s-0038-1634456.

[26] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9. https://doi.org/10.1016/j.radonc.2014.10.001.

[27] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 2016;121:459–67. https://doi.org/10.1016/j.radonc.2016.10.002.

[28] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated Learning: Strategies for Improving Communication Efficiency. ArXiv161005492 Cs 2017.

[29] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[30] Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiother Oncol 2020;144:189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[31] Theophanous S, Choudhury A, Lønne P-I, Samuel R, Guren MG, Berbee M, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study. Radiother Oncol 2021;159:183–9. https://doi.org/10.1016/j.radonc.2021.03.013.

[32] Fish R, Sanders C, Adams R, Brewer J, Brookes ST, DeNardo J, et al. A core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC): a patient and health-care professional consensus. Lancet Gastroenterol Hepatol 2018;3:865–73. https://doi.org/10.1016/S2468-1253(18)30264-4.

[33] Glynne-Jones R, Sebag-Montefiore D, Meadows HM, Cunningham D, Begum R, Adab F, et al. Best time to assess complete clinical response after chemoradiotherapy in squamous cell carcinoma of the anus (ACT II): a post-hoc analysis of randomised controlled phase 3 trial. Lancet Oncol 2017;18:347–56. https://doi.org/10.1016/S1470-2045(17)30071-2.

[34] Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, et al. Prognostic Factors for Patients with Anal Cancer Treated with Conformal Radiotherapy - A Systematic Review. In Review; 2022. https://doi.org/10.21203/rs.3.rs-1324086/v1.

[35] Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2020;53:1487–509. https://doi.org/10.1007/s10462-019-09709-4.

[36] Chang C, Deng Y, Jiang X, Long Q. Multiple imputation for analysis of incomplete data in distributed health data networks. Nat Commun 2020;11:5467. https://doi.org/10.1038/s41467-020-19270-2.

[37] Brink C, Hansen CR, Field M, Price G, Thwaites D, Sarup N, et al. Distributed learning optimisation of Cox models can leak patient data: Risks and solutions 2022.

[38] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med 2016;4:218–218. https://doi.org/10.21037/atm.2016.03.37.

[39] Cunningham P, Delany SJ. k-Nearest Neighbour Classifiers - A Tutorial. ACM Comput Surv 2022;54:1–25. https://doi.org/10.1145/3459665.

[40] Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 2016;16:74. https://doi.org/10.1186/s12911-016-0318-z.

[41] Lantz B. Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications. 1. publ. Birmingham: Packt Publ; 2013.

[42] Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092–101. https://doi.org/10.1016/j.jclinepi.2006.01.009.

[43] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96. https://doi.org/10.1002/sim.7992.

[44] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015;162:W1. https://doi.org/10.7326/M14-0698.

[45] Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: A web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015:ocv083. https://doi.org/10.1093/jamia/ocv083.

[46] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17. https://doi.org/10.1002/sim.4154.

[47] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[48] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010;21:128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[49] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.

[50] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55. https://doi.org/10.7326/M14-0697.

[51] Bazan JG, Hara W, Hsu A, Kunz PA, Ford J, Fisher GA, et al. Intensity-modulated radiation therapy versus conventional radiation therapy for squamous cell carcinoma of the anal canal. Cancer 2011;117:3342–51. https://doi.org/10.1002/cncr.25901.

[52] Kachnic LA, Winter K, Myerson RJ, Goodyear MD, Willins J, Esthappan J, et al. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. Int J Radiat Oncol Biol Phys 2013;86:27–33. https://doi.org/10.1016/j.ijrobp.2012.09.023.

[53] Chuong MD, Freilich JM, Hoffe SE, Fulp W, Weber JM, Almhanna K, et al. Intensity-Modulated Radiation Therapy vs. 3D Conformal Radiation Therapy for Squamous Cell Carcinoma of the Anal Canal. Gastrointest Cancer Res GCR 2013;6:39–45.

[54] Franco P, Arcadipane F, Ragona R, Mistrangelo M, Cassoni P, Munoz F, et al. Volumetric modulated arc therapy (VMAT) in the combined modality treatment of anal cancer patients. Br J Radiol 2016;89:20150832. https://doi.org/10.1259/bjr.20150832.

[55] Hackshaw A. Small studies: strengths and limitations. Eur Respir J 2008;32:1141–3. https://doi.org/10.1183/09031936.00136408.

[56] de Meric de Bellefon M, Lemanski C, Castan F, Samalin E, Mazard T, Lenglet A, et al. Long-term follow-up experience in anal canal cancer treated with Intensity-Modulated Radiation Therapy: Clinical outcomes, patterns of relapse and predictors of failure. Radiother Oncol J Eur Soc Ther Radiol Oncol 2020;144:141–7. https://doi.org/10.1016/j.radonc.2019.11.016.

[57] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91. https://doi.org/10.1016/j.jclinepi.2006.01.014.

[58] Bertsimas D, Pawlowski C, Zhuo YD. From Predictive Methods to Missing Data Imputation: An Optimization Approach. J Mach Learn Res 2017;18:196:1-196:39.

[59] Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010;10:112. https://doi.org/10.1186/1471-2288-10-112.

[60] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017;17:162. https://doi.org/10.1186/s12874-017-0442-1.

[61] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393–b2393. https://doi.org/10.1136/bmj.b2393.

[62] Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Appl Artif Intell 2019;33:913–33. https://doi.org/10.1080/08839514.2019.1637138.

[63] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40–9. https://doi.org/10.1002/mpr.329.

[64] Cardenas ML, Spencer CR, Markovina S, DeWees TA, Mazur TR, Weiner AA, et al. Quantitative FDG-PET/CT predicts local recurrence and survival for squamous cell

carcinoma of the anus. Adv Radiat Oncol 2017;2:281–7. https://doi.org/10.1016/j.adro.2017.04.007.

[65] Brown PJ, Zhong J, Frood R, Currie S, Gilbert A, Appelt AL, et al. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging 2019;46:2790–9. https://doi.org/10.1007/s00259-019-04495-1.

[66] Rusten E, Rekstad BL, Undseth C, Klotz D, Hernes E, Guren MG, et al. Anal cancer chemoradiotherapy outcome prediction using 18F-fluorodeoxyglucose positron emission tomography and clinicopathological factors. Br J Radiol 2019;92:20181006. https://doi.org/10.1259/bjr.20181006.

[67] Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation 2016;133:601–9. https://doi.org/10.1161/CIRCULATIONAHA.115.017719.

[68] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2. https://doi.org/10.1214/08-AOAS169.

[69] Becker BJ, Wu M-J. The Synthesis of Regression Slopes in Meta-Analysis. Stat Sci 2007;22. https://doi.org/10.1214/07-STS243.

[70] Yoneoka D, Henmi M. Synthesis of linear regression coefficients by recovering the within-study covariance matrix from summary statistics: Synthesis of Linear Regression. Res Synth Methods 2017;8:212–9. https://doi.org/10.1002/jrsm.1228.

[71] Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. BMJ 2017:j3932. https://doi.org/10.1136/bmj.j3932.

[72] Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. Diagn Progn Res 2019;3:13. https://doi.org/10.1186/s41512-019-0059-4.

[73] London: Royal College of Radiologists. The timely delivery of radical radiotherapy: Standards and guidelines for the management of unscheduled treatment interruptions. 2008. https://www.rcr.ac.uk/publication/timely-delivery-radical-radiotherapy-guidelines-management-unscheduled-treatment (accessed February 25, 2022).

5.6 Supplementary material

5.6.1 Supplementary material A. Data dictionary

Essential data items are denoted in **bold**. All other data items are optional.

All missing values will be coded as **NA**.

Baseline characteristics

- **Biological sex** [*sex*]: Binary variable
  - 0 – Male
  - 1 – Female
- **Age at the start of radiotherapy** (years) [*age*]: Continuous numerical variable
- **TNM staging**: Categorical variables
  - **T stage** [*t_stage*]
    - 1: T1
    - 2: T2
    - 3: T3
    - 4: T4
  - **N stage** [*n_stage*]
    - for TNM version 7: 0: N0; 1: N1; 2: N2; 3: N3
    - for TNM version 8: 0: N0; 1: N1a; 2: N1b; 3: N1c
  - **M stage** [*m_stage*]
    - 0: M0
    - 1: M1
- **TNM staging version** [*tnm_version*]: Discrete numerical variable
- **Primary tumour GTV** (cm^3) [*pr_tumour_gtv*]: Continuous numerical variable
- **Histology** [*histology*]: Binary variable
  - 0 – SCC
  - 1 – Basaloid SCC
- HPV status [*hpv_status*]: Binary variable
  - 0 – Negative
  - 1 – Positive
- Performance status [*perf_status*]: Categorical variable
  - 0 – Fully active, able to carry on all pre-disease performance without restriction.

    - 1 – Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light housework, office work.
    - 2 – Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours.
    - 3 – Capable of only limited self-care, confined to bed or chair more than 50% of waking hour.
    - 4 – Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair.
- Metastasis site at diagnosis [*met_site_diag*]: Categorical variable
    - 0 – No distant metastasis
    - 1 – Lymph nodes outside pelvis
    - 2 – Viscera or bones
    - 3 – Multiple sites
- GTV delineation definition [*gtv_definition*]: Categorical variable
    - This variable does not need to be assessed for each patient individually, only on a per-centre level.
    - 1 – Primary tumour only
    - 2 – Primary tumour and anal canal in areas of tumour involvement
    - 3 – Primary tumour and entire anal canal
- Differentiation grade [*diff_grade*]: Categorical variable
    - 0 – Well differentiated
    - 1 – Moderately differentiated
    - 2 – Poorly differentiated

Treatment-related factors

- **Radiotherapy technique** [*rt_technique*]**:** Categorical variable
    - 1 – 3D-CRT
    - 2 – IMRT
    - 3 – VMAT
- **Total prescribed dose** (in EQD2$_{\alpha/\beta=10Gy}$): Continuous numerical variable
    - **To primary tumour** [*prescr_dose_prtumour*]
    - **To involved lymph nodes** [*prescr_dose_invnodes1, prescr_dose_invnodes2*]
    - **To elective nodes** [*prescr_dose_elenodes1, prescr_dose_elenodes2*]

- **Concurrent chemotherapy?** [*conc_chemo*]: Binary variable
    - 0 – No
    - 1 – Yes
- **Concurrent chemotherapy – number of cycles** [*conc_chemo_cycles*]: Discrete numerical variable
- **Concurrent chemotherapy – drugs used** [*conc_chemo_drugs*]: Categorical variable
    - 0 – No chemotherapy
    - 1 – Mitomycin C and 5-Fluorouracil
    - 2 – Mitomycin C and Capecitabine
    - 3 – Cisplatin and 5-Fluorouracil
    - 4 – Cisplatin and Capecitabine
    - 5 – Other
- Total number of prescribed treatment fractions [*prescr_fractions*]: Discrete numerical variable
- Total number of delivered treatment fractions [*deliv_fractions*]: Discrete numerical variable
- Overall treatment time (days) [*overall_treatment_time*]: Discrete numerical variable
- Completed radiotherapy treatment? [*compl_treatement*]: Binary variable
    - 0 – No
    - 1 – Yes
- Treatment breaks? [*treatment_breaks*]: Binary variable
    - Defined as any extension to the treatment time of more than 2 days over the planned overall treatment time (as defined by RCR [73]) – extensions due to planned breaks such as holidays should not be included.
    - Estimated from the total number of delivered fractions - gives expected treatment time - compared to overall treatment time
    - We don’t need chart checks for breaks for all patients, but just an estimate of whether treatment time is extended compared to expected
    - 0 – No
    - 1 – Yes
- Simultaneous or sequential boost? [*boost*]: Categorical variable
    - 0 – No boost
    - 1 – Simultaneous boost
    - 2 – Sequential boost

- Total delivered dose (in EQD2$_{\alpha/\beta=10Gy}$): Continuous numerical variable
    - To primary tumour [*deliv_dose_prtumour*]
    - To involved lymph nodes [*deliv_dose_invnodes1, deliv_dose_invnodes2*]
    - To elective nodes [*deliv_dose_elenodes1, deliv_dose_elenodes2*]

Outcomes

- **Overall survival status** [*os_status*]: Binary variable
    - 0 – Alive
    - 1 – Dead
- **Overall survival - follow-up time (days)** [*os_fup*]: Discrete numerical variable
    - Calculated in number of days from the first fraction of radiotherapy to either event or censoring, whichever happens first.
- **Locoregional failure** [*lrf_status*]: Binary variable
    - 0 – No
    - 1 – Yes
- **Site of locoregional failure** [*lrf_site*]: Categorical variable
    - 0 – No locoregional failure
    - 1 – Primary tumour
    - 2 – Pelvic lymph nodes / lymph nodes in the primary treatment volume
    - 3 – Primary tumour and lymph nodes simultaneous
    - 4 – Other
- **Locoregional failure - follow-up time (days)** [*lrf_fup*]: Discrete numerical variable
    - Calculated in number of days from the first fraction of radiotherapy to either event or censoring, whichever happens first.
- **Distant metastasis** [*dm_status*]: Binary variable
    - 0 – No
    - 1 – Yes
- Site of distant metastasis [*dm_site*]: Categorical variable
    - 0 – No distant metastasis
    - 1 – Lymph nodes outside pelvis
    - 2 – Viscera or bones
    - 3 – Multiple sites
- **Distant metastasis - follow-up time (days)** [*dm_fup*]: Discrete numerical variable
    - Calculated in number of days from the first fraction of radiotherapy to either event or censoring, whichever happens first.

Availability of imaging and treatment plans

In future analysis, baseline imaging biomarkers will be explored and incorporated to the models. The availability of the following factors will be assessed in each centre:

- Pre-treatment FDG PET-CT scan available: Binary variable
    - 0 – No
    - 1 – Yes
- Pre-treatment MRI scan available: Binary variable
    - 0 – No
    - 1 – Yes
- Treatment planning CT scan available: Binary variable
    - 0 – No
    - 1 – Yes
- Full 3D dose distributions for all treatment phases available: Binary variable
    - 0 – No
    - 1 – Yes
- Radiotherapy structure set data available: Binary variable
    - 0 – No
    - 1 – Yes
- Dose-volume histogram (DVH) data available: Binary variable
    - 0 – No
    - 1 – Yes

5.6.2 Supplementary material B. Specification of secondary models

Overall survival secondary models

A range of secondary models for overall survival will be fit, in order to test the robustness of the primary model, as well as explore the impact of having a different set of factors or different factor parameterisation on the model fit. The following changes will be made to the primary model in separate secondary models:

1. TNM staging: will replace T stage and N stage. Categories: Low risk (T1-3N0) vs High risk (T4N(any) or T(any)N+).
2. Age: Modelled as a categorical factor instead of continuous. Categories: [18-39] vs [40-59] vs [60-79] vs [80-99].
3. Age: Modelled as a continuous, non-linear factor. Multiple transformations will be tested prior to the analysis and the most appropriate transformation will be applied.

4. Primary tumour GTV (cm^3): Modelled as a categorical factor instead of continuous. Categories: [0-49.99] vs [50-99.99] vs [100-149.99] vs [150-199.99] vs [200+].
5. Chemotherapy regimen: will be included as a factor with 5 categories. Categories: [No chemotherapy] vs [Mitomycin C and 5-Fluorouracil] vs [Mitomycin C and Capecitabine] vs [Cisplatin and 5-Fluorouracil] vs [Cisplatin and Capecitabine] vs [Other].
6. Incomplete/Interrupted treatment will be included as a binary factor: No vs Yes
7. Performance status will be included as a categorical factor: 0 vs 1 vs 2 vs 3 vs 4 (See Supplementary material A for categorisation).

Note: Items 6 and 7 were found to be prognostic for overall survival in univariable analysis in the systematic review, but not in multivariable analysis. These are *optional* data items in atomCAT2. We will assess the amount of data available and if possible, secondary models which include these factors will be fit.

Locoregional control secondary models

A range of secondary models for locoregional control will be fit. The following changes to the primary model will be made in separate secondary models:

1. Performance status will be included as a categorical factor: 0 vs 1 vs 2 vs 3 vs 4 (See Supplementary material A for categorisation).
2. TNM staging: will replace T stage and N stage. Categories: Low risk (T1-3N0) vs High risk (T4N(any) or T(any)N+).
3. Overall treatment time: will be included as a continuous, linear factor.
4. Age: Modelled as a categorical factor instead of continuous. Categories: [18-39] vs [40-59] vs [60-79] vs [80-99].
5. Age: Modelled as a continuous, non-linear factor. Multiple transformations will be tested prior to the analysis and the most appropriate transformation will be applied.
6. Primary tumour GTV (cm^3): Modelled as a categorical factor instead of continuous. Categories: [0-49.99] vs [50-99.99] vs [100-149.99] vs [150-199.99] vs [200+].
7. Incomplete/Interrupted treatment will be included as a binary factor: No vs Yes.

Note: Items 1 and 7 were found to be prognostic for locoregional control in univariable analysis, but not in multivariable analysis. These are *optional* data items in atomCAT2. We will assess the amount of data available and if possible, secondary models which include these factors will be fit.

Freedom from distant metastasis secondary models

A range of secondary models for freedom from distant metastasis will be fit. The following changes to the primary model will be made in separate secondary models:

1. TNM staging: will replace T stage and N stage. Categories: Low risk (T1-3N0) vs High risk (T4N(any) or T(any)N+).
2. Age: Modelled as a categorical factor instead of continuous. Categories: [18-39] vs [40-59] vs [60-79] vs [80-99].
3. Age: Modelled as a continuous, non-linear factor. Multiple transformations will be tested, and the most appropriate transformation will be applied.
4. Primary tumour GTV (cm^3): Modelled as a categorical factor instead of continuous. Categories: [0-49.99] vs [50-99.99] vs [100-149.99] vs [150-199.99] vs [200+].
5. Chemotherapy regimen: will be included as a factor with 5 categories. Categories: [No chemotherapy] vs [Mitomycin C and 5-Fluorouracil] vs [Mitomycin C and Capecitabine] vs [Cisplatin and 5-Fluorouracil] vs [Cisplatin and Capecitabine] vs [Other].

Chapter 6 - Prognostic models for anal cancer using distributed learning: the international multi-centre atomCAT2 study

6.1 Abstract

6.1.1 Background

Anal cancer is a rare disease typically treated with concurrent chemoradiotherapy. Lack of understanding of prognostic factors means that the options for individualisation of treatment are limited. Due to the rarity of the cancer, single centre (or even single country) data are unlikely to be sufficient for robust development of prognostic models. Distributed learning can allow for analysis of datasets from multiple centres, without exchanging sensitive individual-level patient data. The aim of this study was to collaboratively develop and validate prediction models for multiple anal cancer outcomes through distributed learning in an international consortium.

6.1.2 Methods

This was a retrospective cohort analysis of patients treated with radical intent for anal cancer using conformal radiotherapy in 12 treatment centres based across the UK and Europe. A prospective study protocol and a comprehensive statistical analysis plan were collaboratively developed and published. Collected data included baseline patient characteristics and treatment characteristics. Distributed multivariable Cox Proportional Hazards models for overall survival, locoregional control, and freedom from distant metastasis were developed and validated across all participating centres using the Vantage6 distributed learning infrastructure.

6.1.3 Results

Data from 1,099 patients treated from 2004 to 2022 were analysed. Nodal involvement (HR=1.41, 95% confidence interval 1.06-1.89), male sex (HR=1.69, 95% CI 1.30-2.17), older age (HR=1.34 per 10 years, 95% CI 1.18-1.52), and larger primary tumour size were associated with poorer overall survival. Male sex (HR=1.89, 95% CI 1.41-2.56), higher T stage (HR= 1.55 for T3-4 versus T1-2, 95% CI 1.08-2.22), and larger primary tumour size were prognostic for poorer locoregional control. Nodal involvement (HR=

2.59, 95% CI 1.67-4.00) and larger primary tumour size were prognostic for poorer freedom from distant metastasis. All models exhibited satisfactory performance (cross-validated weighted c-indices 0.60, 0.56, 0.56).

6.1.4 Conclusion

Analysis of a large, contemporary, and international anal cancer cohort provided unique insights into the contrasting prognostic effects of various factors on three clinically important outcomes. These results could inform the design of future clinical trials and the stratification of patients into risk groups, with the ultimate aim of improving outcomes for future patients.

6.2 Background

Anal cancer is a rare disease comprising approximately 0.3% of the total number of cancer cases [1,2]. Over the last two decades, incidence rates have been gradually increasing [3]. The current standard therapy for localised anal cancer consists of concurrent chemotherapy and radiotherapy [4–6]. In the majority of treatment centres, radiotherapy is delivered conformally via intensity-modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT).

The current standard treatment with IMRT/VMAT confers favourable outcomes: complete clinical response rates of 86.7%, as well as three-year overall survival rates of 85.6% and three-year disease-free survival rates of 75.6% have been reported in a modern UK multicentre cohort [7]. Notably, the majority of disease relapses (83.4%) have been reported to occur at the site of the primary disease, highlighting the challenge of achieving locoregional tumour control in a subset of patients. Therefore, a better understanding of the prognostic factors for each outcome is needed, which would enable the development of a more stratified approach to treatment. Even though ongoing prospective clinical trials are currently trying to address this issue [8], clinical practice can meanwhile be informed by prognostic models that are developed using clinical data generated from routine practice [9,10].

Training and validation of robust prognostic models relies on the availability of large amounts of high-quality data [11,12]. Therefore, conducting robust analyses of prognostic factors for anal cancer outcomes can be particularly challenging: Due to the rarity of the disease, each centre treats only a small number of patients every year, and as a result it can take years to obtain sufficient data. In addition, sharing data between centres is limited by ethical barriers and data protection regulations, which render traditional multi-centre data analysis unfeasible [13]. A data analysis approach called distributed learning (DL), or federated learning [14], has paved the way towards collaborative data analysis across centres and international borders.

DL can be implemented to develop prognostic models using local datasets originating from multiple centres, without having to exchange any sensitive individual-level patient data between centres [15]. Only non-identifiable aggregated information in the form of mathematical parameters, such as model coefficients, is shared between centres in order to train and validate a distributed model. Therefore, distributed learning strategies do not breach patient data privacy and minimise issues related to data sharing [16]. DL models are trained iteratively through the exchange of local model coefficients and fit errors between each participating centre and a central server. At each iteration, the central server aggregates the model coefficients computed locally in each centre and a single globally-convergent model is determined by minimizing the total error [17]. The process is repeated until the pre-specified convergence criteria are fulfilled.

This methodology is ideal for the prognostic model development for rare cancers, including anal cancer. Through DL, sufficient amounts of data from multiple centres can be analysed to develop generalisable models, whilst avoiding the aforementioned data sharing barriers. The atomCAT1 proof-of-concept study [18] has demonstrated the feasibility of this approach for anal cancer outcome modelling. In atomCAT1, a Cox proportional hazards model for overall survival was trained and validated in a distributed fashion, using data from three international radiotherapy treatment centres. However, this study only addressed overall survival as the outcome of interest.

To extend this work further, we established the international atomCAT consortium (Anal cancer Treatment Outcome Modelling with Computer Aided Theragnostics), which consists of 12 cancer treatment centres based in the UK and across Europe. In the atomCAT2 study, we aimed to develop and validate prediction models for multiple anal cancer outcomes after chemoradiotherapy through distributed learning. These models can be used to identify key prognostic factors for the outcomes explored, and to determine their effect size. The results from this study may guide the design of future clinical trials in anal cancer.

6.3 Methods

6.3.1 Prospective study protocol & statistical analysis plan

A prospective study protocol and a comprehensive statistical analysis plan for the atomCAT2 study were collaboratively developed, published [19] and registered in Open Science Foundation [20].

6.3.2 Institutional data access & data protection approvals

Each centre acquired separate local approval for accessing and collecting patient data for research. Each local coordinating investigator provided a copy of the letter confirming that use of data for research had been approved (e.g. from the Institutional Review Board, IRB), including approval reference number, to the central study coordinator. For UK centres, a central project application was submitted for review by the Health Research Authority (HRA) and the Research Ethics Committee (REC), to allow for coordinated approval across the National Health Service (NHS). The central project application received HRA and REC approval (IRAS project ID: 303103, REC reference: 22/WA/0081). The atomCAT2 study assumed radiotherapy being the standard of care, with no intervention performed specific to this protocol. Therefore, no informed patient consent was needed to collect the data required for the analyses.

6.3.3 Study design & patient population

In this retrospective multi-centre cohort study, outcomes after primary (chemo)radiotherapy for anal squamous cell carcinoma (ASCC) were investigated. Patients were treated according to each participating centre's protocols, which consisted of concurrent radiotherapy and chemotherapy with varying regimens or radiotherapy only.

Patients treated with radical intent external beam radiotherapy for primary ASCC, with or without concomitant chemotherapy were included. Inclusion was restricted to patients treated with conformal radiotherapy techniques (forward-planned 3D conformal radiotherapy (3D-CRT) or IMRT/VMAT). Patients treated with palliative intent, and patients who had received prior pelvic radiotherapy or brachytherapy (either primary or as boost treatment) were excluded.

6.3.4 Outcome definitions

Three outcomes were explored: overall survival, locoregional control and freedom from distant metastasis. The CORMAC initiative has identified these as key outcome research measures in anal cancer [21].

Overall survival was calculated in days from the first fraction of radiotherapy to either event or censoring, whichever happened first. An event was defined as death from any cause at any point during follow-up. Patients were censored at the last clinical follow-up date if alive.

Time to locoregional control was calculated in days from the first fraction of radiotherapy to either event or censoring, whichever happened first. An event was defined as either abdominoperineal resection to control locoregional disease during follow-up, or locoregional disease progression during follow-up (not managed by surgery), or lack of complete response at 26 weeks from the first radiotherapy fraction. Patients were censored at death, at last clinical follow-up, if undergoing abdominoperineal resection for non-disease related reasons, or in case of distant metastases.

Freedom from distant metastasis was calculated in days from the start of radiotherapy to either event or censoring, whichever happened first. An event was defined as distant disease recurrence (previously untreated lymph node metastasis outside the pelvis, or other metastatic sites such as lung, liver, bone) as a first event. Patients were censored at local recurrence, at death, or at last clinical follow-up.

The complete definitions for all three outcomes can be found in the study protocol [19].

6.3.5 Identification of relevant prognostic factors

To identify established prognostic factors for the outcomes in question, a systematic review of the literature was conducted [22]. This review analysed studies which were published after 2000 and reported on anal cancer outcomes after treatment with conformal radiotherapy. Only studies with large cohorts (>100 patients) were investigated. Factors identified as prognostic through multivariable analysis in multiple studies were selected and prioritised (based on the number of studies reporting on them), which formed an initial list of relevant data to be collected. This list was reviewed by three senior clinical oncologists (AG, MGG, MB), who added additional relevant factors. A data dictionary was created and shared between all centres, for standardised data collection and reporting.

6.3.6 Patient data collection

Patient data were identified and extracted from existing research and clinical databases at each participating centre. To ensure good data quality, each centre spot checked all extracted data, ensuring adherence to the data coding system specified in the data dictionary and identifying any outliers. As a prerequisite, each dataset consisted of data from a minimum of 40 patients to ensure a representative sample, to achieve a reasonable balance of patient heterogeneity, and to limit reporting of subgroups with only a few patients. Prior to the analysis, all patient data was pseudonymized and stripped of protected health information, such as treatment dates, birth and death dates, as well as generic medical record numbers.

6.3.7 Missing data

The study protocol [19] specified a framework on how to deal with missing data at individual centres, in multiple different scenarios. According to this framework, since only fewer than 10% of patients at each centre had missing data items for the primary analyses, complete-case analysis was implemented for these analyses. The one exception was the gross tumour volume (GTV, cm^3), which was not routinely delineated in one centre and was systematically missing for the majority of patients in another centre. The mean GTV across all other centres was used to calculate the global “median of means” value. This value was assigned to all patients with missing GTV data in the two centres.

6.3.8 Sample size

A prospective sample size calculation was carried out using the framework set out by Riley et al. [23], and was implemented using the “pmsampsize” package in R, in order to determine the minimum sample size required to fit a Cox proportional hazards model for each of the three outcomes. The number of prognostic factors which were included in the final models was based on the total number of patients available across the consortium. The detailed methodology used to carry out the sample size calculation is provided in the study protocol [19].

6.3.9 Statistical analysis

6.3.9.1 Descriptive data analysis

The descriptive data analysis, which included the calculation of summary statistics and survival statistics from each centre, was conducted according to the framework set out in the study protocol [19]. The overall summary statistics for continuous variables were calculated using weighted means. Two-year, three-year and 5-year overall survival, locoregional control and freedom from distant metastasis were estimated by Kaplan-Meier methods. Potential follow-up times were based on the inverse Kaplan-Meier estimator [24].

6.3.9.2 Cox model development and reporting

The primary analysis consisted of the development and Type 2b internal validation [25] of distributed multivariable Cox Proportional Hazards models [26] across all participating centres, separately for overall survival, locoregional control, and freedom from distant metastasis. The primary models were pre-specified in the study protocol [19]. Based on the available patient numbers, 8 parameters could be included in the primary models. The predictors used were consequently age, sex, T stage, nodal involvement, gross tumour volume (GTV), prescribed dose to primary tumour and histology (as per the prioritisation lists in the study protocol). For GTV, a $\log^{10}$ transformation was applied prior to model inclusion. All analyses were run as multivariable models, with no data driven factor selection or model reduction.

For each model developed, the estimated factor effects were reported as Hazard Ratios (HRs), along with 95% confidence intervals (CIs). Factors were deemed prognostic if their 95% CIs did not overlap with 1. For each of the three outcomes, the baseline outcome rates at 2 years, 3 years and 5 years were calculated. The baseline outcome rate can be defined as the outcome rate when all model factors are set to their baseline value. To calculate the baseline outcome rates, all categorical factors were set to 0, the age at the start of radiotherapy was set to 35 years, the prescribed dose to the primary tumour was set to 40 $EQD2_{\alpha/\beta=10Gy}$, and the $\log^{10}$ of GTV was set to 0.02572. The combination of the baseline outcome rate and the factor effect estimates (HRs) enables the prediction of outcomes for future patients, rendering the model useable for individual patient outcome prediction.

6.3.9.3 Evaluation and visualisation of model performance

Model performance was assessed using Harrell's concordance index (c-index) [27] on a per-centre basis. The global weighted mean c-index was also calculated. A closed-loop internal-external "leave-one-centre-out" cross-validation method [28] was applied to obtain the out-of-sample performance. During this validation phase, the model was trained using data from all but one centres and was then validated on the last centre. The procedure was repeated to cover all possible combinations. The resulting c-indices provided an estimate of the over-optimism of the global model. The weighted mean c-index values were reported.

The model development and validation procedure and results were reported in accordance with the TRIPOD statement and checklist [29].

6.3.9.4 Model calibration

For the model calibration, the baseline outcome rates at three years were calculated at each participating centre. The weighted mean baseline outcome rates across the entire consortium were then calculated. These were used at each participating centre alongside the factor coefficients from each model to calculate the predicted three-year overall survival, locoregional control and freedom from distant metastasis rates for each patient. Subsequently, the mean predicted outcome rates were calculated and were used to split local cohorts into low risk and high risk groups. Then, the mean predicted three-year outcome rates for each risk group were computed. The actual (observed) three-year outcome rates for each risk group were estimated by Kaplan-Meier methods. For each outcome, the mean predicted outcome rates for the low risk and high risk groups were plotted against the respective actual outcome rates to give an indication of model calibration.

6.3.10 Distributed learning architecture

The Vantage6 v2.3.4 software was used to establish the three elements required to carry out an analysis via distributed learning. The first component is a "node", where individual-level patient data is accessed, and local model coefficients are computed. The second component is a trusted coordinating "server", which handles the communication with the nodes and performs the aggregation of coefficients from all nodes. The final component is a “researcher”, which provides a pre-specified model for training and validation.

At each participating centre, the DL node was set up on either a physical or a virtual personal computer running either Windows, MacOS or Ubuntu, with an installation of Python (v3.7 or v3.8), Docker Desktop (personal edition), and the Vantage6 v2.3.4 Python library. The source code for the infrastructure implementation is openly accessible [https://github.com/IKNL/vantage6 - Version 2.3.4]. Network connectivity was fully compliant with local institutional policies, and only one secured network port through the institution firewall was enabled for Vantage6 traffic.

The Distributed Cox algorithm developed by Lu et al. [26] was adapted to the Vantage6 v2.3.4 infrastructure as R scripts (v.3.6.2) and is publicly available on GitHub [https://github.com/IKNL/vtg.coxph].

Medical Data Works BV (MDW, https://medicaldataworks.nl/) provided and maintained the DL infrastructure that was used to conduct the atomCAT2 DL analysis.

Scripts for model coefficient computation and leave-one-centre-out model validation were packaged as application containers via Docker and were locally executed at each centre.

6.4 Results

6.4.1 Summary of patient characteristics and survival statistics

A total of 1,119 patients treated from 2004 to 2022 across 12 participating centres were identified for inclusion in the analysis. A small number of these patients (n=16) had missing data in essential data items (outcome data was missing for 13 patients, histology data for 2 patients, and T stage data for 1 patient) and were therefore excluded, in order to carry out complete-case analysis.

As a result, data from a total of 1,099 patients treated between 2004 and 2022 were analysed during the model training and validation phase. Based on the sample size calculation (Sections 5.3.6 and 6.3.8) and the total number of patients available), 8 parameters could be included in the models. Table 6-1 summarises the patient characteristics of the cohort included in the analysis, stratified by centre. The survival statistics for the three outcomes explored are provided in Table 6-2.

The overall cohort had a weighted mean age of 63 at the start of radiotherapy, with the majority of patients (68%) being females. The patient age was consistent across centres. The primary tumour volume ranged considerably across centres, from a mean of 23.9cm^3 in Centre 6 to a mean of 76.1cm^3 in Centre 3. The tumours were predominantly squamous cell carcinomas (88%). The mean prescribed dose to primary tumour (in equivalent dose in 2Gy per fraction, α/β=10Gy, EQD2$_{\alpha/\beta=10Gy)}$ was relatively consistent across centres, ranging from 49.8Gy to 60.1Gy. The most common chemotherapy regimen was MMC and 5FU, which was prescribed to 73% (n=801) of all patients. Only 7% (n=74) of patients did not receive any chemotherapy.

233 (21%) of patients in the overall cohort died, 174 (16%) failed locoregionally, and 125 (11%) had a distant metastasis. Weighted mean potential follow-up times across the entire cohort were 54.3 months for overall survival, 46.6 months for locoregional control, and 46.3 months for freedom from distant metastasis. The two-year, three-year, and five-year overall survival, locoregional control and freedom from distant metastasis rates, calculated at each centre individually, are presented in Figure 6-1.

The overall cohort had a weighted mean overall survival rate of 88% at 2 years, 84% at 3 years and 77% at 5 years. The weighted mean locoregional control rates were 85% at 2 years, 83% at 3 years, and 81% at 5 years. Weighted mean freedom from distant metastasis rates across the entire cohort were 89% at 2 years, 88% at 3 years, and 87% at 5 years. Overall survival rates varied more than locoregional control and freedom from distant metastasis rates across centres. At 2 years, mean overall survival ranged from 82% to 97% between centres, at 3 years from 75% to 97%, and at 5 years from 66% to 87%. The differences in locoregional control rates between centres were less apparent, ranging from 78% to 91% at 2 years, 73% to 90% at 3 years, and 73% to 89% at 5 years. Lastly, freedom from distant metastasis was more consistent across centres than the other two outcomes, ranging from 84% to 96% at 2 years and 3 years, and from 80% to 94% at 5 years.

Table 6-1. Summary statistics for patient and treatment characteristics. SD: standard deviation; GTV: Gross tumour volume; SCC: Squamous cell carcinoma; MMC: Mitomycin C; 5FU: 5-fluorouracil; Cap: Capecitabine; Cispl: Cisplatin.

	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5	Centre 6	Centre 7	Centre 8	Centre 9	Centre 10	Centre 11	Centre 12	Overall cohort
Number of patients	**197**	**150**	**128**	**112**	**97**	**82**	**77**	**61**	**53**	**50**	**48**	**44**	**1099**
Treatment period	2015-2021	2014-2022	2013-2017	2016-2021	2011-2016	2009-2021	2008-2017	2008-2021	2004-2017	2010-2022	2016-2018	2013-2022	2004-2022
Sex													
Male	61 (31%)	48 (32%)	35 (27%)	27 (24%)	25 (26%)	35 (43%)	33 (43%)	21 (34%)	23 (43%)	13 (26%)	19 (40%)	9 (20%)	349 (32%)
Female	136 (69%)	102 (68%)	93 (73%)	85 (76%)	72 (74%)	47 (57%)	44 (57%)	40 (66%)	30 (57%)	37 (74%)	29 (60%)	35 (80%)	750 (68%)
Age at the start of radiotherapy (years)													
Mean	61.5	63.3	62.8	62.8	61.9	59.5	61.3	62.0	63.6	63.9	62.1	64.8	63.3
(sd, range)	(11.0, 29-87)	(12.3, 29-90)	(10.6, 40-89)	(11.0, 34-86)	(10.6, 40-87)	(12.2, 35-86)	(10.4, 29-85)	(9.7, 44-83)	(10.5, 39-84)	(10.3, 42-84)	(9.5, 38-78)	(12.6, 39-90)	(11.1, 29-90)
T stage													
T1-2	109 (55%)	99 (66%)	73 (57%)	60 (54%)	58 (60%)	53 (65%)	56 (73%)	23 (38%)	21 (40%)	20 (40%)	34 (71%)	26 (59%)	632 (57%)
T3-4	88 (45%)	51 (34%)	55 (43%)	52 (46%)	39 (40%)	29 (35%)	21 (27%)	38 (62%)	32 (60%)	30 (60%)	14 (29%)	18 (41%)	467 (43%)
N stage													
N0	95 (48%)	79 (53%)	69 (54%)	45 (40%)	57 (59%)	41 (50%)	33 (43%)	36 (59%)	21 (40%)	15 (30%)	28 (58%)	19 (43%)	538 (49%)
N+	102 (52%)	71 (47%)	59 (46%)	67 (60%)	40 (41%)	41 (50%)	44 (57%)	25 (41%)	32 (60%)	35 (70%)	20 (42%)	25 (57%)	561 (51%)
Primary tumour GTV (cm3)													
Mean	55.7	68.6	76.1	64.4	91.0	23.9	59.2	73.6	62.0	50.3	56.5	43.6	62.3
(sd, range)	(88.3, 1.1-974.4)	(68.6, 1.79-446.0)	(67.5, 4.1-459.4)	(68.7, 0.3-314.6)	(101.9, 3.9-651.2)	(32.2, 1.1-212.0)	(73.8, 0.8-433.0)	(67.9, 1.8-328.3)	(0, 62.0-62.0)	(34.1, 8.8-143.8)	(14.7, 10.0-85.0)	(69.60, 1.9-357.4)	(71.7, 0.3-974.4)
GTV delineation													
	Primary tumour only	Primary tumour only	Primary tumour and anal canal at the level of the tumour	Primary tumour only	Primary tumour only for the majority. For some patients, primary tumour and anal canal at the level of the tumour	Primary tumour and anal canal at the level of the tumour	Primary tumour only	Primary tumour only	N/A - Used consortium mean GTV	Primary tumour only	Used consortium mean GTV for most patients. Where GTV is available: Primary tumour only.	Primary tumour only	
Histology													
SCC	180 (92%)	132 (88%)	107 (84%)	95 (85%)	90 (93%)	80 (98%)	76 (99%)	57 (93%)	37 (70%)	41 (82%)	42 (88%)	33 (75%)	970 (88%)
Basaloid SCC	17 (8%)	18 (12%)	21 (16%)	17 (15%)	7 (7%)	2 (2%)	1 (1%)	4 (7%)	16 (30%)	9 (18%)	6 (12%)	11 (25%)	129 (12%)
Primary tumour dose (EQD2 α/β=10)													
Mean	52.04	51.8	56.4	51.4	49.8	55.3	60.1	53.5	52.5	54.0	59.6	56.9	53.8
(sd, range)	(1.7, 52.0-65.2)	(3.83, 40.7-63.6)	(2.0, 54.0-58.1),	(4.5, 40.7-62.6)	(4.2, 28.3-54.9)	(2.0, 49.6-60.0)	(2.5, 54-66)	(3.4, 49.6-60)	(2.1, 54.6-58.4)	(3.0, 49.6-60.0)	(2.3, 58.4-63.7)	(4.2, 44.3-62.0)	(4.3, 28.3-66.0)
Chemotherapy regimen													
No chemotherapy	1 (1%)	13 (9%)	9 (7%)	20 (18%)	0 (0%)	6 (7%)	12 (16%)	6 (10%)	0 (0%)	0 (0%)	1 (2%)	6 (14%)	74 (7%)
MMC and 5FU	176 (89%)	100 (67%)	114 (88%)	77 (69%)	94 (97%)	67 (82%)	0 (0%)	49 (80%)	48 (90%)	44 (88%)	0 (0%)	32 (73%)	801 (73%)
MMC and Cap	14 (7%)	33 (22%)	1 (1%)	14 (13%)	0 (0%)	7 (9%)	64 (83%)	0 (0%)	3 (6%)	0 (0%)	47 (98%)	5 (11%)	188 (17%)
Cispl and 5FU	1 (1%)	0 (0%)	4 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (8%)	0 (0%)	2 (4%)	0 (0%)	0 (0%)	12 (1%)
Cispl and Cap	0 (0%)	1 (1%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	3 (<1%)
Other	5 (3%)	3 (2%)	0 (0%)	0 (0%)	3 (3%)	2 2%)	1 (1%)	1 (2%)	2 (4%)	4 (8%)	0 (0%)	0 (0%)	21 (2%)

Table 6-2. Summary of survival statistics, including potential follow-up times, by outcome and centre.

		Centre 1	Centre 2	Centre 3	Centre 4	Centre 5	Centre 6	Centre 7	Centre 8	Centre 9	Centre 10	Centre 11	Centre 12	Overall cohort
Number of patients		**197**	**150**	**128**	**112**	**97**	**82**	**77**	**61**	**53**	**50**	**48**	**44**	**1099**
Number of events														
	Deaths	43 (22%)	35 (23%)	20 (16%)	13 (12%)	26 27%)	14 (17%)	21 (27%)	17 (28%)	22 (42%)	5 (10%)	11 (23%)	6 1(4%)	233 (21%)
	Locoregional failures	33 (17%)	25 (17%)	13 (10%)	15 (13%)	18 (19%)	13 (16%)	13 (17%)	10 (16%)	10 (19%)	8 (16%)	12 (25%)	4 (9%)	174 (16%)
	Distant metastases	25 (13%)	17 (11%)	7 (5%)	5 (4%)	16 (16%)	12 (15%)	10 (13%)	5 (8%)	9 (17%)	5 (10%)	7 (15%)	7 (16%)	125 (11%)
Overall survival - potential follow-up time (months)														
	Median	56.7	42.5	62.3	42.3	79.8	46.4	47.8	44.5	110.4	37.6	36.9	43.9	54.3
Estimated 3-year overall survival														
	Survival	83%	78%	92%	88%	86%	82%	75%	80%	75%	97%	78%	86%	83%
	(std error, 95% CI)	(3%, 78%-89%)	(4%, 71%-86%)	(2%, 88%-97%)	(3%, 82%-95%)	(4%, 79%-93%)	(5%, 72%-92%)	(5%, 66%-82%)	(5%, 71%-92%)	(6%, 64%-88%)	(3%, 92%-100%)	(6%, 67%-91%)	(6%, 75%-98%)	(2%, 80%-87%)
Locoregional control - potential follow-up time (months)														
	Median	43.1	31.3	61.5	39.3	73.1	43.6	39.6	37.6	83.1	36.0	36.9	40.9	46.6
Estimated 3-year locoregional control														
	Survival	82%	80%	90%	85%	82%	81%	75%	83%	83%	80%	73%	89%	82%
	(std error, 95% CI)	(3%, 77%-88%)	(4%, 73%-89%)	(3%, 85%-96%)	(4%, 78%-92%)	(4%, 75%-90%)	(5%, 72%-92%)	(4$, 82%-96%)	(5%, 74%-94%)	(5%, 73%-94%)	(7%, 68%-93%)	(7%, 61%-87%)	(5%, 79%-100%)	(2%, 80%-86%)
Freedom from distant metastasis - potential follow-up time (months)														
	Median	42.2	28.8	61.4	39.3	75.8	42.6	39.6	36.6	86.8	35.6	34.7	42.0	46.3
Estimated 3-year freedom from distant metastasis														
	Survival	87%	86%	95%	96%	86%%	85%	86%	91%	84%	87%	84%	86%	81%
	(std error, 95% CI)	(3%, 82%-92%	(3%, 79%-93%)	(2%, 91%-99%)	(2%, 93%-100%)	(4%, 84%-96%	(5%, 76%-95%)	(4%, 78%-94%)	(4%, 84%-99%)	(5%, 74%-95%)	(6%, 77%-98%)	(6%, 73%-96%)	(6%, 76%-99%)	(1%, 85%-90%)

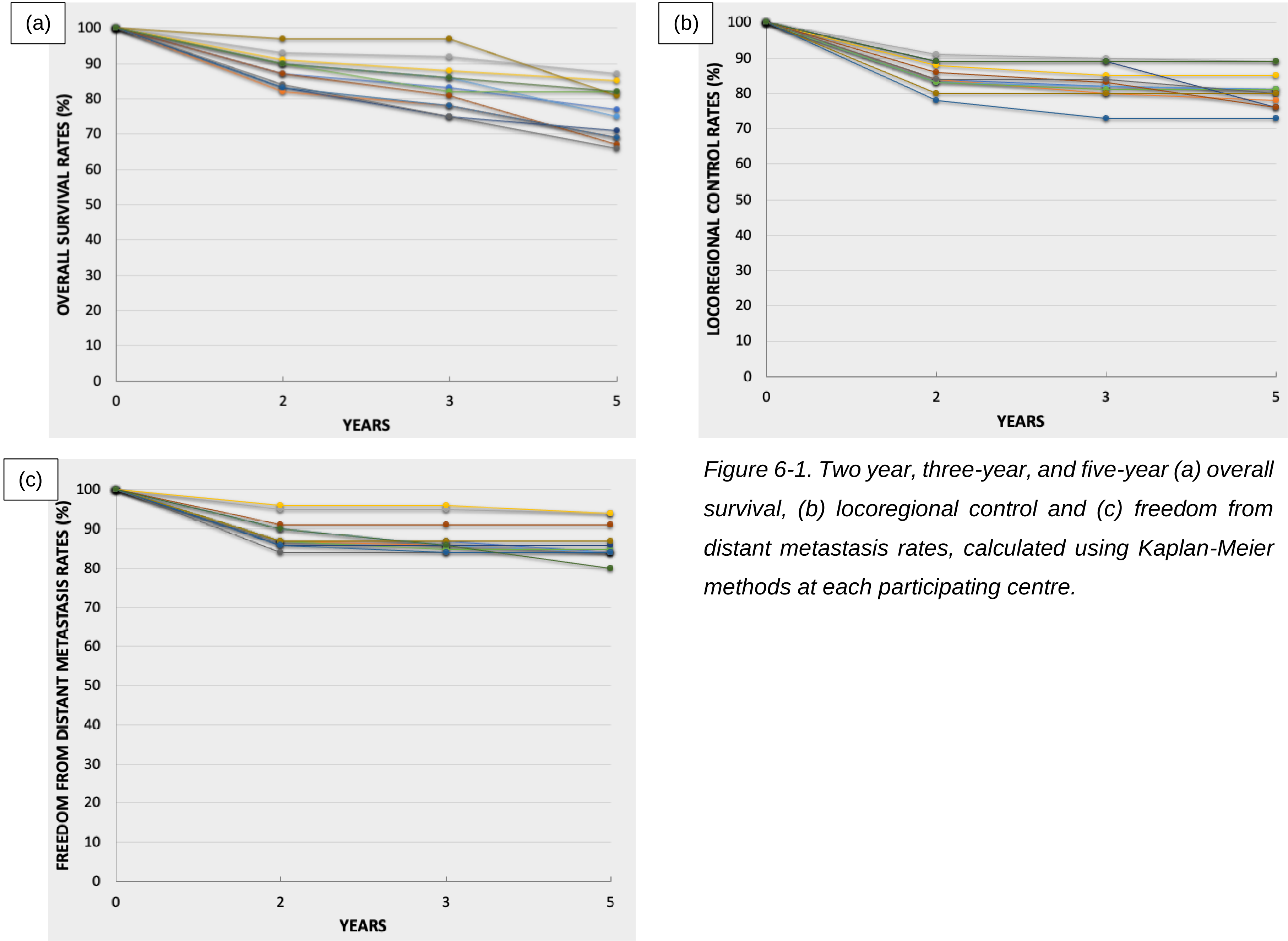


Figure 6-1. Two year, three-year, and five-year (a) overall survival, (b) locoregional control and (c) freedom from distant metastasis rates, calculated using Kaplan-Meier methods at each participating centre.

6.4.2 Multivariable models for overall survival, locoregional control and freedom from distant metastasis

The results from the overall survival, locoregional control and freedom from distant metastasis global models, trained using all available data from all centres, are presented in Table 6-3. The age at the start of radiotherapy, the primary tumour size and the prescribed dose to the primary tumour were modelled as continuous variables. The factor effects are expressed in the form of HR estimates.

Table 6-3. Summary of results from the overall survival, locoregional control, and freedom from distant metastasis models trained on all 12 cohorts.

Factor	Hazard ratio (95% CI)		
	Overall survival	**Locoregional control**	**Freedom from distant metastasis**
Nodal involvement (N+ relative to N0)	1.41 (1.06 - 1.89)	1.35 (0.97 - 1.89)	2.59 (1.67 - 4.00)
T stage (T3-4 relative to T1-2)	1.27 (0.94 - 1.73)	1.55 (1.08 - 2.22)	1.19 (0.79 - 1.79)
Sex (Female relative to male)	0.59 (0.46 - 0.77)	0.53 (0.39 - 0.71)	0.89 (0.60 - 1.30)
Age at the start of radiotherapy (per 10 years)	1.34 (1.18 - 1.52)	1.04 (0.10 - 1.20)	1.07 (0.91 - 1.27)
Gross tumour volume (log10)	1.98 (1.39 - 2.84)	2.40 (1.59 - 3.63)	1.77 (1.11 - 2.83)
Prescribed dose to primary tumour (per 10 Gy)	1.11 (0.79 - 1.58)	1.31 (0.88 - 1.95)	1.28 (0.80 - 2.03)
Histology (Basaloid SCC relative to SCC)	1.02 (0.69 - 1.52)	0.70 (0.41 - 1.22)	0.80 (0.44 - 1.45)

Nodal involvement, male sex, older age, and larger primary tumour size were associated with poorer overall survival. Moreover, male sex, higher T stage and larger primary tumour size were associated with worse locoregional control. Lastly, nodal involvement and larger primary tumour size were associated with worse freedom from distant metastasis.

6.4.3 Global model performance and leave-one-centre-out validation

The performance of each global model was assessed on each node, yielding a weighted mean c-index of 0.603 for the overall survival model, 0.564 for the locoregional control model and 0.563 for the freedom from distant metastasis model. As indicated in Table 6-4, the performance of all three models varied considerably across centres for all outcomes explored, ranging from 0.45 to 0.72 for overall survival, from 0.37 to 0.85 for locoregional control and from 0.36 to 0.78 for freedom from distant metastasis.

The leave-one-centre-out validation c-indices are also summarised in Table 6-4. The weighted mean leave-one-centre-out validation c-index values for all three models were very similar to the respective global model validation c-index values, with a difference of less than 0.002 in all three cases. This suggests that model performance remains stable when model training is carried out using data from all but one completely independent dataset, which was used for validation.

Table 6-4. Summary of results from the global validation and the leave-one-centre-out validation of the overall survival, locoregional control, and freedom from distant metastasis models. For the leave-one-centre-out validation, each model was trained on all but one cohort, and subsequently validated on the last, independent cohort.

Centre	**Number of patients**	**Overall survival**		**Locoregional control**		**Freedom from distant metastasis**	
		Global model c-index	**LOCOV c-index**	**Global model c-index**	**LOCOV c-index**	**Global model c-index**	**LOCOV c-index**
1	197	0.57	0.57	0.60	0.59	0.49	0.49
2	150	0.72	0.72	0.56	0.55	0.52	0.52
3	128	0.63	0.62	0.58	0.58	0.59	0.60
4	112	0.51	0.51	0.60	0.61	0.68	0.68
5	97	0.55	0.55	0.49	0.49	0.59	0.59
6	82	0.66	0.66	0.53	0.53	0.55	0.55
7	77	0.59	0.59	0.47	0.47	0.61	0.60
8	61	0.63	0.63	0.62	0.62	0.50	0.50
9	54	0.55	0.55	0.50	0.50	0.51	0.51
10	50	0.45	0.45	0.37	0.37	0.36	0.36
11	48	0.71	0.71	0.61	0.61	0.71	0.71
12	43	0.65	0.65	0.85	0.85	0.78	0.78
Weighted mean		**0.60**	**0.60**	**0.56**	**0.56**	**0.56**	**0.56**

6.4.4 Model calibration

Figure 6-2 presents the calibration plots for three-year overall survival, locoregional control, and freedom from distant metastasis. The closer the points are to the dashed reference line, the stronger the calibration, signifying that the predicted outcome rates correspond more closely to the actual (observed) outcome rates. The overall survival model calibration plot (Figure 6-2a) and the locoregional control model calibration plot (Figure 6-2b) indicate a moderate calibration for the low risk groups, but substantially weaker calibration for the high risk groups, especially for the smaller cohorts. The calibration appears to be considerably stronger for freedom from distant metastasis (Figure 6-2c) for both the low risk and high risk groups. Overall, the calibration of all three models appears to be weaker for the smaller cohorts compared to the larger cohorts.

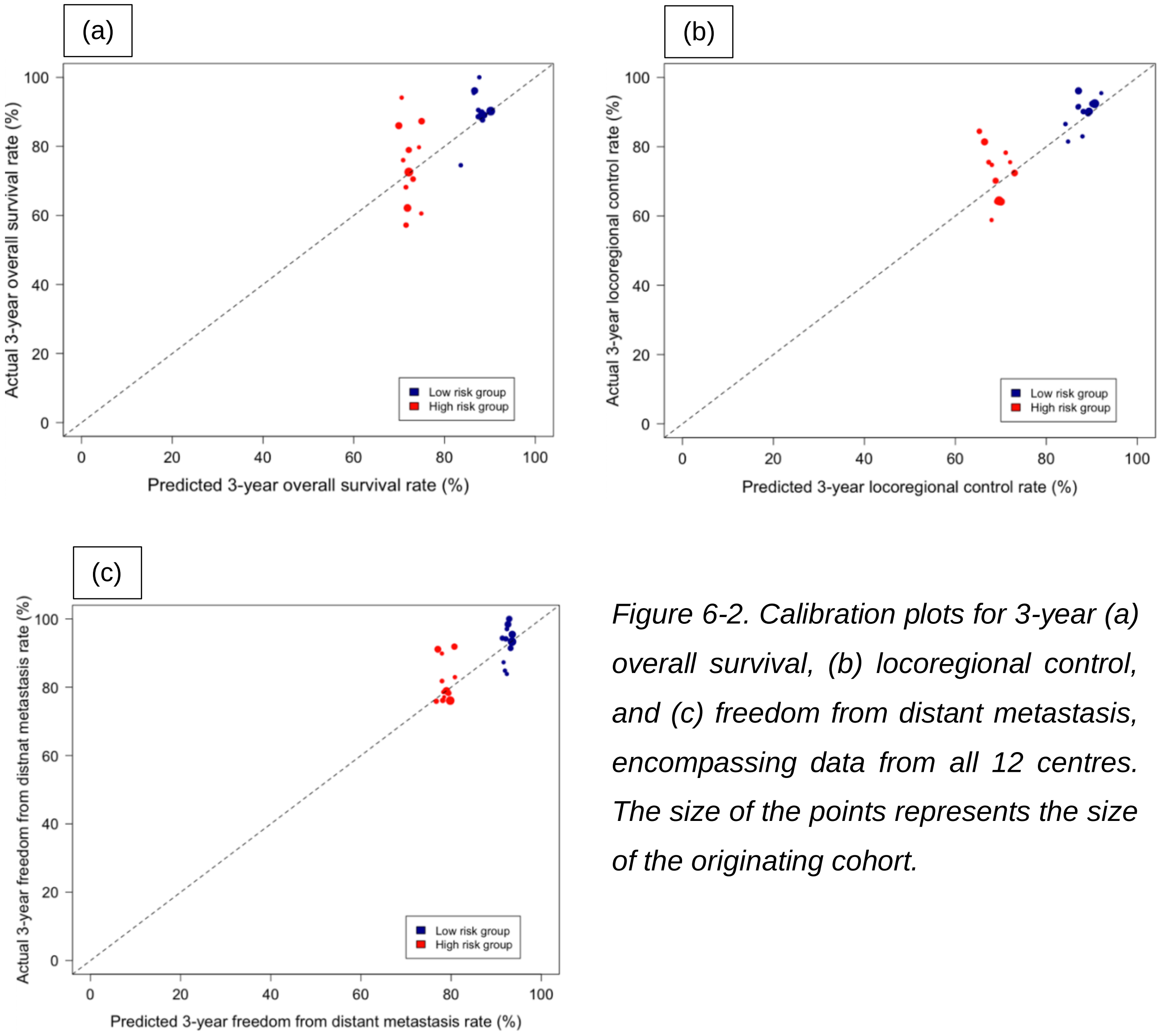


Figure 6-2. Calibration plots for 3-year (a) overall survival, (b) locoregional control, and (c) freedom from distant metastasis, encompassing data from all 12 centres. The size of the points represents the size of the originating cohort.

6.5 Discussion

We developed and validated models for three clinically important outcomes in anal cancer via DL in a large, multi-centre cohort of 1,099 patients. Nodal involvement, male sex, older age, and larger primary tumour size were identified as prognostic for poorer overall survival. Male sex, higher T stage, and larger primary tumour size were found to be associated with poorer locoregional control. Nodal involvement and larger primary tumour size were deemed prognostic for poorer freedom from distant metastasis. All three models exhibited satisfactory performance and were not over-optimistic, as demonstrated by the leave-one-centre-out validation.

To our knowledge, the multi-national atomCAT2 cohort is the largest contemporary international anal cancer cohort ever analysed. The outcome prediction models were developed using real-world data and reflect the real-world heterogeneity of outcomes observed across centres. Analysing a cohort of 1,099 patients with anal cancer meant that a large number of parameters (n=8) could be included in the models, enabling us to investigate their effect separately on the three outcomes of interest. By prospectively carrying out a sample size calculation, we aimed to reduce the chance of model overfitting and to ensure that the overall risk of each outcome is estimated precisely [23]. As suggested by the results from the leave-one-centre-out validation (Table 4), none of the models developed suffer from overfitting. Since the weighted mean c-indices from the global models and from the leave-one-centre-out validation models are very similar, we can assume that the models can be used to carry out predictions in new, completely independent datasets, without exhibiting a significant drop in performance. Additionally, since the models were trained using data from multiple centres, which employed different treatment techniques and protocols, the results can be considered generalisable.

Other similar studies which carried out multivariable modelling to identify prognostic factors for anal cancer outcomes after conformal radiotherapy included data from 1,015 patients [30] (abstract only), 987 patients [31] and 385 patients [7] in their analysis. All of these were multi-centre, although none analysed international cohort data. This highlights the importance of multi-centre collaborations in the field of anal cancer outcome modelling. The DL methodology was the key to establishing the atomCAT2 collaboration, as it eliminated the need for complex cross-border data sharing agreements, which could have taken years to get in place.

The results from this study largely confirm results from previous, smaller studies. A higher N stage or lymph node involvement [31,31–33], male sex [7,30,34–37], and older age [37,38] were previously established as prognostic for poorer overall survival in multiple studies. In the locoregional control model, male sex was found to be prognostic, further confirming the findings from previous studies [7,34–36]. For freedom from distant metastasis, the negative effect of higher N stage was established, which confirmed the findings from two other studies [34,39]. However, a number of previously established prognostic factors were not confirmed as prognostic in this study, including higher T stage for overall survival [37,40,41], higher N stage or nodal involvement [7,34,39] for locoregional control, as well as higher T stage [34,40] and male sex [34,42] for freedom from distant metastasis. Importantly, our results suggest that the prescribed radiotherapy dose is not prognostic for any of the three outcomes that were explored. This contradicts the findings from previous studies which have developed tumour control probability (TCP) models using literature-based data [43] and through the analysis of a large Nordic database [44]. These studies explored the relationship between radiotherapy dose and treatment response and their results demonstrated a clear dose-response relationship. More specifically, both studies concluded that a lower radiotherapy dose should be delivered to small tumours (dose de-escalation), and a higher radiotherapy dose should be delivered to large tumours (dose escalation). This hypothesis is currently being investigated in the ongoing PLATO trial [8], which aims to investigate the role of dose de-escalation and escalation according to how advanced the anal cancer is at diagnosis. The results from this trial will provide vital information on the effect of the radiotherapy dose on patient outcomes that could lead to the establishment of treatment stratification approaches. Lastly, a novel finding from the atomCAT2 results is that the primary tumour size (GTV in cm3) was found to be prognostic for all three outcomes explored. This emphasises the need to explore the effect of imaging-specific factors on various anal cancer outcomes, in order to establish their prognostic value.

To our knowledge, only two phase III prospective clinical trials have reported prognostic factors on anal cancer outcomes; the RTOG 98-11 [45] and ACT II trials [46], both of which were multicentre and were characterised by high quality data curation, but were conducted in the pre-IMRT era. The prognostic factors identified in this study largely agree with the results from these two trials. Nodal involvement and male sex were found to be prognostic for worse overall survival in both trials [45,46]. Older age was associated with poorer overall survival in the ACT II trial [46], and a larger primary tumour size was associated with poorer overall survival in the RTOG 98-11 trial [45]. Male sex was also deemed prognostic for worse locoregional control in the ACT II trial [46]. However, the results from the ACT II trial also suggest that nodal involvement is prognostic for poorer locoregional control, a finding that could not be confirmed by the results from the atomCAT2 study. The above suggest that, overall, prognostic factors identified in the pre-IMRT era are mostly consistent with the prognostic factors identified by analysing a cohort treated with modern radiotherapy techniques. Additionally, even though the atomCAT2 study analysed data from a much more heterogeneous population, the overall conclusions are very similar to those of the aforementioned trials.

A number of study limitations are related to the technical aspects of the DL implementation. Since the DL methodology and infrastructure is relatively new, there are still a number of prognostic modelling aspects that cannot be carried out in a fully distributed fashion. For instance, feature selection in the context of DL has only been sparsely explored. Only a small number of studies have tried to adapt centralised feature selection algorithms for use in a distributed setting [47]. This is also true for missing data imputation techniques. Even though a large number of feasible centralised approaches to missing data imputation are currently available [48,49] only a few of these have been adapted for use via the DL infrastructure [50,51], and additional evidence is needed to confirm their validity. Therefore, a simple approach of complete-case analysis, and overall mean imputation was carried out to address the data missingness in the atomCAT2 cohort. Mean imputation is suboptimal, and as a result, the imputed data could have negatively affected the accuracy of the results. Moreover, the GTV delineation was not uniform across centres (Table 1); in the majority of centres, the GTV included the primary tumour only, whereas for a small number of centres the GTV included the primary tumour and the anal canal on level with the tumour. Due to the retrospective nature of this study, this limitation could not be addressed. Finally, potential correlations between the various factors analysed could not be assessed across the entire cohort due to technical constraints with the DL infrastructure. This is important, since the primary tumour size (GTV) is likely to be correlated to T stage, although this could not be explicitly tested, apart from on a per-centre basis. However, GTV and T stage are related but distinct measures in the assessment of anal cancer. GTV is a direct measurement of the size of the primary tumour, whereas the T stage classification accounts for the size of the primary tumour as well as the extent to which it has grown into nearby structures. Tumours of any size can be classified as T4, even small tumours with a small GTV, in cases where the tumour has invaded nearby structures. Therefore, including both GTV and T stage as factors in an outcome prediction model may improve the performance of the model by capturing slightly different aspects of the disease. This notion is supported by the results from the locoregional control model (Table 6-3), which indicate that both GTV and T stage were associated with worse locoregional control. However, only GTV was associated with poorer overall survival and freedom from distant metastasis. In order to robustly test whether it is informative to use both factors in our models, sensitivity analyses can be undertaken. In these analyses, models can be developed without GTV or T stage, in order to investigate whether this change affects the effect of the other factor, as well as the overall performance of the model. For instance, if GTV is removed from the overall survival model and T stage becomes prognostic, this would suggest that the two factors are partly providing the same information and only one should be included in the model. In addition, the performance of the sensitivity analysis models (in terms of c-index) can be compared with the performance of the primary models to provide further evidence as to whether both factors should be included in the model, or only one.

Various plans have been devised for future analyses, to extend the work conducted. A range of secondary and exploratory analyses will be carried out, in order to explore alternative coding/parameterisation of factors, and to assess the robustness of the results to the choices made for the primary models. To further evaluate the model robustness, they will also be trained only on datasets comprising of more than 20 events, to evaluate whether the number of events per centre affects the behaviour of the models. Due to time limitations, the calibration of the models could not be assessed fully. In the work presented above, the mean predicted outcome rate from each individual centre was used to stratify local cohorts into risk groups. This means that risk groups were defined differently in each centre, and hence, this does not provide a risk stratification that could be applied to a future patient diagnosed with anal cancer. However, this approach can still provide an indication of model calibration and performance. To improve upon this approach, the weighted mean outcome rates across the entire consortium need to be calculated and subsequently used to stratify patients into low risk and high risk groups. This analysis will take place before the work is submitted for publication. Future work beyond atomCAT2 includes exploring more complex research questions in anal cancer via the established consortium, using DL.

This may involve the inclusion of biomarker data, as well as additional radiotherapy-specific data in our models, such as imaging and radiomics data.

In conclusion, the atomCAT2 study analysed the largest contemporary, international anal cancer cohorts, with all patients treated with conformal radiotherapy. It has provided unique insights into the distinct prognostic effect of different patient and disease characteristics on overall survival, locoregional control and freedom from distant metastasis. The results from the atomCAT2 study could inform the design of future clinical trials and the stratification of patients into risk groups, with the ultimate aim of improving outcomes for future patients.

6.6 References

[1] Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. Int J Epidemiol 2016:dyw276. https://doi.org/10.1093/ije/dyw276.

[2] Salati SA. Anal Cancer : A Review. Int J Health Sci 2012;6:206–30. https://doi.org/10.12816/0006000.

[3] van der Zee RP, Richel O, de Vries HJC, Prins JM. The increasing incidence of anal cancer: can it be explained by trends in risk groups? Neth J Med 2013;71:401–11.

[4] Nigro ND, Vaitkevicius VK, Considine B. Combined therapy for cancer of the anal canal: A preliminary report. Dis Colon Rectum 1974;17:354–6. https://doi.org/10.1007/BF02586980.

[5] Rao S, Guren MG, Khan K, Brown G, Renehan AG, Steigen SE, et al. Anal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up☆. Ann Oncol 2021;32:1087–100. https://doi.org/10.1016/j.annonc.2021.06.015.

[6] Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol 2014;40:1165–76. https://doi.org/10.1016/j.ejso.2014.07.030.

[7] Shakir R, Adams R, Cooper R, Downing A, Geh I, Gilbert D, et al. Patterns and Predictors of Relapse Following Radical Chemoradiation Therapy Delivered Using Intensity Modulated Radiation Therapy With a Simultaneous Integrated Boost in Anal Squamous Cell Carcinoma. Int J Radiat Oncol 2020;106:329–39. https://doi.org/10.1016/j.ijrobp.2019.10.016.

[8] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[9] Sturdza A, Pötter R, Fokdal LU, Haie-Meder C, Tan LT, Mazeron R, et al. Image guided brachytherapy in locally advanced cervical cancer: Improved pelvic control and survival in RetroEMBRACE, a multicenter cohort study. Radiother Oncol 2016;120:428–33. https://doi.org/10.1016/j.radonc.2016.03.011.

[10] Tanderup K, Fokdal LU, Sturdza A, Haie-Meder C, Mazeron R, van Limbergen E, et al. Effect of tumor dose, volume and overall treatment time on local control after

radiochemotherapy including MRI guided brachytherapy of locally advanced cervical cancer. Radiother Oncol 2016;120:441–6. https://doi.org/10.1016/j.radonc.2016.05.014.

[11] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ 2013;346:e5595–e5595. https://doi.org/10.1136/bmj.e5595.

[12] Kent P, Cancelliere C, Boyle E, Cassidy JD, Kongsted A. A conceptual framework for prognostic research. BMC Med Res Methodol 2020;20:172. https://doi.org/10.1186/s12874-020-01050-7.

[13] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9. https://doi.org/10.1016/j.radonc.2014.10.001.

[14] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 2016;121:459–67. https://doi.org/10.1016/j.radonc.2016.10.002.

[15] Kirienko M, Sollini M, Ninatti G, Loiacono D, Giacomello E, Gozzi N, et al. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. Eur J Nucl Med Mol Imaging 2021;48:3791–804. https://doi.org/10.1007/s00259-021-05339-7.

[16] Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform 2020:184–200. https://doi.org/10.1200/CCI.19.00047.

[17] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[18] Theophanous S, Choudhury A, Lønne P-I, Samuel R, Guren MG, Berbee M, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study. Radiother Oncol 2021;159:183–9. https://doi.org/10.1016/j.radonc.2021.03.013.

[19] Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Dennis K, et al. Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study. Diagn Progn Res 2022;6:14. https://doi.org/10.1186/s41512-022-00128-8.

[20] Theophanous, Stelios. atomCAT2 - A multicentre study of overall survival, locoregional control and distant metastasis in anal cancer utilising distributed learning 2021. https://doi.org/10.17605/OSF.IO/J7AUH.

[21] Fish R, Sanders C, Adams R, Brewer J, Brookes ST, DeNardo J, et al. A core outcome set for clinical trials of chemoradiotherapy interventions for anal cancer (CORMAC): a patient and health-care professional consensus. Lancet Gastroenterol Hepatol 2018;3:865–73. https://doi.org/10.1016/S2468-1253(18)30264-4.

[22] Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, et al. Prognostic Factors for Patients with Anal Cancer Treated with Conformal Radiotherapy - A Systematic Review. In Review; 2022. https://doi.org/10.21203/rs.3.rs-1324086/v1.

[23] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96. https://doi.org/10.1002/sim.7992.

[24] Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343–6. https://doi.org/10.1016/0197-2456(96)00075-X.

[25] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015;162:W1. https://doi.org/10.7326/M14-0698.

[26] Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: A web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015:ocv083. https://doi.org/10.1093/jamia/ocv083.

[27] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17. https://doi.org/10.1002/sim.4154.

[28] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[29] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55. https://doi.org/10.7326/M14-0697.

[30] Vendrely V, Lemanksi C, Pommier P, Le Malicot K, Francois E, De La Rochefordiere A, et al. OC-0270 - Final results of the French national cohort ANABASE : treatment and outcome in anal cancer. Radiother Oncol 2022;170:S226–7.

[31] Caravatta L, Mantello G, Valvo F, Franco P, Gasparini L, Rosa C, et al. Radiotherapy with Intensity-Modulated (IMRT) Techniques in the Treatment of Anal Carcinoma (RAINSTORM): A Multicenter Study on Behalf of AIRO (Italian Association of Radiotherapy and Clinical Oncology) Gastrointestinal Study Group. Cancers 2021;13:1902. https://doi.org/10.3390/cancers13081902.

[32] Schernberg A, Escande A, Rivin Del Campo E, Ducreux M, Nguyen F, Goere D, et al. Leukocytosis and neutrophilia predicts outcome in anal cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2017;122:137–45. https://doi.org/10.1016/j.radonc.2016.12.009.

[33] Call JA, Haddock MG, Quevedo JF, Larson DW, Miller RC. Intensity-modulated radiotherapy for squamous cell carcinoma of the anal canal: efficacy of a low daily dose to clinically negative regions. Radiat Oncol Lond Engl 2011;6:134. https://doi.org/10.1186/1748-717X-6-134.

[34] Martin D, Rödel F, von der Grün J, Rödel C, Fokas E. Acute organ toxicity correlates with better clinical outcome after chemoradiotherapy in patients with anal carcinoma. Radiother Oncol 2020;149:168–73. https://doi.org/10.1016/j.radonc.2020.05.016.

[35] Balermpas P, Martin D, Wieland U, Rave-Fränk M, Strebhardt K, Rödel C, et al. Human papilloma virus load and PD-1/PD-L1, CD8+ and FOXP3 in anal cancer patients

treated with chemoradiotherapy: Rationale for immunotherapy. OncoImmunology 2017;6:e1288331. https://doi.org/10.1080/2162402X.2017.1288331.

[36] Schernberg A, Huguet F, Moureau-Zabotto L, Chargari C, Rivin Del Campo E, Schlienger M, et al. External validation of leukocytosis and neutrophilia as a prognostic marker in anal carcinoma treated with definitive chemoradiation. Radiother Oncol J Eur Soc Ther Radiol Oncol 2017;124:110–7. https://doi.org/10.1016/j.radonc.2017.06.009.

[37] Hosni A, Han K, Le LW, Ringash J, Brierley J, Wong R, et al. The ongoing challenge of large anal cancers: prospective long term outcomes of intensity-modulated radiation therapy with concurrent chemotherapy. Oncotarget 2018;9:20439–50. https://doi.org/10.18632/oncotarget.24926.

[38] Rouard N, Peiffert D, Rio E, Mahé M-A, Delpon G, Marchesi V, et al. Intensity-modulated radiation therapy of anal squamous cell carcinoma: Relationship between delineation quality and regional recurrence. Radiother Oncol J Eur Soc Ther Radiol Oncol 2019;131:93–100. https://doi.org/10.1016/j.radonc.2018.10.021.

[39] Martin D, Rödel F, Balermpas P, Winkelmann R, Fokas E, Rödel C. C-Reactive Protein-to-Albumin Ratio as Prognostic Marker for Anal Squamous Cell Carcinoma Treated With Chemoradiotherapy. Front Oncol 2019;9:1200. https://doi.org/10.3389/fonc.2019.01200.

[40] de Meric de Bellefon M, Lemanski C, Castan F, Samalin E, Mazard T, Lenglet A, et al. Long-term follow-up experience in anal canal cancer treated with Intensity-Modulated Radiation Therapy: Clinical outcomes, patterns of relapse and predictors of failure. Radiother Oncol 2020;144:141–7. https://doi.org/10.1016/j.radonc.2019.11.016.

[41] Bitterman DS, Grew D, Gu P, Cohen RF, Sanfilippo NJ, Leichman CG, et al. Comparison of anal cancer outcomes in public and private hospital patients treated at a single radiation oncology center. J Gastrointest Oncol 2015;6:524–33. https://doi.org/10.3978/j.issn.2078-6891.2015.061.

[42] Brown PJ, Zhong J, Frood R, Currie S, Gilbert A, Appelt AL, et al. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging 2019;46:2790–9. https://doi.org/10.1007/s00259-019-04495-1.

[43] Muirhead R, Partridge M, Hawkins MA. A tumor control probability model for anal squamous cell carcinoma. Radiother Oncol 2015;116:192–6. https://doi.org/10.1016/j.radonc.2015.07.014.

[44] Johnsson A, Leon O, Gunnlaugsson A, Nilsson P, Höglund P. Determinants for local tumour control probability after radiotherapy of anal cancer. Radiother Oncol 2018;128:380–6. https://doi.org/10.1016/j.radonc.2018.06.007.

[45] Ajani JA, Winter KA, Gunderson LL, Pedersen J, Benson AB, Thomas CR, et al. Prognostic factors derived from a prospective database dictate clinical biology of anal cancer: The intergroup trial (RTOG 98-11). Cancer 2010;116:4007–13. https://doi.org/10.1002/cncr.25188.

[46] Glynne-Jones R, Sebag-Montefiore D, Adams R, Gollins S, Harrison M, Meadows HM, et al. Prognostic factors for recurrence and survival in anal cancer: generating hypotheses from the mature outcomes of the first United Kingdom Coordinating Committee on Cancer Research Anal Cancer Trial (ACT I). Cancer 2013;119:748–55. https://doi.org/10.1002/cncr.27825.

[47] He Y, Zhou Y, Feng Y. Distributed Feature Selection for High-dimensional Additive Models 2022. https://doi.org/10.48550/ARXIV.2205.07932.

[48] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393–b2393. https://doi.org/10.1136/bmj.b2393.

[49] Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus 2013;2:222. https://doi.org/10.1186/2193-1801-2-222.

[50] Chang C, Deng Y, Jiang X, Long Q. Multiple imputation for analysis of incomplete data in distributed health data networks. Nat Commun 2020;11:5467. https://doi.org/10.1038/s41467-020-19270-2.

[51] Brink C, Hansen CR, Field M, Price G, Thwaites D, Sarup N, et al. Distributed learning optimisation of Cox models can leak patient data: Risks and solutions 2022.

Chapter 7 - Discussion

7.1 Summary of aims

This thesis project aimed to determine prognostic factors for patients with anal cancer treated with chemoradiotherapy using real-world data, across an international consortium. To achieve this, it was necessary to:

- Establish a local structure for data collection and use within Leeds Cancer Centre by incorporating anal cancer patient data from various disparate sources and collating them into a single data warehouse (Chapter 2).
- Systematically review and summarise the literature on prognostic factors to identify established prognostic factors for a range of clinically important disease-related outcomes for use within the prognostic models (Chapter 3).
- Establish the feasibility and effectiveness of a distributed learning approach through the atomCAT1 proof-of-concept study (Chapter 4).
- Prospectively develop and publish a study protocol and data analysis plan for the atomCAT2 study (Chapter 5).
- Implement this approach in a larger international consortium, in order to develop prognostic models for anal cancer outcomes (Chapter 6).

Through the analysis of a large cohort of anal cancer patients treated at multiple international centres, it was possible to collaboratively develop distributed prediction models for multiple anal cancer outcomes and identify distinct prognostic factors for overall survival, locoregional control and freedom from distant metastasis. Specifically, nodal involvement, male sex, older age, and larger primary tumour size were identified as prognostic for poorer overall survival. Male sex, higher T stage, and larger primary tumour size were found to be associated with poorer locoregional control. Nodal involvement and larger primary tumour size were deemed prognostic for poorer freedom from distant metastasis. The results from this work can feed into the design of future clinical trials, which may ultimately guide the establishment of more personalised approaches to anal cancer radiotherapy.

In the following sections, the main results, limitations, and potential future directions of each study are discussed in more detail.

7.2 Leeds anal cancer data warehouse (Chapter 2)

7.2.1 Summary of results

In this study, an institutional data warehouse was created, consisting of comprehensive information on patients with anal cancer treated in LCC, including high quality data for patients treated with VMAT. A comprehensive data dictionary was generated, in order to document all relevant data items and to facilitate the use of the data warehouse in the future. Four LCC databases and clinical systems were identified as the main sources of data, and additional data was sourced from existing research datasets. The amount of automatic data extraction, as well as manual data collection and manual review required to develop the data warehouse was evaluated. Subsequently, robust data quality evaluation was conducted, identifying data items of high quality, and highlighting areas where data quality needs to be improved. A plan was devised on how to update the data warehouse at regular intervals by executing a semi-automatic pipeline. The data warehouse can be accessed by authorised researchers and clinicians at LCC for research and audit.

7.2.2 Limitations

Currently, the data included in the data warehouse cannot be continually updated in a fully automated manner, as some manual data collection and review is required. This represents a major limitation of the work. The current version of the data warehouse also lacks some data items which are considered to be important for anal cancer research, including HPV and HIV status. HPV infection has previously been established as a prognostic factor for worse overall survival and locoregional control [1,2]. However, the HPV status of patients is currently not routinely assessed at LCC and therefore this data is unavailable. The HIV status of patients is routinely assessed, however, accessing the relevant data through automated methods was challenging and will be an area of future development within the warehouse. The source of other important data items, such as baseline comorbidity data and treatment toxicity data has been identified, but the data could not be easily collected due to being stored in free-text form. The automatic extraction of free-text data from their source could not be implemented since NLP algorithms [3] were not available at the time of the data warehouse development. Patient reported outcome data were available for some patients, however, they are not yet routinely collected and therefore were not included in this version.

In addition, the quality of a number of data items is lacking for various reasons that are documented in the data dictionary. The main reasons for low data quality include lack of data provenance, i.e. lack of meta-data specifying the origin of the data and how they were originally recorded; lack of clear ontologies, i.e. lack of standardised data definitions; and finally, lack of interoperability across the different databases and clinical systems from which the data were sourced from [4].

7.2.3 Perspectives and future work

Potential future work on the Leeds anal cancer data warehouse involves devising and implementing a pipeline for continuous updating of the data for existing and new patients at regular intervals, with as little manual data entry as possible. To achieve this, several objectives need to be achieved. Firstly, in collaboration with clinicians, a set of essential data items that can be prospectively collected at the point of diagnosis, at the MDT discussion meeting, and during follow-up need to be identified. These data items should be recorded directly by the clinical care team using a system that will be tailored to their preferences. The prospective data should then be integrated into the data warehouse automatically. Moreover, we aim to update the existing automatic data extraction pipeline, in order to be able to pull data from PPM backend. In order to do this, the necessary access permissions need be obtained, and a new data extraction pipeline needs to be set up. For data items that are only recorded and stored in free-text form in PPM (TNM staging, outcome data, baseline comorbidity data and treatment toxicity data), we aim to implement natural language processing techniques to enable automatic extraction. In order to ensure that the extracted data is accurate and of high quality, a proportion of the data will require manual review. Data items with an unidentified source may be further explored to confirm whether they can be extracted automatically.

Additional future work could include extending the range of data items that are included in the data warehouse. For instance, data on baseline tissue biomarkers can be collected from histopathology samples and incorporated in the data warehouse. As discussed in Chapter 3, only a small number of biomarkers have been identified as prognostic for anal cancer outcomes. Therefore, through the collection and analysis of biomarkers that predict treatment response, valuable insights may be gained. For instance, in patients with treatment-sensitive disease, dose reduction may reduce side-effects, whilst in patients with treatment-resistant disease, additional targeted combination therapies could potentially improve outcomes. Furthermore, a range of complex radiotherapy-specific data can be added to the data warehouse. A pipeline which extracts radiomics features from diagnostic imaging (pre-treatment PET-CT scan, pre-treatment MRI scan, treatment planning CT) and automatically updates the data warehouse can be developed. Other types of radiotherapy-specific data that can be added include 3D dose distributions for all treatment phases [5], radiotherapy structure set data [6] and dose-volume histograms [7]. These data may be analysed in future atomCAT studies, as discussed further in Chapter 5.

Importantly, the work described in this thesis has primarily focused on survival and cancer-related outcomes, as well as on prognostic factors for these. The next step would be to consider treatment-related toxicity and quality of life. To do so, patient-reported outcome data should be incorporated into the data warehouse and into future atomCAT work. As previous studies have highlighted, efforts in collecting and analysing PROs from patients with anal cancer should be maximised, since they are fundamental in understanding the disease from a patient's perspective, which may in turn aid in the improvement of treatment for future patients. A prospective study conducted in two Danish treatment centres has reported that toxicity scores as assessed by patients themselves only weakly agree with the equivalent scores from a clinician [8]. More recently, a number of studies analysed PROs relating to toxicity during and after chemoradiotherapy for anal cancer. One study reported that nine out of ten patients felt that their quality of life had deteriorated since their diagnosis [9]. Additionally, nine out of ten patients did not feel comfortable discussing their diagnosis with friends and family. In terms of psychological outcomes, the majority of patients reported feeling anxious (81%), fearful (78%) and depressed (73%) on a daily basis. Physical outcomes resembled clinician-assessed outcomes more closely, with the majority of patients reporting faecal incontinence (71%), radiation proctitis (68%) and urinary incontinence (65%). The study by Gilbert et al. analysed PROs from 121 patients across 40 UK radiotherapy treatment centres at baseline and at one-year follow-up [10]. The PRO analysis indicates that the majority of side effects significantly improve one year after the end of treatment, including pain, anxiety, appetite loss, blood and mucus in stools, and buttock pain. Similar results were reported in a study carried out in Texas, USA, although with a much smaller patient cohort of 21 patients [11]. The above highlight the importance of collecting and analysing PROs, instead of relying solely on the clinician's assessment, in order to achieve a patient-centred approach to treatment.

7.3 Systematic review of prognostic factors in anal cancer (Chapter 3)

7.3.1 Summary of results

In this systematic review of the literature, 19 published studies which analysed large cohorts of patients treated for anal cancer with conformal radiotherapy and which identified prognostic factors on disease-related outcomes through univariable and multivariable analysis were reviewed. The most prevalent prognostic factors identified were T stage, N stage, sex, pre-treatment biopsy HPV load, as well as the presence of baseline leukocytosis, neutrophilia, and anaemia. A different set of prognostic factors was identified for each of the outcomes explored. No imaging factors were identified as prognostic by more than one of the analysed studies. Moreover, all identified factors were relatively well-known, 'classic' prognostic factors, such as age and disease stage, with limited information about underlying tumour biology or identifying histopathological subgroups. This highlights the need for further prognostic factor research in anal cancer.

7.3.2 Limitations

The main limitations of this systematic review are discussed in detail in Chapter 3, with additional limitations discussed here. Even though the literature search was limited to studies with large cohorts of 100 patients or more, this number was slightly arbitrary and does not guarantee that the results from the included studies are robust. Prior to the initiation of any multivariable modelling study, a prospective sample size calculation should be carried out. This should be ideally conducted using robust statistical techniques [12], instead of the commonly used 10 events per variable rule of thumb [13–15]. Depending on the number of patients that are available for analysis, the appropriate number of factors to be included in the model can be selected. If a larger number of factors are included in the model, the study may be underpowered, resulting in small sample bias [16] and the identification of non-generalisable prognostic factors, which are only prognostic in the analysed cohort. The majority of studies explored in this systematic review did not report whether they had carried out a prospective sample size calculation. Therefore, it was impossible to fully assess the quality of the results conferred from these studies. Lastly, the substantial variation between studies, especially in terms of outcome definitions, staging version used, treatment regimens and methodology employed, renders them less comparable. The effect size of the identified prognostic factors was also not universally reported; therefore, it was not possible to carry out a meta-analysis.

7.3.3 Perspectives

To render the results from prognostic factor research robust and generalisable, large cohorts of patients should be analysed. This systematic review has highlighted the lack of large studies exploring the effect of prognostic factors on survival and disease-related outcomes after conformal radiotherapy for anal cancer in cohorts of more than 100 patients. This was addressed in the atomCAT2 study (Chapter 6), where 1,099 patients treated across multiple international radiotherapy centres were jointly analysed and the prognostic value of multiple factors was confirmed. Furthermore, additional prospectively planned studies, as well as external validation studies are needed to validate the results from prognostic factor studies. These validated factors can then be used for the stratification of patients in risk groups, which may guide the design of future RCTs in anal cancer. This systematic review has also emphasised the lack of large prognostic factor studies evaluating imaging factors and biomarkers. In the future, this could be addressed by the established atomCAT consortium. As discussed in Section 7.5.3, future atomCAT outcome models could include imaging factors and biomarkers, in order to determine their prognostic value in a large international cohort of more than 1,000 patients.

7.4 atomCAT1 proof-of-concept study (Chapter 4)

7.4.1 Summary of results

The atomCAT1 study demonstrated that it is feasible to implement distributed learning in order to analyse data from patients treated for a rare cancer in multiple international centres, and to collaboratively develop outcome prediction models without exchanging individual level patient data between centres. The distributed models demonstrated good performance that was stable between centres and yielded clinically expected factor effects. This study enabled us to expand the collaboration and establish the atomCAT consortium.

7.4.2 Limitations

In atomCAT1, only overall survival was explored as the outcome of interest. This is a limitation, since a number of disease-related outcomes have been identified as clinically important in the literature [17,18], which should be addressed. Additionally, the systematic review (Chapter 3) indicated that the relevant prognostic factors vary between different outcomes, and as a result, these outcomes need to be analysed separately. Thus, in order to consider the individualisation of treatment for future patients, we need to separately identify patients who are at risk of locoregional failure and of distant disease progression. This limitation was addressed in the atomCAT2 study, which not only analysed overall survival, but also locoregional control and freedom from distant metastases.

Unlike for atomCAT2, no prospective sample size calculation was carried out. This was due to the proof-of-concept nature of the study, where the model performance achieved was of secondary importance. In order to develop a robust outcome model with the number of factors included in the global atomCAT1 model (n=5), a minimum sample size of 641 would be required [12,19], which was unrealistic at for a study solely demonstrating the feasibility of the approach.

Even though the factors analysed in the atomCAT1 models were pre-specified based on expected clinical relevance, the correlation between factors was not investigated *a priori*. However, the resulting model performance might have been affected by the inclusion of correlated variables. This is part of a larger issue relating to feature selection in the distributed learning setting, which is discussed in more detail in Section 7.5.2. There is currently no feasible way to test for factor correlation across multiple datasets in a fully distributed fashion.

In terms of the disease staging factor, T3N0 disease was classified as low risk in the main atomCAT1 analysis, according to the AJCC (v8) staging for anal cancer [20]. However, in the PLATO trial [21] T3N0 disease is considered as high risk. Albeit few studies have provided the exact oncological outcome rates according to TNM staging, the findings of RTOG 98-11 [22], ACT I [23] and ACT II [24] trials indicate significantly worse disease-free survival and progression-free survival rates for T3 tumours compared to smaller tumours. Therefore, the choice of classifying T3N0 disease as low risk might have impacted the overall model performance.

Lastly, providing the baseline survival curve is best practise when reporting the results from prediction models, to allow the prediction of survival probability estimates for individual patients. Unfortunately, the implementation of Vantage6 used to carry out the atomCAT1 analysis did not support the calculation of the baseline survival function in a fully distributed manner. Consequently, it was not possible to provide the baseline survival curve without sharing individual-level patient data between centres. This was recognised as a significant limitation. While the atomCAT1 study represents proof-of-principle work, and the resulting model should thus not be considered ready to use for patient outcome prediction, this information will be necessary for larger scale work.

7.4.3 Future work

Future work is described in Chapters 5 and 6 and discussed further in the next section (7.5).

7.5 atomCAT2 study protocol and results (Chapters 5 and 6)

7.5.1 Summary of results

A comprehensive prospective study protocol was developed for the atomCAT2 study, which included robust outcome definitions, data items to be collected and their definitions, a plan on how to handle missing data in different scenarios, and a prospective sample size calculation. Moreover, a robust prospective statistical analysis plan was devised, which specified how the descriptive data analysis would be carried out, which factors would be included in the primary and secondary models for each outcome, how the models would be developed and validated, how model performance should be evaluated and how the results would be reported. The protocol also described the distributed learning architecture to be implemented and what tasks each participating centre had to complete in order to prepare for the analysis phase. During the atomCAT consortium recruitment phase, the protocol was shared with a large number of international anal cancer treatment centres. Through this approach, 11 new centres were recruited, and the collaboration was expanded to a total of 14 centres across the UK and Europe, which formed the international atomCAT consortium (Figure 7-1). From these, a total of 12 centres participated in the atomCAT2 analysis within the timescale allowed for completing my PhD.

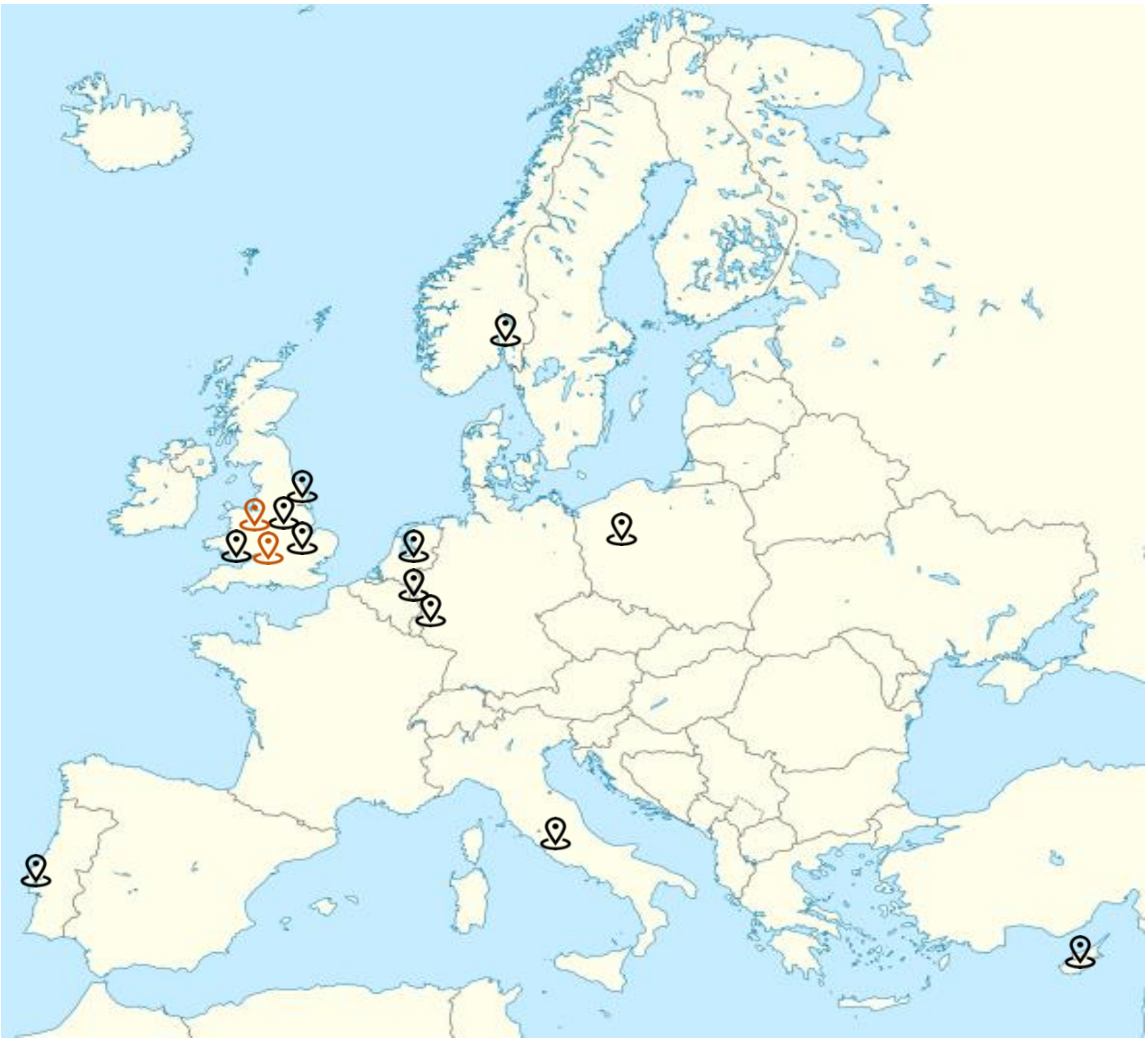


Figure 7-1. Map of Europe showing the location of all centres that are part of the international atomCAT consortium. The two centres marked in red did not participate in the first round of atomCAT2 analysis.

In the atomCAT2 study, models for three clinically important outcomes in anal cancer were developed and validated by analysing a large, multi-centre cohort of 1,099 patients via distributed learning. In this cohort, a different set of factors were identified as prognostic for each of the outcomes explored (nodal involvement, male sex, older age, and larger primary tumour size were found prognostic for poorer overall survival; male sex, higher T stage, and larger primary tumour size for poorer locoregional control; and nodal involvement and larger primary tumour size for poorer freedom from distant metastasis). All three models exhibited satisfactory performance and were not over-optimistic.

7.5.2 Limitations

The most prevalent limitations of the atomCAT2 analysis relate to the current limitations of the distributed learning infrastructure and technology. More specifically, distributed learning in oncology and prognostic research is still an emerging field with lots of knowledge gaps. To this day, only a limited number of algorithms have been adapted for use in a distributed setting. For instance, feature selection is an essential part of outcome model development, which guides the identification of the most appropriate prognostic factors for the outcome of interest [25]. Established feature selection methods are necessary when developing models using high-dimensional data that consists of a large pool of candidate factors. Even though numerous centralised feature selection algorithms exist [26], only few studies have explored feature selection implementation in the distributed learning framework [27]. In atomCAT2, the identification of relevant prognostic factors for the outcomes explored was guided by a systematic review of the literature (Chapter 3) and expert opinion from clinicians. Even though this process was robust, it could be further informed by employing distributed feature selection methodology, as this would allow a larger number of candidate features to be considered for model development.

Moreover, handling missing data when conducting outcome modelling using clinical datasets can be a complex and challenging task [28]. This challenge is particularly prevalent in the context of model development through distributed learning, where sharing of individual-level patient data between centres is not permissible. In cases where data is missing at random (MAR) or missing not at random (MNAR) [29] in individual centres, there is no consensus on which is the most robust approach to handle missing data without negatively affecting the performance of the resulting distributed outcome model. The simple techniques used to handle missing data in atomCAT2, such as complete case analysis and overall mean imputation are rarely optimal options, as they produce biased results, significantly reducing the statistical power of the resulting models [30]. Instead, more complex approaches such as regression-based imputation and multiple imputation are likely to perform better [28]. However, these approaches have not yet been adapted for use in a distributed setting. It is therefore important that future research focuses on exploring different options for data imputation, in order to evaluate whether they can be applied in a distributed setting and identify which have the smallest negative effect on the performance of distributed outcome models.

Currently, missing data imputation has only been sparsely explored in the context of distributed learning and there is only limited precedence to guide best practise [31,32].

Additional aspects of the atomCAT2 analysis were limited by the current implementation of the distributed learning infrastructure. Robust outcome modelling involves calculating of the baseline hazard function. However, this calculation has not yet been implemented in the distributed learning infrastructure, and consequently the atomCAT2 analysis could not cover this. Instead, only baseline outcome rates at specific timepoints of interest were calculated. Moreover, the discrimination of the atomCAT2 models was evaluated using Harrell’s concordance index [33], which is not the ideal performance metric when a time range is of primary interest, since it has been shown to be over-optimistic and sensitive to the study censoring distribution [33]. Other performance metrics, such as Gönen and Heller's unbiased concordance statistic K [34], the Royston-Sauerbrei D statistic [35] and the Brier score [36] might be more appropriate for this analysis. However, these have not yet been implemented and further work is needed to test their convergence and accuracy in a distributed setting. Lastly, for the model calibration phase, the patient cohorts at each participating centre were stratified into two risk groups (low risk, high risk). However, in the centres with smaller cohorts, it is likely that there was insufficient information to create meaningful calibration due to the small number of events for each outcome (deaths for overall survival, locoregional failures for locoregional control and distant metastases for freedom from distant metastasis) in each of the risk groups.

7.5.3 Future work

The work carried out for the atomCAT2 study has the potential to be extended in several directions. Firstly, even though the atomCAT2 analysis is privacy-preserving, an independent trusted third party could be invited to re-create the distributed atomCAT2 models via a centralised approach, in order to test the accuracy of the distributed models and the reported results. This can further confirm that the distributed algorithms and their centralised counterparts converge at the same point.

Furthermore, as the atomCAT consortium has now been fully established, it can be further expanded, with the aim of providing a forum for discussion and collaboration on key research questions in anal cancer, with a focus on improving patient outcomes and quality of life after treatment. Using the infrastructure of existing research groups, including the International Rare Cancers Initiative [37], may be one way to expand the range of the international centres involved. To achieve a more patient-centred approach, patients that have previously been diagnosed or treated for anal cancer, as well as members of the public who have had first-hand experience with the disease, can be engaged to guide the future direction that the atomCAT research will take. Patient and public involvement (PPI) groups can be involved in various stages of future research, including the definition and prioritisation of key research questions from the patient's perspective, the design of future studies, and the dissemination of the findings from future research [38,39]. Through this approach, future atomCAT research will be more relevant to the needs of the patients.

Future atomCAT outcome models may incorporate more complex data, such as biomarkers, imaging, and radiotherapy-specific treatment data. Moreover, these data can be analysed in conjunction with patient-reported outcomes (PROs; see Section 7.2.3), as they can yield essential insights on the disease from the patient's perspective, which may in turn guide the design of improved treatment strategies. Lastly, use of distributed learning methodology in atomCAT2 provides an exemplar for other rare cancers where single-centre datasets are limited and where international, multi-centre analyses are necessary for bringing the field forward.

7.6 Challenges in setting up the atomCAT consortium and the atomCAT2 study

Forming the international atomCAT consortium and setting up the international multi-centre atomCAT2 study presented numerous challenges that are not fully addressed in the study protocol publication (Chapter 5) nor in the atomCAT2 paper (Chapter 6).

The centre recruitment phase involved reaching out to as many centres as possible, via both open and closed recruitment. This phase was time consuming and took more than three months to complete. Even after the official recruitment phase was concluded, more centres were recruited through networking at the IMACC conference [40], where the results of the atomCAT1 study were presented and the atomCAT2 study was promoted. The initial list of interested centres included centres from Canada and Australia, which had a 15-hour difference between them. As a result, finding a suitable time for the consortium meetings was problematic. To overcome this, consortium meetings were arranged at alternating times which would work for the Australian centre and the Canadian centre, respectively. Moreover, all meetings were recorded, and comprehensive meeting notes were taken. These were made available to all centres, including centres that could not join the meetings.

Keeping all interested centres engaged and tracking their progress on the project set up was also particularly challenging. The main tasks to be carried out by all participating centres included: collecting the relevant patient data and adding them to the atomCAT2 template dataset, gaining a local approval for data collection and for use of the data for research (for example from an Institutional Review Board, IRB, or a local Research Ethics Committee, REC), reviewing and signing an infrastructure user agreement and a collaboration agreement, and finally setting up the distributed learning infrastructure on a local computer. Some centres progressed quicker than others on these tasks, and as a result, a comprehensive spreadsheet was created to track each centre's progress. This spreadsheet was regularly updated and was used as a guide for keeping with the project timelines and aided the progression to the next phases of the project.

In terms of the local ethics approvals, a number of UK centres have established governance structures for use of radiotherapy and oncology data for research (e.g. LeedsCAT, see Chapter 2; and the corresponding ukCAT in Manchester [41]), which could approve the use of data for the purposes of this project. Therefore, no further approvals were required for these centres. In Leeds, in addition to the LeedsCAT governance board approval, the project was reviewed and approved by the Caldicott Guardian at LTHT and by The Trust Information Governance Department. For UK centres which did not have this infrastructure in place and could not gain a local approval to access and collect patient data for atomCAT2, a central project application was submitted for review by the Health Research Authority (HRA) and the Research Ethics Committee (REC). Even though no individual level patient data had to leave the originating centre at any time, the study was considered research (as patient data were accessed and used beyond individual patient care) and thus had to undergo full HRA and REC review. In order to achieve this, an Integrated Research Application System (IRAS) form had to be filled in and fully reviewed and approved by the study sponsor (University of Leeds). Due to the novelty of distributed learning, it was difficult to effectively communicate how the study was going to achieve its goals whilst ensuring that no patient data would leave the originating organisation. Overall, this was a lengthy process which took approximately seven months to complete, from the date of providing the sponsor with the study protocol to the date of the first site being approved.

Due to the privacy-preserving nature of the study, no data sharing agreement was needed. However, a collaboration agreement had to be reviewed and signed by all participating centres prior to the analysis. Initially, the collaboration agreement from another distributed learning project led by MAASTRO clinic was used as a template, since it had already undergone many rounds of review by multiple international centres, some of which were also participating in atomCAT2. The first draft of the agreement was circulated to all centres in April 2021, at which point centres were asked to identify the point of contact for their legal teams and to begin the review process. Getting all participating centres to agree to the terms of the agreement was particularly challenging. Multiple rounds of review by all centres and subsequent updating of the agreement according to feedback were carried out in order to generate the final version. One particular issue with the collaboration agreement was the difference between European Union (EU) and United Kingdom (UK) laws and regulations regarding data protection and privacy, and more specifically General Data Protection Regulation (GDPR). The participating centres from the UK asked to remove any clauses related to GDPR, whereas European centres insisted that GDPR should be mentioned in the agreement. This was complicated further when individual European centres reworded the GDPR clause according to their default definitions, which in some cases contradicted the wording from other European centres. Overall, it took more than a year to get the collaboration agreement approved by all centres, and a further five months to collect all signatures. The final version of the agreement was fully signed in November 2022.

Installing and setting up the distributed learning infrastructure in all centres was another challenging aspect of the study. This task involved installing new software and opening a network port in each centre’s IT system. Instructions were provided to all centres detailing how to do this, but technical issues were still present, mainly due to restricted user permissions. Lastly, various technical issues with the distributed Cox regression algorithm [42] and the validation algorithm [43] had to be managed prior to the analysis, primarily related to the Vantage6 version used for atomCAT2. After multiple rounds of debugging by the software developers at the Netherlands Comprehensive Cancer Organisation (IKNL) meetings, the algorithms were ready to be implemented in Vantage6 v2.3.4.

7.7 Practicing open science

Open science can be defined as “a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding” [44]. Even though open science practices have not yet adopted by the entire scientific community, their popularity and necessity has been growing over the past decade [45,46].

Open science principles have been implemented throughout the set up and execution of the atomCAT2 project. Despite not being able to openly share individual-level patient data due to patient confidentiality, other aspects of open science other than “open data” have been engaged. Through the implementation of the distributed learning infrastructure, the aim was to facilitate knowledge sharing between international cancer treatment centres. All software used to carry out atomCAT2 is freely and openly available online and can be accessed by anyone, including the Vantage6 distributed learning infrastructure [47], the distributed Cox regression model [42], and the distributed model validation algorithm [43]. For statistical data analysis, the R language was used [48], which is also open source. The sample size calculation, and all other data analysis plans were documented in a way that facilitates reproducibility of results. Finally, the atomCAT2 study protocol was published in an open access journal [19] and has been pre-registered in Open Science Foundation [49].

7.7 Conclusion

This thesis has addressed many aspects of anal cancer radiotherapy, both on a local and an international scale. A comprehensive local anal cancer data warehouse was developed, which is now accessible by clinicians and researchers at LCC. The literature was systematically reviewed and established prognostic factors for a range of anal cancer outcomes after conformal radiotherapy were identified. The feasibility of implementing distributed learning for outcome modelling in a rare cancer was demonstrated in a multicentre collaboration. The collaboration was subsequently extended, and an international consortium of radiotherapy treatment centres was formed. Through the atomCAT consortium and by applying a distributed learning approach, robust and generalisable outcome prediction models were developed using real-world data from the largest cohort of anal cancer patients treated with conformal radiotherapy that has ever been analysed. Finally, prognostic factors for the three outcomes explored were determined. The results from the atomCAT2 study have the potential to significantly impact the way anal cancer is managed in the future by guiding the design of future RCTs in anal cancer, and ultimately aiding the personalisation of treatment. Further research beyond atomCAT2 may entail the exploration of more complex research questions in anal cancer through the established consortium and via the distributed learning methodology. Specifically, this may involve the analysis of biomarker data, as well as additional radiotherapy-specific data, such as imaging and radiomics data.

Importantly, the work undertaken has demonstrated the potential of distributed learning as a powerful tool for analysing datasets across hospitals and across country borders without compromising patient privacy. This is particularly relevant in the medical field, where patient privacy is paramount, and large-scale collaborations are required to further advance research. Distributed learning enables the access to a large pool of diverse real-world data that can be analysed to yield robust insights. The results obtained from such analyses can be validated and compared across borders and healthcare systems, rendering them generalisable to a greater proportion of the relevant population.

The distributed learning approach has the potential to revolutionise medical research on a global scale in the coming years, not only in the context of rare cancers, where single-centre datasets are limited in size, but in the context of other cancers as well. Advancements in the field of precision medicine have led to the recognition that each tumour is biologically unique, and as a result there is a need for the development of novel personalised treatment approaches. Cancer subtyping based on molecular and genetic tumour characteristics is becoming increasingly important for effective cancer treatment, and consequently an increasing number of cancers are expected to be divided into subtypes in the coming years. Ultimately, this will extend the potential and applicability of the distributed learning approach, since patient cohorts with specific cancer subtypes available in individual centres might no longer be large enough for meaningful analyses. New avenues for distributed analyses will open up as the methodology will become applicable to large number of cohorts that were not previously considered.

In conclusion, by enabling researchers to work together across multiple centres and countries, distributed learning can help overcome the challenges posed by limited resources and data silos. This can, in turn, enable healthcare professionals to make more informed decisions which can lead to improved patient outcomes. It is hoped that the work in this thesis will inspire further collaborations and advancements in medical research, leading to a better understanding of disease outcomes and ultimately significant improvements in patient care across the world.

7.8 References

[1] Balermpas P, Martin D, Wieland U, Rave-Fränk M, Strebhardt K, Rödel C, et al. Human papilloma virus load and PD-1/PD-L1, CD8+ and FOXP3 in anal cancer patients treated with chemoradiotherapy: Rationale for immunotherapy. OncoImmunology 2017;6:e1288331. https://doi.org/10.1080/2162402X.2017.1288331.

[2] Rödel F, Steinhäuser K, Kreis N-N, Friemel A, Martin D, Wieland U, et al. Prognostic impact of RITA expression in patients with anal squamous cell carcinoma treated with chemoradiotherapy. Radiother Oncol 2018;126:214–21. https://doi.org/10.1016/j.radonc.2017.10.028.

[3] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 2022. https://doi.org/10.1007/s11042-022-13428-4.

[4] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[5] Dapper H, Rodríguez I, Münch S, Peeken JC, Borm K, Combs SE, et al. Impact of VMAT-IMRT compared to 3D conformal radiotherapy on anal sphincter dose distribution in neoadjuvant chemoradiation of rectal cancer. Radiat Oncol Lond Engl 2018;13:237. https://doi.org/10.1186/s13014-018-1187-7.

[6] Innolytics. DICOM Standard Browser 2020. https://dicom.innolitics.com/ciods/rt-dose/structure-set (accessed November 11, 2022).

[7] Jiao S xiu, Wang M li, Chen L xin, Liu X. Evaluation of dose-volume histogram prediction for organ-at risk and planning target volume based on machine learning. Sci Rep 2021;11:3117. https://doi.org/10.1038/s41598-021-82749-5.

[8] Kronborg C, Serup-Hansen E, Lefevre A, Wilken EE, Petersen JB, Hansen J, et al. Prospective evaluation of acute toxicity and patient reported outcomes in anal cancer and plan optimization. Radiother Oncol 2018;128:375–9. https://doi.org/10.1016/j.radonc.2018.06.006.

[9] Raymond M, Simonetta M-A. Patient-reported outcomes: The anal cancer patient lived experience. J Clin Oncol 2022;40:2–2. https://doi.org/10.1200/JCO.2022.40.4_suppl.002.

[10] Gilbert A, Drinkwater K, McParland L, Adams R, Glynne-Jones R, Harrison M, et al. UK national cohort of anal cancer treated with intensity-modulated radiotherapy: One-

year oncological and patient-reported outcomes. Eur J Cancer 2020;128:7–16. https://doi.org/10.1016/j.ejca.2019.12.022.

[11] Kouzy R, Abi Jaoude J, Lin D, El Alam MB, Minsky BD, Koay EJ, et al. Patient-Reported GI Outcomes in Patients With Anal Cancer Receiving Modern Chemoradiation. JCO Oncol Pract 2020;16:e1524–31. https://doi.org/10.1200/OP.20.00122.

[12] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96. https://doi.org/10.1002/sim.7992.

[13] Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol 2011;64:993–1000. https://doi.org/10.1016/j.jclinepi.2010.11.012.

[14] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2017;26:796–808. https://doi.org/10.1177/0962280214558972.

[15] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res 2019;28:2455–74. https://doi.org/10.1177/0962280218784726.

[16] Nayak B. Understanding the relevance of sample size calculation. Indian J Ophthalmol 2010;58:469. https://doi.org/10.4103/0301-4738.71673.

[17] Fish R, Sanders C, Ryan N, der Veer SV, Renehan AG, Williamson PR. Systematic review of outcome measures following chemoradiotherapy for the treatment of anal cancer (CORMAC). Colorectal Dis Off J Assoc Coloproctology G B Irel 2018;20:371–82. https://doi.org/10.1111/codi.14103.

[18] Theophanous S, Samuel R, Lilley J, Henry A, Sebag-Montefiore D, Gilbert A, et al. Prognostic factors for patients with anal cancer treated with conformal radiotherapy—a systematic review. BMC Cancer 2022;22:607. https://doi.org/10.1186/s12885-022-09729-4.

[19] Theophanous S, Lønne P-I, Choudhury A, Berbee M, Dekker A, Dennis K, et al. Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study. Diagn Progn Res 2022;6:14. https://doi.org/10.1186/s41512-022-00128-8.

[20] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging: The Eighth Edition AJCC Cancer Staging Manual. CA Cancer J Clin 2017;67:93–9. https://doi.org/10.3322/caac.21388.

[21] ISRCTN registry [Internet]. London: BMC. ISRCTN88455282, PLATO - Personalising anal cancer radiotherapy dose 2016. https://doi.org/10.1186/ISRCTN88455282.

[22] Ajani JA, Winter KA, Gunderson LL, Pedersen J, Benson AB, Thomas C, et al. Intergroup RTOG 98–11: A phase III randomized study of 5-fluorouracil (5-FU), mitomycin, and radiotherapy versus 5-fluorouracil, cisplatin and radiotherapy in

carcinoma of the anal canal. J Clin Oncol 2006;24:4009–4009. https://doi.org/10.1200/jco.2006.24.18_suppl.4009.

[23] UKCCCR Anal Cancer Trial Working Party. Epidermoid anal cancer: results from the UKCCCR randomised trial of radiotherapy alone versus radiotherapy, 5-fluorouracil, and mitomycin. The Lancet 1996;348:1049–54. https://doi.org/10.1016/S0140-6736(96)03409-5.

[24] James RD, Glynne-Jones R, Meadows HM, Cunningham D, Myint AS, Saunders MP, et al. Mitomycin or cisplatin chemoradiation with or without maintenance chemotherapy for treatment of squamous-cell carcinoma of the anus (ACT II): a randomised, phase 3, open-label, 2×2 factorial trial. Lancet Oncol 2013;14:516–24. https://doi.org/10.1016/S1470-2045(13)70086-X.

[25] Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol 2016;71:76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002.

[26] Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinforma 2015;2015:1–13. https://doi.org/10.1155/2015/198363.

[27] He Y, Zhou Y, Feng Y. Distributed Feature Selection for High-dimensional Additive Models 2022. https://doi.org/10.48550/ARXIV.2205.07932.

[28] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393–b2393. https://doi.org/10.1136/bmj.b2393.

[29] Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus 2013;2:222. https://doi.org/10.1186/2193-1801-2-222.

[30] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017;17:162. https://doi.org/10.1186/s12874-017-0442-1.

[31] Brink C, Hansen CR, Field M, Price G, Thwaites D, Sarup N, et al. Distributed learning optimisation of Cox models can leak patient data: Risks and solutions 2022.

[32] Chang C, Deng Y, Jiang X, Long Q. Multiple imputation for analysis of incomplete data in distributed health data networks. Nat Commun 2020;11:5467. https://doi.org/10.1038/s41467-020-19270-2.

[33] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17. https://doi.org/10.1002/sim.4154.

[34] Heller G, Mo Q. Estimating the concordance probability in a survival analysis with a discrete number of risk groups. Lifetime Data Anal 2016;22:263–79. https://doi.org/10.1007/s10985-015-9330-3.

[35] Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23:723–48. https://doi.org/10.1002/sim.1621.

[36] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010;21:128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[37] Keat N, Law K, McConnell A, Seymour M, Welch J, Trimble T, et al. International Rare Cancers Initiative (IRCI). Ecancermedicalscience 2013;7:ed20. https://doi.org/10.3332/ecancer.2013.ed20.

[38] NIHR, Research Design Service South Central. What is Patient and Public Involvement in health and social care research? 2017. https://www.rds-sc.nihr.ac.uk/ppi-information-resources/ (accessed November 7, 2022).

[39] Hoddinott P, Pollock A, O'Cathain A, Boyer I, Taylor J, MacDonald C, et al. How to incorporate patient and public perspectives into the design and conduct of research. F1000Research 2018;7:752. https://doi.org/10.12688/f1000research.15162.1.

[40] IMACC Faculty. The First International Multidisciplinary Anal Cancer Conference 2021. https://events.au.dk/imacc2021/conference (accessed November 7, 2022).

[41] Price G, van Herk M, Faivre-Finn C. Data Mining in Oncology: The ukCAT Project and the Practicalities of Working with Routine Patient Data. Clin Oncol 2017;29:814–7. https://doi.org/10.1016/j.clon.2017.07.011.

[42] IKNL. Distributed Cox regression algorithm 2019.

[43] Maastricht University. Distributed validation (c-index) algorithm 2022.

[44] Ramachandran R, Bugbee K, Murphy K. From Open Data to Open Science. Earth Space Sci 2021;8. https://doi.org/10.1029/2020EA001562.

[45] McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. How open science helps researchers succeed. ELife 2016;5:e16800. https://doi.org/10.7554/eLife.16800.

[46] Burgelman J-C, Pascu C, Szkuta K, Von Schomberg R, Karalopoulos A, Repanas K, et al. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. Front Big Data 2019;2:43. https://doi.org/10.3389/fdata.2019.00043.

[47] Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Gelijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. AMIA Annu Symp Proc 2020.

[48] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2014.

[49] Theophanous, Stelios. atomCAT2 - A multicentre study of overall survival, locoregional control and distant metastasis in anal cancer utilising distributed learning 2021. https://doi.org/10.17605/OSF.IO/J7AUH.

Appendix A – Full list of atomCAT consortium members

Study coordinator

Stelios Theophanous (University of Leeds, UK)

Senior investigators

Ane Appelt, PhD (University of Leeds, UK)

Leonard Wee, PhD (MAASTRO, The Netherlands)

Prof Eirik Malinen, PhD (Oslo University Hospital, Norway)

Co-investigators

Stelios Theophanous (University of Leeds, UK)

Ananya Choudhury (MAASTRO, The Netherlands)

Per-Ivar Lønne, PhD (Oslo University Hospital, Norway)

Alexandra Gilbert, MD PhD (University of Leeds, UK)

Maaike Berbee, MD (MAASTRO, The Netherlands)

Marianne Grønlie Guren, MD PhD (Oslo University Hospital, Norway)

Prof Andre Dekker, PhD (MAASTRO, The Netherlands)

Prof Andrew Scarsbrook, MD (Leeds Teaching Hospitals NHS Trust, UK)

Prof David Sebag-Montefiore, MD (University of Leeds, UK)

Partner investigators

Leeds Teaching Hospitals NHS Trust, UK

Alexandra Gilbert, MD, PhD

MAASTRO, The Netherlands

Maaike Berbee, MD

Oslo University Hospital, Norway

Marianne Grønlie Guren, MD, PhD

Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Rebecca Muirhead, MD

Greater Poland Cancer Centre, Poznań, Poland

Łukasz Raszewski, MD

Bank of Cyprus Oncology Centre, Nicosia, Cyprus

Vassilios Vassiliou MD, PhD

Cardiff University, UK

Emiliano Spezi, PhD

Hull University Teaching Hospitals NHS Trust, Hull, UK

Rajarshi Roy, MD

Fondazione Policlinico Universitario A.Gemelli IRCCS, Università Cattolica S.Cuore, Rome, Italy

Prof Maria Antonietta Gambacorta, MD

The Netherlands Cancer Institute - Antoni van Leeuwenhoek (NKI-AVL), The Netherlands

Baukelien van Triest, MD, PhD

Addenbrookes' Hospital, Cambridge, UK

Rashmi Jadon, MBBS, MRCP, FRCR, MD

The Christie NHS Foundation Trust, Manchester, UK

Rohit Kochhar, MD

RWTH Aachen University Medical Center, Aachen, Germany

Ahmed Allam Mohamed, MBBCh, MSc, MD

Champalimaud Foundation, Lisbon, PT

Oriol Parés, MD

Additional investigators

Leeds Teaching Hospitals NHS Trust, UK

Ann Henry

John Lilley

Greater Poland Cancer Centre, Poznań, Poland

Maciej Trojanowski

Bank of Cyprus Oncology Centre, Nicosia, Cyprus

Elisavet Papageorgiou

Ioannis Stylianou

Antri Demetriou

Loukia Georgiou

Cardiff University, UK

Richard Adams

Velindre University NHS Trust, Cardiff, United Kingdom

Muhammad Amin

Thomas Rackley

Hull University Teaching Hospitals NHS Trust, Hull, UK

Athina Sdrolia

Gisela Lima

Peter Colley

Jenny Marsden

Nilesh Tambe

Fondazione Policlinico Universitario A.Gemelli IRCCS, Università Cattolica S.Cuore, Rome, Italy

Mariachiara Savino

Nikola Dino Capocchiano

Andrea Damiani

Viola de Luca

Stefania Manfrida

Carlotta Masciocchi

Vincenzo Valentini

The Netherlands Cancer Institute - Antoni van Leeuwenhoek (NKI-AVL), The Netherlands

Charlotte Deijen

Rens van Haveren

Tomas Janssen

Addenbrookes' Hospital, Cambridge, UK

Joanna Lau

Adam Loveday

The Christie NHS Foundation Trust, Manchester, UK

Gareth Price

Joseph Mercer

RWTH Aachen University Medical Center, Aachen, Germany

Michael J Eble

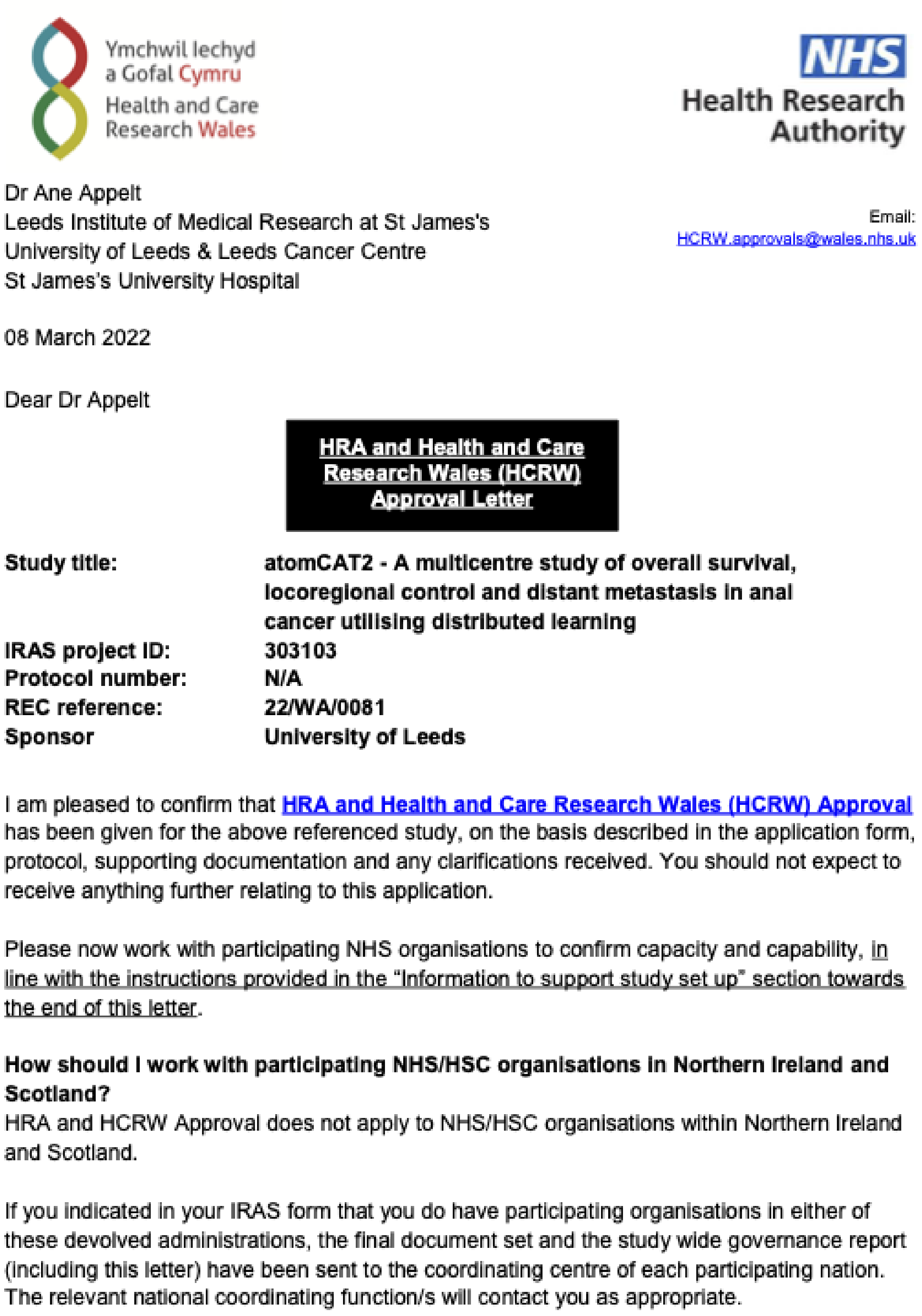
Champalimaud Foundation, Lisbon, PT

Sandra Vieira

Joep Stroom

Patricia Lopes

Appendix B – HRA approval letter for the atomCAT2 study



Ymchwil Iechyd
a Gofal Cymru
Health and Care
Research Wales

NHS
Health Research
Authority

Dr Ane Appelt
Leeds Institute of Medical Research at St James's
University of Leeds & Leeds Cancer Centre
St James's University Hospital

Email:
HCRW.approvals@wales.nhs.uk

08 March 2022

Dear Dr Appelt

HRA and Health and Care Research Wales (HCRW) Approval Letter

Study title:	**atomCAT2 - A multicentre study of overall survival, locoregional control and distant metastasis in anal cancer utilising distributed learning**
IRAS project ID:	**303103**
Protocol number:	**N/A**
REC reference:	**22/WA/0081**
Sponsor	**University of Leeds**

I am pleased to confirm that **HRA and Health and Care Research Wales (HCRW) Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?
HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see IRAS Help for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?
HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to obtain local agreement in accordance with their procedures.

What are my notification responsibilities during the study?

The standard conditions document "*After Ethical Review – guidance for sponsors and investigators*", issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- Registration of research
- Notifying amendments
- Notifying the end of the study

The HRA website also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?
Please do not hesitate to contact me for assistance with this application. My contact details are below.

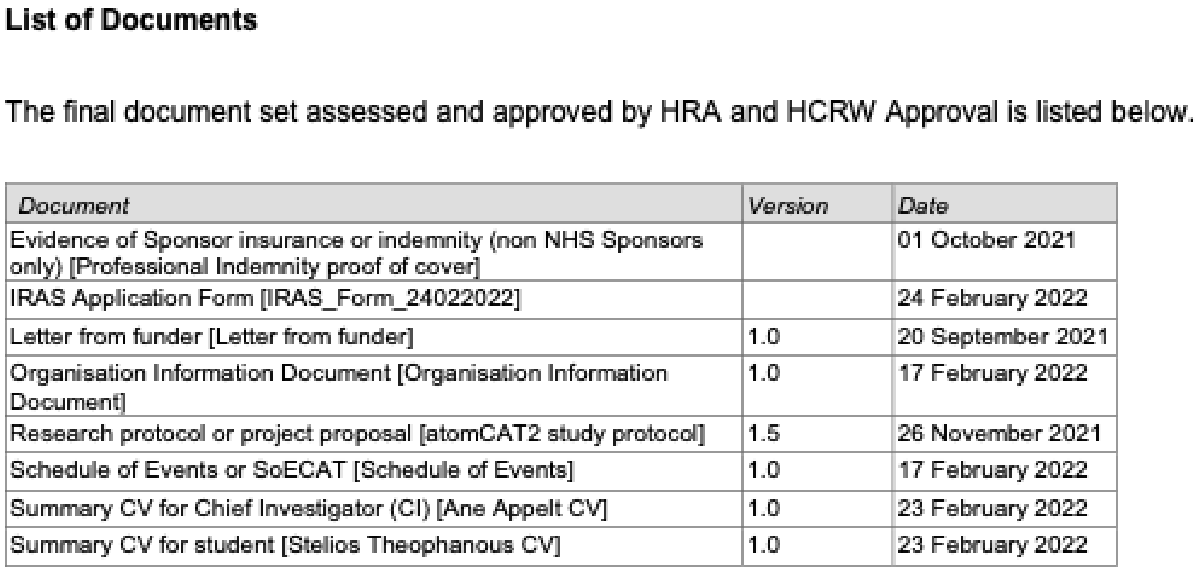
Your IRAS project ID is **303103**. Please quote this on all correspondence.

Yours sincerely,
Sue Byng

Approvals Specialist

Email: HCRW.approvals@wales.nhs.uk

Copy to: *Ms Jean Unlacke*



List of Documents

The final document set assessed and approved by HRA and HCRW Approval is listed below.

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Professional Indemnity proof of cover]		01 October 2021
IRAS Application Form [IRAS_Form_24022022]		24 February 2022
Letter from funder [Letter from funder]	1.0	20 September 2021
Organisation Information Document [Organisation Information Document]	1.0	17 February 2022
Research protocol or project proposal [atomCAT2 study protocol]	1.5	26 November 2021
Schedule of Events or SoECAT [Schedule of Events]	1.0	17 February 2022
Summary CV for Chief Investigator (CI) [Ane Appelt CV]	1.0	23 February 2022
Summary CV for student [Stelios Theophanous CV]	1.0	23 February 2022

IRAS project ID	303103

Information to support study set up

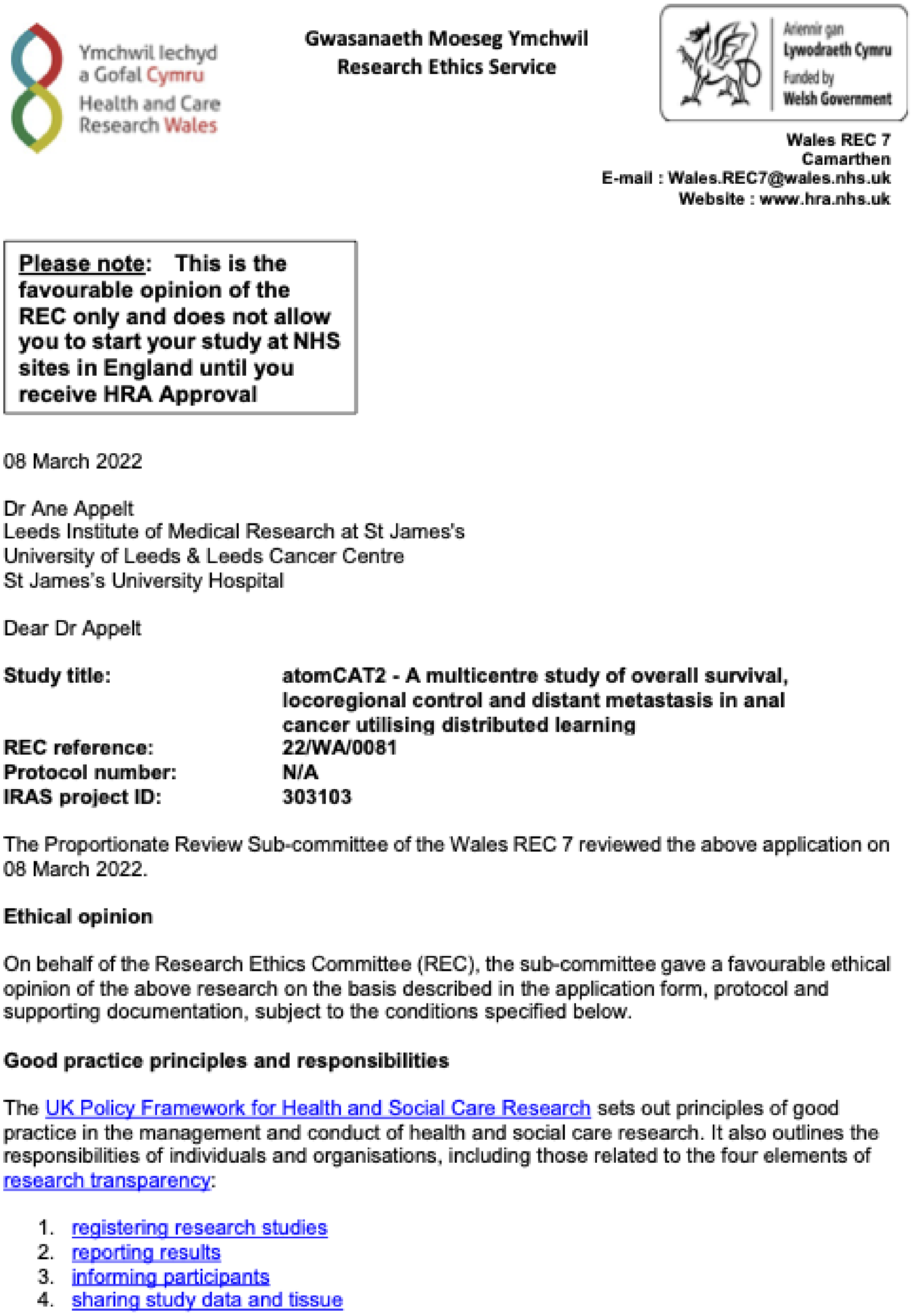
The below provides all parties with information to support the arranging and confirming of capacity and capability with participating NHS organisations in England and Wales. This is intended to be an accurate reflection of the study at the time of issue of this letter.

Types of participating NHS organisation	Expectations related to confirmation of capacity and capability	Agreement to be used	Funding arrangements	Oversight expectations	HR Good Practice Resource Pack expectations
All sites will perform the same research activities therefore there is only one site type	Research activities should not commence at participating NHS organisations in England or Wales prior to their formal confirmation of capacity and capability to deliver the study	An Organisation Information Document has been submitted and the sponsor is not requesting and does not expect any other site agreement to be used.	Please submit evidence of funding from Cancer Research UK. No study funding will be provided to sites as per the Organisation Information Document.	A Principal Investigator should be appointed at study sites.	Use of identifiable patient records held by an NHS organisation to identify potential participants should be undertaken by a member of the direct care team for the patient, so it would not normally be acceptable for this to be done by staff not employed by that organisation.

Other information to aid study set-up and delivery

This details any other information that may be helpful to sponsors and participating NHS organisations in England and Wales in study set-up.
The applicant has indicated they intend to apply for inclusion on the NIHR CRNPortfolio.

Appendix C – REC approval letter for the atomCAT2 study



Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

Gwasanaeth Moeseg Ymchwil
Research Ethics Service

Ariennir gan Lywodraeth Cymru
Funded by Welsh Government

Wales REC 7
Camarthen
E-mail : Wales.REC7@wales.nhs.uk
Website : www.hra.nhs.uk

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

08 March 2022

Dr Ane Appelt
Leeds Institute of Medical Research at St James's
University of Leeds & Leeds Cancer Centre
St James's University Hospital

Dear Dr Appelt

Study title:	**atomCAT2 - A multicentre study of overall survival, locoregional control and distant metastasis in anal cancer utilising distributed learning**
REC reference:	**22/WA/0081**
Protocol number:	**N/A**
IRAS project ID:	**303103**

The Proportionate Review Sub-committee of the Wales REC 7 reviewed the above application on 08 March 2022.

Ethical opinion

On behalf of the Research Ethics Committee (REC), the sub-committee gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Good practice principles and responsibilities

The UK Policy Framework for Health and Social Care Research sets out principles of good practice in the management and conduct of health and social care research. It also outlines the responsibilities of individuals and organisations, including those related to the four elements of research transparency:

1. registering research studies
2. reporting results
3. informing participants
4. sharing study data and tissue

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA and HCRW Approval (England and Wales)/ NHS permission for research is available in the Integrated Research Application System.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

All research should be registered in a publicly accessible database and we expect all researchers, research sponsors and others to meet this fundamental best practice standard.

It is a condition of the REC favourable opinion that **all clinical trials are registered** on a publicly accessible database within six weeks of recruiting the first research participant. For this purpose, 'clinical trials' are defined as:

- clinical trial of an investigational medicinal product
- clinical investigation or other study of a medical device
- combined trial of an investigational medicinal product and an investigational medical device
- other clinical trial to study a novel intervention or randomised clinical trial to compare interventions in clinical practice.

Failure to register a clinical trial is a breach of these approval conditions, unless a deferral has been agreed by the HRA (for more information on registration and requesting a deferral see: Research registration and research project identifiers).

If you have not already included registration details in your IRAS application form you should notify the REC of the registration details as soon as possible.

Publication of Your Research Summary

We will publish your research summary for the above study on the research summaries section of our website, together with your contact details, no earlier than three months from the date of this favourable opinion letter.
Should you wish to provide a substitute contact point, make a request to defer, or require further information, please visit:

https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/

N.B. If your study is related to COVID-19 we will aim to publish your research summary within 3 days rather than three months.

During this public health emergency, it is vital that everyone can promptly identify all relevant research related to COVID-19 that is taking place globally. If you haven't already done so, please register your study on a public registry as soon as possible and provide the REC with the registration detail, which will be posted alongside other information relating to your project. We are also asking sponsors not to request deferral of publication of research summary for any projects relating to COVID-19. In addition, to facilitate finding and extracting studies related to COVID-19 from public databases, please enter the WHO official acronym for the coronavirus disease (COVID-19) in the full title of your study. Approved COVID-19 studies can be found at: https://www.hra.nhs.uk/covid-19-research/approved-covid-19-research/

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

After ethical review: Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study, including early termination of the study
- Final report
- Reporting results

The latest guidance on these topics can be found at https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/.

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion").

Approved documents

The documents reviewed and approved were:

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Professional Indemnity proof of cover]		01 October 2021
IRAS Application Form [IRAS_Form_24022022]		24 February 2022
IRAS Checklist XML [Checklist_02032022]		02 March 2022

Letter from funder [Letter from funder]	1.0	20 September 2021
Research protocol or project proposal [atomCAT2 study protocol]	1.5	26 November 2021
Summary CV for Chief Investigator (CI) [Ane Appelt CV]	1.0	23 February 2022
Summary CV for student [Stelios Theophanous CV]	1.0	23 February 2022

Membership of the Proportionate Review Sub-Committee

The members of the Sub-Committee who took part in the review are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Learning

We are pleased to welcome researchers and research staff to our HRA Learning Events and online learning opportunities– see details at:
https://www.hra.nhs.uk/planning-and-improving-research/learning/

With the Committee's best wishes for the success of this project.

IRAS project ID: **303103**	**Please quote this number on all correspondence**

Yours sincerely

KHorley

PP: Katie Horley
Dr John Buchan
Chair

E-mail: Wales.REC7@wales.nhs.uk

Enclosures: List of names and professions of members who took part in the review

Copy to: Ms Jean Uniacke

Wales REC 7

Attendance at PRS Sub-Committee of the REC meeting on 08 March 2022

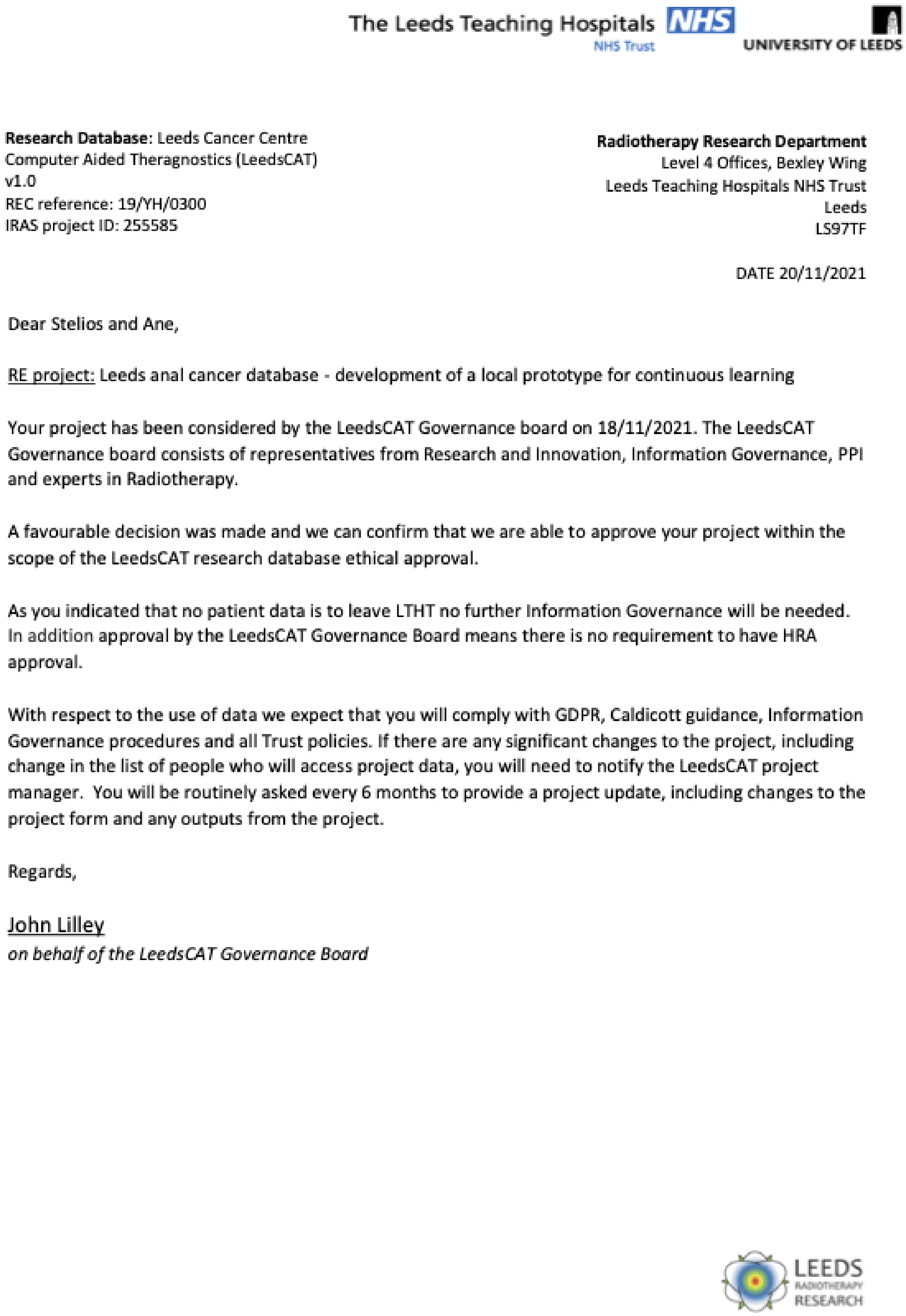
Committee Members:

Name	*Profession*	*Present*	*Notes*
Dr John Buchan	Retired Medical Practitioner	Yes	
Mrs Julie Edwards	Cardiac Physiologist	Yes	
Dr Chaitanya Lanka	Anaesthesiologist	Yes	

Also in attendance:

Name	*Position (or reason for attending)*
Ms Sue Byng	Approvals Specialist
Miss Katie Horley	Approvals Administrator

Appendix D – LeedsCAT approval letter for the data warehouse



The Leeds Teaching Hospitals NHS
NHS Trust

UNIVERSITY OF LEEDS

Research Database: Leeds Cancer Centre
Computer Aided Theragnostics (LeedsCAT)
v1.0
REC reference: 19/YH/0300
IRAS project ID: 255585

Radiotherapy Research Department
Level 4 Offices, Bexley Wing
Leeds Teaching Hospitals NHS Trust
Leeds
LS97TF

DATE 20/11/2021

Dear Stelios and Ane,

RE project: Leeds anal cancer database - development of a local prototype for continuous learning

Your project has been considered by the LeedsCAT Governance board on 18/11/2021. The LeedsCAT Governance board consists of representatives from Research and Innovation, Information Governance, PPI and experts in Radiotherapy.

A favourable decision was made and we can confirm that we are able to approve your project within the scope of the LeedsCAT research database ethical approval.

As you indicated that no patient data is to leave LTHT no further Information Governance will be needed. In addition approval by the LeedsCAT Governance Board means there is no requirement to have HRA approval.

With respect to the use of data we expect that you will comply with GDPR, Caldicott guidance, Information Governance procedures and all Trust policies. If there are any significant changes to the project, including change in the list of people who will access project data, you will need to notify the LeedsCAT project manager. You will be routinely asked every 6 months to provide a project update, including changes to the project form and any outputs from the project.

Regards,

John Lilley
on behalf of the LeedsCAT Governance Board

LEEDS
RADIOTHERAPY
RESEARCH

Appendix E – Data dictionary for the Leeds anal cancer data warehouse

Table 1. Identifiable patient data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Patient_INT_IDENT	IDENTITY		No	Primary/Foreign Key - DO NOT EXPOSE		AutoIncrement Number	N/A	N/A
Patient_StartingDate	datetime		No	Date used for calculating relative time	Patient_FirstTxStartDateTime	MosClinDose_Hst	10	Depends on radiotherapy start date
Patient_NHSNumber	nchar	10	No	Patient NHS number		MosClinIdent	9	Not yet nationally validated
Patient_HospitalNumber	nchar	16	No	Patient Hospital Number		MosClinIdent	10	
Patient_Surname	nvarchar	64	Yes			MosClinPatient	9	Because there could be a delay in updating name changes
Patient_Forename	nvarchar	64	Yes			MosClinPatient	9	Because there could be a delay in updating name changes
Patient_BirthDate	date		Yes			MosClinPatient	10	
Patient_DeathDate	date		Yes			MosClinAdmin	9	Relies on PPM being updated. MANUALLY REVIEWED.
Patient_OriginHospital	nvarchar	64	Yes			Manual extraction	7	Potentially prone to human error. Some missing data
Patient_DiagnosisDate	date		Yes	Date of diagnosis		MosClinMedical	5	Prone to human error and linkage errors. MANUALLY REVIEWED.
Patient_FirstTxStartDateTime	datetime		Yes	Radiotherapy start date		dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and collected
Patient_FirstTxEndDateTime	datetime		Yes	Radiotherapy end date		dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and collected
Patient_PermanentColostomyDate	date		Yes	Date of permanent colostomy procedure		Manual Extraction	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured
Patient_LocoregionalRecurrenceDate	date		Yes	Date of recurrence diagnosis		Manual extraction	6	Prone to human error and subjective interpretation
Patient_DistantMetastasisDate	date		Yes	Date of distant metastasis diagnosis		Manual extraction	5	Prone to human error and subjective interpretation
Patient_3monthClinicalRespAssessDate	date		Yes	Date of 3 month treatment response assessment		Manual extraction	5	Prone to human error and subjective interpretation
Patient_6MonthClinicalRespAssessDate	date		Yes	Date of 6 month treatment response assessment		Manual extraction	5	Prone to human error and subjective interpretation
Patient_PFSdate	date		Yes	Progression free survival at last clinical contact		Date only available if Anon_PFS=1. Patient_PFSdate = (SELECT MIN(Dates) FROM (VALUES (Patient_DeathDate), (Patient_LocoregionalRecurrenceDate), (Patient_DistantMetastasisDate)) AS value(Dates))	5	Prone to human error and subjective interpretation. Calculated using other data items
Patient_FirstStagingScan	date		Yes	Date of staging scan		Manual extraction	N/A	Unidentified source, missing data

Table 2. Demographics, pre-existing comorbidities, diagnostic data, follow-up data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Anon_INT_IDENT	int		No	Primary Key - DO NOT EXPOSE		Autoincrement	N/A	N/A
Anon_MethodOfRelativeTime	nvarchar	50	No	What is the starting date when calculating time passed?	Treatment Start Date	N/A	N/A	N/A
Anon_Historic	bit		Yes	Is the data from previous datasets	0 = No 1 = Yes	N/A	N/A	N/A
Anon_Static	bit		Yes	Is it data locked and static	0 = No 1 = Yes	N/A	N/A	N/A
Anon_Reviewed	bit		Yes	Has this data been reviewed?	0 = No 1 = Yes	N/A	N/A	N/A
Anon_Gender	int		Yes		0 = Male 1 = Female 2 = Other -1 = Not available -2 = Not assessed / Not relevant	MosClinAdmin	8	Gender could be changed or incorrecly recorded
Anon_HIVstatus	int		Yes		0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Unidentified source, missing data
Anon_HPVstatus	int		Yes		0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Not routinely assessed, missing data
Anon_Smoking	int		Yes	Is the patient a smoker?	0 = Never 1 = Previous 2 = Current -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data

Anon_AbSurgeryA	int		Yes	Comorbidities - Surgery - Appendiectomy	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AbSurgeryCS	int		Yes	Comorbidities - Surgery - C-Section	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AbSurgeryH	int		Yes	Comorbidities - Surgery - Hysterectomy	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationAA	int		Yes	Comorbidities - Medication - Ant-acids (e.g. omeprazole & lansoprazole)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationACE	int		Yes	Comorbidities - Medication - ACE inhibitors (e.g. ramipril)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationAD	int		Yes	Comorbidities - Medication - Antidepressants (e.g. citalopram, venlafaxine, prozac/fluoxetine, amitryptilline)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationAP	int		Yes	Comorbidities - Medication - Antiplatelets (e.g. aspirin, clopidogrel)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationI	int		Yes	Comorbidities - Medication - Inhalers	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationNSAID	int		Yes	Comorbidities - Medication - NSAIDs (e.g. Ibuprofen)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationP	int		Yes	Comorbidities - Medication - Paracetamol	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationS	int		Yes	Comorbidities - Medication - Statins (e.g. simvastatin)	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationWT	int		Yes	Comorbidities - Medication - Warfarin/tinzaparin	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MedicationO	int		Yes	Comorbidities - Medication - Other	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_DiabetesTypeOne	int		Yes	Comorbidities - Type.1.Diabetes	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_DiabetesTypeTwo	int		Yes	Comorbidities - Type.2.Diabetes	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data

Anon_DiabetesOther	int		Yes	Comorbidities - Diabetes.Other	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MIAccute	int		Yes	Comorbidities - Acute.MI	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_CCF	int		Yes	Comorbidities - CCF	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Angina	int		Yes	Comorbidities - Angina.CAD	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Hypertension	int		Yes	Comorbidities - Hypertension	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Arrhythmia	int		Yes	Comorbidities - Arrhythmia	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_VenousInsuffieciency	int		Yes	Comorbidities - Venous.Insuffieciency	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_VenousVaricoseVeins	int		Yes	Comorbidities - Varicose.Veins	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_VenousTromboembolic	int		Yes	Comorbidities - Venous.Thromboembolic.Disease	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Cardiomyopathy	int		Yes	Comorbidities - Cardiomyopathy	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Respiratory	int		Yes	Comorbidities - Respiratory.System	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Asthma	int		Yes	Comorbidities - Asthma	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_LungDisease	int		Yes	Comorbidities - Restrictive.Lung.Disease	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_COPD	int		Yes	Comorbidities - COPD	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Maladsorption	int		Yes	Comorbidities - Malabsorption	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data

Anon_IBD	int		Yes	Comorbidities - IBD	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_PeticUlcers	int		Yes	Comorbidities - Peptic.Ulcers	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Pancreatitis	int		Yes	Comorbidities - Pancreatitis	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_LiverDysfunction	int		Yes	Comorbidities - Liver.Dysfunction	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Renal	int		Yes	Comorbidities - Renal	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Paraplegia	int		Yes	Comorbidities - Paraplegia	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Neuromuscuar	int		Yes	Comorbidities - Neuromuscuar	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Parkinsons	int		Yes	Comorbidities - Parkinsons	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Demyelination	int		Yes	Comorbidities - Demyelination	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MND	int		Yes	Comorbidities - MND	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_TIA	int		Yes	Comorbidities - TIA	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Stroke	int		Yes	Comorbidities - Stroke	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Dementia	int		Yes	Comorbidities - Dementia	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Rheumatological	int		Yes	Comorbidities - Rheumatological	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_RA	int		Yes	Comorbidities - RA	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data

Anon_PsoriaticArthritis	int		Yes	Comorbidities - Psoriatic.Arthritis	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Gout	int		Yes	Comorbidities - Gout	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AnkSpond	int		Yes	Comorbidities - Ank.Spond	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Maligancy	int		Yes	Comorbidities - Malignancy	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Obesity	int		Yes	Comorbidities - Obesity	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_Hyperlipidaemia	int		Yes	Comorbidities - hyperlipidaemia	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_PAD	int		Yes	Comorbidities - PAD	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_SpinalCordInjury	int		Yes	Comorbidities - Spinal.Cord.Injury	0 = No 1 = Yes -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuDiarrhoea	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuConstipation	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuFaecalIncont	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuFlatulence	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuAbdomBloat	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuAnalPain	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuNausea	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuAnorexia	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuUrinaryFreq	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuUrinaryInco	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data

Anon_AcuUniraryReten	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuHaematuria	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuProctitis	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuHaemorrhoid	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuVaginalMucos	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuLowerGIMucos	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuOralMucos	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuSkinToxicity	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuGIHaemorrhage	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuNeutropenia	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuAnaemia	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuThrombocyt	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_AcuFatigue	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_PainCramping	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_OtherAcuteTox	int		Yes	Acute toxicity	Scored from 0 to 5 -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_MethodOfAcuTox	int		Yes	Method for determining Acute Toxicity	0 = Prospective 1 = Retrospective -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	N/A	Only available in free-text form, missing data
Anon_ToxicityScoring	nvarchar	255	Yes	Toxicity scoring system used		Manual Extraction	N/A	Only available in free-text form, missing data

Anon_PerfStatus	int		Yes	WHO performance status	0 = Fully active, able to carry on all pre-disease performance without restriction 1 = Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work 2 = Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours 3 = Capable of only limited selfcare, confined to bed or chair more than 50% of waking hours 4 = Completely disabled. Cannot carry on any selfcare. Totally confined to bed or chair 5 = Dead -1 = Not available -2 = Not assessed / Not relevant	ChemoCare	6	ChemoCare schema very complicated. Data extracted may correspond to a different treatment/diagnosis.
Anon_ICD10Label	nchar	100	Yes	ICD10 disease classification		[dbo].[MosClinMedical].[TPG_ID]	8	This is manually interpreted by non clinical staff and is error prone and very generic. Quality will be improved further in the coming months
Anon_DiagnosisSite	nvarchar	100	Yes	Site of diagnosis		[dbo].[MosClinMedical].[TPG_ID]	8	This is manually interpreted by non clinical staff and is error prone and very generic. Quality will be improved further in the coming months
Anon_DaysToFirstTxStart	int		Yes		fn_getTimeBetweenEvents_NoDefault(Year,Patient_StartingDate, Patient_FirstTxStartDateTime)	dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Depends on radiotherapy start date, which is automatically recorded and collected.
Anon_DaysToFirstTxEnd	int		Yes		fn_getTimeBetweenEvents_NoDefault(Year,Patient_StartingDate, Patient_FirstTxEndDateTime)	dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Depends on radiotherapy end date, which is automatically recorded and collected.
Anon_FirstTreatmentIntent	int		Yes	Intent of the first treatment	0 = Pallative 1 = Adjuvant 2 = Radical -1 = Not available	[MosClinPatCPlan].[TX_Intent]	8	Potentially prone to human error during data recording
Anon_AgeAtFirstStart	int		Yes	Age at the starting date	fn_getTimeBetweenEvents_NoDefault(Year,Patient_StartingDate, MosClinPatient.Birth_DtTm)	dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Depends on radiotherapy start date and date of birth. Both are automatically recorded and collected.
Anon_FirstStagingMod	int		Yes	Was staging assessed with a CT scan?	0 = CT 1 = MRI 2 = PET -1 = Not available	Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_FirstPETInjectedDose	int		Yes	Injected dose during PET scan		Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_FirstPETFastingBloodSugar	int		Yes	Did patient fast before PET scan?	0 = No 1 = Yes -1 = Not available	Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_DaysToFirstStagingScan	int		Yes	Number of days between starting date and staging scan	fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_FirstStagingScan)	Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_FirstStagingScanner	nvarchar	100	Yes	Staging scanner ID		Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_FirstStageLabel	nvarchar	255	Yes	Method of assessing TNM staging		Manual Extraction	6	Staging versions changed at specific timepoints. However, patients may have been classified using the previous version
Anon_FirstT	nchar	2	Yes	T staging at diagnosis		Manual Extraction	6	Collected from clinic notes. Can vary from note to note (replans) and prone to human error. Missing data for pre-VMAT patients
Anon_FirstN	nchar	2	Yes	N staging at diagnosis		Manual Extraction	6	Collected from clinic notes. Can vary from note to note (replans) and prone to human error. Missing data for pre-VMAT patients

Anon_FirstM	nchar	2	Yes	M staging at diagnosis		Manual Extraction	6	Collected from clinic notes. Can vary from note to note (replans) and prone to human error. Missing data for pre-VMAT patients
Anon_MetastasisSiteAtDiag	int		Yes	Site of distant metastasis prior to anal cancer treatment	0 = No distant metastasis 1 = Lymph nodes outside pelvis 2 = Viscera or bones 3 = Multiple sites -1 = Not available	Manual Extraction	6	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_FirstHistology	nvarchar	255	Yes	Tumour histology		Automatic Extraction	9	MANUALLY REVIEWED.
Anon_FirstHistologyGrade	int		Yes		0 = Tumour grade was assessed but could not be identified 1 = low/good diff 2 = medium/mod diff 3 = high/poor diff -1 = Not available -2 = Not assessed / Not relevant	Automatic Extraction	9	MANUALLY REVIEWED.
Anon_FirstPrimaryTumourSize	int		Yes	Primary tumour size (Gross Tumor Volume)	If --999 = Not assessed / Not relevant	Manual Extraction	8	Collected from treatment plans. Prone to human error (but not very likely). Only available for VMAT patients.
Anon_FirstAnalMargin	int		Yes		0 = Anal margin 1 = Anal canal -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	2	Unidentified source, missing data
Anon_PermanentColostomy	int		Yes	Has the patient undergone a permanent colostomy procedure?	0 = No 1 = Yes -1 = Not available	Manual Extraction	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured
Anon_ResectionSurgery	int		Yes	Did the patient undergo resection surgery?	0 = No 1 = Yes -1 = Not available	Manual Extraction	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured
Anon_LocoregionalRecurrence	int		Yes	Has the patient been diagnosed with a locoregional recurrence?	0 = No 1 = Yes -1 = Not available	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_LocoregionalRecurrenceSite	nvarchar	100	Yes	Site of recurrence		Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_DaysToLocoregionalRecurrence	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_LocoregionalRecurrenceDate) -99999 = No locoregional recurrence	Automatically calculated	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_MetastasisSiteAfterTreatment	int		Yes	Site of distant metastasis post anal cancer treatment	0 – No distant metastasis 1 – Lymph nodes outside pelvis 2 – Viscera or bones 3 – Multiple sites -1 = Not available	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_DaysToDistantMetastasis	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_DistantMetastasisDate) -99999 = No locoregional recurrence	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_Death	int		Yes	Has the patient died?	0 = No 1 = Yes -1 = Not available	MosClinAdmin	9	Relies on PPM being updated
Anon_DeathDueToAnalCancer	int		Yes	Was the death due to anal cancer?	0 = No 1 = Yes -1 = Not available	Manual Extraction	3	Unidentified source, missing data. Very difficult to confirm whether death was due to anal cancer
Anon_DaysToDeath	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_BirthDate ,Patient_DeathDate) -99999 = Patient is alive	MosClinAdmin	9	Relies on PPM being updated. MANUALLY REVEWIED
Anon_3MonthClinicalResponse	nvarchar	255	Yes	Clinical response at 3 months after end of treatment		Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients

Anon_3MonthImageResponse	int		Yes	Imaging method of assessing clinical response 3 months after end of treatment	0=CT 1=PET 2=MRI 0.1 = CT and PET 0.2 = CT and MRI 1.2 = PET and MRI 3 = PET/CT and MRI -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_DaysTo3monthClinicalRespAssess	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_3monthClinicalRespAssessDate)	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_6MonthClinicalResponse	nvarchar	255	Yes	Clinical response at 6 months after end of treatment		Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_6MonthImageResponse	int		Yes	Imaging method of assessing clinical response 6 months after end of treatment	0=CT 1=PET 2=MRI 0.1 = CT and PET 0.2 = CT and MRI 1.2 = PET and MRI 3 = PET/CT and MRI -1 = Not available -2 = Not assessed / Not relevant	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_DaysTo6MonthClinicalRespAssess	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_6MonthClinicalRespAssessDate)	Manual Extraction	5	Collected from clinic notes. Prone to human error. Missing data for pre-VMAT patients
Anon_TotalAdmissionDaysPerPatient	int		Yes	Total number of admission days for each patient		Automatic calculation	7	Admission(s) could be unrelated to anal cancer. Only includes admissions to LTHT
Anon_Trial	int		Yes	Did the patient participate in a trial?	0 = No 1 = Yes -1 = Not available	[dbo].[MosClinPatTrial]	8	Only radiotherapy-related trials included. MOSAIQ data available 2014 onwards
Anon_PFS	int		Yes	Progression free survival (no failure/mets or nil at death)	If Anon_Death =1 OR if Anon_LocoregionalRecurrence=1 OR if Anon_DistantMetastasis =1 then Anon_PFS=1. Otherwise Anon_PFS=0	Manual Extraction	5	Calculated using other fields.
Anon_DaysToPFS	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_PFSdate) -99999 = Patient has not had disease progression	Manual Extraction	5	Calculated using other fields.

Table 3. Radiotherapy data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Patient_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE		INTERNAL	N/A	N/A
Patient_bPETscanDate	date		Yes	Date of PET Staging Scan		Manual Extraction	N/A	Nuclear medicine data item, missing data
Patient_MRIscanDate	date		Yes	Date of MRI Staging Scan		Manual Extraction	N/A	Nuclear medicine data item, missing data
Patient_TxStartDateTime	datetime		Yes	Radiotherapy start date		dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and extracted
Patient_TxEndDateTime	datetime		Yes	Radiotherapy end date		dbo.fn_getLastTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and extracted

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Anon_DaysTobPETscan	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_bPETscanDate)	Manual Extraction	N/A	Nuclear medicine data item, missing data

Anon_DaysToMRIscan	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_MRIscanDate)	Manual Extraction	N/A	Nuclear medicine data item, missing data
Anon_DaysToTxStart	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_TxStartDateTime)	dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and extracted
Anon_DaysToTxEnd	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_TxEndDateTime)	dbo.fn_getLastTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and extracted
Anon_OverallTreatmentTime	int		Yes		Anon_DaysToTxEnd - Anon_DaysToTxStart	dbo.fn_getLastTreatmentDtTmFromSiteID_JPCB(SIT_ID) minus dbo.fn_getFirstTreatmentDtTmFromSiteID_JPCB(SIT_ID)	10	Automatically recorded and extracted
Anon_CompletedTreatment	int		Yes		0 = No 1 = Yes -1 = Not available	IS [dbo].[MosClinSite][Fractions] = [dbo].[fn_getLastDeliveredFractionFromSiteID_JPCB]	10	Automatically recorded and extracted
Anon_TreatmentIntent	int		Yes	Intent of the first treatment	0 = Pallative 1 = Adjuvant 2 = Radical -1 = Not available	[MosClinPatCPlan].[TX_Intent]	8	Potentially missing and prone to human error during data recording
Anon_RTsiteName	nvarchar	255	Yes	Site of radiotherapy		[dbo].[MosClinSite].[Site_Name]	8	Prone to human error during data recording
Anon_PxEnergy	float		Yes	Energy per fractions prescribed	MosClinSite.dose	([MosClinSite].Dose_Ttl /100)/MosClinSite.fractions	9	Automatically recorded and extracted
Anon_PxFractionNumber	int		Yes	Number of fractions prescribed	MosClinSite.Fractions	MosClinSite.Fractions	7	MANUALLY REVIEWED
Anon_PxPrimaryTumourDose	float		Yes	Planned dose to tumour site		[MosClinSite].Dose_Ttl /100	5	MANUALLY REVIEWED
Anon_PxInvNodesDose	floar		Yes	Planned dose to inv nodes		Manual Extraction	3	Unidentified source. Can be assumed from dose to primary tumour. Missing data
Anon_PxEleNodesDose	float		Yes	Planned dose to ele nodes		Manual Extraction	3	Unidentified source. Can be assumed from dose to primary tumour. Missing data
Anon_TxEnergy	float		Yes	Energy delivered	fn_getEnergyDeliveredFromSiteID_jpcb(MosClinSite.SIT_ID)	fn_getEnergyDeliveredFromSiteID_jpcb(MosClinSite.SIT_ID)	9	Automatically recorded and extracted
Anon_TxFractionNumber	int		Yes	Number of fractions delivered	fn_getLastDeliveredFractionFromSiteID_jpcb(MosClinSite.SIT_ID)	fn_getLastDeliveredFractionFromSiteID_jpcb(MosClinSite.SIT_ID)	7	MANUALLY REVIEWED
Anon_TxPrimaryTumourDose	float		Yes	Delivered dose to tumour site		fn_getDeliveredDoseFromSiteID_jpcb(MosClinSite.SIT_ID)/100	5	MANUALLY REVIEWED
Anon_TxInvNodesDose	float		Yes	Delivered dose to inv nodes		Manual Extraction	3	Unidentified source. Can be assumed from dose to primary tumour. Missing data
Anon_TxEleNodesDose	float		Yes	Delivered dose to ele nodes		Manual Extraction	3	Unidentified source. Can be assumed from dose to primary tumour. Missing data
Anon_RTTechnique	int		Yes	Radiotherapy technique received	0 = 3D-CRT 1 = IMRT 2 = VMAT 3 = Planned Pelvis 4 = Parallel Opposed 5 = EBRT Planned 6 = EBRT Sim Planned -1 = Not available	MosClinSite.[Technique]	8	Prone to human error during data recording
Anon_RTbreakUnplanned	int		Yes	Did the patient have an unplanned radiotherapy break?	0 = No 1 = Yes -1 = Not available	Manual Extraction	3	May be assumed from overall treatment time. Very difficult to confirm break was unplanned. Missing data
Anon_Boost	int		Yes		0 – No boost 1 – Simultaneous boost 2 – Sequential boost -1 = Not available	Manual Extraction	3	Missing data. Can be assumed from clinical protocol. If sequential boost, then patient must have more than 1 prescription

Table 4. Chemotherapy data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Patient_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE		Auto Incrementing	N/A	N/A
Patient_ChemoStartDate	date		Yes	Start date of chemotherapy treatment		[dbo].[PPMLeedsChemoRegimens]	8	Depends on ChemoCare being up to date
Patient_ChemoCycleDate	Datetime		Yes			[PPMLeedsChemoCycles]	8	Depends on PPM being up to date

Field	DataType	Length	NULL?	Description	Values	Source	Quality Score	Justification for Quality Score
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Anon_DaysToChemoStart	int		Yes		fn_getTimeBetweenEvents_NoDefault(Year,Patient_StartingDate,[dbo].[PPMLeedsChemoRegimens][ec_RegimenStartDate])	[dbo].[PPMLeedsChemoRegimens]	8	Depends on the chemotherapy start date
Anon_ChemoRegime	nvarchar	255	Yes			[dbo].[PPMLeedsChemoRegimens][ec_RegimenLabel]	9	MANUALLY REVIEWED
Anon_MissedChemo	int		Yes	Did the patient miss a chemotherapy appointment?	0 = No 1 = Yes -1 = Not available	Manual Extraction ?	1	Very poor quality data: missed chemo for toxicity vs missed chemo error. Mising data
Anon_TotalChemoDose	int		Yes	Chemotherapy dose delivered		Manual Extraction - Maybe from actual ChemoCare tables. Need to read data dictionary	N/A	Unidentified source, missing data
Anon_TotalChemoCycles	int		Yes	Number of chemotherapy cycles delivered		[PPMLeedsChemoCycles] + [dbo].[PPMLeedsChemoRegimens]	8	Depends on ChemoCare being up to date
Anon_CycleComplete	int		Yes	As the cycle completed?	0 = No 1 = Yes -1 = Not available	[dbo].[PPMLeedsChemoRegimens][ec_ActionStatusLabel]	7	Depends on ChemoCare being up to date
Anon_DaysToChemoCycle	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_ChemoCycleDate)	[PPMLeedsChemoCycles]	8	Depends on PPM being up to date

Table 5. Surgery data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality Score
Patient_SurgeryDate	date		Yes	Date of surgery		Manual Extraction or PPM surgery table.	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality score
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Anon_DaysToSurgery	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_SurgeryDate)	Manual Extraction or PPM surgery table.	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured
Anon_SurgeryIntent	int		Yes	Intent of surgical procedure	0 = First Definitive Treatment 1 = Subsequent Treatment 2 = Continuing Treatment (Excluded) -1 = Not available	Manual Extraction or PPM surgery table.	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured
Anon_SurgeryType	nvarchar		Yes	Description of surgery		Manual Extraction or PPM surgery table.	4	Surgery data in PPM is not reliable. Surgeries outside LTHT are not captured

Table 6. Hospital admissions data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality score
Patient_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Patient_AdmissionDate	date		Yes	Date of admission		PPMLeedsAdmissions	8	Admission might not be related to anal cancer. Only admissions to LTHT captured.
Patient_DischargeDate	date		Yes	Date of discharge after admission		PPMLeedsAdmissions	8	Admission might not be related to anal cancer. Only admissions to LTHT captured.

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality score
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Anon_DaysToAdmission	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_AdmissionDate)	PPMLeedsAdmissions	7	Admission might not be related to anal cancer. Only admissions to LTHT captured.
Anon_TotalDaysOfAdmission	int		Yes	Number of days of admission	fn_getTimeBetweenEvents(Year, Patient_Patient_AdmissionDate, Patient_DischargeDate)	PPMLeedsAdmissions	7	Depends on admission and discharge dates
Anon_ReasonsForAdmission	nvarchar		Yes	Reason for admission		PPMLeedsAdmissions	6	Admission might not be related to anal cancer. Only admissions to LTHT captured.

Table 7. Clinical trial data

Field	DataType	Length	NULL?	Description	Values	Data Source	Quality Score	Justification for Quality score
Patient_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Patient_TrialDateColumnName	nvarchar	255	Yes		Patient_TrialDate_[TrialNumber]			
Patient_TrialDate	date		Yes	Date of the trial the patient participated in		MosClinPATTrials.DtTm_Reg	10	Automatically extracted. No information on non-radiotherapy trials
Field	DataType	Length	NULL?	Description	Values	Source	QualityScore	Justification for quality score
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE			N/A	N/A
Anon_TrialColumnName	nvarchar	255	Yes		Anon_Trial_[TrialNumber]			
Anon_TrialName	nvarchar		Yes	Name of the trial the patient participated in	MosClinTrial.Trial_id	MosClinPatTrial.trl_id	10	Automatically extracted. No information on non-radiotherapy trials
Anon_TrialTypeColumnName	nvarchar	255	Yes		Anon_TrialType_[TrialNumber]			
Anon_DaysToTrialColumnName	nvarchar	255	Yes		Anon_DaysToTrial_[TrialNumber]			
Anon_DaysToTrial	int		Yes		fn_getTimeBetweenEvents(Year, Patient_StartingDate, Patient_TrialDate)	MosClinPATTrials.DtTm_Reg	10	Automatically extracted
Anon_OutcomeTrialColumnName	nvarchar	255	Yes		Anon_OutcomeTrial_[TrialNumber]			
Anon_OutcomeTrial	nvarchar		Yes	Outcome of the trial the patient participated in	1 = On Trial, On Treatment 2 = On Trial, On Treatment, Enter Next Phase of Trial 3 = On Trial, In Follow Up 4 = Off Trial, Records Retained	MosClinPatTrials.Trial_sts	9	Automatically extracted. No information on non-radiotherapy trials

Table 8. Link table

Field	DataType	Length	NULL?	Description	Values
Patient_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE	
Anon_INT_IDENT	int		No	Foreign Key - DO NOT EXPOSE	
Anon_ID	nchar	16	Yes	Randomly unique genrated ID	prefixed with ATOM
Anon_StudyID	nchar	4	Yes	Study ID	ATOM