Improving pressure ulcer prevention trial design and analysis using multi-state modelling of existing data

Isabelle Louise Smith

Submitted in accordance with the requirements for the degree

of

Doctor of Philosophy

The University of Leeds

School of Medicine

February 2022

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her 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.

Content from chapters 3, 4 and 5 were published as an output in the journal Statistics in Medicine, with my supervisors as co-authors:

Isabelle L Smith, Jane E Nixon, Linda Sharples. Power and Sample size for multistate model analysis of longitudinal discrete outcomes in disease prevention trials. *Statistics in Medicine*, 38(27) : 5182 − 5196, 2020.

The concept for the work was developed in discussion with the candidates supervisors. The candidate directed the design of the project, conducted all analyses, drafted the original publication and took responsibility for submitting to the journal. The candidate's supervisors provided oversight and discussion in line with usual supervisory arrangements, and they reviewed and contributed to editing of the paper.

Acknowledgements

I would like to express my sincere thanks to my supervisors, Linda Sharples and Jane Nixon for their support and encouragement throughout the whole project. I am grateful to the Clinical Trials Research Unit at the University of Leeds, with particular thanks to Deborah Stocken for her support. Thanks also to the numerous people who participate in trials or who design and conduct trials, and those who allowed the resulting data to be used for this thesis. Many thanks also to Christopher Jackson, Elizabeth McGinnis and Lisette Schoonhoven who provided methodological and clinical guidance on the project direction, and the Pressure Ulcer Research Service User Network who provided a patient perspective throughout. Thanks also to Ardo Van den Hout for sharing his code which is referenced in Chapter 8.

This project is funded by the National Institute for Health Research (NIHR) (DRF-2016-09-085). The views expressed are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care.

I would like to thank Mum and Charlotte for their unconditional love and support. Finally, I dedicate this thesis to Dad, who is missed every day.

Abstract

Pressure ulcer (PU) prevention trials are challenging due to low incidence leading to large sample size requirements. Longitudinal data at multiple skin sites per patient are collected, but commonly aggregated to a single endpoint. Multi-state models (MSM) have potential to improve efficiency of trials but there is little published on MSM as the primary analysis method.

The aim was to understand PU development natural history and better use longitudinal data for PU research design and analysis.

After fitting a 4-state progression MSM to existing trial datasets, a simulation study explored impact on power of using MSM instead of methods based on a single endpoint. This required a hypothesis test definition for multiple effect estimates in the MSM setting. State misclassification was explored using Hidden Markov Models (HMM) applied to trial data, with impact on power and bias of misclassified states assessed through simulations. Candidate state definitions in the presence of missing data were proposed and analysed using a selection model.

MSM led to increased power in some situations. When the intervention was effective in reducing onset and development across all states, follow-up could be halved from 60 to 30 days and assessments reduced from daily to every $2-3$ days compared to the base case. State misclassification, when analysed appropriately, led to little loss of power and unbiased effect estimates, but there were convergence and identifiability concerns. Selection models were shown to be a special case of HMM and can be implemented using readily available software. Descriptive summaries of trial data suggested non-ignorable missing data, however analysis results were insensitive to different state definitions.

For disease prevention trials where participants pass through a series of health states, MSM may lead to efficient trial designs. Missing data is easily accommodated. Further work is required to develop robust modelling strategies for misclassified data.

Contents

Tables

Figures

Abbreviations

AHCPR	Agency for Healthcare Policy and Research
AIC	Akaike's Information Criteria
AMD	Age related macular degeneration
APM	Alternating Pressure Mattress
BOS	Bronchiolitis Obliterans Syndrome
CI	Confidence interval
CRN	Clinical Research Nurse (Gold standard)
CTRU	Clinical Trials Research Unit
EPUAP	European Pressure Ulcer Advisory Panel
FWER	Family Wise Error Rate
HAQ	Health Assessment Questionnaire
HMM	Hidden Markov Model
HR	Hazard ratio
HSF	High Specification Foam
HTA	Health Technology Assessment
IAD	Incontinence Associated Dermatitis
IQR	Inter-Quartile Range
LOCF	Last Observation Carried Forward
LRT	Likelihood Ratio Test
MAMS	Multi-arm Multi-stage
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
MSM	Multi state model
NIHR	National Institute for Health Research
NHS	National Health Service
NPUAP	National Pressure Ulcer Advisory Panel
NRLS	National Reporting and Learning System
PH	Proportional Hazards

PPPIA	Pan Pacific Pressure Injury Alliance
PSA	Psoriatic arthritis
PU	Pressure ulcer
PURSUN	Pressure Ulcer Research Service User Network
OR	Odds ratio
RCT	Randomised controlled trial
SD	Standard deviation
TTE	Time to event
TVN	Tissue Viability Nurse
VPC	Variance Partition Coefficient
WN	Ward Nurse

Notation glossary

Notation	**Description**
N	Total number of subjects
i	Individual $i, i = 1, 2, ..., N$
t	Time since randomisation
W	Number of assessments per subject
K	Number of skin sites, or components
$\boldsymbol{\theta}$	Vector of model parameters
$\boldsymbol{Y} = \{Y_t \mid t \in (0, \infty)\}$	Stochastic disease process, $Y_t \in S = \{1, 2, ..., D\}$
$\boldsymbol{Y^*} = \{Y_t^* \mid t \in (0, \infty)\}$	Observed disease process, , $Y_t^* \in S^* = \{1, 2, ..., D\}$
$S = \{1, 2, ..., D\}$	State space for $\boldsymbol{Y}$
$S^* = \{1, 2, ..., D\}$	State space for $\boldsymbol{Y^*}$
$p_{rs}(t)$	Transition probability $p_{rs}(t) = P(Y_{u+t} = s \mid Y_u = r)$
$\mathbf{P}$	DXD transition probability matrix where $p_{rs}(t)$ is the (r, s) entry
$q_{rs}(t)$	Transition intensity $q_{rs}(t) = \lim_{\delta \to 0} \left\{ \frac{P(Y_{t+\delta}=s \mid Y_t=r)}{\delta} \right\}$, $r, s \in S, r \neq s, t \geq 0$
$\mathbf{Q}$	DXD transition intensity matrix where $q_{rs}(t)$ is the (r, s) entry
e_{rs}	Misclassification probability $P(Y_t^* = s \mid Y_t = r)$
$\mathbf{E}$	DXD misclassification matrix where e_{rs} is the (r, s) entry
$L_i(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x})$	Contribution of individual i to the likelihood function
$\boldsymbol{E}$	$D \times D$ misclassification probability matrix where the (r, s) entry is equal to e_{rs}
Ω_w	Set of possible paths of latent states at times $t_1, ..., t_w$
d_{iw}	number of skin sites observed at time t_w for patient i
$X_k(t_{iw})$	Disease status for skin site k for patient i at the w^{th} timepoint

$\boldsymbol{x}_i$	p-vector of independent variables $(x_{1i}, ..., x_{pi})$ for patient i
$\boldsymbol{\nu}_{ik}$	q-vector of independent variables $(\nu_{1i}, ..., \nu_{pi})$ for skin site k within patient i
$\boldsymbol{\beta}$	p-vector of coefficients corresponding to independent variables $\boldsymbol{x}_i$
$\boldsymbol{\psi}$	q-vector of coefficients corresponding to independent variables $\boldsymbol{\nu}_i$
H_0	Null hypothesis
H_A	Alternative hypothesis
u_i	Patient level random effect
$\mathbf{R}$	Vector of indicator values that take the value 1 if an outcome is observed, and 0 otherwise

Chapter 1

Introduction

The aim of this PhD is to improve the design and analysis of disease prevention clinical trials with discrete longitudinal assessments, informed by re-analysis of two clinical trial datasets of pressure ulcer prevention interventions.

The National Institute for Health Research (NIHR) operational priorities, published in June 2021, state that as part of their work to reduce research waste, their priority is to ensure that the research they fund is well-designed, efficiently delivered, unbiased, published in full, widely disseminated, and usable [2].

The trial forge initiative was set up in 2014 to provide a systematic approach for improving clinical trial efficiency [3]. They identified 17 areas of trials for which efficiencies could be made including; choosing the right research question, choosing the right design, feasibility and pilot work, obtaining funding, logistical planning for trial delivery, data management, writing and publishing the trial protocol, training trial staff, motivating trial staff, identifying trial sites, managing and monitoring trial sites, recruitment, data collection, retention, analysis, dissemination of findings and close down. Some important areas are discussed briefly here before introducing the motivating problem.

First, it is important to optimise processes for recruiting patients from the target population. The Prioritising Recruitment in Randomised Trials (PRioRiTy) study used a priority setting partnership based on the methods of the James Lind Alliance to establish research priorities for improving recruitment to clinical trials [4]. Recruitment is a well-known challenge for clinical trials with approximately

55% of trials funded by the NIHR Health Technology Assessment (HTA) achieving their recruitment target [5]. Feasibility and pilot work may include using a smaller clinical trial to establish the feasibility of running a larger definitive trial, with recruitment being the most common measure of feasibility [6]. The use of a pilot study may be more efficient if conducted within the main trial (internal pilot) rather than before the main trial (external pilot) because centres can remain open and recruitment can continue without interruption if the internal pilot is considered successful [7]. Success is commonly assessed using pre-specified trial progression criteria [8]. Progression based on recruitment could be defined using a stop/go rule or using a traffic light system where green represents no recruitment problems and to continue as planned, red represents infeasible recruitment for the main trial, and amber represents recruitment challenges for which methods to improve recruitment should be explored [7]. A qualitative approach can be taken to understand clinical trial specific recruitment processes and develop a plan of action to address any identified challenges [9]. Such pilot studies can increase the ultimate success of the full trial, or prevent further resources being invested in a trial that is unlikely to provide a definitive answer to the research question.

Second, efficiencies in trial conduct can be made when important trial outcomes are collected routinely as part of clinical practice, a situation that has been explored by McCord *et al* [10]. The authors provided an overview of the potential benefits, challenges, and potential barriers for using routinely collected data in clinical trials. The main conclusion was that routinely collected data have considerable potential to make clinical trials more efficient through streamlining data collection. However, there may be challenges depending on the availability, completeness and accuracy of routine data sources for the specific research question. Careful consideration is therefore required before adopting a routine data source to replace more traditional data collection methods [10]. The Core Outcome Measures in Effectiveness Trials (COMET) Initiative was set up so that researchers of specific conditions can agree a set of standardised outcomes to be collected and reported as a minimum in order for the trial findings to be relevant to key stakeholders, including patients and healthcare decision makers. [11]. Findings from COMET studies may be used to inform relevant routine data sources, which could further lead to efficiency gains for clinical trials. An example of utilising both routine data sources and a core outcome set is the WHITE cohort study of patients who present with a fractured hip [12]. The cohort was designed to collect data on patients whose data are also recorded on the National Hip Fracture Database, complimented by additional data collection according to a core outcome set for hip fractures [13]. Furthermore, trials of interventions for this patient population may be evaluated by running a trial within the WHITE cohort, which provides efficiency gains through utilising an existing infrastructure [12,14,15].

Third, efficiencies can be made through trial designs particularly when there are multiple potential treatments to be assessed for the same patient population, or there are several populations that may benefit from the same treatment. The most common trial design is a parallel group design where the trial is designed based on a fixed number of (typically 2) comparator interventions, data are collected on a pre-planned number of participants and the data are analysed at one fixed time point at the end of the trial [16]. Adaptive trials, including group sequential designs, are an alternative to this fixed trial structure [17]. Pallman *et al* [16] published a guide for adaptive clinical trials, describing their use and providing advice on their conduct and reporting. The guide describes an adaptive design such that there can be reviews of the data and subsequent adaptations to the trial conduct prior to the final analysis. These adaptations may include re-estimating the sample size, dropping intervention arms if there are multiple comparators, or stopping the trial early due to evidence of efficacy or futility. They outline how trials with adaptive designs can be more efficient for many reasons, including that they may require fewer participants, futile treatment groups may be dropped early and a definitive conclusion may be obtained earlier compared to a traditional fixed design. For example, where new treatment options arise during a trial, or a treatment is very unlikely to show a positive effect, a Multi-Arm Multi-Stage (MAMS) trial design can be used. These designs, also described as platform trials, evaluate several interventions against a common control group and have pre-specified adaptation rules to allow dropping of ineffective intervention(s) and the flexibility of adding new intervention(s) during the trial. Alternatively, umbrella trials can be used to assess multiple interventions for a single disease where the patients can be stratified into clinical subgroups, or basket trials can be used to investigate a common intervention for use in multiple disease types [18]. Umbrella, basket and platform trials often have a master protocol with multiple sub studies depending on the research question, which means there may be standardized trial operational structures, patient recruitment and selection, and data collection, which lead to efficiency gains in trial delivery.

Fourth, efficiencies could be made through trial analysis. Trials should be designed so that the number of participants recruited in the trial sample is sufficient to correctly conclude a target treatment difference with an acceptable probability (power), whilst minimising the probability of incorrectly concluding a treatment difference if one does not exist (type I error). Researchers often design a clinical trial based on detecting this difference at a single time point and conduct the analysis accordingly. For example, if the trial was designed to compare differences in a continuous or binary outcome at 3 months post randomisation a generalised linear model may be used. If the data are collected at baseline (randomisation), and longitudinally for a pre-specified length of time a generalised linear mixed model could be used to explicitly model the longitudinal data. Contrasts can be used to estimate the treatment effect at the time point of interest so that the estimand of interest does not change but the use of a generalised linear mixed model may increase the power compared to analysing the outcome at a single timepoint. The use of a generalised linear mixed model has been shown to lead to an increase in power by up to 25% compared to a t-test for continuous outcomes [19] or up to 42.7% compared to Pearson's Chi-Squared test for binary outcomes [20] in the presence of missing data.

Motivating Clinical Problem

This thesis is motivated by trials of pressure ulcer (PU) prevention interventions conducted in the Leeds Clinical Trials Research Unit (CTRU). PUs are a significant problem in populations with impaired mobility and are categorised on a scale of 0 (no changes in skin) to 4 (deep ulceration - see Chapter 2 for further details). Two trials funded by the NIHR HTA used development of a Category 2+ PU as the primary endpoint [21,22]. The first trial published in 2006 was the PRESSURE trial and compared two interventions (alternating overlay and replacement mattresses), with the primary outcome being incidence of PUs. The primary statistical analysis used Pearson's Chi-square test for proportions [21]. The required sample size was $2,100$ participants to have 80% power to detect a 50% reduction in the proportion of patients who developed a new PU, using a 2-sided 5% significance level and anticipated 5% loss to follow-up. The trial under-recruited with an Intention To Treat (ITT) population sample size of $1,971$, where the ITT population consisted all randomised participants in their randomly allocated group, regardless of the treatment they actually received. Despite patients at high risk of PU development being recruited, the proportion of patients developing new Category 2+ PUs was lower than assumed in the original sample size calculation at 10.5% (95% Confidence Interval (CI) $(9.2\%, 11.4\%)$).

The second trial published in 2019 was the PRESSURE2 trial, which was designed to compare two mattresses (Alternating Pressure Mattress (APM) and High Specification Foam (HSF)) in terms of the difference in time to development of a new PU [22]. In order to maximise the incidence of new PUs, and therefore minimise the sample size, the trial team adapted the patient eligibility criteria from the PRESSURE trial by excluding patients who were admitted to hospital for elective surgery, and only including patients who were acutely ill. Based on incidence of PUs in acutely ill patients in the PRESSURE trial, the incidence of new PUs for PRESSURE2 was assumed to be 23% in the control group (HSF). The trial was powered at 90% according to a log-rank test to detect a hazard ratio of 0.759 (assuming an incidence of 18% in the intervention group (APM)). Under a traditional fixed design, the required sample size would have been $2,914$ assuming a 6% loss to follow-up rate. However, informed by the design of an earlier trial of PU prevention interventions [23], an adaptive design was used due to large sample

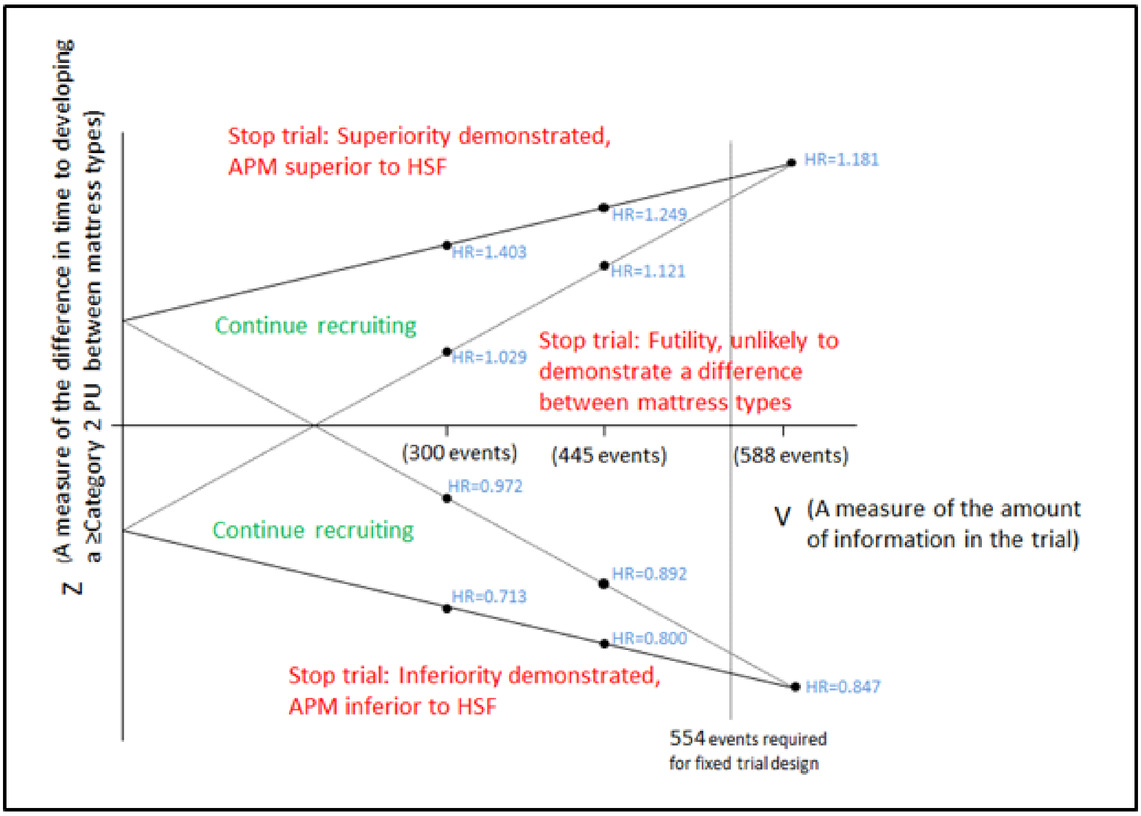

Z (A measure of the difference in time to developing a ≥Category 2 PU between mattress types)
Stop trial: Superiority demonstrated, APM superior to HSF
HR=1.181
HR=1.249
HR=1.403
HR=1.121
Continue recruiting
HR=1.029
Stop trial: Futility, unlikely to demonstrate a difference between mattress types
(300 events)
(445 events)
(588 events)
HR=0.972
V (A measure of the amount of information in the trial)
Continue recruiting
HR=0.892
HR=0.713
HR=0.800
Stop trial: Inferiority demonstrated, APM inferior to HSF
HR=0.847
554 events required for fixed trial design

Figure 1.1: PRESSURE2 trial design, reproduced with permission from Brown *et al* [1]

size requirements and anticipated recruitment challenges. Specifically, a double-triangular group-sequential design [24] was adopted where a maximum of 2 formal interim analyses were planned with stopping boundaries corresponding to safety, futility or efficacy (Figure 1.1). The maximum sample size in a group sequential trial of 2,954 was larger compared to the sample size required for a traditional fixed design but the chance of stopping early appealed to patients, the trial team and the funder.

Unfortunately, recruitment was far slower than originally anticipated (Figure 1.2), and the proportion of participants who developed a PU was also lower than anticipated with an overall incidence at the end of the trial of 7.9% (95% CI (6.7%, 9.1%)). Due to the recruitment rate being lower than expected, the trial had a no cost extension approved and stopped recruitment 6 months after the original planned recruitment end date with a final total of 2,029 participants randomised.

PUs are a key quality indicator for the National Health Service (NHS) [25] and the top 5 priority questions from the James Lind Alliance for PUs are focused on

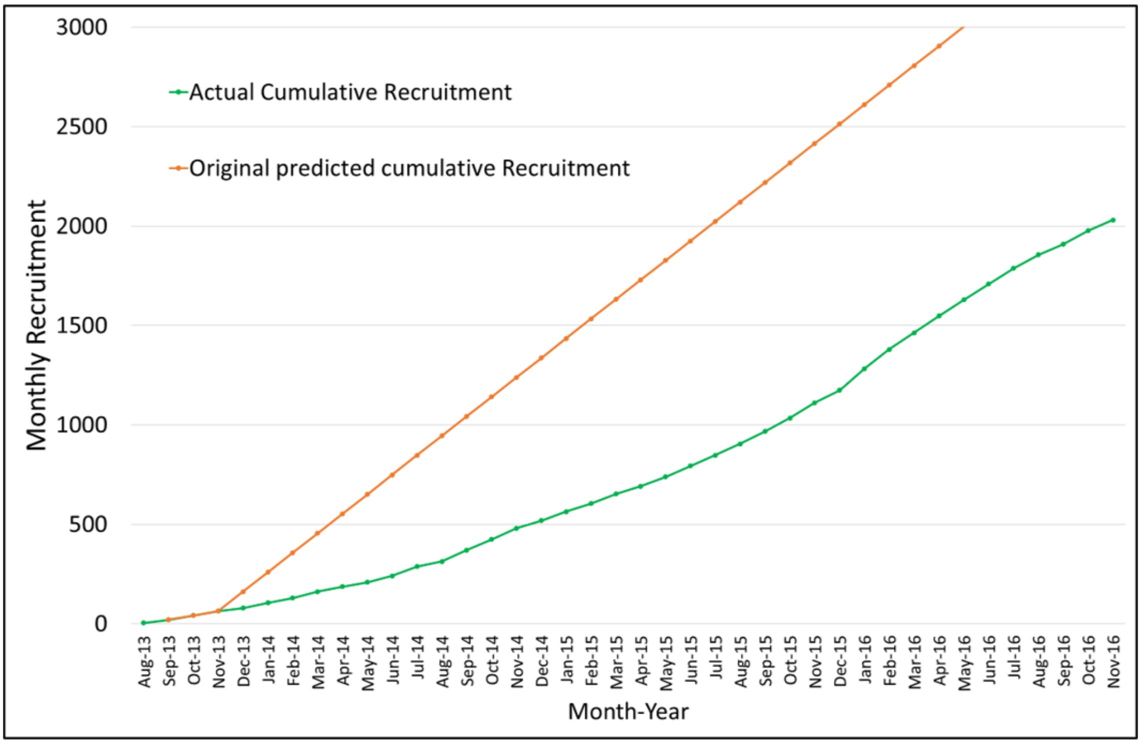


Figure 1.2: PRESSURE2 trial recruitment

PU prevention [26]. Therefore, trials of PU prevention interventions are important, but despite the efforts to improve efficiency via patient recruitment and sequential monitoring of results, the challenges faced by the PRESSURE and PRESSURE2 trials suggests that further improvements in efficiency need to be made. Without solutions, the challenges make trials of PU prevention interventions prohibitive for funders.

Whilst the PRESSURE2 trial was ongoing, the use of routinely collected data was explored. PU data are recorded in multiple sources including the Safety Thermometer [27], and National Reporting and Learning System (NRLS) [28], which are in place to monitor prevalence and incidence of PUs across NHS England Trusts. However, an audit conducted by investigators in Leeds demonstrated that there was a high level of under-reporting of PUs, which was considered unacceptable for research use [29, 30]. Therefore, routinely collected PU data are not an option for making PU prevention trial delivery more efficient.

One area that could be explored further is the planned primary analysis. The PRESSURE and PRESSURE2 trials focused on a single event (or time to event) analysis. However, investigators assessed multiple skin sites at multiple time points, recording whether skin was healthy or not and the classification if there was pressure damage. During analysis repeated assessments of skin sites were combined for each patient to identify whether they developed at least one PU during trial participation, which means that considerable data were not utilised. Greater use of these data could lead to further efficiency gains in the trial design.

Multi-state models

Multi-state models (MSM) represent different disease categories (states) and movement of patients between these disease categories (transitions). They are convenient representations of diseases that can be classified into distinct categories, with clear definitions, and where onset, progression and regression of the disease correspond to transitions between states in the model. MSM have the potential to utilise more of the data collected in disease prevention clinical trials and could lead to increased power by analysing the longitudinal data structure. Note that the treatment effect is defined as the set of hazard ratios estimated for each transition of interest.

In the PU research setting, assessment is based on clinical appearance of the skin, which can lead to potential misclassification of skin status especially if the assessor does not have specialist training [31]. Expert assessors may be more expensive and therefore a trade-off must be made between the frequency of assessments and the accuracy of assessments must be made. In addition to misclassification, missing data are common for trials with outcomes measured longitudinally. It is particularly an issue in the populations at high risk of PU because they are typically elderly and have problems with mobility, which may prevent high risk skin sites being assessed. Additionally, they may have bandages and dressings covering areas that are at risk of PU, or they may have had amputations resulting from their underlying health conditions (e.g. diabetes). Overall, the issues of aggregation, misclassification and missing data can lead to both bias and imprecision of estimated treatment effects and so should be addressed in design and analysis of trials. Thus, this thesis will provide insight into the use of MSM for the design and analysis in disease settings with discrete states measured longitudinally, taking into account the challenges of misclassification and missing data in the PU setting.

1.1 Aim and objectives

The aim of this PhD is to understand the natural history of PU development and to improve the design and analysis of PU research by making better use of all data collected during repeated assessments at multiple skin sites.

Objectives

The objectives of this thesis are to:

1. Conduct a targeted literature review of existing PU prevention trial research to understand current methods of design and analysis.

2. Develop a better understanding of PU onset and development through statistical models that make full use of longitudinal assessment of skin site level data.

3. Assess the impact on power and sample size for disease prevention trials designed using multi-state models compared to commonly used methods of analysis.

4. Assess the impact of misclassified outcomes on power, bias and coverage for a trial designed using multi-state models.

5. Assess the missing data mechanism in the PU setting and apply a selection model to jointly model the disease process and the missing data mechanism.

The aim and objectives are addressed throughout the chapters of this thesis; Chapter 2 describes the motivating problem in more detail and presents a targeted review of the PU literature conducted to understand the common methods and their limitations for the design and analysis of PU prevention clinical trials. The common methods of analysis are applied to two case studies in Chapter 3. This is followed by an introduction to MSM including notation and an application to motivating datasets in Chapter 4. A simulation study to assess the potential impact on power and sample size for disease prevention trials is presented in Chapter 5. Chapter 6 explores how misclassification of outcomes can be incorporated in the analysis of a case study dataset. The impact of misclassification of outcomes in MSM in terms of power, bias and coverage is assessed in Chapter 7. An assessment of missing data mechanisms in a motivating dataset and implementation of a selection model is presented in Chapter 8 and a final discussion including recommendations for practice is provided in Chapter 9. Whilst motivated by clinical trials of PU prevention measures, the findings are relevant to any disease where discrete outcome data are collected longitudinally, the disease status cannot improve and interest lies in prevention of disease progression.

Chapter 2

Literature review

2.1 Introduction

Pressure ulcers (PUs) are defined as *localized injury to the skin and/or underlying tissue usually over a bony prominence, as a result of pressure, or pressure in combination with shear (lateral pressure)* [32]. Skin sites susceptible to pressure injury and ulcer formation are those exposed to pressure and which are not able to tolerate pressure, such as buttocks and heels in patients with very limited activity and mobility. PUs are commonly categorised using the International NPUAP/EPUAP/PPPIA PU Classification System [32]; classification consists of an ordered scale from 'Category 1:Non-blanchable Erythema', 'Category 2:Partial Thickness Skin Loss', 'Category 3:Full Thickness Skin Loss' and 'Category 4:Full Thickness Tissue Loss'. Some PUs are classified as 'Unstageable:Depth Unknown' until enough slough and/or eschar is removed to expose the wound base [32]. Also, rarely, some PUs present as 'Deep Tissue Injury', and the category of PU may not be determined until the epidermis sloughs off [32].

A systematic review of prevalence studies reported point estimates of PU prevalence in 'at risk' populations in the UK to range from 5.1% to 32.1% for hospitals, 4.4% to 6.8% for community settings and 4.6% to 7.5% for nursing homes [33]. In line with these results, a 2013 study suggested 14.8% (95% CI 13.6%−16.0%) of hospital patients (excluding paediatrics, obstetrics and psychiatric care settings) in the UK have a Category 1 or more severe PU [34]. Corresponding prevalence in community settings has been reported as 0.74 (95% CI $0.6 - 0.8$) per $1,000$ adult population according to the results of a wound care survey conducted to assess prevalence of PUs within the population receiving health care in Bradford, UK [35]. Similarly, a cross-sectional observational study of two community NHS sites (including; community nursing services, residential homes, rehabilitation units, specialist palliative care units, nursing homes and General Practitioners) in the North of England reported PU prevalence of 0.58 (95% CI $0.56 - 0.60$) per $1,000$ [36] adult population. PUs represent a significant cost burden to UK healthcare providers with a Category 1 PU estimated to cost £$1,214$ and the most severe PU estimated to cost £$14,108$ with increasing costs due to longer healing times and increased number of complications as PUs increase in severity [37]. Furthermore, PUs cause major problems to affected patients, impacting on physical, social and psychological quality of life domains through increased risk of hospital admission, physical restrictions and lifestyle changes required for the treatment and prevention of PUs. Distressing symptoms include pain, inflammation, exudate and wound odour [38]. PUs are a key quality indicator for the Department of Health [25] and measures, including Safety Thermometer [27], and National Reporting and Learning System (NRLS) [28] are in place to monitor prevalence and incidence of PUs across NHS England Trusts. Despite the scale of the problem there are few high quality randomised controlled trials (RCTs) assessing effectiveness of preventative strategies [39].

Methodological issues

Several methodological issues arise in research on PU prevention, the main statistical ones addressed in this thesis are: inefficiency due to aggregation of longitudinal measurements, misclassification of the true PU category and failure to capture all scheduled measurements (missing data). I identified these issues through conducting the analysis of an RCT [22] and observational cohort study [40]. In research studies (2 RCTs and an observational cohort) led by Leeds CTRU, investigators assessed multiple high-risk pressure-area skin sites at multiple time points for each patient (including sacrum, buttocks, heels), recording whether skin was healthy or not and the PU classification [21,22,40]. The populations varied within each study according to the objective: in the PRESSURE trial, acutely ill or elective surgery patients were recruited in hospital and followed up for up to 60 days [21]; in the PRESSURE2 trial, acutely ill patients were recruited from an inpatient setting and followed for up to 90 days [22] and in the observational cohort study acutely ill patients were recruited from a hospital or community setting and followed for up to 30 days [40]. In each case, longitudinal assessments were conducted twice weekly for the first few weeks after recruitment with a reduction to once weekly until study completion (defined as no longer at high risk of PU development, transferred to non-participating centre, death or the end of the study follow-up period) providing repeated measures for each skin site assessed. For the purposes of analysis, repeated assessments of skin sites were reviewed to identify whether a PU developed at any skin site at any point during follow up; these data were then combined for each patient to identify whether patients developed at least one Category 2 or more severe PU during their study participation. Thus, a large amount of potentially useful data was aggregated to a single binary outcome to indicate whether a patient developed a (Category 2+) PU.

Although recommended assessment times were specified in the protocol, in practice assessments were conducted when it was appropriate to approach the patient and when there was research nurse capacity. Therefore, assessments were not necessarily conducted at the same time points for all patients, and the time interval between assessments were variable. Follow-up visits were not conducted for a number of reasons including the patients being too unwell or, the research nurse being unavailable. Some patients had partial skin assessments at some time points due to bandages or dressings being in situ or being unable to move, for example. Therefore, missing data may be at a patient or skin site level and may not be missing at random. Furthermore, even if assessments were conducted as per protocol, skin changes can occur very quickly [41] and may resolve or deteriorate between assessments, which can mean part of the disease process is not recorded (interval censored).

The current method of skin assessment is based on clinical appearance, which inherently means there is an element of subjectivity in the assessment. This leads to potential misclassification of skin status even when staff are specifically trained and is of particular concern for the assessment of Category 1 pressure injury [31]. A systematic review of risk factors for PU development identified two large studies that reported that, if a patient has a Category 1 PU, the odds of developing a Category 2+ PU were $2-3$ times that of a patient with healthy, intact skin [42] (published odds ratios (95% CI) of 3.1 $(2.4-4.1)$ [43] and 2.0 $(1.3-2.9)$ [21]. Due to misclassification of Category 1 PUs, NIHR Health Technology Assessment (HTA) funded trials of PU prevention strategies run by the Leeds CTRU have used development of a Category 2+ PU as the primary endpoint [21,22]. However, despite patients at high risk of PU development being recruited, the proportion of patients developing new Category 2+ PUs are much lower than for Category 1+ PUs and lead to large sample size requirements [21,22,40]. This issue makes conducting both early and late phase trials problematic as this outcome requires large samples before being confident of taking, say, a phase II trial to phase III or in powering phase III trials. These issues of aggregation, misclassification and missing data can lead to both bias and imprecision of estimated treatment effects and so should be addressed in design and analysis of trials.

2.2 Aim

To establish whether these issues are common for other researchers of PU prevention strategies outside of Leeds, a structured literature review was conducted with the aim of identifying common themes for:

1. Recruitment setting, sample size and patient risk of PU development
2. Skin sites assessed
3. Assessment intervals and maximum length of follow-up
4. Assessor expertise
5. PU classification system and endpoints
6. Analysis methods

2.3 Methods

In order to understand the design and analysis approaches used in published RCTs of PU prevention strategies, a structured review of the existing PU prevention literature was conducted using a pearl growing approach to search for literature [44]. The justification for using a pearl growing approach rather than a "systematic review" is that the aim was to identify common methods of design and analysis rather than, for example, identifying every paper to draw conclusions about particular treatment effects. This method is considered effective in identifying high quality literature in more obscure locations that wouldn't necessarily be identified in a traditional systematic review [45]. This method of review uses initial literature, through which additional references are identified and this process continues until no further relevant references are identified. The initial literature used in this review were literature A and B detailed below, with further literature identified as described under literature C:

- **Literature A: Published Cochrane reviews** The Cochrane library for systematic reviews was searched for the term "PRESSURE ULCERS" and reviews found in this search were screened for relevance to prevention of PUs. The articles identified in each of these reviews were screened and duplicate articles were deleted. This search of Cochrane reviews was expected to include all key trials in pressure ulcer prevention.

- **Literature B: Published systematic review** Based on advice from supervisors, a systematic review of risk factors for PU development published by Coleman *et al* [42] was recommended as a pearl that might identify additional RCTs not included in a published Cochrane review. The articles identified in this review were screened and duplicate articles were deleted.

- **Literature C: Other published literature** In addition to the Cochrane reviews and the systematic review, advice was sought from Professor Jane Nixon (Clinical supervisor), and clinical members of the external advisory group to identify any key trials in PU prevention that may not have been included in the reviews. Additional relevant articles were also identified through review of the articles identified in Literature A and B.

Articles identified through Literature A and B were screened for relevance against the following criteria:

Inclusion criteria

- Adult study populations in any setting
- Randomised controlled trial
- English language
- Full text available

Exclusion criteria

- Duplicate or reviewed as part of another reference
- Did not report PU incidence or skin deterioration
- Conference abstract
- Incomplete study report

Data extraction

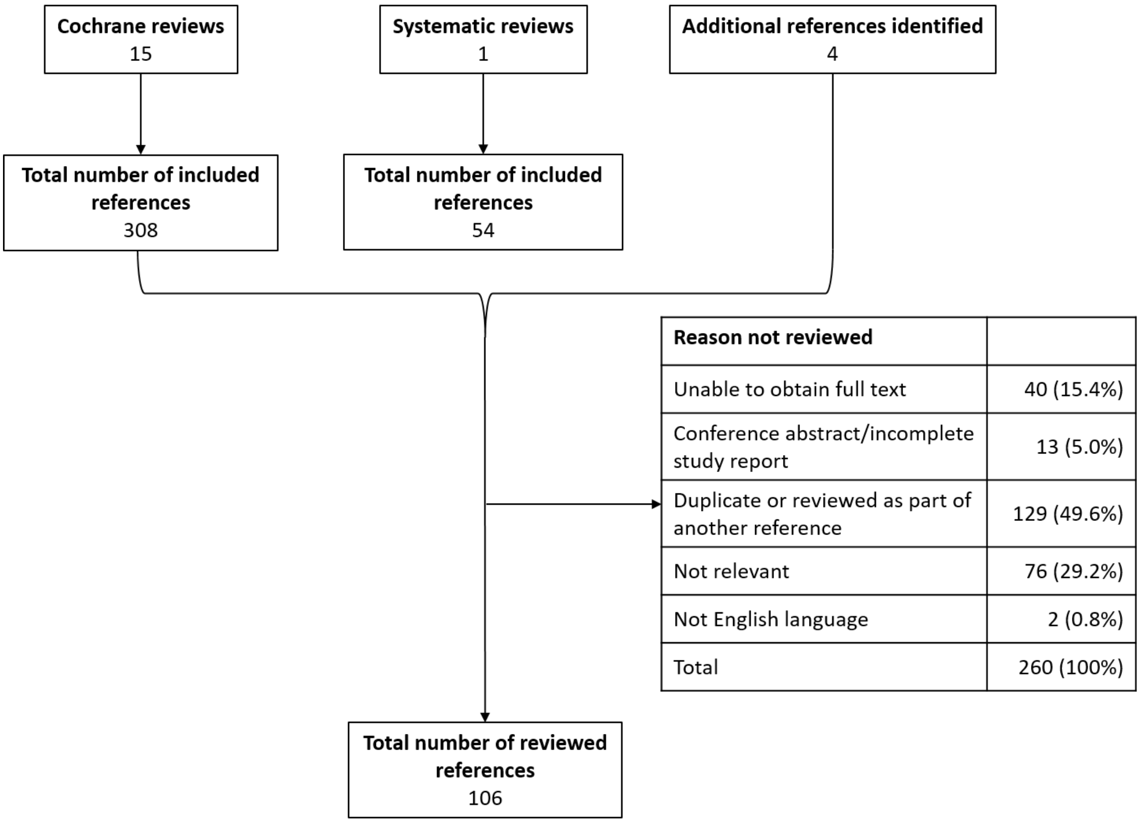
The review of each paper was conducted by Isabelle Smith and the following data items were extracted. No checking was conducted.

- Maximum length of follow-up
- Number of patients recruited
- Recruitment setting
- Key eligibility criteria for patients (particularly whether there were any eligibility criteria relating to skin status)
- Skin sites assessed

- Assessment schedule
- Assessor expertise
- PU classification scale
- Endpoints relating to skin status
- Analysis methods
- Other information deemed relevant

2.4 Results

The first search was conducted 31/10/2016 and identified 7 Cochrane reviews [46–52]. A second search conducted 22/06/2021 led to a total of 16 Cochrane reviews [47, 49, 50, 52–63]. This final 16 included 9 new reviews, 5 of the reviews identified in the first search and 2 updated Cochrane reviews from the initial search [53, 57]. In addition to the references included in the Cochrane reviews, RCTs referenced in the systematic review of risk factors [42] were obtained. A total of 362 references were included in the Cochrane and systematic reviews, and a further 4 references were identified as part of literature search C [64–67]. In total, 260 references were excluded providing a final total of 106 references for the review (Figure 2.1). The main reason for exclusion was duplication (129 (49.6%)) and a third were not relevant to the review if the focus was on healing rather than prevention or if PUs were not reported (76 (29.2%)). Note that the systematic review of risk factors for PU development identified just 2 additional RCTs [68, 69] compared to the Cochrane reviews suggesting that although the search was not exhaustive, there were unlikely to be many additional "key" trials that had been omitted from the review. The full table of reviewed references and data extracted is provided in Appendix A.


Cochrane reviews
15
Systematic reviews
1
Additional references identified
4
Total number of included references
308
Total number of included references
54
Total number of reviewed references
106

Reason not reviewed	
Unable to obtain full text	40 (15.4%)
Conference abstract/incomplete study report	13 (5.0%)
Duplicate or reviewed as part of another reference	129 (49.6%)
Not relevant	76 (29.2%)
Not English language	2 (0.8%)
Total	260 (100%)

Figure 2.1: Summary of literature review paper identification

2.4.1 Recruitment setting, sample size and patient risk of PU development

Of the 106 papers included in the review, 65 (61.3%) were conducted in the acute or hospital setting, 31 (29.2%) were in the community or long term care settings, with 6 (5.7%) conducted in both settings and 4 (3.8%) where the setting was unclear. Patients were commonly eligible if they were at risk of developing a PU with 50% of the references using a risk assessment score to define high risk. The Braden scale [70] was used in 34 (32.1%) cases, and the Norton [71] and Waterlow [72] were used in 10 (9.4%) and 7 (6.6%) references respectively. Pre-existing skin condition was commonly used to assess eligibility with 71 (67.0%) specifying at least one criterion. Of these, 38 (53.5%) studies only included patients who were PU free or had intact skin at baseline, 11 (15.5%) accepted patients with a Category 1 PU or less, and 4 (5.6%) accepted patients with a Category 2 PU or less severe. The remaining 18 (25.4%) studies were more study specific, for example, Bliss [73] recruited patients with Grade 2 or 3 PUs and excluded patients with sores $>$ 5cm or patients with discoloured areas > 2cm.

The number of patients recruited in the studies ranged from 10 to $4,023$, with a median (Inter Quartile Range (IQR)) number recruited of 114 $(62, 380)$ and mean of 337.

2.4.2 Skin sites assessed

The skin sites assessed were commonly not pre-specified according to the study methods, however many did report the locations where PUs were observed. Typically, these included the sacral area, buttocks and heels among others. In some cases, the skin sites were restricted to include specific sites such as the heels for the evaluation of, say, offloading devices or cushions [74]. Note that the absence of pre-specified skin sites in the included studies does not necessarily mean that skin sites were not pre-specified in the protocol.

Assessment intervals and maximum length of follow-up

The frequency of skin assessments varied, with 40 (37.7%) studies specifying at most, daily skin assessments, 18 (17.0%) specified less frequent assessments than daily but more frequent than weekly, and 16 (15.1%) were weekly. There were 9 (8.5%) studies that assessed the skin multiple times per day but these were typically around surgery or for short observation periods in intensive care units. There were also 5 (4.7%) studies for which skin status was assessed less frequently than monthly but these were for studies in long term care settings. In 2 cases, the frequency of assessments were directed by the patient's condition where the frequency might increase for patients with deteriorating health, increasingly limited mobility or changes in skin. [75, 76].

The minimum length of follow-up was 1 day, for example for studies assessing PU prevention during surgery [23, 77], and the maximum length of follow-up was 2 years for a trial where a complex intervention was delivered at a centre level in the community setting and included, for example, feedback to those providing direct care on how to prevent skin breakdown [78]. Of those where data were available, the median (IQR) length of follow-up was 28 $(14, 90)$ days.

It is clear from the literature, that there is no consensus on the length of follow-up or the assessment intervals, but a trade-off is required between observing timely changes in the disease process and the cost of employing specialist researchers to assess patients.

2.4.3 Assessor expertise

A total of 57 (53.8%) studies reported the use of trained staff or researchers, 22 (20.8%) utilised attending health care professionals with no specialist training, whilst 33 (31.1%) did not specify who conducted the trial outcome assessments. There is a common concern within the literature that assessors might lack the experience or expertise required for accurate PU classification within a research context. For example, Beeckman *et al* developed the PUCLAS tool to standardise the training offered to healthcare professionals in the identification and classification of PUs in their RCT [79]. A total of 64 (60.4%) studies incorporated some form of verification of PU classification using additional assessors to check the skin status, or assessed the inter-rater reliability.

2.4.4 PU classification and skin site endpoints

The most common PU classification scale could be categorised as the National Pressure Ulcer Advisory Panel (NPUAP), European Pressure Ulcer Advisory Panel (EPUAP) or Pan Pacific Pressure Injury Alliance (PPPIA) guidelines, which were consolidated to the joint NPUAP/EPUAP/PPPIA guidelines in 2009 and were updated in 2014 [32]. A total of 50 (47.2%) studies specified a variation of this classification, with the Torrance [80], Exton-Smith [81], Shea [82], and Agency for Healthcare Policy and Research (AHCPR) [83] being used in more historical studies. Whilst there are differences in terminology, for example describing PU severity in terms of “Category”, “Stage” or “Grade”, the common PU classification scales can be mapped onto the NPUAP/EPUAP/PPPIA classification scale. The PU classification scale was unclear or not specified in 24 (22.6%) studies, and 18 (17.0%) studies used a bespoke method of classifying pressure damage.

In 54 (50.9%) studies, the equivalent of a Category 1 PU was an endpoint of interest (primary or secondary), whilst in 26 (24.5%) studies the equivalent of a Category 2 PU was an endpoint of interest. There were 16 (15.1%) instances where the definition was bespoke, such as the deterioration of skin status [23, 65, 73, 76, 84–86]. These were not always clearly defined in the methods, but there were some examples where skin deterioration was defined according to a change in reference to baseline skin status, for example Kathirvel *et al* defined a PU event as a patient who moved from PU free to the equivalent of a Category 1+ PU or from Category 1 PU to Category 2+ PU [65]. There were 19 (17.9%) studies where the PU grade of interest could not be ascertained.

Throughout the literature there were concerns about the accuracy of assessing a Category 1 PU, with some studies using verification of the endpoint by another assessor, including some that used a transparent disc method to assess blanching [87–89] and there was one example where a Category 1 PU was confirmed if observed two days in a row [90]. Some studies use the equivalent of a Category 2+ PU as the primary endpoint due to concerns about the reliability of diagnosis with two studies explicitly stating that they did not analyse the incidence of Category 1 PUs due to these concerns [91, 92]. A further issue with using a Category 1 PU as an endpoint, was that some studies [23, 66, 67, 93, 94] excluded patients with darkly pigmented skin due to the challenges in identifying early pressure damage. However, exclusion of patients based on the colour of their skin will lead to ungeneralisable trial results and exclusion of already under-represented groups from research [95, 96].

2.4.5 Analysis methods

In statistical analysis, PU incidence at any skin site (binary response) or the time to incidence of a PU at any skin site were the most frequently used outcomes of interest. In 64 (60.4%) studies identified in the literature, the incidence of PUs were compared using univariate analysis techniques such as Fisher's exact test and tests for contingency tables for example Pearson's Chi-Squared test for proportions [97]. These methods can be generalised to the situation in which the response has m

levels, for example PU categories or trials with more than 2 arms, and related tests can also accommodate the ordinal nature of PU classification [97]. Such simple comparisons are limited as they do not provide an estimate of the treatment effect, are not adjusted for important clinical factors and ignore the timing of PU onset. When applied at a skin site level, rather than for a patient level summary, they did not account for within patient correlation. That is, the effect of intervention was assessed for individual skin sites [79,87,98,99]. There was one exception with a study that compared a dressing applied to one trochanter to no dressing on the other trochanter within the same patient; generalised estimating equations were used to account for the within patient correlation to estimate the relative risk [100].

A commonly used model-based analysis for the incidence of a PU as a binary outcome was the logistic model, which was used in 23 (21.7%) studies in the literature review. This is a generalised linear model that conditions on treatment and other independent variables. Although this approach provides an estimate of the effect size point estimate and associated precision, and allows adjustment for other independent variables, it ignores how treatments influence the timing of new PU onset, the total time at risk for an individual and the sampling times for assessment.

Time to event (TTE) methods were identified in the literature to take time at risk and timing of PU onset into account. A total of 24 (22.6%) studies described outcomes using Kaplan-Meier (product-limit) methods. The log-rank test was used by 17 (16.0%) studies to formally compare the survivor functions of two treatment groups. The log-rank test can be generalised to compare more than 2 survival functions under the null hypothesis that they are all equal. Limitations of these univariate methods are that they assume that the endpoint is observed at the time it occurs, so they ignore interval censoring, do not provide an estimate of the treatment effect and are not adjusted for independent variables.

The most common model-based analysis used for time to event outcomes in the PU literature was the Cox proportional hazards (PH) model [101] used by 14 (13.2%) studies. This multivariate method can provide an estimate of the effect size and its precision, and allows adjustment for other independent variables. As

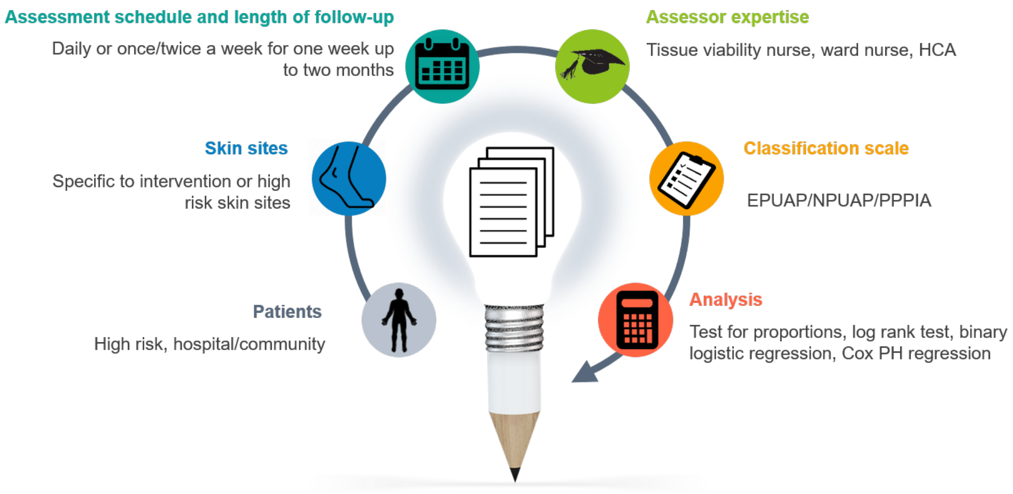

Assessment schedule and length of follow-up
Daily or once/twice a week for one week up to two months
Assessor expertise
Tissue viability nurse, ward nurse, HCA
Skin sites
Specific to intervention or high risk skin sites
Classification scale
EPUAP/NPUAP/PPPIA
Patients
High risk, hospital/community
Analysis
Test for proportions, log rank test, binary logistic regression, Cox PH regression

Figure 2.2: Summary of findings from literature review of PU prevention intervention RCTs

with the binary outcome, TTE analyses are limited since there is no information on the stage at which treatments influence the onset of a new PU. Furthermore, the sampling scheme is ignored, the censoring mechanism is often assumed to be ignorable and PH regression models may be used without checking the assumption of PH.

2.5 Discussion

The literature review of 106 RCTs of PU prevention interventions assessed the reporting of recruitment setting, patient eligibility, skin site assessments, assessor expertise, skin status endpoints and analysis methods (Figure 2.2). The recruitment setting and patient eligibility were generally well reported, and whilst multiple classification scales were reported, the EPUAP/NPUAP/PPPIA guidelines are widely used as the international classification [32].

The endpoints and corresponding analysis methods described in the literature review were almost invariably applied to patients as the unit of analysis. This means that longitudinal data for each skin site collected from each patient were aggregated to provide a patient level outcome of PU onset. Potentially important information therefore was not taken into account in the analysis; for example intermediate changes in skin condition between baseline and PU development, and the number and location of new PUs subsequent to the first one for each patient are often ignored. This aggregated analysis is not only inefficient but also precludes assessment of the correlation between related skin sites within a patient. Aggregating repeated measures data for each skin site to provide a patient level outcome of PU development therefore means that potentially important information fails to be taken into account in the analysis. Multiple PU classifications were reported as endpoints of interest, including the incidence (or time to occurrence) of a Category 1 PU, Category 2 PU or some form of deterioration of skin status. Therefore, the transitions through skin states (Healthy, Categories $1 - 4$) are clinically important. The hazard rate for each transition may be related to skin site as some may have a higher propensity to develop clinical symptoms for pressure damage than others.

The literature review identified various assessment frequencies, such as once daily, multiple times weekly or once weekly. The frequency of assessments should be optimised to minimise the data collection burden for both patients and research staff whilst providing high quality data for PU research. Despite pre-specified assessment frequencies it is unlikely that assessments were conducted at the same time point for all participants in a particular trial; time intervals between assessments may be variable for patient-related factors (e.g. more or less frequent assessments due to patient condition) or missed assessments due to logistical issues unrelated to the patient. Data of this type whereby only snapshots of the process are obtained are called panel data [102]. Panel data are common in observational studies or routine data sources with less structured follow up regimes, but the focus of this thesis is on panel data in RCTs.

A further issue was that the subjective nature of PU classification means that potential misclassification should be considered when analysing and interpreting results. It was clear from the literature that misclassification was of concern to PU researchers with many studies conducting inter-rater reliability studies and quality control checks for skin assessments (Section 2.4.3). However, these inter-rater reliability studies often served as a discussion point or reassurance in the accuracy or lack of bias in the endpoint assessments rather than being incorporated into the analysis. More than half of the studies reported the use of trained staff or researchers, which is in line with recommendations that specialist staff should be used for PU assessment due to expected levels of misclassification when non-specialist ward staff collect data [31, 103]. The PUCLAS tool [79] was developed to standardise the training offered to healthcare professionals in the identification and classification of PUs. However, the assessment of PUs remains subjective and misclassification will therefore continue to be an issue in trials of PU prevention interventions. Failure to accommodate this additional measurement error in the analysis may lead to less precise or biased treatment effect estimates and/or reduced precision in the estimates [104]. Although inter-rater reliability studies were frequently conducted, the sensitivity of analysis results to the additional misclassification when using non-specialist staff was rarely considered.

There were concerns that high levels of misclassification occur in the assessment of Category 1 PUs, however these skin changes are more common and have been shown to have a prognostic relationship with the development of Category ≥ 2 PUs [42,105]. Reliance on development of Category ≥ 2 PUs as an endpoint means that PU prevention trials face a challenge in terms of delivery due to low incidence requiring large sample sizes to detect a treatment difference.

Overall, the methods highlight limitations which could be addressed at least in part by using longitudinal data. There is a need to establish recommendations for the length of assessment intervals and length of follow-up, and to understand how experience of assessors impacts on the reliability of skin assessments and subsequent analyses.

MSM have the potential to address some of these problems and their potential impact on the design and analysis of clinical trials are explored throughout this thesis. In order to understand their impact, two illustrative datasets are available to be re-analysed and used as case studies for this research. In order to assess the use of MSM compared to the common methods of analysis used in the PU prevention literature, these datasets will first be re-analysed using binary and TTE methods in the next chapter.

Chapter 3

Analysis of existing datasets

3.1 Introduction

The most common methods of analysis identified in the literature review were for binary and TTE endpoints. Binary endpoints were typically the incidence of a new PU and analysed using Pearson's Chi-squared test or binary logistic regression, whilst TTE endpoints were typically the time to onset of a new PU and were analysed using a log-rank test, Kaplan-Meier methods or the Cox proportional hazards model.

Within the Leeds Clinical Trials Research Unit, the PRESSURE and PRESSURE2 trials contain longitudinal outcome data for the purposes of measuring PU development and form case studies for this research [21, 106]. Ethical approval was received by the sponsor (University of Leeds) to re-analyse these datasets for the purposes of this thesis, confirming that additional approval was not required because research forms for the datasets included consent for secondary analyses of the data. The characteristics of the studies are summarised in Table 3.1, with detail provided here:

- PRESSURE [21]: An RCT comparing two types of mattress: alternating pressure mattress (APM) overlay and APM replacement, in acute and elective hospital patients. The trial consisted of 1, 971 patients randomised using a 1:1 allocation ratio and followed up for a maximum of 60 days post randomisation. Research nurses assessed 7 skin sites per patient at up to 13 time points.

Table 3.1: Characteristics of illustrative datasets

	PRESSURE	**PRESSURE2**
Patient population	Aged ≥ 55 years, admitted in the previous 24 hours to vascular, orthopaedic, medical, or care of elderly people wards, acute or elective admissions. Expected stay ≥ 7 days. Braden scale activity and mobility scores of 1 or 2, or an existing Grade 2 PU. Elective surgical patients without limitation of activity and mobility or an existing PU were eligible if the average length of hospital stay for their surgical procedure was ≥ 7 days or they were expected to have Braden scale activity and mobility scores of 1 or 2 for ≥ 3 days postoperatively	Aged ≥ 18 years, in-patient with evidence of acute illness recruited from adult secondary care and community in-patient acute admission facilities. Expected stay ≥ 5 days. Braden Activity and Mobility scores of 1 or 2, or an existing Category 1 PU or localised skin pain on a Category ≤ 1 pressure area
Intervention	Alternating pressure mattress overlay	Alternating pressure mattress
Control	Alternating pressure mattress replacement	High specification foam mattress
Primary outcome	Incidence of new Grade 2+ PU	Time to onset of new Category 2+ PU
Number of patients	$1,971$	$2,029$
Maximum length of follow-up	60 days	90 days
Assessment frequency	Twice weekly	Twice weekly for 30 days, once weekly until 60 days (defined as the treatment phase), 1 visit 30 days post discharge or post end of treatment phase, which ever occurred soonest
Number of skin sites per patient per assessment	7	14

- PRESSURE2 [106]: An RCT comparing APM and high specification foam (HSF) mattresses in acutely ill inpatients. A total of $2,029$ patients were randomised using a 1:1 allocation ratio and followed up for a maximum of 60 days as an inpatient (defined as the treatment phase), and had a final visit 30 days after discharge or after the end of their treatment phase, whichever occurred soonest. Patients were assessed twice weekly for the first 4 weeks as inpatients, and once weekly until discharge, providing a maximum of 14 assessments including baseline and the final post discharge visit. At each assessment, there were 14 pre-specified skin sites which had the PU status recorded.

The most common approaches to analysis in the PU literature have been on a patient basis (Chapter 2.3). From a clinical perspective it is sensible to consider this for a number of reasons; firstly, UK NHS quality indicator criteria are often based on the number of patients who have a PU [25, 27, 28]. Second, there may be measurement error in recording of the skin sites. For example, consider an individual with 4 assessments who has a PU on the left heel observed at time 2, such that the observed PU classification are $Y_{LH} = \{1, 3, 3, 4\}$ but has a healthy right heel ($Y_{RH} = \{1, 1, 1, 1\}$). If the heel skin site classifications are interchanged at the second assessment, that is the right heel skin state is recorded for the left heel and vice versa, then the observed data would be $Y^*_{LH} = \{1, 1, 3, 4\}$ and $Y^*_{RH} = \{1, 3, 1, 1\}$ incorrectly recording PU damage at the right heel. Combining the component level data at each timepoint to derive the patient's most severe PU classification avoids this potential error. However, aggregating skin site level data to provide a patient level outcome may mean that potentially important information, such as multiple PUs, fails to be taken into account in the analysis. In order to assess differences in patient and skin site level analyses, this chapter applies the common methods of analysis identified in the literature to patient level endpoints, and illustrates extensions to the common methods to analyse skin site level data accounting for the correlation of outcomes within patients.

3.2 Aim

The aim of this chapter is to re-analyse the PRESSURE and PRESSURE2 datasets using binary and time to event methods of analysis.

Objectives

1. Define binary and time to event outcomes at the patient and skin site level
2. Test for a difference in patient level outcomes using the Pearson's Chi-Squared test and log-rank test
3. Use binary logistic regression and Cox proportional hazards regression to analyse patient level outcomes
4. Use binary logistic regression and Cox proportional hazards regression with and without patient random effects to analyse skin site level outcomes

3.3 Methods

In each of the trial datasets the outcome classification was according to the relevant international guidelines at that time, which have since been assimilated to the current EPUAP/NPUAP/PPPIA guidelines [32]. In both trials an additional "1a" grade or "altered" category was added to the classification scale to denote pressure-related skin changes that were present but did not yet meet the criteria of a Category 1 PU (Table 3.2). In each dataset, individual skin sites were assessed and a PU classification was assigned. These were aggregated to analyse the data at the patient level.

Due to the long interval between the last hospital assessment and the final visit 30 days later in the PRESSURE2 trial, the analysis dataset was restricted to that collected during hospital stay for the first 60 days. There were concerns that factors outwith the trial protocol, such as discharge plans, could have affected PU development during the 30 day interval between discharge and the final visit and may therefore confound the assessment of interventions on the development of PUs [106].

Table 3.2: Definitions of PU classes used in the original PRESSURE and PRESSURE2 trials

PRESSURE	**PRESSURE2**
Grade 0 - No skin changes	Category 0 - Healthy intact skin
Grade 1*a* - redness to skin (blanching)	Category *A* - Alterations to intact skin
Grade 1*b* - redness to skin (non-blanching)	Category 1 - Non-blanchable erythema of intact skin. Intact skin with non-blanchable erythema of a localised area usually over a bony prominence. Discolouration of the skin, warmth, oedema, hardness or pain may also be present. Darkly pigmented skin may not have visible blanching.
Grade 2 - partial thickness wound involving epidermis or dermis only	Category 2 - Partial-thickness skin loss or blister. Partial-thickness loss of dermis presenting as a shallow open ulcer with a red-pink wound bed, without slough. May also present as an intact or open/ruptured serum or serosanguinous-filled blister
Grade 3 - full thickness wound involving subcutaneous tissue	Category 3 - Full-thickness skin loss. Full-thickness tissue loss. Subcutaneous fat may be visible but bone, tendon or muscle are not exposed. Some slough may be present. May include undermining and tunnelling.
Grade 4 - full thickness wound through subcutaneous tissue to muscle or bone	Category 4 - Full-thickness tissue loss. Full-thickness tissue loss with exposed bone, tendon or muscle. Slough or eschar may be present. Often includes undermining or tunnelling..
Grade 5 - black eschar	Unstageable - Full-thickness skin loss in which actual depth of the ulcer is completely obscured by slough (yellow, tan, grey, green or brown) and/or eschar (tan, brown or black) in the wound bed.

Furthermore, the clinical opinion of the PRESSURE2 trial management group was that restricting the analysis to the first 60 days may have been more clinically meaningful [106].

For consistency with later analysis using MSM, patients were excluded from the analysis dataset in both trials for the following reasons:

1. Category 2+ on any skin site at baseline (randomisation)

2. No follow-up assessments

Note that in both trials, baseline skin status was a randomisation factor with the presence of a pre-existing Category 2+ PU as a specific level. Therefore the exclusion of patients with pre-existing Category 2+ on any skin site at baseline is unlikely to lead to an imbalance of patients across the arms of each trial.

3.3.1 Endpoint definition

In this section, we define the binary endpoint observed for patients and for each skin site in the two illustrative datasets. Let k denote the index for the k^{th} skin site, where $k = 1, ..., K$. Note that $K = 7$ for the PRESSURE trial, and $K = 14$ for the PRESSURE2 trial. Let w index the assessment number $w = 1, ..., W$. Then $X_k(t_{iw})$ denotes the PU classification for skin site k for patient i at the w^{th} assessment time, t_{iw}. Let Z_{ik} denote the binary response variable for skin site k for patient i such that

$$Z_{ik} = \begin{cases} 1, & \text{if} \quad X_k(t_{iw}) \in \{2, 3, 4, 5\} \text{ for any } w \in \{1, ...W\} \\ 0, & \text{otherwise} \end{cases} \tag{3.1a}$$

The skin site level endpoint, Z_{ik} can be used to define the patient level binary endpoint, Y_i such that

$$Y_i = \begin{cases} 1, & \text{if} \quad Z_{ik} = 1 \text{ for any } k \in \{1, ...K\} \\ 0, & \text{if} \quad Z_{ik} = 0 \,\forall\, k \in \{1, ...K\} \end{cases} \tag{3.2a}$$

3.3.2 Methods to analyse binary response

First, consider the case where patients have been assigned using a random process to one of two groups. In an RCT, these would be intervention and control arms.

In the review of published PU prevention trials, a commonly used model-based analysis for the incidence of a PU as a binary outcome was the logistic model. This is a generalised linear model that conditions on treatment and other independent variables. Let π_i denote the conditional probability of patient i having the event of interest, i.e. $\pi_i = P(Y_i = 1 \mid \boldsymbol{x}_i)$. Assuming a logit link function, the logistic model is then given by

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \boldsymbol{\beta}^T\boldsymbol{x}_i \quad i = 1, ..., n \tag{3.3}$$

where $\boldsymbol{x}_i$ is a p-vector of independent variables $(x_{1i}, ..., x_{pi})$ for patient i, with coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ and α_0 denotes the log odds of developing a PU when all covariates take the value zero. Without loss of generality, we take x_{1i} to be the treatment allocation for patient i, with value 1 for the intervention and 0 for the control. Therefore, β_1 is the log-odds ratio of a PU in the intervention group relative to the control, all else being equal. The treatment effect is assessed using the null hypothesis $H_0 : \beta_1 = 0$.

3.3.3 Methods to analyse time to event outcome

In order to take time at risk and timing of PU onset into account, let t denote the time since randomisation at which a patient is observed to develop a PU. The observation t is a realisation of the random variable T. We define the survivor function, $S(t)$, as the probability that the time without a new PU is greater than or equal to t, i.e.

$$S(t) = P(T \geq t), \quad t \geq 0 \tag{3.4}$$

The hazard function, $h(t)$, is a rate defined as the instantaneous probability of PU onset in continuous time and is given by:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} \tag{3.5}$$

The cumulative hazard function, $H(t)$ is defined as:

$$H(t) = \int_0^t h(u)du = -\log S(t) \tag{3.6}$$

Note that the specifications in equations 3.4 to 3.6 are interchangeable in the sense that specification of one will permit specification of others.

During an assessment period, not all subjects will be observed to develop a PU, this could be due to patient death, the end of a follow up schedule (administrative reasons) or because the patient is lost to follow-up, for example. The survival time for a patient is only known up until the point of their last observation, therefore the survival status is said to be right censored at the last time point at which the patient was known to be alive. Formally we define c_i, the censoring time for individual i, such that we observe $t_i^* = min(t_i, c_i)$. Right censoring, when the observed survival time (t_i^*) is less than the true survival time (t_i), is common in survival analysis. Less commonly, left censoring can occur when the actual survival time occurs before the observation period. Since this thesis is primarily concerned with RCTs, for which time zero is the point of randomisation, left truncation will not be discussed further. A third type of censoring (interval censoring) occurs when a patient is not under observation between 2 time points. For example, at follow up time t_1 the patient may not have any new PUs but at follow up time t_2 the patient has a new PU, the time of onset is known to be between t_1 and t_2 but the exact time is unknown. Data arising in this way is also known as panel data.

For many survival analysis methods the censoring mechanism is assumed to be independent of the survival time of an individual. That is, a patient whose survival time is censored at time c_i is representative of the patients who remain in the risk set at time c_i. Note that in trials the observation schedule is specified in advance and therefore it is usually clear when a measurement has been missed and the reason for missing data may be known, however in purely observational data it may not be possible to know when or why measurements have been missed. Methods for handling missing (censored) data that are not independent of the disease status are discussed in Chapter 8.

The most common model-based analysis used for time to event outcomes in the PU literature was the Cox proportional hazards (PH) model [101]. In this model, let $h_1(t)$ and $h_2(t)$ denote the hazard functions for two otherwise identical patients in treatment groups 1 and 2 respectively. The PH assumption is such that $h_1(t) = \phi h_2(t)$ for some non-negative constant, ϕ. If the PH assumption holds $S_1(t) = S_2(t)^\phi$ and it follows that, $S_1(t) \leq S_2(t)$ if $0 \leq \phi \leq 1$, else $S_1(t) > S_2(t)$ if $\phi > 1$. The general PH model can be written as:

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \quad i = 1, ..., n \tag{3.7}$$

Where $h_i(t)$ denotes the hazard of individual i developing a new PU at time t. Cox showed that the PH coefficients $\boldsymbol{\beta}^T$ can be estimated using maximum partial likelihood estimation and from these, hazard ratios and corresponding 95% confidence intervals can be obtained [101, 107]. Taking x_{1i} to be the treatment allocation as before, then $h_0(t)$ is the baseline hazard in the control group when all covariates take the value zero and does not have a parametric form in the Cox model. The hazard ratio of treatment versus control is given by $\exp(\beta_1)$ all else being equal. The partial likelihood ratio test (LRT) can be used to test the null hypothesis that $\beta_1 = 0$. Note that in this analysis, some elements of $\boldsymbol{x_i}$ may be time varying.

3.3.4 Extensions to common methods of analysis

So far, the methods described have focused on patient level data, but detailed skin site level data were collected in the PRESSURE and PRESSURE2 datasets. Methods to analyse skin site level data with a binary response or TTE outcome are outlined here.

Binary response

The simplest model for the skin site level data would be a logistic fixed effects regression model in line with 3.3. Let η_{ik} denote the conditional probability that skin site k, within patient i develops a PU, i.e. $\eta_{ik} = P(Z_{ik} = 1 \mid x_i, \nu_{ik})$. Ignoring the hierarchical structure, the logistic model is given by:

$$logit(\eta_{ik}) = \log\left(\frac{\eta_{ik}}{1-\eta_{ik}}\right) = \sum_{j=1}^{K} \alpha_j \mathrm{I}_{\mathrm{j=k}} + \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\psi}^T \boldsymbol{\nu}_{ik}, \quad i = 1, ..., n, \quad k = 1, ..., K \tag{3.8}$$

where α_k, $k = 1, ..., K$ denotes the log odds of a PU developing at skin site k when all covariates are equal to zero. Note that $\mathrm{I}_{\mathrm{j=k}}$ is the indicator function such that

$$\mathrm{I}_{\mathrm{j=k}} = \begin{cases} 1, & \text{if} \quad j = k \\ 0, & \text{otherwise} \end{cases} \tag{3.9a}$$

The vectors $\boldsymbol{\beta}$ and $\boldsymbol{x_i}$ are as defined in equation 3.2a, with β_1 denoting the log-odds ratio for intervention relative to control, all else being equal. Equation 3.8 includes a q-vector of skin site specific covariates, $\boldsymbol{\nu}_{ik}$, with coefficients $\boldsymbol{\psi}$ for completeness, but these are not considered in this thesis.

The limitation of this model is that it assumes the observations are independent and therefore fails to take into account the correlation between skin sites in the same patient. This may lead to underestimated standard errors and biased parameter estimates [104]. Alternatively, the logistic mixed model can explicitly account for this correlation by incorporating patient random effects as follows:

$$logit(\eta_{ik}) = \log(\frac{\xi_{ik}}{1-\xi_{ik}}) = \sum_{j=1}^{K} \alpha_j \mathrm{I}_{\mathrm{j=k}} + \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\psi}^T \boldsymbol{\nu}_{ik} + u_i, \quad i = 1, ..., n, \quad k = 1, ..., K \tag{3.10}$$

where $u_i \sim N(0, \sigma_u^2)$ denotes the patient level random effects, or equivalently, the variability in outcome that is due to patient differences. In most cases, as here, the random effects are assumed to follow a normal distribution, although this is not strictly necessary. However, normal random effects models can be estimated easily in most widely used statistical packages. The variance partition coefficient (VPC) is a measure of the proportion of the total variance due to patient variability and can be calculated as [108]:

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}} \tag{3.11}$$

where $\frac{\pi^2}{3}$ is the variance of a standard logistic distribution and represents the variance of the level 1 residuals in 3.11 [108].

Time to event outcome

Skin site level data can be analysed using TTE methods through fixed and random effects. The PH model including skin site as a fixed effect can be written as:

$$h_{ik}(t) = h_0(t)\exp(\sum_{j=2}^{K} \alpha_j I_{j=k} + \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\psi}^T \boldsymbol{\nu}_{ik}), \quad i = 1, ..., n, \quad k = 1, ..., K \tag{3.12}$$

Where $h_{ik}(t)$ denotes the hazard of developing a new PU at time t for skin site k within individual i. Note that in this model, $h_0(t)$ is the baseline hazard for skin site 1 (sacrum) when all covariates take the value zero and proportional hazards are assumed for the other skin sites. The parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{x_i}$ are as defined in equation 3.7, with β_1 denoting the log-hazard ratio for intervention relative to control, all else being equal. As with the logistic regression with random effects, Equation 3.12 includes a q-vector of skin site specific covariates, $\boldsymbol{\nu}_{ik}$, with coefficients $\boldsymbol{\psi}$ for completeness.

For TTE data, clustering of skin sites within patients can be accounted for through the use of a shared frailty term. The hazard function can be extended to

$$h_{ik}(t) = h_0(t)\exp(\sum_{j=2}^{K}\alpha_j I_{j=k} + \boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{\psi}^T\boldsymbol{\nu}_{ik} + u_i), \quad i = 1, ..., n, \quad k = 1, ..., K \tag{3.13}$$

where $u_i = \log(v_i)$, such that v_i is a realisation of the frailty random variable $V_i \sim \Gamma(\theta, \theta)$. Alternative distributions can be used as appropriate but the gamma distribution is considered the simplest and most well understood frailty model [109]. Alternatively, the log-Normal frailty model is also considered one of the more simple methods and it may be useful to assess the sensitivity of results to different frailty distributions. Note that the baseline hazard will remain non-parametric in the models fitted to the case study datasets, although alternative parametric approaches can be considered.

3.3.5 Independent variables and hypothesis testing

Multiple independent variables could have been included in the analysis such as variables used to inform the treatment allocation, and variables collected at baseline with a prognostic relationship to the outcome. However, for simplicity, only the treatment variable was assessed in the re-analysis of the patient level case study datasets: Overlay vs replacement for the PRESSURE trial, and APM vs HSF for the PRESSURE2 trial. In the analysis of the skin site level data, both the treatment variable and skin site level variables will be assessed. To understand how beneficial it is to conduct a skin site level analysis compared to a patient level analysis, the VPC will be examined for binary response data, and for both binary and TTE methods, the treatment effect estimates will be examined. If the effect of treatment is similar for both the patient level and skin site level analyses, and if the proportion of the total variance due to patient variability suggests that skin site data do not contribute much additional information to the analysis, then it may be reasonable to conduct analyses of PU data at the patient level rather than at the skin site level.

Statistical significance of independent variables in the logistic regression and Cox PH models were assessed at the 5% level. To test the overall significance of categorical variables, LRT were used. Formal tests could be conducted to assess contrasts between individual skin sites and the average of other skin sites but a multiple testing correction would be required. The primary purpose of the analysis in this section is to assess the value of conducting the analysis using skin site level or patient level data to estimate the effect of treatment. Therefore, the comparison of specific levels of independent variables was examined using point estimates of the relevant estimand (Odds ratio (OR) or hazard ratio (HR)) and corresponding Wald type 95% confidence intervals.

All analyses were conducted in R including use of the "*glm*", "*glmer*" and "*coxph*" functions.

3.4 Results

3.4.1 Patient level analysis

Of 1971 patients in the PRESSURE trial, a total of $1,659$ (84.2%) patients were in the analysis dataset of which 153 (9.2%) developed a new PU; with 73 (8.8%) in the intervention (overlay) group and 80 (9.6%) in the control (replacement) group (Tables 3.3 and 3.4). Of 2029 patients in the PRESSURE2 trial, a total of $1,729$(85.2%) were in the PRESSURE2 analysis dataset of which 127 (7.3%) developed a new PU; with 47 (5.4%) in the intervention (APM) group and 80 (9.2%) in the control (HSF) group (Tables 3.3 and 3.4). Analysis results are shown in Table 3.4 for a Chi-squared test for proportions and the estimated odds ratio obtained from a logistic regression model where the outcome is regressed on the treatment allocation only. There was no evidence of a treatment effect on the incidence of a Category 2+ PU in the PRESSURE dataset, where the odds of developing a Category 2+ PU in the intervention group was was 0.91 (95% CI $0.65, 1.27$) times the odds in the control group. Meanwhile, a statistically significant treatment effect was observed for the PRESSURE2 dataset with the intervention resulting in a decrease in the probability of developing a Category 2+ PU compared to the control group (OR (95% CI)$=$ 0.57 $(0.39, 0.82)$).

TTE analyses were also applied to these datasets, with KM plots presented in

Table 3.3: PU incidence (patient level) in illustrative datasets for patients who do not have a Category 2+ PU at baseline

Dataset	**Number of patients**	**Incidence of Category 2+ PU**
PRESSURE	$1,659$	153 (9.2%)
PRESSURE2	$1,729$	127 (7.3%)

Table 3.4: Analysis of patient level binary outcomes in illustrative datasets

Dataset	**Variable**	**Incidence of new PU**		**Analysis results**	
		Yes	No	χ^2 test p-value	OR (95% CI)
PRESSURE	Overlay	73 (8.8%)	754 (91.2%)	0.6384	0.91 $(0.65, 1.27)$
	Replacement	80 (9.6%)	752 (90.4%)		
PRESSURE2	APM	47 (5.4%)	816 (94.6%)	0.0034	0.57 $(0.39, 0.82)$
	HSF	80 (9.2%)	786 (90.8%)		

Figure 3.1. Inspection of the KM-plots suggests that the treatment effect does not appear until after the first week at risk. That is, there is a delayed treatment effect which is common in trials of prevention interventions where the treatment can take time to work [110]. Results of the of the log-rank test are presented in Table 3.5 alongside hazard ratios estimated from the Cox regression model. Although the endpoints and analysis method differ, there are similar conclusions to the analysis of the binary endpoints. That is, there is no evidence of a treatment effect in the PRESSURE dataset in terms of time to development of a new Category 2+ PU, with the hazard of developing a Category 2+ PU in the intervention group equal to 0.84 (95% CI $0.61, 1.15$)) times the odds in the control group. Whereas a statistically significant treatment effect was observed for the PRESSURE2 dataset with the intervention resulting in an estimated hazard of developing a Category 2+ PU 0.61 (95% CI $0.43, 0.88$) times the control. That is, in the PRESSURE2 dataset there

was evidence that the intervention provided a benefit to patients. Note however that inspection of the KM-plots suggests that the proportional hazards assumption is not valid for either dataset due to the delayed treatment effect (Figure 3.1).

Table 3.5: Analysis of patient level TTE outcomes

Dataset	**Variable**	**Analysis results**	
		log-rank test p-value	HR (95% CI)
PRESSURE	Overlay	0.27	0.84 (0.61, 1.15)
	Replacement		
PRESSURE2	APM	0.0071	0.61 (0.43, 0.88)
	HSF		

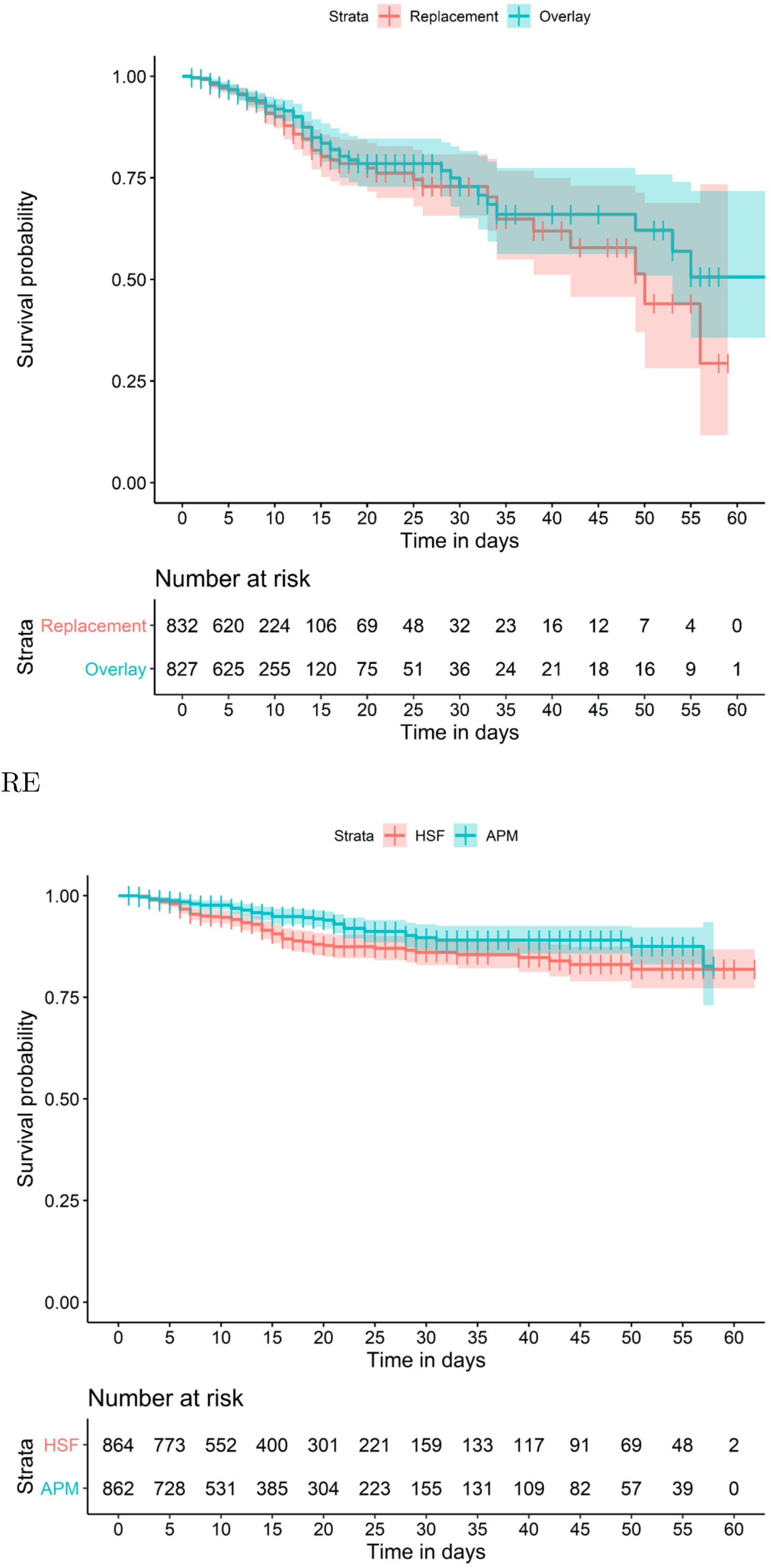
3.4.2 Skin site level analysis

The incidence of a new Category 2+ PU for individual skin sites in the two trial datasets are presented in Table 3.6. In the PRESSURE trial there were a total of $10,241$ skin site assessments, of which 205 (2.0%) were observed to develop a new Category 2+ PU, and in the PRESSURE2 trial there were a total of $24,742$ skin sites of which 183 (0.7%) were observed to develop a new Category 2+ PU. Note that the incidence was expected to be higher in the PRESSURE trial because there were half as many skin sites per patient and the skin sites were selected because they were considered at highest risk of PU development. The observed PU incidence for individual skin sites indicate that there were different probabilities of PU incidence for different skin sites with similar patterns for both datasets. For example, in the PRESSURE2 trial the incidence of Category 2+ PUs at the sacrum, buttocks and heels accounted for the majority of new PUs with 146 (79.8%) observed at these skin sites.

Results are shown in Table 3.7 for the estimated odds ratio obtained through a logistic regression model where the outcome is regressed on the variable of interest (treatment allocation) and skin site as a fixed effect. Including skin site as a 7 level

Table 3.6: Incidence of Category 2+ PUs at skin site level for the PRESSURE and PRESSURE2 trials

Variable	**PRESSURE**		**PRESSURE2**	
	Yes	No	Yes	No
Intervention				
Intervention	96 (1.9%)	5,045 (98.1%)	67 (0.5%)	12,285 (99.5%)
Control	109 (2.1%)	4,991 (97.9%)	116 (0.9%)	12,274 (99.1%)
Skin sites				
Sacrum	51 (3.5%)	1,415 (96.5%)	36 (2.1%)	1,699 (97.9%)
Back	-	-	6 (0.3%)	1,787 (99.7%)
Left buttock	51 (3.4%)	1,446 (96.6%)	38 (2.2%)	1,693 (97.8%)
Right buttock	59 (3.9%)	1,446 (96.1%)	34 (2.0%)	1,705 (98.0%)
Left ischial	-	-	4 (0.2%)	1,811 (99.8%)
Right ischial	-	-	4 (0.2%)	1,812 (99.8%)
Left hip	2 (0.2%)	1,274 (99.8%)	2 (0.1%)	1,776 (99.9%)
Right hip	2 (0.2%)	1,268 (99.8%)	1 (0.06%)	1,781 (99.94%)
Left heel	22 (1.4%)	1,592 (98.6%)	20 (1.1%)	1,727 (98.9%)
Right heel	18 (1.1%)	1,595 (98.9%)	18 (1.0%)	1,708 (99.0%)
Left ankle	-	-	5 (0.3%)	1,734 (99.7%)
Right ankle	-	-	4 (0.2%)	1,717 (99.8%)
Left elbow	-	-	2 (0.1%)	1,805 (99.9%)
Right elbow	-	-	9 (0.5%)	1,804 (99.5%)



(a) PRESSURE

(b) PRESSURE2

Figure 3.1: Kaplan-Meier Plots for the time to development of a Category 2+ PU by randomised treatment for the PRESSURE and PRESSURE2 trials (patient level)

fixed effect in the analysis of the PRESSURE data, the odds of developing a Category 2+ PU at any site in the intervention group were 0.87 $(0.66, 1.15)$ times the odds in the control group, all else being equal. Augmenting this model to include a random intercept for patients results in a similar (common) treatment effect, with an OR of 0.90 $(0.41, 1.96)$. Note that there was a slight change in the treatment effect and wider confidence intervals due to the between-patient variation. Meanwhile, including skin site as a 14 level fixed effect in the analysis of the PRESSURE2 data, the odds of developing a Category 2+ PU at any site in the intervention group were 0.58 $(0.42, 0.78)$ times the odds in the control group, all else being equal. After incorporating patient random effect, the OR was estimated as 0.64 $(0.30, 1.37)$ suggesting that there was no evidence of a treatment effect on the incidence of a Category 2+ PU. Again note that there was an increase in variance due to the variation between patients around the treatment effect. For both datasets, the inclusion of skin site as a categorical fixed effect was statistically significant (LRT, $p < 0.0001$). The point estimate for each level of the skin site variable was relative to the sacrum, which was one of the skin sites considered at high risk of PU development. The estimates were in line with clinical expectations, with the buttocks having a similar odds of developing a PU to the sacrum. The heels were less likely to develop a PU compared to the sacrum, but more likely than any of the other observed skin sites (ischial tuberosities, back, hips, ankles and elbows) when they were compared to the sacrum. These conclusions were consistent across both the fixed and random effects models.

As discussed in Section 3.3.4, the limitation of the fixed effects model is that it does not account for the correlation of outcomes within patients. A logistic model incorporating patient as a random effect was fitted to the trial datasets with results shown in Table 3.7. The point estimates and confidence intervals were similar across both fixed and random effects models. The between patient variance was estimated as $\hat{\sigma}_u^2 = 41.9$ and $\hat{\sigma}_u^2 = 36.4$ on the logistic scale for the PRESSURE and PRESSURE2 datasets respectively, which means that, according to the VPC, approximately 90.7% and 91.7% of the total variance was due to differences between patients rather than between skin sites. This suggests that skin site data do not contribute much additional information to the analysis in the PU setting and therefore, patient level analysis may be sufficient in the PU setting.

For completeness, skin site level analyses were also conducted for TTE outcomes using both fixed effects and random effects models. The parameter estimates for these models are shown in Table 3.8. As with the logistic regression applied to these data, the point estimates and confidence intervals were similar for the fixed and random effects models. Including skin site as a 7 level fixed effect in the analysis of the PRESSURE data, the hazard of developing a Category 2+ PU at any site in the intervention group was 0.79 $(0.60, 1.04)$ times the hazard in the control group, all else being equal. Including a shared Gamma frailty for patients led to a similar (common) treatment effect, with a HR of 0.78 $(0.55, 1.10)$. For the PRESSURE2 trial, including skin site as a 14 level fixed effect in the TTE analysis showed that the hazard of developing a Category 2+ PU at any site in the intervention group was 0.60 $(0.45, 0.82)$ times the hazard in the control group, all else being equal, with a similar estimate of treatment effect after incorporating a shared Gamma frailty for patient, with a HR of 0.54 $(0.37, 0.81)$. Note, as with the analysis of the binary outcome, that for both datasets there were wider confidence intervals due to the between-patient variation. For both datasets, the inclusion of skin site as a categorical fixed effect was statistically significant (LRT, $p < 0.0001$). The point estimate for each level of the skin site variable was relative to the sacrum, which was one of the skin sites considered at high risk of PU development. The point estimates for each skin site relative to the sacrum led to similar conclusions as the logistic regression models with the buttocks having a similar hazard for developing a PU compared to the sacrum. The heels were observed to have a lower hazard for developing a PU compared to the sacrum, but the estimated hazard ratio was higher than any of the other observed skin sites (ischial tuberosities, back, hips, ankles and elbows) when they were compared to the sacrum. These conclusions were consistent across both the fixed and random effects models.

Table 3.7: Logistic regression applied to the skin site level binary outcome for the PRESSURE and PRESSURE2 trials (Fixed effects and random intercept accounting for patient)

Variable	**PRESSURE**		**PRESSURE2**	
	OR (95% CI)		OR (95% CI)	
	Fixed effects	Random effects	Fixed effects	Random effects
Intervention				
Intervention	0.87 (0.66, 1.15)	0.90 (0.41, 1.96)	0.58 (0.42, 0.78)	0.64 (0.30, 1.37)
Control (reference)	-	-	-	-
Skin sites				
Sacrum (reference)	-	-	-	-
Back	-	-	0.16 (0.06, 0.35)	0.10 (0.04, 0.25)
Left buttock	0.98 (0.66, 1.46)	0.95 (0.55, 1.64)	1.06 (0.67, 1.68)	1.05 (0.59, 1.86)
Right buttock	1.13 (0.77, 1.66)	1.27 (0.75, 2.17)	0.94 (0.58, 1.51)	0.87 (0.49, 1.56)
Left ischial	-	-	0.10 (0.03, 0.26)	0.06 (0.02, 0.19)
Right ischial	-	-	0.10 (0.03, 0.26)	0.06 (0.02, 0.19)
Left hip	0.04 (0.01, 0.14)	0.01 (0.003, 0.06)	0.05 (0.01, 0.17)	0.03 (0.01, 0.13)
Right hip	0.04 (0.01, 0.14)	0.01 (0.003, 0.06)	0.03 (0.001, 0.12)	0.01 (0.002, 0.11)
Left heel	0.38 (0.23, 0.63)	0.25 (0.13, 0.49)	0.55 (0.31, 0.94)	0.42 (0.21, 0.80)
Right heel	0.31 (0.18, 0.53)	0.19 (0.10, 0.38)	0.50 (0.27, 0.87)	0.36 (0.18, 0.71)
Left ankle	-	-	0.14 (0.04, 0.32)	0.09 (0.03, 0.24)
Right ankle	-	-	0.11 (0.03, 0.28)	0.07 (0.02, 0.20)
Left elbow	-	-	0.05 (0.01, 0.17)	0.03 (0.01, 0.13)
Right elbow	-	-	0.24 (0.11, 0.47)	0.15 (0.07, 0.34)
		$\hat{\sigma}_u^2 = 41.9$		$\hat{\sigma}_u^2 = 36.4$

Table 3.8: Cox regression applied to the skin site level TTE outcome for the PRESSURE and PRESSURE2 trials (Fixed effects and frailty accounting for patient)

Variable	**PRESSURE**		**PRESSURE2**	
	HR (95% CI)			
	Fixed effects	Random effects	Fixed effects	Random effects
Intervention				
Intervention	0.79 (0.60, 1.04)	0.78 (0.55, 1.10)	0.60 (0.45, 0.82)	0.54 (0.37, 0.81)
Control (reference)	-	-	-	-
Skin sites				
Sacrum (reference)	-	-	-	-
Back	-	-	0.16 (0.07, 0.37)	0.15 (0.06, 0.34)
Left buttock	0.99 (0.67, 1.46)	1.04 (0.70, 1.54)	1.05 (0.67, 1.66)	1.09 (0.69, 1.73)
Right buttock	1.13 (0.78, 1.64)	1.16 (0.79, 1.69)	0.93 (0.58, 1.49)	0.94 (0.58, 1.50)
Left ischial	-	-	0.10 (0.04, 0.29)	0.10 (0.03, 0.27)
Right ischial	-	-	0.10 (0.04, 0.29)	0.10 (0.03, 0.27)
Left hip	0.04 (0.01, 0.17)	0.04 (0.01, 0.15)	0.05 (0.01, 0.22)	0.05 (0.01, 0.20)
Right hip	0.04 (0.01, 0.17)	0.04 (0.01, 0.15)	0.03 (0.004, 0.19)	0.02 (0.003, 0.18)
Left heel	0.41 (0.25, 0.68)	0.41 (0.25, 0.67)	0.55 (0.32, 0.94)	0.53 (0.31, 0.93)
Right heel	0.33 (0.20, 0.57)	0.33 (0.19, 0.57)	0.49 (0.28, 0.87)	0.48 (0.27, 0.85)
Left ankle	-	-	0.14 (0.05, 0.35)	0.13 (0.05, 0.33)
Right ankle	-	-	0.11 (0.04, 0.31)	0.10 (0.04, 0.29)
Left elbow	-	-	0.05 (0.01, 0.22)	0.05 (0.01, 0.20)
Right elbow	-	-	0.23 (0.11, 0.48)	0.21 (0.10, 0.45)

3.5 Discussion

This chapter re-analysed two existing datasets using common binary and TTE methods identified in the literature. Firstly, the methods were applied to patient level data and the results from both the binary and TTE analyses were shown to be consistent within each dataset. The estimands in the model based analyses were the odds ratio for incidence of, and the hazard ratio for time to onset of severe disease respectively. However, there are limitations of each, as highlighted in Chapter 2. The binary outcome does not take into account the length of time a patient is in the trial before discharge and there is evidence from the Kaplan-Meier curves in Figure 3.1 that the proportional hazards assumption is not valid for either dataset. In the PRESSURE2 example, both the binary and TTE analyses concluded a statistically significant treatment effect suggesting that the intervention does provide a benefit to patients in terms of the onset of Category 2+ PU. However, inspection of the KM-plot suggests that the treatment effect does not appear until after the first week at risk; such a delayed effect might be expected in a PU prevention trial, in which Category 2+ ulcers take some time to develop, with a corresponding delay in evidence of prevention.

After applying methods to the patient level dataset, methods were used to analyse skin site level data. The findings from the analysis of the binary skin site level data suggested that at least 90% of the total variance is due to between patient variability and the small incidence of Category 2+ PU at most skin sites suggests that patient level analysis is likely to be adequate for estimating the effect of PU prevention interventions. Therefore, analyses will predominantly be conducted on a patient level for the remainder of this thesis.

Analysing PU trial data using methods for longitudinal data may help to understand the natural history of the disease and to identify where treatment may have most benefit.

Chapter 4

Multi-state models

4.1 Introduction

This thesis is motivated by trials of PU prevention strategies, where the endpoint of interest in published trial reports is typically the incidence of a PU or time to new PU. These endpoints are often calculated from longitudinal measurements of PU category using an ordinal classification scale. Such discrete longitudinal outcomes have been used in many other disease areas such as psoriatic arthritis [111, 112] and in cancer settings where both overall survival and progression-free survival are of interest [113]. The methods used in this thesis will therefore be relevant to settings beyond PU prevention.

When designing a trial where the incidence of an event (or time to event) is the primary endpoint, the length of follow-up needs to be considered. Follow-up should be long enough that a sufficient number of clinically relevant events are observed, but not so long that the follow-up burden is excessive for both participants and trial resources, particularly when the rate of new events decreases. The frequency of assessments may coincide with standard clinic visits, or may be set such that changes in disease status can be observed.

MSM have been used to explore the natural history of diseases in a range of conditions and settings, including applications to data arising through cohort studies or registries, and secondary or exploratory analysis of RCT data. Such analyses have been used in some cases to inform trial design features such as the patient population

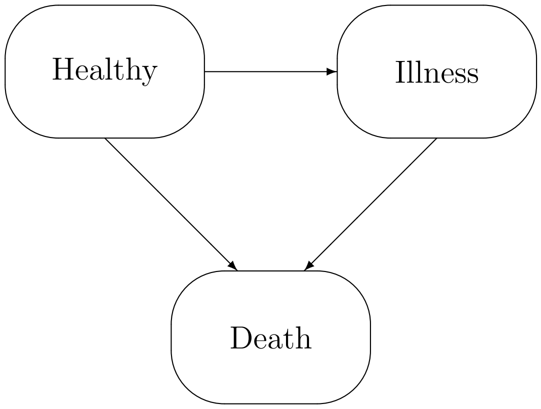


Figure 4.1: Illness-death model, progression only

or assessment schedules, or to determine future research questions. In addition to informing design features, MSM may be useful for redefining endpoints for trials of disease prevention, where a discrete outcome is collected longitudinally and where more than one level is of interest to researchers.

In this chapter, some examples of MSM are discussed to demonstrate how they have been used in medical research to provide deeper insights into natural history of disease or treatment effects and overcome some of the limitations of traditional analysis methods discussed in Chapter 2.3. In addition, some of the key decisions about the model structure and assumptions that must be made are discussed. This is followed by Section 4.3.1 which outlines notation for MSM and state definition for the motivating datasets. The results when applied to the trial datasets are presented in Section 4.4. A final discussion and plan for further investigation of MSM for the design and analysis of trials with a discrete longitudinal outcome is presented in Section 4.5.

4.1.1 Model structure

There are a many considerations to be made before fitting a MSM. In the first instance, the number of disease states and the number of transitions should be determined [114]. The simplest MSM is a standard survival model which has 2 states; Alive and Dead with a single transition from Alive to Dead. The Alive state

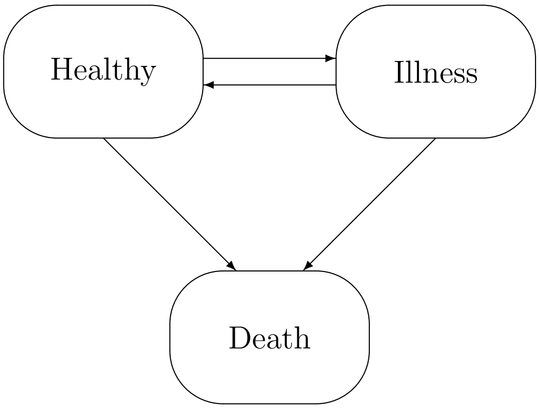


Figure 4.2: Illness-death model with regression

is the initial state and the Dead state is an absorbing state because patients cannot exit this state once they have entered it. For the case studies presented in Chapter 3, these two states were “Free of Category 2+ PU” and “Category 2+ PU”. In some cases, there may be events that prevent the Dead state from being observed, for example if a patient dies before the event of interest is observed. Here, there is a single Alive state and multiple absorbing states of which one is usually of primary interest. In this case, a competing risks model can be used to estimate covariate effects on an event of interest in the presence of competing events [115].

The focus of this thesis is MSM where the initial (Alive) state is split into one or more intermediate or transient states, and a single absorbing state. One of the most common and simplest examples of such a MSM is the illness-death model with uni-directional transitions. A simple example of such a model, shown in Figure 4.1, has 3 states to represent Healthy, Illness and Death and patients can move from Healthy to Illness, Healthy to Death and Illness to Death [115]. This model has been used in a range of disease areas including bladder cancer [116], lung transplantation [117, 118] and has led to greater understanding of the natural history of disease. The illness-death model with uni-directional transitions has been widely used, however the number of states and transitions may not be appropriate for all disease settings. Firstly, the number of transitions in the illness-death MSM could be decreased to give a 3-state progressive model where transitions can only occur

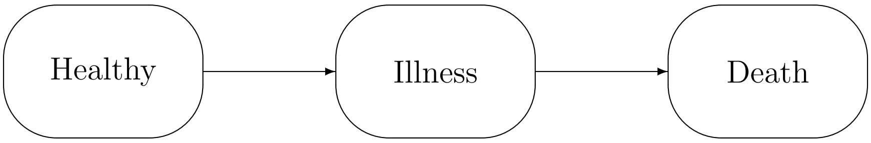


Figure 4.3: 3-progressive model

in order of disease severity in line with Figure 4.3 [119]. The illness-death model can be extended to form a more general disease progression model if there is more than one level of disease severity for example, by adding additional states. Transitions between these transient states may be uni-directional, particularly if disease prevention trials are of interest, or bi-directional if the natural history of disease more generally is of interest. For example, the COACH trial was an RCT comparing standard care to basic or intensive additional support for patients hospitalised with a primary diagnosis of heart failure. The trial was designed to detect differences in the number of hospitalisations due to heart failure or death from any cause [120]. This composite endpoint is common for heart failure research. Postmus *et al* recognised limitations with using a composite endpoint and re-analysed the COACH trial data using a 3 state MSM where the states represented 1: discharged from hospital, 2: hospitalisation because of heart failure and 3 death. There were 4 possible transitions; $1 \rightarrow 2$, $1 \rightarrow 3$, $2 \rightarrow 1$, and $2 \rightarrow 3$ [121]. The aim of the model was to predict overall survival and recurrent hospitalisation due to heart failure. The model was externally validated using a prospective cohort study and was shown to perform well in terms of prediction [121]. This is an example of how MSM can be a useful method for analysing data where multiple endpoints are of interest and is flexible to the number and direction of transitions. Whilst the COACH trial utilised an illness-death model with bi-directional transitions to accommodate recurrent hospitalisation due to heart failure [121], an alternative approach taken by Ieva *et al* was to include an additional state for each additional hospitalisation and discharge experienced by an individual, and one absorbing state to represent death with uni-directional transitions [122]. This was possible through the use of a large administrative dataset with a total of $35,224$ records from $15,298$ patients.

Compared to a traditional survival analysis, the number of parameters to be estimated in a MSM can increase rapidly according to the number of states, number of transitions and any covariate effects. Whilst the model structure should primarily be determined based on the clinical problem, the availability of data is critical to ensuring model fit and certain decisions or assumptions may be required. Firstly, if there is a small number of transitions observed between a particular pair of states, a decision could be made to not model those data. For example Ieva *et al*, did not include events for patients who experienced 6 or more hospital admissions due to lack of data [122]. This is similar to grouping disease states together but decisions of this type should be made jointly with clinical experts. Secondly, it may be reasonable to impose constraints on covariate effects for particular transitions so that those transitions with a small number of observations can be estimated using data from other observed transitions. As with grouping states together, this decision should be made jointly with clinical experts to ensure the assumption of equal covariate effects is clinically plausible. Alternatively, it may be appropriate to estimate covariate effects on a subset of transitions as agreed with clinical experts. For example, the FOGT -2 trial was re-analysed using a MSM. The FOGT-2 trial consisted 796 participants with rectal cancer allocated to 1 of 3 treatment groups after their primary surgery; the primary endpoint of the trial was overall survival [123]. With improving prognosis for patients with rectal cancer, Manzini *et al* proposed using an MSM to assess covariate effects on different stages of the disease process in order to obtain more accurate predictions of long-term survival. The MSM had 8 states denoting different stages of the chemotherapy schedule, local recurrence, distant metastasis and death, with a total of 21 transitions [124]. However, covariate effects were not modelled for 11 transitions where there were 20 or fewer observations because of the quantity of available data and these transitions were considered less clinically relevant.

Due to the size of their dataset, Ieva *et al* explored a variety of such model features including both semi-parametric and fully parametric TTE models, inclusion of factors thought to affect outcomes, and differences in the time scale (patient age, time since study entry and time since entry to the previous state). Patient specific random effects (frailties) in the semi-parametric models were also explored to account for patients who may have a higher propensity for re-admissions, although the inclusion of random effects did not change the conclusions of the analysis, suggesting that a model with fixed effects only was adequate. The authors acknowledged that due to the size of the available dataset, they were able to explore the effect of a range of covariates on outcomes which might not have been possible with a smaller dataset. They also highlighted the need to fully understand the model assumptions made and to evaluate whether they are appropriate for the dataset, through sensitivity analyses and assessment of model fit.

Random effects can be incorporated into MSM if appropriate for the clinical problem and data structure such as that explored above by Ieva *et al.* In some settings such as the PU case studies, multiple measurements may be recorded for an individual at any one time leading to potential correlations between measurements on the same individuals. An example of this is the data collected on psoriatic arthritis (PSA) by the University of Toronto PsA clinic which have been analysed using MSM [111]. The dataset included data from 510 participants who had no damage in the joints of their hands at entry to the clinic. As part of the data collection, each patient had 14 joints on each hand for which damage may be reported. These data were combined to determine which of 4 states each joint was in: State 1 - Damage in neither joint, State 2 - Damage in the left hand joint only, State 3 - Damage in the right hand joint only, State 4 - Damage in both the left and right joints. A 4 state MSM, with 4 permitted transitions, was fitted for each joint with a patient-specific random effect to account for correlation of joint outcomes within each patient. A further use of random effects was also considered by this research group to account for patients who did not develop any joint damage, described as 'stayers' and patients who would develop joint damage 'movers'. This mover-stayer model was appropriate because a large proportion of patients, 71%, did not develop any damage in any of their hand joints throughout the data collection period. As with the approach taken by Ieva *et al* [122] the aim and subsequent recommendations from the PsA study was to explore a variety of different models to determine the most appropriate statistical model alongside clinical plausibility.

A recognised benefit of MSM, when there are sufficient data, is the ability to provide a deeper understanding of the primary analysis results for RCTs demonstrated through published secondary analyses of trial data. Le Rademacher *et al* compared an illness-death model with a time-dependent Cox model in cancer clinical trials using simulation [125]. The simulation study was informed by a re-analysis of an existing dataset and 4 different combinations of treatment effects were explored. One of the combinations specified a treatment effect on the transition from Illness state to death but no treatment effect on the other transitions. The results of the simulation study demonstrated that in this scenario the MSM was unbiased, however the treatment effect for overall survival estimated by the time dependent Cox model was biased towards the null until an interaction of treatment with entering the Illness state was included in the model. The authors demonstrated that correct model specification is critical and explained that the choice of model is dependent on the research question of interest. For example, MSM were able to estimate the effect of treatment on the transition from the Healthy to Illness state, but were unable to estimate the effect of illness on survival, whereas the time dependent Cox model was able to quantify the impact of illness on overall survival beyond the treatment effect [125]. In addition to deciding the MSM model structure (states and possible transitions), it is important to ensure the most appropriate statistical model is fitted.

4.1.2 Statistical model choices

A common assumption in MSM is the Markov property, where the transition intensities are assumed to depend on the history of the disease process only through the current disease state and the time since the origin. This assumption can be restrictive but is common for panel data to enable the likelihood to be computed [102,126]. If the model fit is poor, it is possible that the Markov assumption has been violated. Note that even if the model fits well, it is still possible that the Markov assumption may not hold, and the plausibility of the Markov assumption should be considered within the clinical context. An alternative assumption is that the transition intensities depend on the current disease state and the length of time spent there (semi-Markov model), however these models are challenging to fit to panel data because there is uncertainty on time of entry to each state, and therefore the duration spent in that state. Furthermore, in a semi-Markov model the time scale is usually set to zero (clock-reset) on entry to each state, rather than modelling the time since study entry (clock-forward) which may not be appropriate for randomised clinical trials where time from randomisation should be accounted for. There is little advice on which time-scale to use but should be informed by the clinical context [115]. Methods also exist when the Markov assumption does not hold, however these are less researched [119].

For some of the examples discussed here all transition times were observed exactly. However, in RCTs generally where the outcome requires detection of disease onset or a particular stage of disease, there is often a period of sub-clinical disease, before symptoms and signs are overt and the data are therefore interval censored [102]. For example, cancer trials may be designed based on assessing the effect of treatment on progression-free survival. Progression may be assessed via imaging or other tests and is assessed intermittently which means the exact time of progression is unknown. Among others, Zeng *et al* have researched the impact of ignoring this interval censoring on the design and analysis of cancer clinical trials [127,128]. They showed that a Cox model used to analyse the ‘true’ progression time and an MSM accounting for interval censored data both led to unbiased treatment effect estimates. However, it was noted that it is unrealistic to be able to model the true progression time. The authors proposed sample size criteria for cancer trials assessing progression-free survival taking into account interval censoring by using an illness-death model. A simulation study showed that ignoring interval censoring and designing the trial based on a Cox model of the 'true' progression times led to sample size estimates up to 16.5% lower than required for the stated power under an MSM design [128]. Therefore, the appropriate analysis for a trial should consider interval censoring, and be determined at the trial design stage so that the sample size estimation will provide adequate power for the final analysis.

The frequency of assessments for a clinical trial must be pre-specified as part of the protocol. In addition to exploring the impact of ignoring interval censoring Zeng *et al* [128] conducted a simulation study to explore potential efficiency gains of increasing the frequency of patient assessments. Using a 3 state illness-death model, they concluded that, in their context, the gain in power from increasing frequency of measurements was small in comparison to increasing the sample size. For example, doubling the frequency of assessments from 4 to 8 within the same length of follow-up led to approximately 5% increased power, whereas increasing the sample size by 33% led to approximately 10% increase. Therefore, for patient populations that are small or difficult to reach, increasing frequency of measurements may be appropriate, otherwise it may be more efficient to increase the sample size. The assessment of power was based on constraining the treatment effect on the 2 transitions out of healthy state to be equal, with time exiting the well state taken to be the estimand of interest. However, choice of the frequency of assessments in the design of a trial should take into account the intended analysis model, cost and available resources, all of which will be informed by the clinical setting.

Grüger [129] described four observation schedules and examined whether they were informative or not in a simulation study as follows:

1. Examination at regular intervals: This is common for RCTs where the observation scheme is set up in advance according to a protocol. Even if there is departure from the visit schedule this scheme can remain non-informative

because it was specified in advance. Note this scenario is the case for most RCTs, however reasons for delayed or missed assessments, or dropout from the observation scheme should be examined to assess the likely missing data mechanism. If the data are thought to be missing not at random (MNAR) the sensitivity of results to different assumptions about the missing data mechanism needs to be assessed.

2. Random sampling: This situation is less common to RCTs and is more applicable to observational studies. This observation schedule is non-informative providing selection of patients is independent of their disease history.

3. "Doctor's care": This may be present in some trials, particularly if the end of follow-up is based on the patient's health state. Provided that an observation does not depend on the health status at the time of the observation (although it may depend on the health status at previous observations), the sampling scheme is not informative.

4. Patient self-selection: In this observation scheme patients may direct whether or not they are assessed informed by their clinical condition. This observation scheme is informative and therefore bias may be present and needs to be accounted for in the analysis [129]. In this case it would be necessary to analyse the sampling and disease process simultaneously

.

If data are missing not at random, methods to jointly model the observation scheme and MSM can be used; these are discussed in more detail in Chapter 8.

The models discussed so far have been continuous-time MSM but discrete-time MSM may be appropriate in some settings. In this case, transitions are timed on a uniform grid where the time between one grid point and the next corresponds to a fixed time interval within which only one transition can occur [130]. This might be reasonable in an RCT where there is a protocol schedule for assessments, however, assessments are unlikely to be conducted at the same time point for all participants, for logistical reasons, or patient factors. Furthermore, if the time between pre-specified assessments is too long, changes in disease status may be missed. Continuous-time MSM can estimate unbiased transition rates and treatment effects when assessment time intervals vary, provided that the measurements themselves are independent of the fact that a measurement was taken. Williams *et al* compared the use of a discrete time MSM to a continuous time MSM for a cost effectiveness analysis and demonstrated that the results were sensitive to the choice of model highlighting the importance of checking that the assumptions for both clinical and cost-effectiveness analyses are reasonable [131].

4.1.3 Implementation

The distributions of the transition times may be semi-parametric where the baseline hazard for each individual transition intensities may be unspecified or fully parametric. Maximum likelihood estimation of transition rates and covariates is commonly used and is available in a range of software, the choice of which depends on the features of the MSM. Examples of readily available software packages include: the *msm* package in *R* which can be used to fit continuous-time Markov models and Hidden Markov models for panel data, the *flexsurv* and *mstate* packages in *R* can be used to fit non-parametric and semi-parametric MSM to data with exactly observed transition times, the *multistate* package in *Stata* can be used to fit parametric MSM to data with exactly observed transition times. Bayesian methods can also be used to estimate model parameters [132, 133] however the currently available software is limited and a Bayesian model may be more computationally intensive to fit compared to Frequentist methods because they require simulation methods to estimate model parameters which can take a longer time to converge [134].

4.1.4 Inference

Although MSM is a recognised method to provide a deeper insight into disease natural history and corresponding covariate effects, the interpretation of their results can be difficult [125]. An example was encountered in an application of MSM to predict disease recurrence and progression patterns in chronic myeloid leukaemia [135]. The authors used an illness-death model with transitions from initial state to remission, initial state to progression and remission to progression. The results showed that one treatment had a benefit in terms of progression when participants were in remission, however it was harmful in terms of progression from the initial state. This was an exploratory analysis and therefore underpowered, with few participants entering the progression state. In addition to low power for at least some transitions, multiple testing concerns have also been highlighted when covariate effects are tested on multiple transitions [124]. Cassarly *et al* conducted a simulation study to assess type I error and power, using a LRT to assess the overall effect of treatment on the disease process modelled using a MSM structure [136]. In this example, the authors used data from trials in the stroke setting and considered MSM with 4, 5, 6 and 7 states compared to repeated logistic regression. When the treatment effect was the same for all transitions bar one, MSM provided increased power compared to repeated logistic regression. However, when the treatment effects differed across all transitions, repeated logistic regression models were more powerful. Le Radermacher *et al* calculated power and type I error in their simulation study comparing the illness-death model to time dependent Cox models [125]. They concluded that type I errors were close to 0.05 for the MSM, however type I error and power were reported for individual transitions rather than the overall model. Therefore, a gap remains in understanding the impact on the overall type I error and power in concluding treatment effects when tested at the transition specific level.

In addition to interpretation challenges, caution should be exercised in making causal conclusions using MSM. The focus of this thesis is on the design of clinical trials where interest lies in designing trials according to hypothesis tests and ensuring that the type I and type II errors overall are of a specific size. If all patients start in a particular state at time zero, the time of randomisation in trials, then causal inference can be made for transitions from the starting state. Otherwise the risk set from other states will have altered because it will include patients who started in the later state and patients who have transitioned to that state after randomisation. If the treatment has an effect on early transitions then the remaining risk sets in treatment and control arms will no longer be consistent with the original randomisation. Therefore we cannot make causal claims for treatment effect estimates from states which are not the starting state. Causal inference methods such as those described by Gran *et al* should be considered if interest lies in determining the effect of intervening at specific stages of the multi-state disease process [137]. These approaches included artificially changing specific transition intensities, inverse probability of treatment weighting and G-computation. Causal inference methods will not be explored further in this thesis, but results of any MSM analyses should bear these considerations in mind.

4.1.5 Summary

Overall, general statistical methods and software to fit models to observed data [102, 138, 139], accessible resources including books [130, 140] and tutorials [115] are available to support those using MSM. However, challenges may be encountered in determining the number of states based on the clinical problem, and the availability of data [114]. A lack of data may lead to challenges in the interpretation of results and the appropriate hypothesis test(s). The next sections illustrate an application of MSM to motivating data to determine whether it is an appropriate method for the data and to inform a later simulation study evaluating the impact on power and sample size of using MSM compared to common binary or TTE methods. Note that because the thesis is focused on trials of prevention interventions, the models will be assumed to be progressive throughout.

4.2 Aim

The aim of this chapter is to re-analyse the PRESSURE and PRESSURE2 datasets using multi-state models.

Objectives

1. Define disease states at the patient and skin site level

2. Develop multi-state models to analyse patient level outcomes
3. Apply multi-state models to analysis skin site level outcomes without patient random effects

4.3 Methods

4.3.1 Notation

Let $\mathbf{Y}$ denote the disease process, which is defined by a stochastic process consisting of multiple random variables Y_t such that $\mathbf{Y} = \{Y_t | t \in (0, \infty)\}$, $Y_t \in S = \{1, 2, ...D\}$. S is the state space for Y_t and consists of all D possible values, or states, that could be occupied. Note that $t \in (0, \infty)$ denotes a process in continuous time, but discrete time could also be specified.

The stochastic process, $\mathbf{Y}$, can be represented through probabilities for transitions between state r to state s between time u and time $u + t$. That is,

$$P(Y_{u+t} = s \mid Y_u = r, \mathcal{H}_t) \tag{4.1}$$

where $r, s \in S$, $t, u \geq 0$ and the history of the process up to time t is denoted by $\mathcal{H}_t$.

Throughout this thesis the process is assumed to be time homogeneous. That is,

$$P(Y_{u+t} = s \mid Y_u = r, \mathcal{H}_t) = P(Y_t = s \mid Y_0 = r, \mathcal{H}_t) = p_{rs}(t). \tag{4.2}$$

Note that the assumption of time homogeneity can be assessed through model checks and if inappropriate, piecewise constant intensity models or non-parametric models can be considered [119].

These probabilities representing the stochastic process, also called transition probabilities, correspond to the (r, s) entry of a $D \times D$ transition probability matrix, $\mathbf{P}(t)$ such that

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1D} \\ p_{21} & p_{22} & \cdots & p_{2D} \\ \cdots & \cdots & \cdots & \cdots \\ p_{D1} & p_{D2} & \cdots & p_{DD} \end{pmatrix}, \tag{4.3}$$

where each row sums to 1, and the dependence on t has been suppressed. The stochastic process, $\mathbf{Y}$, is Markov if the probability of moving from state r to state s between time u and time $u+t$ does not depend on the history of the process, only the current state and the time interval. That is,

$$P(Y_{u+t} = s \mid Y_u = r, \mathcal{H}_t) = P(Y_{u+t} = s \mid Y_u = r). \tag{4.4}$$

Transition intensities for a Markov process are given by

$$q_{rs} = \lim_{\delta \to 0} \frac{P(Y_{t+\delta} = s \mid Y_t = r)}{\delta}, \tag{4.5}$$

where $r, s \in S$, $r \neq s$ and $t \geq 0$. These transition intensities correspond to the (r, s) entry of a $D \times D$ transition intensity matrix, $\mathbf{Q}$ such that

$$\mathbf{Q} = \begin{pmatrix} -\sum_{s \neq 1} q_{1s} & q_{12} & \cdots & q_{1D} \\ q_{21} & -\sum_{s \neq 2} q_{2s} & \cdots & q_{2D} \\ \cdots & \cdots & \cdots & \cdots \\ q_{D1} & q_{D2} & \cdots & -\sum_{s \neq D} q_{Ds} \end{pmatrix}, \tag{4.6}$$

where each row sums to 0.

The transition probabilities can be obtained from the transition intensities using results from matrix algebra, with

$$\mathbf{P}(t) = \exp(t\mathbf{Q}). \tag{4.7}$$

Eigenvalue decomposition may be used to derive $\mathbf{P}(t)$. To illustrate this, a simple 3-state MSM with only forward transitions will be used (see Figure 4.1). In this case, the transition intensity matrix, assumed to have constant transition intensities, is a 3 x 3 matrix with 3 distinct eigenvalues λ_1, λ_2 and λ_3, which satisfy the determinant equation $|\boldsymbol{Q} - \lambda \boldsymbol{I}| = 0$ where I denotes the 3 x 3 identity matrix. There is an eigenvector, $\boldsymbol{c}_i$, corresponding to each eigenvalue such that $(\boldsymbol{Q} - \lambda_i \boldsymbol{I})c_i = 0$ for $i = 1, 2, 3$. The probability matrix is derived as $\boldsymbol{P}(t) = \boldsymbol{U} \exp(\boldsymbol{D}t)\boldsymbol{U}^{-1}$, where $\boldsymbol{U}$ is the matrix of eigenvectors, and $\boldsymbol{D}$ is the diagonal matrix containing the eigenvalues. In the illness-death model, the eigenvalues are $\lambda_1 = -(q_{12} + q_{13}), \lambda_2 = -q_{23}, \lambda_3 = 0$, with corresponding eigenvectors $\boldsymbol{c}_1^T = (1, 0, 0), \boldsymbol{c}_2^T = (\kappa, 1, 0), \boldsymbol{c}_3^T = (1, 1, 1)$, where $\kappa = \frac{q_{12}}{q_{12}+q_{13}-q_{23}}$. Therefore, the probability matrix for the progressive illness-death model is given by

$$\boldsymbol{P}(t) = \begin{pmatrix} \exp(-(q_{12} + q_{13})t) & p_{12} & p_{13} \\ 0 & \exp(-q_{23}t) & 1 - \exp(-q_{23}t) \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.8}$$

where

$p_{12} = \kappa[\exp(-q_{23}t) - \exp(-(q_{12} + q_{13})t)]$

$p_{13} = 1 - (1 - \kappa) \exp(-(q_{12} + q_{13})t) - \kappa \exp(-q_{23}t)$.

However, when the models become more complicated, either through additional states, transitions or time dependent transition intensities, although eigenvalue decomposition can still be used, the transition probabilities cannot be expressed in closed form.

Covariates may be incorporated into transition-specific regression models as

$$q_{rs}(t) = q_{rs}((t|\boldsymbol{x}(t)) = q_{rs.0}(t) \exp(\boldsymbol{\beta}_{rs}^T \boldsymbol{x}(t)), \tag{4.9}$$

where $q_{rs.0}(t)$ denotes the baseline hazard, $\boldsymbol{\beta}_{rs}$ is a parameter vector of length p corresponding to the covariate vector $x(t)$ also of length p. A common assumption is that transition intensities are constant through time although piecewise constant hazards are useful for exploring whether this assumption is valid. As discussed in the MSM literature, there are situations where there are sparse data on specific transitions and solutions to this problem have been to combine states, but an alternative approach could be to constrain particular parameters to be equal to each other [141].

Suppose that individual i is observed at W timepoints, dropping the i for simplicity, the observed disease states are $\mathbf{y} = (y_1, y_2, ...y_W)$. Under the Markov assumption, the contribution of individual i to the likelihood function conditional on the first state is given by

$$\begin{aligned} L_i(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{x}) &= P(Y_2 = y_2, ..., Y_W = y_W \mid Y_1 = y_1, \boldsymbol{\theta}, \boldsymbol{x}) \\ &= (\prod_{w=2}^{W-1} P(Y_w = y_w \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}))C(y_W \mid y_{W-1}, \boldsymbol{\theta}, \boldsymbol{x}) \end{aligned}$$

Where the definition of $C(y_W \mid y_{W-1}, \boldsymbol{\theta}, \boldsymbol{x})$ depends on what state is observed at the W^{th} time point and whether censoring needs to be accounted for [130].

If the state is known at the W^{th} time point then

$$C(y_W \mid y_{W-1}, \boldsymbol{\theta}, \boldsymbol{x}) = P(Y_w = y_w \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}) \tag{4.10}$$

If the exact time of entry to the absorbing state, D, is observed at t_W then

$$C(y_W \mid y_{W-1}, \boldsymbol{\theta}, \boldsymbol{x}) = \sum_{s=1}^{D-1} P(Y_w = s \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}) q_{sD}(t_{W-1} \mid \boldsymbol{\theta}, \boldsymbol{x}) \tag{4.11}$$

Finally, if the state is right-censored at t_W then

$$C(y_W \mid y_{W-1}, \boldsymbol{\theta}, \boldsymbol{x}) = \sum_{s \in C} P(Y_w = s \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}) \tag{4.12}$$

where C denotes the set of possible states [102, 130].

The full likelihood function is given by

$$L = \prod_{i=1}^{N} L_i(\boldsymbol{\theta} \mid \mathbf{y}, x)$$

where θ is a vector of all model parameters. These transition specific regression models may be estimated by maximising the log likelihood. Specifically, Jackson developed the *msm* package within R, which uses eigenvalue decomposition, as described previously, to maximise the likelihood in terms of $\log(q_{rs})$ (to enhance convergence of the log-likelihood) to obtain estimates of the parameters that define q_{rs} [102].

4.3.2 State definition for the pressure ulcer data in PRESSURE and PRESSURE2

Outcome assessment

In each of these datasets the outcome assessment scales may be mapped onto a common set of states. Category 2+ PUs were a common endpoint of interest in the literature and will be defined as the absorbing state (Severe disease). Note that it is clinically appropriate to group Category 2+ PUs because development of a Category 2 PU will often prompt intensive therapy and further development is less common. Based on the international classification scale, Healthy, Altered and Category 1 PU will be represented by 3 transient PU states (Healthy, Pre-clinical, Mild disease). These are defined in Table 4.1 according to the classification used in each study. Each skin site is assessed and a PU classification is assigned. These are combined for the patient level analysis.

State definition

In this section, we define the empirical states observed for patients and for each skin site in the two illustrative datasets, including treatment of missing data. We start by defining multiple component outcomes for the case of PU data. For each patient we define a composite outcome with the k^{th} component representing the PU classification for the k^{th} skin site, $k = 1, ..., K$. Let w index the assessment number $w = 1, ..., W$. As in Section 3.3.1, $X_k(t_{iw})$ denotes the observed value for component k for patient i at the w^{th} assessment time, t_{iw}. Note that for simplicity of notation, we assume that all patients have the same number of assessments, although this is easily generalised to different numbers of assessments for each patient. The observed state for participant i at time t_{iw}, denoted by $Y(t_{iw})$ is then defined through

Table 4.1: States used in MSM and their associated PU classes used in the original PRESSURE and PRESSURE2 trials

State	**PRESSURE**	**PRESSURE2**
1 Healthy	Grade 0	Category 0
2 Pre-clinical	Grade 1a	Category A
3 Mild disease	Grade 1b	Category 1
4 Severe disease	Grade 2	Category 2
	Grade 3	Category 3
	Grade 4	Category 4
	Grade 5	Unstageable

a function, g, of the K components.

$$Y(t_{iw}) = g(X_k(t_{iw})) \tag{4.13}$$

For PU prevention trials, each participant has a maximum of $K = 14$ possible skin site assessments at each assessment time. The state space for $X_k(t_{iw})$ is $S_k = \{1, 2, 3, 4\}$ and the overall state for participant i at time t_{iw} is defined by taking the most severe state of observed skin sites

$$Y(t_{iw}) = \max_k (X_k(t_{iw})). \tag{4.14}$$

However, not all patients have complete data for all skin sites. The number of components that are observed at time t_{iw} for participant i can be denoted by $d_{iw}, d_{iw} \leq K$. There are several options for dealing with missing data depending on the assumptions we are prepared to make and we discuss these in detail in Chapter 8. At this stage we make the simplifying assumption that skin sites were healthy unless an assessment was recorded. That is,

$$Y(t_{iw}) = \begin{cases} \mathit{missing}, & \text{if } d_{iw} = 0 \\ \max_k \left(X_k(t_{iw})\right), & \text{otherwise} \end{cases} \tag{4.15a}$$

Since the focus of this thesis is in trials of prevention interventions, the disease process is assumed to be strictly progressive in line with Figure 4.4. Therefore, $Y(t_{iw})$ takes the most severe category observed up to time t_{iw}. This results in a 4 state progressive model with a single absorbing state (Severe disease). Patients start in the Healthy or Pre-clinical skin states at time zero (date of randomisation) and a Markov model is assumed so that transitions depend only on current disease stage, time since randomisation and covariates. This is a simplifying assumption that will be assessed by inspecting model fit. Note that in these case studies treatment started immediately and early skin changes could occur quickly, therefore the delayed treatment effect observed in the analyses of TTE outcomes (Chapter 3) are modelled explicitly by analysing the earlier skin changes through a multi-state model. This adds to the confidence in the assumption of time homogeneity for each transition. Patients were followed up for a fixed length of time or until discharge, death or onset of Severe disease.

Note that although not detailed here, methods are available for incorporating random effects into MSM. For example in the PsA setting a multi-level MSM was used to analyse the states hand joint locations with patient-specific random effects to account for the clustering of joints within patients [111]. Whilst these methods have been researched, readily available general software packages are not currently available. Given the findings from the analysis of the binary skin site level data, which suggested that at least 90% of the total variance was due to between patient variability and the small incidence of the Severe disease state at most skin sites, multi-level MSM has not been explored in this thesis.

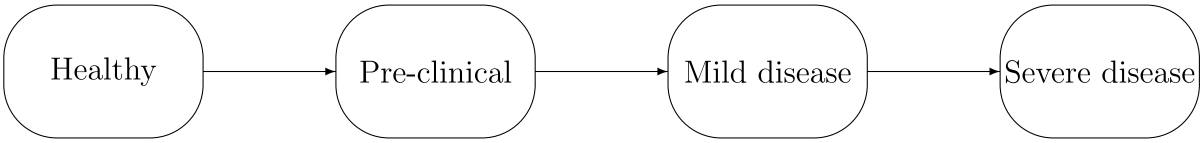


Figure 4.4: 4-state model

4.3.3 Independent variables and hypothesis testing

A 4-state progression MSM (Figure 4.4) was applied to the PRESSURE and PRESSURE2 patient level datasets adjusting for treatment (intervention vs control) as an independent variable on each transition. Similarly, a 4-state MSM was applied to both trial skin site level datasets. For both the patient and skin site level analysis, treatment was included as a patient level independent variable. For the skin site level analysis, skin sites were included as fixed effects; these were categorical variables with 7 levels in PRESSURE, and 14 in PRESSURE2.

The MSM analyses of the trial datasets was exploratory in order to establish whether MSM are a sensible method of analysis for these types of data. For the purposes of this chapter, statistical significance of independent variables were assessed at the 5% level. To test the overall significance of categorical variables, LRTs were used, whilst the effect of specific levels of independent variables on individual transitions was examined using point estimates of hazard ratios and corresponding Wald type 95% confidence intervals. Note that for confirmatory analysis multiple testing should be considered when assessing the effect of a variable on individual transitions, particularly for the primary analysis of a trial. An approach to multiple testing for MSM is described in Chapter 5.

4.4 Results

4.4.1 Patient level analysis

The number of observed state occupancies are presented in Table 4.2. Note that there are no backwards transitions observed by definition of the disease state where the most severe observation is carried forwards. The results from fitting a MSM to these data are presented in Table 4.3. The results are consistent with those observed in the analyses of the single binary or TTE endpoints, however the MSM provides further insight into the effect of treatment on PU development. In the PRESSURE trial, for the control group there was a relatively high transition rate from Healthy skin to Pre-clinical changes of 0.15 $(0.13, 0.16)$ per day, with only a very slight increase in transition rates in the intervention group (HR (95% CI)= 1.05 $(0.85, 1.30)$). Transition rates to more severe skin states were lower and, as a result, the HRs for the treatment effects had wide confidence intervals and were not significantly different from one (no difference) at the 5% level. There were more events observed in the PRESSURE2 trial, providing greater power to assess treatment effects, especially those affecting later transitions. In this trial, transition rates in the control group were generally lower than in the original PRESSURE trial from Healthy to Pre-clinical changes AT 0.06 $(0.05, 0.07)$ per day (Table 4.3). There was a non-significant decrease in transition rates for the intervention group between Pre-clinical and Mild PU in PRESSURE2, but a significant decrease in Severe PU onset conditional on prior development of a Mild PU (HR (95% CI)= 0.5 $(0.35, 0.71)$). Inspection of the expected versus observed prevalence for each state in Figure 4.5 suggests that the model fit is adequate for both datasets.

4.4.2 Skin site level

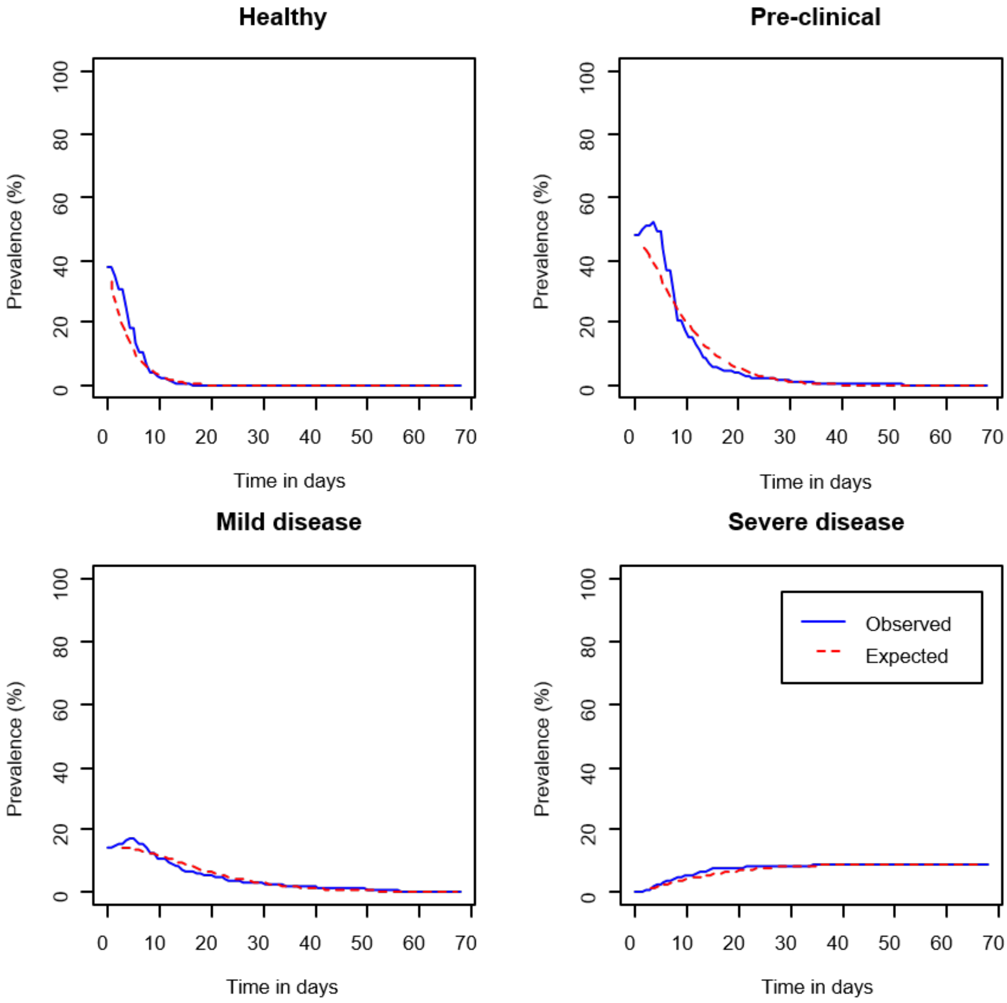
The observed state occupancies for all skin sites in the PRESSURE and PRESSURE2 trial are presented in Table B.2 and by individual skin sites in Appendix B. There were a total of $115, 574$ transitions compared to $4, 843$ in the patient level dataset. The observed state occupancies by individual skin site indicate that

Table 4.2: Observed state occupancies in illustrative datasets (patient level)

Dataset	**From state ↓**	**To state →**			
PRESSURE		1	2	3	4
	1	684	320	23	8
	2	0	$2,237$	195	56
	3	0	0	1234	89
	4	0	0	0	0
PRESSURE2		1	2	3	4
	1	595	138	11	7
	2	0	$5,365$	152	78
	3	0	0	$1,195$	42
	4	0	0	0	0

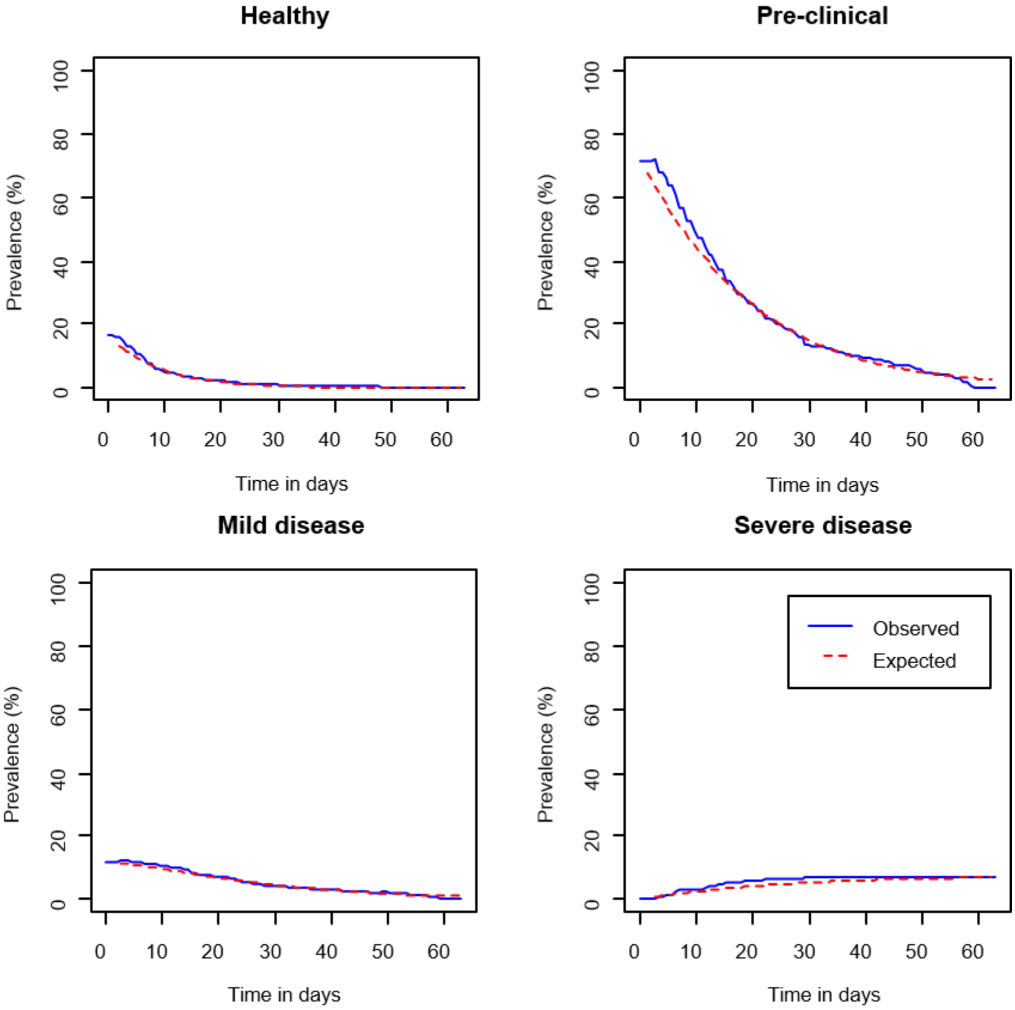
Table 4.3: MSM applied to patient level data from the PRESSURE and PRESSURE2 trials

Dataset	**Transition**	**Baseline transition intensity (95% CI)**	**HR (95% CI) (Intervention vs control)**
PRESSURE	$1 \to 2$	$0.15\ (0.13, 0.16)$	$1.05\ (0.85, 1.30)$
	$2 \to 3$	$0.04\ (0.03, 0.04)$	$0.94\ (0.75, 1.19)$
	$3 \to 4$	$0.03\ (0.03, 0.04)$	$0.76\ (0.55, 1.05)$
PRESSURE2	$1 \to 2$	$0.06\ (0.05, 0.07)$	$1.09\ (0.79, 1.50)$
	$2 \to 3$	$0.01\ (0.01, 0.01)$	$0.85\ (0.66, 1.09)$
	$3 \to 4$	$0.02\ (0.02, 0.03)$	$0.50\ (0.35, 0.71)$



(a) PRESSURE

Figure 4.5: Expected vs observed prevalence for fitted MSM



(b) PRESSURE2

Figure 4.5: Expected vs observed prevalence for fitted MSM

Table 4.4: Observed state occupancies for all skin sites in the PRESSURE and PRESSURE2 datasets

Dataset	**From state** ↓	**To state** →			
PRESSURE		1	2	3	4
	1	16,261	1,465	117	22
	2	0	9,517	476	71
	3	0	0	3,749	112
	4	0	0	0	0
PRESSURE2		1	2	3	4
	1	62,776	3,874	96	40
	1	0	42,078	260	93
	3	0	0	2,305	50
	4	0	0	0	0

there were different patterns of disease progression for different skin sites. For example, the back, ischial tuberosities and hips were more likely to remain in the Healthy state compared to the other skin sites, whilst the sacrum, buttocks and heels were more likely to develop a Category 2+ PU with 146 (79.8%) observed at these skin sites.

The results of a MSM for the skin site level data are reported in Table 4.5 and Table 4.6. Including skin site as a categorical fixed effect in the analysis of the skin site level trial datasets led to similar hazard ratios (intervention compared to control) for all transitions when compared to the analysis of the patient level data, but with narrower confidence intervals due to the increased sample size. For example, including skin site as a 14 level fixed effect in the analysis of the PRESSURE2 data, there was a significant reduction in the transition from Mild to Severe disease (HR (95% CI)= 0.54 (0.40, 0.74)), all else being equal.

For both datasets, the sacrum is the reference category against which other skin sites are compared. Transition rates for the buttocks were not significantly different from the sacrum across all transitions. For the PRESSURE dataset, in the heels there was a higher rate of transition from the Healthy to Pre-clinical state than the sacrum (HR (95% CI)= 1.55 (1.32, 1.82)), but this reduces for the later transitions to Mild and Severe disease states. The results for the heels are similar for the PRESSURE2 dataset. All other skin sites were observed to have lower transition rates between the Healthy, Pre-clinical and Mild disease states when compared to the sacrum. However, the point estimates and precision of the skin site fixed effects for the transition from the Mild state to the Severe state highlight the lack of data available, with wide confidence intervals observed (see the left hip for example). All of the confidence intervals straddled 1 indicating a lack of evidence to conclude a difference in the probability of moving out of state 3 for any skin site compared to the sacrum, apart from the left ankle.

4.5 Discussion

Summary of results

A 4-state MSM was fitted to longitudinal datasets for both trials and indicated that, for PRESSURE2, the treatment effect was not statistically significant on transitions between Healthy and Pre-clinical disease, and between Pre-clinical and Mild disease, but there was a substantial and significant treatment effect for the transition between Mild and Severe disease. This finding is consistent with the Kaplan-Meier estimates in Figure 3.1 that suggested there was a delayed treatment effect, as it shows that the treatment effect was mainly on the later transition and was only evident when patients passed through the intermediate states. Note that the confidence intervals were derived using asymptotically unbiased standard errors, however the estimated variance is downwardly biased for small samples and therefore alternative methods such as bootstrapping could be used to check the results of an important analysis such as the primary analysis of a trial dataset [102]. The plots of the observed and model-fitted prevalence in each of the four states in the illustrative datasets demonstrate reassuring agreement for the patient level analysis, suggesting that the Markov assumption (transitions depend only on current disease stage, time since

Table 4.5: MSM analysis of skin site level data in PRESSURE (fixed effects only)

Variable	**Analysis results**		
	HR (95% CI)		
	$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$
Intervention			
Intervention	1.01 $(0.91, 1.11)$	0.95 $(0.82, 1.11)$	0.78 $(0.59, 1.03)$
Control (reference)	-	-	-
Skin sites			
Sacrum (reference)	-	-	-
Left buttock	1.03 $(0.87, 1.22)$	1.03 $(0.82, 1.30)$	1.06 $(0.72, 1.57)$
Right buttock	1.04 $(0.88, 1.23)$	0.94 $(0.74, 1.19)$	1.26 $(0.87, 1.84)$
Left hip	0.05 $(0.03, 0.07)$	0.34 $(0.14, 0.82)$	2.46 $(0.59, 10.21)$
Right hip	0.06 $(0.04, 0.09)$	0.44 $(0.22, 0.90)$	1.08 $(0.26, 4.44)$
Left heel	1.55 $(1.32, 1.82)$	0.77 $(0.61, 0.98)$	0.32 $(0.20, 0.53)$
Right heel	1.55 $(1.32, 1.82)$	0.68 $(0.53, 0.87)$	0.28 $(0.16, 0.47)$

Table 4.6: MSM analysis of skin site level data in PRESSURE2 (fixed effects only)

Variable	**Analysis results**		
	HR (95% CI)		
	$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$
Intervention			
APM	1.07 (1.01, 1.14)	0.77 (0.65, 0.92)	0.54 (0.40, 0.74)
HSF (reference)	-	-	-
Skin sites			
Sacrum (reference)	-	-	-
Back	0.20 (0.17, 0.24)	0.28 (0.15, 0.51)	2.25 (0.94, 5.37)
Left buttock	0.89 (0.77, 1.04)	0.90 (0.68, 1.19)	1.27 (0.80, 2.00)
Right buttock	0.89 (0.77, 1.03)	0.85 (0.64, 1.12)	1.21 (0.76, 1.93)
Left ischial	0.37 (0.32, 0.43)	0.19 (0.10, 0.37)	1.44 (0.51, 4.06)
Right ischial	0.35 (0.30, 0.41)	0.18 (0.09, 0.35)	1.11 (0.39, 3.13)
Left hip	0.10 (0.08, 0.12)	0.12 (0.04, 0.37)	3.78 (0.90, 15.86)
Right hip	0.12 (0.09, 0.14)	0.14 (0.05, 0.38)	0.80 (0.11, 5.83)
Left heel	1.43 (1.24, 1.65)	0.43 (0.31, 0.59)	1.13 (0.66, 1.96)
Right heel	1.43 (1.23, 1.65)	0.33 (0.23, 0.46)	1.02 (0.58, 1.81)
Left ankle	0.67 (0.58, 0.78)	0.10 (0.05, 0.19)	2.64 (1.03, 6.77)
Right ankle	0.66 (0.57, 0.77)	0.17 (0.10, 0.29)	0.78 (0.28, 2.19)
Left elbow	0.69 (0.59, 0.79)	0.11 (0.06, 0.21)	0.47 (0.11, 1.94)
Right elbow	0.69 (0.60, 0.80)	0.21 (0.13, 0.34)	1.81 (0.87, 3.76)

randomisation and covariates) holds over the duration of each study. This may not be the case for studies with a longer period of study, in which case semi-Markov models could be considered where transition out of disease states depends on the length of time spent in the state itself [130]. However, there are challenges in fitting semi-Markov models to interval censored data because the length of time spent in the state is unknown (because the time of entry and exist is unknown), and a number of assumptions may be required to simplify the model [130].

After applying methods to the patient level dataset, an MSM was used to analyse skin site level data. This model did not account for patient random effects but demonstrated that the effect of treatment was estimated to be similar in magnitude and the treatment effect obtained from the patient level analysis. Furthermore, the transitions through the disease process were largely consistent for individual skin sites, with the exception of the heels, which may have a higher propensity to move from the Healthy to Pre-clinical disease states. As noted in Chapter 3, the analyses for the remainder of the thesis are at the patient level, but information provided for individual skin sites is considered further in the context of missing data in Chapter 8.

The results from these analyses indicate that there is merit in using the longitudinal data to understand the natural history of the disease and to identify where treatment may have most benefit. There are differences in the estimated treatment effect for different transitions, which are obscured by the use of a model with a single outcome, such as TTE. It is of interest to understand how a MSM, which is able to estimate treatment effects at different stages of the disease process, could be used to inform the design and analysis of a future RCT.

Design of RCTs

Despite potential improved trial efficiency and greater understanding of treatment mechanisms for MSM, possible barriers to their use for primary analysis of RCTs have been raised [125]. For instance, MSM have a more complicated structure than simple regression models, so that a number of estimands may be of interest.

Although MSM can be used to calculate traditional endpoints, such as incidence of a particular event or disease category, choice of the specific structure of the model is not necessarily clear-cut. Further, Manzini et al [124] highlighted the need for sufficient numbers of observed state occupancies throughout the MSM structure and difficulties in dealing with missing data in this context.

Previous applications of MSM have generated research questions or helped to refine the patient population to be studied [116]. Applying MSM to the PU trial datasets suggested that the benefit of intervention was starting to emerge for the transition from Pre-Clinical to Mild disease with a stronger treatment effect on the transition from Mild to Severe disease. The results of the MSM could therefore suggest that clinical trials of PU prevention interventions should be conducted in patients who are already in the Mild disease state. This would lead to a higher proportion of patients entering the severe disease state, but with only 15% of patients recruited in the Mild disease state, the trial would take a longer time to recruit compared to a trial recruiting high risk patients in the Healthy and Pre-Clinical disease states. Therefore, the benefit of a smaller sample size may be outweighed by the length of time it would take to identify eligible patients. Whilst it may not be sensible to restrict the patient population to those who are in the Mild disease state, it may be sensible to recruit those who are at least in the Pre-clinical disease state. However, whilst the results have indicated where there is an effect of treatment, the effects on individual transitions are conditional on reaching each state and cannot be interpreted as a causal relationship [140]. Thus, before refining the patient population, additional analysis to explore the direct treatment effect on individual transitions should be conducted such as those proposed by Gran *et al* [137].

Incidence of death or severe disease may be easier to define and is often the estimand of choice in RCTs, but such endpoints may occur rarely, resulting in the need for very large trials. Assessing treatment effects on intermediate health states by using MSM may result in smaller trials, however the impact of using MSM to inform treatment effects and increase power of a RCT is unclear. Furthermore, health technology assessment often requires economic evaluation in addition to clinical effectiveness in order to guide decision making [142]. Aligning the primary analysis model with models used in the health economics analyses may aid the interpretation of the two analyses together [143].

Chapter 2 identified various decisions for characteristics of PU prevention trials including: overall size of the trial, the length of patient follow up and the intervals between patient assessments. MSM may provide insight into how to specify these characteristics for future research, but also how they might be incorporated into the analysis using MSM. These are explored through a simulation study in Chapter 5.

Chapter 5

Power and sample size requirements

5.1 Introduction

The application of MSM to the illustrative PU datasets in Chapter 4 demonstrated that MSM are an appropriate analysis method for ordinal outcome data collected longitudinally at pre-specified time points. The analyses were shown to provide a deeper insight into the effectiveness of treatment on disease progression and utilised more of the data collected during the trial compared to methods based on models for binary or TTE outcomes. It is therefore of interest to explore the potential use of MSM as the primary analysis method at the design stage of the trial.

In order to design a clinical trial the estimands and corresponding treatment effects must be defined in advance. The International Consortium for Harmonisation published an addendum to the existing $E9$ statistical principles for clinical trials, which included the importance of estimands and sensitivity analysis in clinical trials [144]. An estimand is defined as *a precise description of the treatment effect reflecting the clinical question posed by a given clinical trial objective* and is defined through the following components [144] described within the context of PU prevention trials.

- The **treatments** to be assessed, which form the 'arms' of the clinical trial. In PU prevention trials the treatments may be devices that relieve pressure either for the whole body, such as mattress provision or, for specific skin sites such as offloading devices for the heels.

- The **population** of patients for which the clinical question is relevant, and therefore the population who will benefit from the treatments being implemented in practice if recommended based on the results of the trial. In PU prevention trials this is commonly defined as those patients who are at high risk of developing a PU based on criteria such as a PU specific risk assessment tool identified in the literature review in Chapter 2.

- The **variable (or endpoint)** collected for each patient in order to answer the clinical question. It is critical for the research team to consider how often endpoints should be collected, and the length of follow-up. In Phase III trials, it may be appropriate for the frequency of assessments to coincide with usual clinical practice for pragmatic reasons and to minimise burden for participants. Similarly, the length of follow-up should be justified based on the clinical problem. However, the assessment schedule should also be such that clinically relevant changes in the endpoint are observed. The literature review showed that for PU prevention trials, the endpoint is commonly the occurrence of a new PU with the variable (PU classification) collected longitudinally for each patient at each skin site.

- **Intercurrent events** must be considered in the description of the clinical question to provide context for the treatment effect to be estimated. Examples of such events are treatment switching or discontinuation, which is a risk in the motivating PU prevention trials because the interventions may both be in routine use [22]. Treatment switching may lead to bias in the estimated treatment effect and should be considered in either the analysis method or at least the interpretation of the treatment effects supported by sensitivity analyses [145]. Compliance with the interventions is outwith the scope of this thesis and was therefore not assessed in the re-analysis of the case studies but an understanding of any likely non-adherence is critical to the design of the trial and interpretation of trial results.

- A **population-level summary** that estimates the treatment effect based on the endpoint is also pre-specified. This will be informed by the selected analysis method. For methods based on models for binary or TTE outcomes the population level summary is the odds ratio and hazard ratio respectively. For MSM the population-level summary consists of multiple ratios of transition intensities or transition hazards [130]. For the purposes of this thesis, the treatment effect obtained from MSM is defined as the set of hazard ratios for each of the transitions of interest.

Once an estimand has been defined, the required sample size for the trial can be determined. For the binary and TTE analysis methods described in Chapter 3, there are readily available formulae to calculate the required sample size for a clinical trial [97, 107]. For MSM there are relatively few examples of sample size calculations for multi-dimensional treatment effects. Wu and Cook proposed sample size formulae for the design of trials using a continuous time MSM to assess the effect of treatment on recurrent and terminal events [146]. The state space was given by $S = \{0, 1, ..., D\}$ where states $0, 1, 2, ...$ denoted the number of recurrent events and D denoted the absorbing terminal event (death). There were two parameters on which the sample size calculations were based; the log hazard of a recurrent event, and the log hazard of a terminal event denoted by β and θ respectively. Note that β was assumed to be the same for each recurrent event. A partial score statistic was used to derive formulae for the sample size required for each comparison and the sample size required for the trial was the maximum of the two calculations. They demonstrated an example trial design to evaluate the effectiveness of a new treatment for the prevention of skeletal complications in breast cancer patients with skeletal metastases. The hypothetical trial was designed to detect whether a new treatment was superior in terms of the skeletal complication occurrence and/or whether it was superior in terms of mortality. They assumed an overall type I error rate of 5% but due to multiplicity concerns arising from having two comparisons, they made a Bonferroni adjustment so that the type I error rate for each comparison was 2.5%. The sample size calculations yielded sample sizes of 700 and 707 to detect log hazard ratios of $\beta = -0.22$ and $\theta = -0.11$ respectively with 90% power, therefore concluding that 707 was the minimum required sample size for the trial. The proposed methods assumed that all patients started in state 0 at the point of randomisation, which would not be appropriate for the PU trial case studies. Furthermore, their methods were based on the assumption that all event times were observed and they recommended that further work be conducted for the design of trials with panel data.

Zeng *et al* proposed sample size criteria for cancer trials assessing progression-free survival taking into account that the assessment of progression is subject interval censoring [127, 128]. Illness-death models were used with a constraint on the treatment effect on the 2 transitions out of healthy state to be equal, with time exiting the well state taken to be the estimand of interest. A simulation study showed that ignoring interval censoring and designing the trial based on a Cox model of the 'true' progression times led to sample size estimates up to 16.5% lower than required for the stated power under an MSM design [128]. They also showed that the gain in power from increasing frequency of measurements was small in comparison to increasing the sample size. For example, doubling the frequency of assessments from 4 to 8 led to approximately 5% increased power, whereas increasing the sample size by 33% led to approximately 10% increase. However, choice of the frequency of assessments in the design of a trial should take into account the intended analysis model, cost and available resources, all of which will be informed by the clinical setting.

To date, there is no analytical solution to calculate the sample size for a clinical trial where a k state progression MSM for panel data is the intended primary analysis method. In the absence of a formula to determine the required sample size for such trials, simulations can be used to explore sample size estimates [147]. Some examples have been considered in the MSM literature such as those published by Cassarly *et al* [136] and Le Radermacher *et al* [125] described in Chapter 4, but a gap remains in understanding the impact on the overall type I error and power in concluding treatment effects when tested at the transition specific level for MSM applied to panel data.

5.2 Aim

The aim of this chapter is to conduct a simulation study to assess the impact on bias, coverage, power and sample size requirements of using different statistical models and methods to analyse data collected in disease prevention trials.

Objectives

The objectives of the simulation study are to compare logistic regression, Cox PH regression and 4 state progression multi-state Markov models in terms of power, bias and coverage for the following components:

1. Length of follow-up.
2. Assessment intervals.
3. Baseline transition intensities.
4. Treatment effects.

5.3 Methods

The ADEMP general framework for the design of simulation studies has been proposed by Morris *et al* and is widely used by statisticians and clinical trialists [148]. The framework comprises an Aim , Data generating mechanism, Estimand and target, Methods to be evaluated and Performance measures (ADEMP). The aim is described in Section 5.2 and the remaining components are described in Sections 5.3.1 to 5.3.5. It is first important to define the hypothesis testing procedure for the MSM, which follows in the next section.

5.3.1 Hypothesis testing procedure

The conclusion of a statistically significant treatment effect is based on a hypothesis test. For binary and TTE methods, the hypothesis test is based on a single measure of the treatment effect. However, for MSM the testing procedure needs to encompass multiple testing considerations due to the multiple population level summaries.

In a two arm trial with a single treatment effect such as an odds ratio or hazard ratio, denoted by Δ, the null hypothesis is defined as

$$H_o : \Delta = 1 \tag{5.1}$$

The alternative hypothesis is given by

$$H_A : \Delta \neq 1 \tag{5.2}$$

The appropriate hypothesis test is conducted by calculating the relevant test statistic and the probability that the value of the test statistic, or one more extreme, would have been observed under the null hypothesis. If the probability is less than a nominal significance level, often 5% then the treatment effect is described as statistically significant.

For PU prevention trials where a progression model has been proposed, interest lies in detecting an improvement for any transition. Suppose there are d transitions for which a treatment effect is of clinical interest. There will then be a global null hypothesis with d comparisons such that

$$H_0 : \Delta_i = 1, \forall \quad i \in \{1, 2, ..., d\} \tag{5.3}$$

where Δ_i denotes the treatment effect (hazard ratio) for the i^{th} transition. The alternative hypothesis is then

$$H_A : \Delta_i \neq 1, \text{for at least one} \quad i \in \{1, 2, ..., d\} \tag{5.4}$$

If d comparisons were conducted it would present a multiple testing problem.

Multiple testing would lead to an inflated type I error if not accounted for, because there are more opportunities to incorrectly conclude a significant treatment effect. There are various adjustments that could be made depending on the multiplicity concerns [149]. The Bonferroni adjustment is considered the simplest, where each comparison is tested against the overall significance level (e.g. 0.05) divided by the total number of comparisons. This is the approach taken by by Wu and Cook [146] who had 2 comparisons and assessed each according to a 2.5% significance level [146]. This method guarantees that the family wise error rate is less than the overall significance level, but may be overly conservative as the number of comparisons increases and as the correlation between test statistics increases [149]. Holm developed a procedure based on the Bonferroni correction where the p-values from each comparison are assessed based on a closed testing principle [150]. Place the p-values in order of smallest to largest

$$p_1 \leq p_2 \leq ... \leq p_d \tag{5.5}$$

Statistical significance is concluded by examining the p-values in order compared to a reference value such that

$$p_i \leq \frac{\alpha}{d - i + 1} \tag{5.6}$$

Therefore if $p_1 \leq 0.05/d$ statistical significance will be concluded, and the p-values are assessed sequentially until the first i such that $p_i > \frac{\alpha}{d-i+1}$ at which point no further comparisons under the null hypothesis are rejected. Hochberg later developed a further testing procedure based on Bonferroni correction also based on sequential ordering of the p-values from each comparison that is considered more powerful than both the Bonferroni and Holm procedures [151]. In this procedure, the p-values are assessed in descending order of magnitude such that

$$p_d \geq ... \geq p_2 \geq p_1 \tag{5.7}$$

Statistical significance is concluded by examining the p-values compared to a reference value in line with 5.6.

Therefore, if all d p-values are less than 0.05, statistical significance is concluded, but if $p_d > 0.05$ then statistical significance will only be concluded if $p_{d-1} \leq 0.05/2$ and so on. The Hochberg procedure will be adopted throughout the simulation study because it is considered more powerful and the family wise error rate will be assessed to ensure that it is equal to approximately 0.05. The Hochberg procedure only maintains the desired family wise error rate (FWER) if the comparisons are independent or conditionally independent [152].

Note that a LRT comparing the model with and without treatment effects could be considered, which was the approach taken by Cassarly *et al* [136], however this is less intuitive when designing a RCT where a minimally clinically important difference should be specified for each transition, or for a subset of transitions that are of clinical interest.

5.3.2 Data Generating Mechanism

The data were generated from a 4-state progression model as shown in Figure 4.4 at the patient level rather than the skin site level for simplicity but the simulations could be extended to skin site level in future work.

Exponential survival times were randomly generated using baseline transition intensities informed by the illustrative datasets, and varying treatment effects. Exponential censoring times were randomly generated at a rate of 5% per unit time (day) from each disease state to reflect loss to follow-up, independent of treatment allocation. In addition to the baseline transition intensities, and treatment effects, length of follow-up and assessment frequency were also varied with each factor assessed for the following total sample sizes for a 2 arm trial with equal allocation ratio: $N = 100, 200, 500, 1000, 2000$. There was a total of 115 scenarios considered as presented in Table 5.1.

The *base case scenario* assumed that: patients were followed up for a maximum of 60 days, with a moderate treatment effect on each transition ($e^{\beta_{12}} = e^{\beta_{23}} = e^{\beta_{34}} = 0.67$). The baseline transition intensities were informed by the analysis of the case study datasets so that there was a high risk of transitions 1 → 2 and 2 → 3 and a moderate risk of transition 3 → 4 ($q_{12.0} = q_{23.0} = 0.05, q_{34.0} = 0.03$). The base case assumed daily assessments rather than every 2 or 3 days as this is expected to provide more power, and will therefore be a useful scenario to compare alternative assessment schedules with. In all scenarios, the proportions of patients in states 1 (Healthy), 2 (Pre-clinical) and 3 (Mild disease) at baseline ($t = 0$) were 15%, 70% and 15% respectively to reflect starting states observed in PRESSURE2. In each scenario, patients were allocated in a 1:1 ratio to one of two treatment groups (intervention and control). The null model where $e^{\beta_{12}} = e^{\beta_{23}} = e^{\beta_{34}} = 1$ was assessed to check that the type I error was equal to 5% and to serve as confirmation that the simulation code was working as expected.

5.3.3 Estimand and target

The estimand is defined as the estimated coefficients for treatment. For logistic regression it is the odds ratio, for the Cox PH model it is the hazard ratio, and for MSM it is the set of hazard ratios for each of the transitions.

5.3.4 Methods to be evaluated

The methods evaluated under each scenario were the logistic regression model, Cox PH model, and 4 state MSM whereby the treatment effects were either,

- Model A: unconstrained, i.e. $\beta_{12} \neq \beta_{23} \neq \beta_{34}$,
- Model B: completely constrained, i.e. $\beta_{12} = \beta_{23} = \beta_{34}$,
- Model C: partially constrained for early transitions, i.e. $\beta_{12} = \beta_{23} \neq \beta_{34}$,or
- Model D: partially constrained for later transitions, i.e. $\beta_{12} \neq \beta_{23} = \beta_{34}$.

Models B to D use constraints, which can simplify models, and may lead to increased power and precision if they are correct. Model B may be clinically plausible if the treatment is expected to have a uniform effect across all transitions, for example if PU onset and progression were part of a smooth progressive process. Model

C represents a similar treatment effect on early transitions, which changes once a Mild PU develops. One example might be a treatment that delays onset of a PU, but has little benefit once a Mild PU has developed. Model D represents a different effect during initial skin changes but once a Mild PU has developed the treatment affects later transitions to the same extent. For example, a treatment with a strong preventative effect but little value for intervention might fit this model.

5.3.5 Performance

Treatment effects in the logistic regression and Cox PH model were assessed using the Wald statistic and significance concluded at the 5% level. Similarly, for the completely constrained MSM, which has a single common treatment effect, the Wald statistic from the maximum likelihood estimation was calculated. The unconstrained and partially constrained MSM had three and two treatment effects respectively, so that Hochberg's multiple testing procedure based on Bonferroni corrections was adopted in order to maintain the overall 5% type I error as described in Section 5.3.1. The comparisons were assumed to be independent given the Markov assumption, however the FWER was tested under the null hypothesis. In this case, empirical power was reported overall by examining the Wald statistic for the treatment effect on each transition; for example, for the unconstrained model 5% significance was concluded if either (*i*) all three transitions were significant at the 5% level, or (*ii*) at least two treatment effects were significant at the 2.5% level, or (*iii*) at least one treatment effect was statistically significant at the 1.67% level. Empirical power was calculated as the proportion of times a statistically significant treatment effect was concluded. Bias of the estimates was examined by comparing the distribution of point estimates to the true value. Coverage was also assessed by assessing the proportion of times the 95% CI contained the true parameter value; adequate coverage was concluded if the CI included the true value 95% of the time. The Monte Carlo standard error was calculated for each performance measure in line with recommendations for simulation studies [148].

A formal sample size calculation for the number of simulations was not con-

Table 5.1: Factors varied in simulation study with base case settings indicated by an asterisk

N	**Follow-up length**	**Assessment frequency**	**Baseline transition intensities,** $\boldsymbol{q}_0$	**Treatment effects (hazard ratios),** $\exp(\boldsymbol{\beta})$
100	60 days*	Daily*	$(0.05, 0.05, 0.03)^*$	$(1.00, 1.00, 1.00)$
200	30 days	Every 2 days	$(0.01, 0.01, 0.01)$	$(0.67, 0.67, 0.67)^*$
500	14 days	Every 3 days	$(0.01, 0.01, 0.05)$	$(0.50, 0.50, 0.67)$
$1,000$	7 days	Every 7 days	$(0.01, 0.05, 0.01)$	$(0.67, 0.67, 0.50)$
$2,000$		Every 14 days	$(0.05, 0.01, 0.01)$	$(0.90, 0.90, 0.67)$
			$(0.01, 0.05, 0.05)$	$(0.67, 0.67, 0.90)$
			$(0.05, 0.01, 0.05)$	
			$(0.05, 0.05, 0.01)$	
			$(0.05, 0.05, 0.05)$	

N=Total sample size, $\boldsymbol{q}_0 = (q_{12.0}, q_{23.0}, q_{34.0})$, $\exp(\boldsymbol{\beta}) = (e^{\beta_{12}}, e^{\beta_{23}}, e^{\beta_{34}})$
* denotes the base case

ducted, but a total of $1,000$ simulations were run for each scenario. The same datasets were used to compare statistical methods but different datasets were generated for each scenario being considered.

The code to generate the datasets and apply the methods to be evaluated is presented in Appendix C.1.

5.4 Results

5.4.1 Power and Type I error

For the null case, datasets were generated according to the base case, with the exception of the treatment effect $\exp(\boldsymbol{\beta}) = (1, 1, 1)$. Results for the 6 models (4 MSM, logistic and Cox PH regression) applied to the null data are overlaid in Figure 5.1 and indicate that the type I error was close to 5%, as expected, provided the sample size is at least 100.

For the base case, where the treatment effect was equal to 0.67 on each transition, all MSM had greater power compared to the binary logistic regression model and the Cox PH model. For example, with 500 patients the binary and Cox models provide power of 57.5% and 68% respectively, the MSM with no constraint on the treatment effect provides power of 72.5% and MSM with some constraint(s) applied to the treatment effect provide a minimum of 80% power in this case (Figure 5.1). Note that throughout the results in this chapter, the Monte Carlo Standard Error for the estimates of power were considered sufficiently small at < 0.016 and are provided in Appendix C.3.

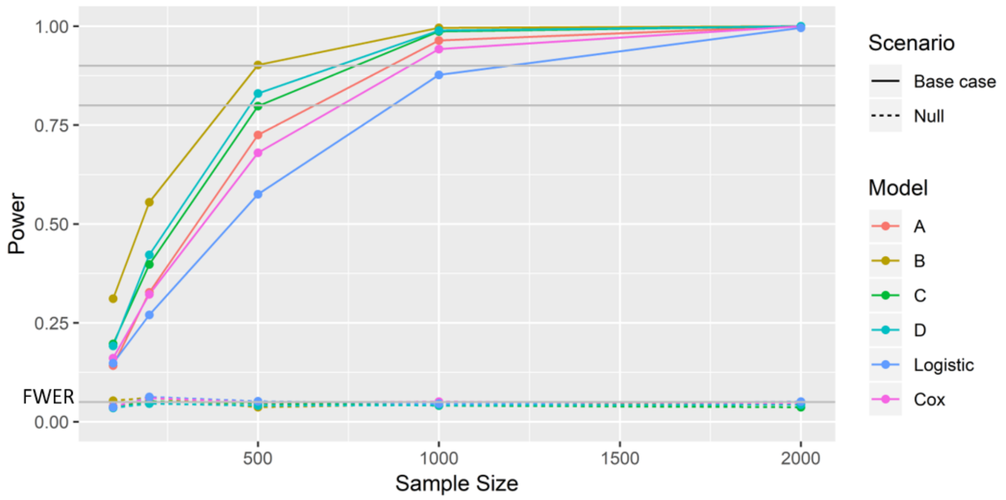


Figure 5.1: Power of detecting a significant treatment effect overall according to sample size for the base case (Maximum follow-up=60 days, Assessment frequency=daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$) and Family Wiser Error Rate (FWER) under the Null ($\exp(\boldsymbol{\beta}) = (1, 1, 1)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Length of follow-up

The simulation study explored maximum lengths of follow-up of 7 days, 14 days, 30 days and 60 days (the base case) with all other parameters remaining as in the base case. For all proposed trial durations, the MSM had greater power than the corresponding Cox and logistic regression analyses when applied to data with the same follow up periods. Figure 5.2 shows results for the unconstrained MSM with various durations of follow up compared to logistic and Cox models with 60 days follow-up. The results indicated that, when fitting an unconstrained MSM, a follow-up period of 60 days provided some additional efficiency compared to a follow-up period of 30 days, whilst a follow-up period of 7 or 14 days led to substantially reduced power, largely due to the low number of transitions to the absorbing state. Notably, the unconstrained MSM with 30 day follow up had similar power to a Cox model with data collected for 60 days (Figure 5.2).

In Appendix C.2.1 plots for all types of MSM explored in addition to the Cox and logistic regression models for 30 days, 14 days and 7 days are presented. In each case, the Cox and logistic regression models consistently led to lower power than all of the MSM models, and demonstrated that shorter follow-up periods can lead to substantial impacts on power and sample size requirements. For example, for this simulation study, at least 80% power was observed for both the Cox and logistic regression models with a sample size of $1,000$ when the follow-up period was a maximum of 30 days. In comparison, a follow-up period of 14 days reduced the power to less than 60% for these models, or $1,750$ participants would be required to ensure approximately 80% power. Meanwhile, the results suggested that an MSM would require approximately $1,000$ to provide 80% power using data collected for 14 days.

Assessment intervals

Assessment intervals of daily, every 2 days, every 3 days, every 7 days and every 14 days were considered with all other parameters remaining as in the base case including planned follow-up of 60 days. The results indicated that MSM fitted to assessments taken daily, every 2 days or every 3 days, performed at least as well as Cox models applied to data collected daily. There was a large improvement in efficiency from using a MSM compared to a logistic regression model in these scenarios. For example, to achieve 80% power, Model A would require around 650

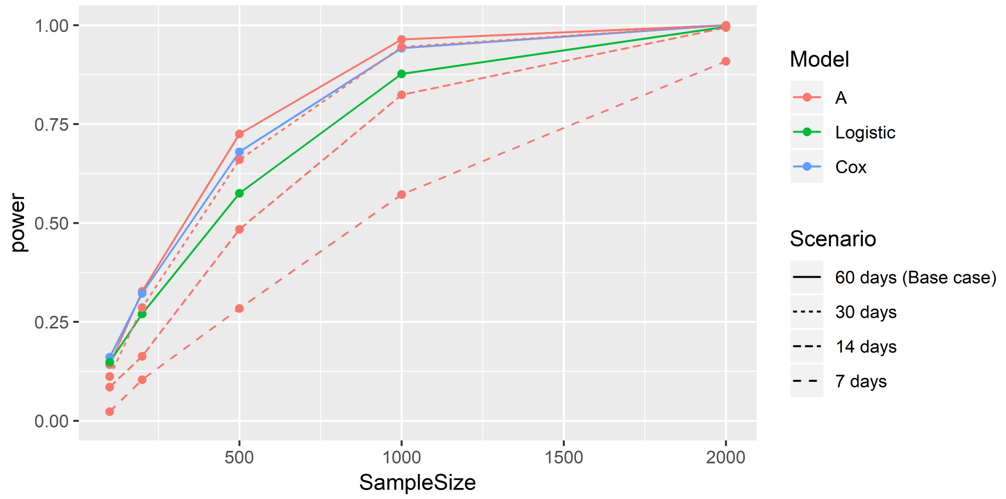


Figure 5.2: Power of detecting a significant treatment effect overall according to sample size for different lengths of follow-up (Assessment frequency=daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$)

patients with data collected daily or every 2 to 3 days, whereas data would need to be collected daily for an additional 200 (approximately) patients to provide similar levels of power using logistic regression (Figure 5.3).

Plots for all types of MSM explored in addition to the Cox and logistic regression models for each level of assessment frequency are presented in Appendix C.2.2. When assessments were conducted every 2, 3, or 7 days, the unconstrained MSM performed similarly to the Cox model in terms of power. However, when the assessment frequency reduced to 14 days, the unconstrained MSM was the worst performing model in terms of power, whilst the Cox model provided a substantial improvement. Specifically, the Cox model would require approximately 250 fewer participants to achieve 80% power. The models with constraints imposed on the final transition (i.e. Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, or Model C: $\beta_{12} \neq \beta_{23} = \beta_{34}$) were the only two models in the scenario with length of follow-up of 14 days that led to improved power over the Cox model. This may be because the length of the intervals mean that intermediate transitions are missed, which reduces the size of the dataset available to estimate the model parameters for each of the 3 transitions.

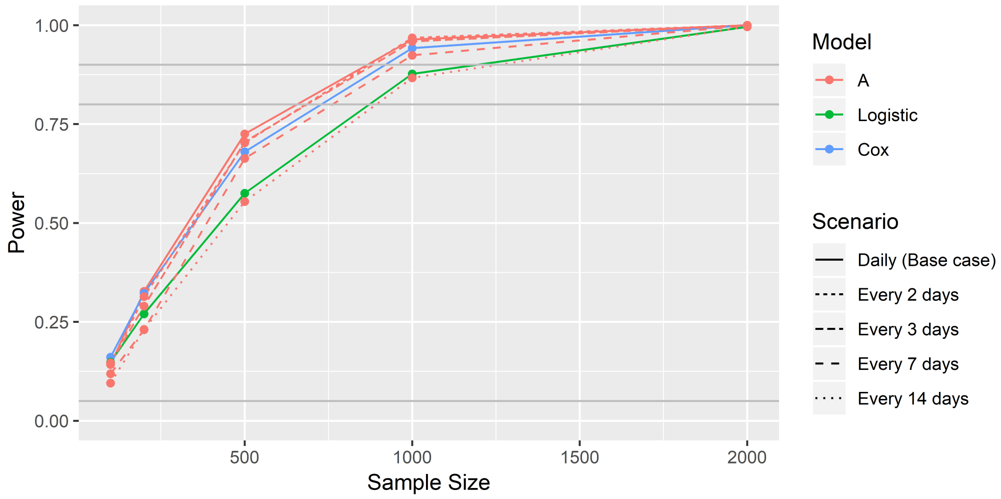
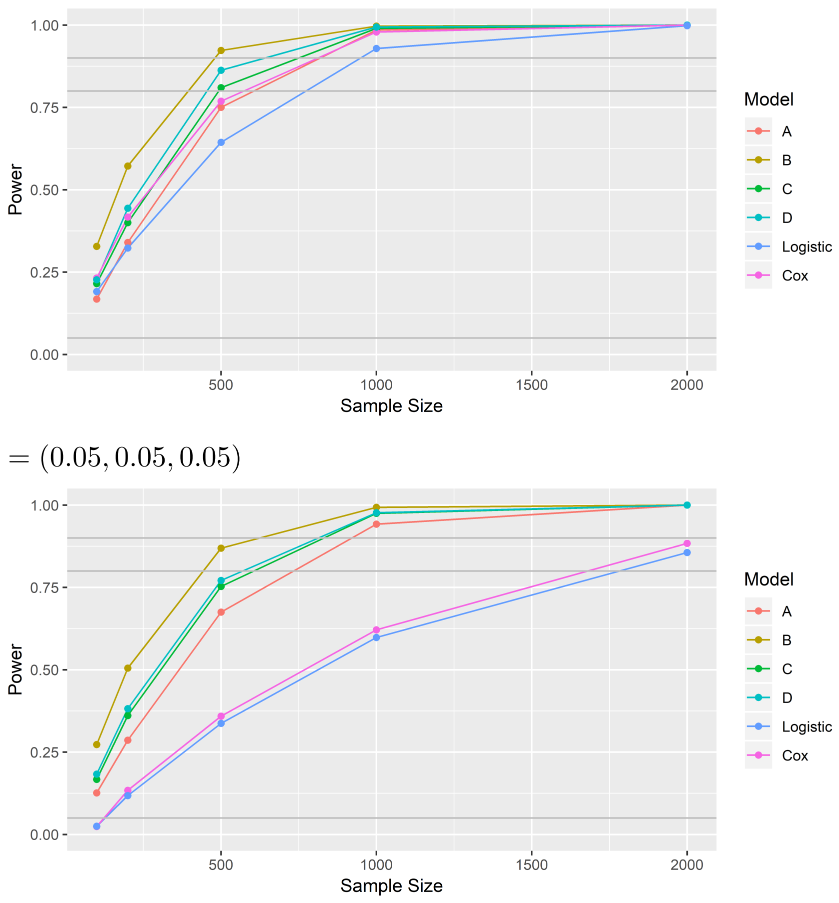


Figure 5.3: Power of detecting a significant treatment effect overall according to sample size for different assessment intervals (Maximum length of follow-up= 60 days, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$)

Baseline transition intensities

MSM led to substantially increased power compared to the logistic and Cox PH regression models when the baseline intensity for the transition from state 3 to state 4 was low ($q_{34.0} = 0.01$). A consistent increase in power was observed in all scenarios where $q_{34.0} = 0.01$, whereas there were similar levels of power observed for each model under scenarios where the baseline transition intensity from state 3 to state 4 was high (Mild to Severe disease, $q_{34.0} = 0.05$) (Figure 5.4). In some cases (e.g $N = 500$, $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.05)$) lower power was observed for the MSM compared to the logistic regression and Cox PH models. Note, for example, that the Cox PH model estimated the treatment effect on the transition from any of the states 1, 2 or 3 to state 4 and significance testing was conducted at the 5% level. In contrast, the MSM estimated the treatment effect on individual transitions (i.e. $1 \to 2$, $2 \to 3$, and $3 \to 4$) and significance testing was conducted according to Hochberg's method for multiple testing. Therefore, it is expected that the Cox PH model would perform at least as well as the overall MSM in situations when the baseline transition intensity to the absorbing state was high and may therefore be the preferred method for primary analysis as it requires less computing power, and is widely understood.



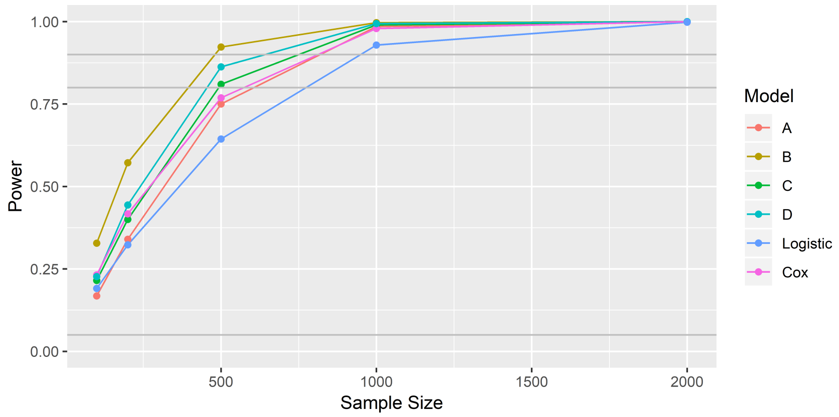
(a) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.05)$

(b) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.01)$

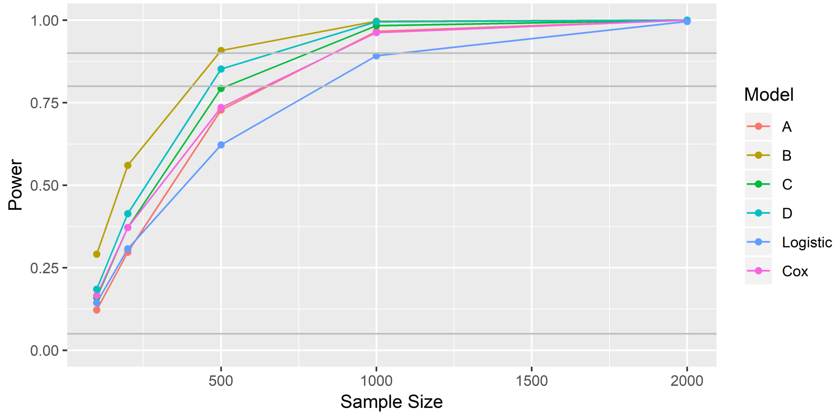
Figure 5.4: Power of detecting a significant treatment effect overall according to sample size for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment intervals = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Treatment effects

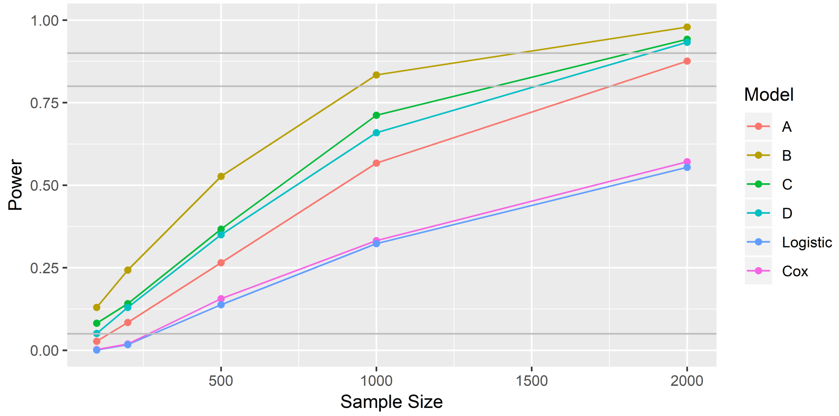
When the treatment effect was high for the early transitions, and moderate for the final transition, that is, $\exp(\beta) = (0.5, 0.5, 0.67)$, all of the MSM's provided increased power compared to the Cox and logistic regression models (Figure 5.5). This is expected given the high treatment effects observed on the earlier transitions were ignored in the models of a single endpoint. Further, MSM provided a greater advantage in terms of power when the treatment effects were moderate on the early transitions and low on the transition to the absorbing state, i.e. $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.9)$. In this situation, the MSM provided approximately 90% power with 1,000 participants, compared to the Cox model, which had 50% power, and the logistic regression model with 40% power.



(c) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.05)$

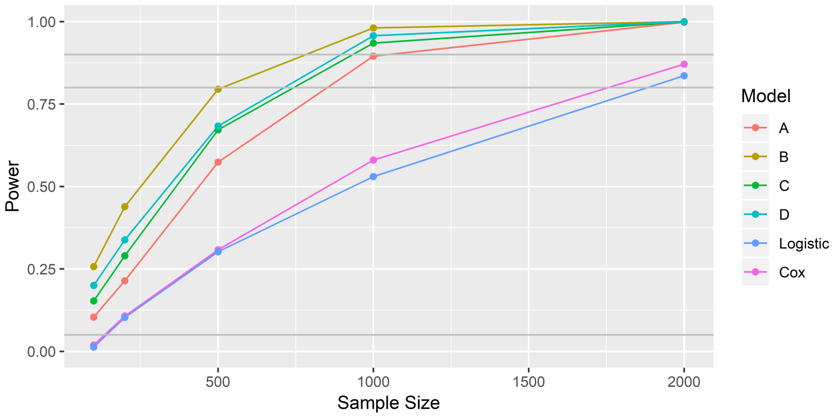


(d) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.05)$

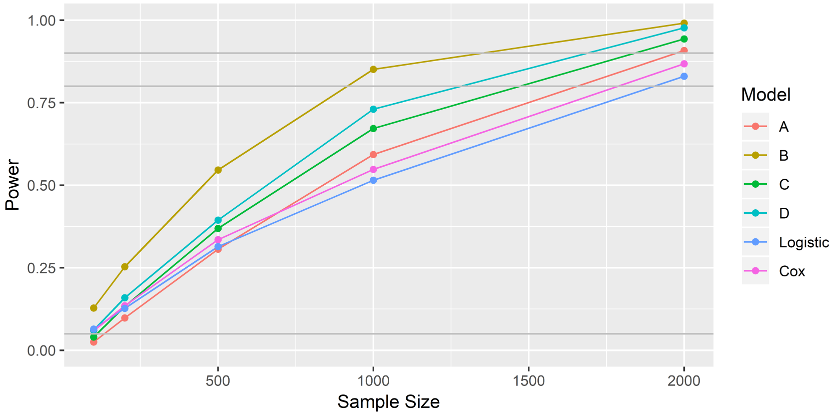


(e) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.01)$

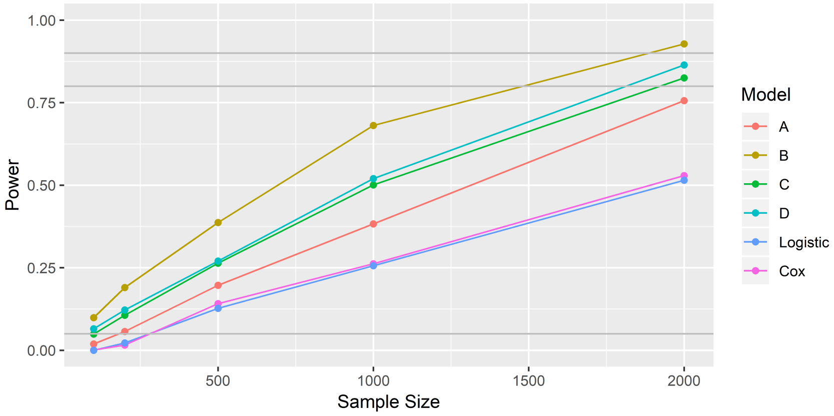
Figure 5.4: Power of detecting a significant treatment effect overall according to sample size for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment intervals = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)



(f) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.01)$



(g) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.05)$



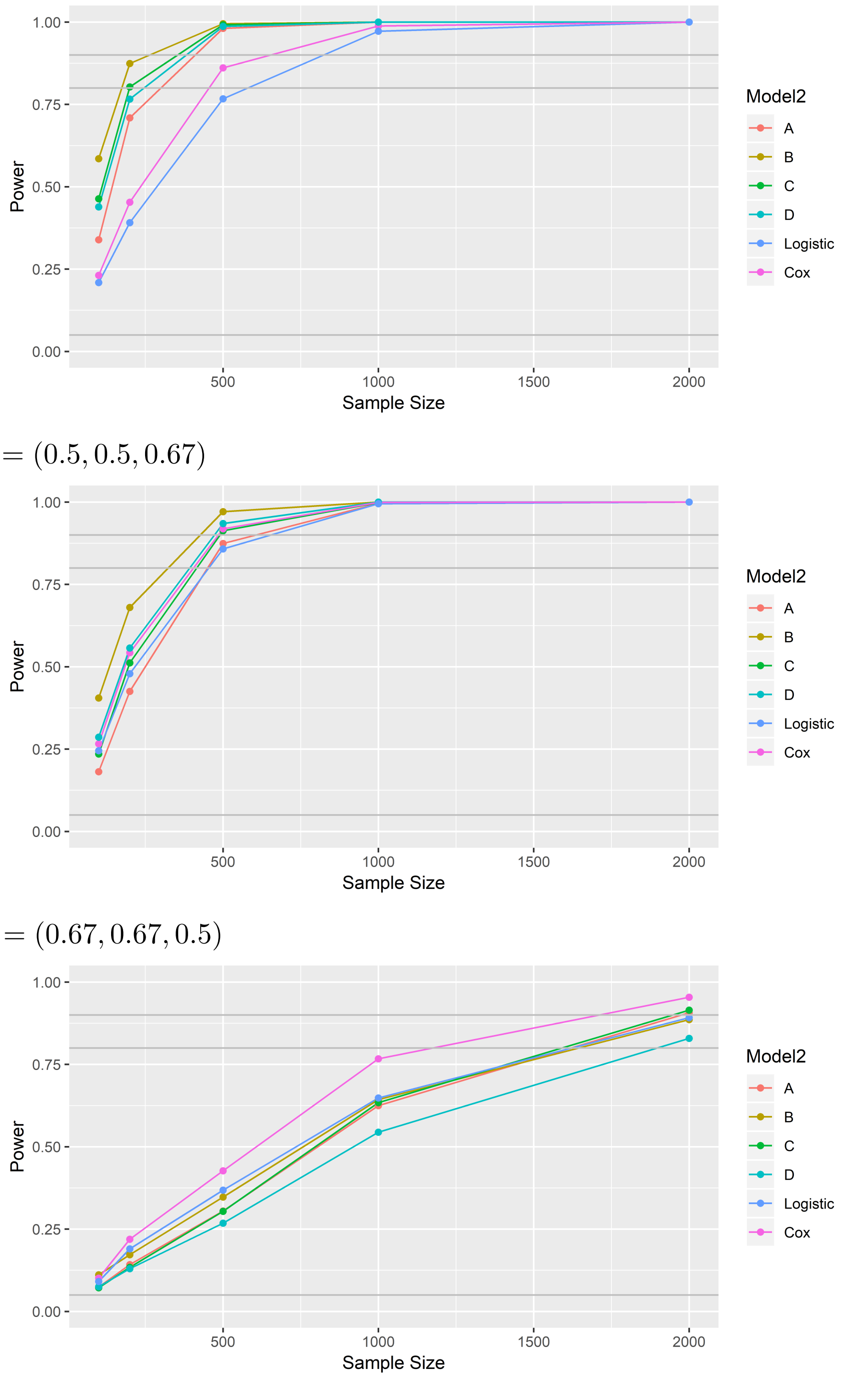
(h) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.01)$

Figure 5.4: Power of detecting a significant treatment effect overall according to sample size for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment intervals = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

In situations where the treatment effect on the early transitions was smaller compared to the final transition, there was little gain in power using an MSM. In the scenario where $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.5)$, the completely constrained model performed best in terms of power, however this model would be inappropriate to use in this scenario because the treatment effects did in fact differ and issues of bias and poor coverage may arise (see Sections 5.4.2 and 5.4.3 for further details). The unconstrained MSM in this scenario actually provided slightly reduced power compared to a Cox model. Furthermore, as with low intensities and low treatment effects on early transitions, the Cox PH model would be expected to perform as well as the overall MSM and may therefore be the preferred method for primary analysis in these scenarios. In this case the advantage of an MSM would be in providing additional insight into the treatment effects on different stages of the disease pathway, and may be an appropriate alternative if the assumption of non-proportional hazards over the whole follow-up period is violated. A further example of this was the scenario where $\exp(\boldsymbol{\beta}) = (0.9, 0.9, 0.67)$. Here, the Cox model provided an advantage in efficiency compared to the MSM with approximately $1,200$ participants required for 80% power with the Cox model compared to $1,600$ for MSM. This scenario most reflects the results of the motivating example where a treatment effect was observed on the final transition but not on the first two. Despite the reduction in power, the illustrative example demonstrated that MSM may be more appropriate due to a violation of the PH assumption in the Cox model.

5.4.2 Bias

Point estimates of the treatment effect (hazard ratio) estimated from the MSM models have been examined and were shown to be unbiased for the unconstrained model in all scenarios, and for all models when the treatment effects were equal. Figure 5.6 illustrates the base case, where all treatment effects were equal to 0.67; the point estimates were centered around 0.67 as expected. There was more variability in the estimate of $e^{\beta_{12}}$ for models A (no constraints) and D (later transition hazard ratios constrained to be equal), compared to those obtained from models B and



(a) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$

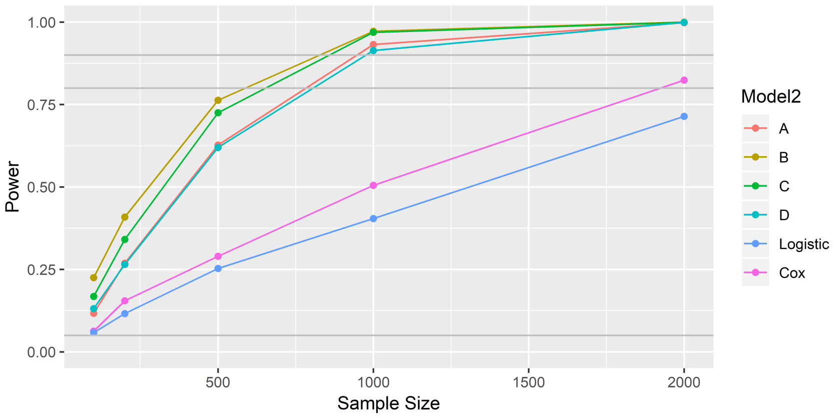
(b) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.5)$

(c) $\exp(\boldsymbol{\beta}) = (0.9, 0.9, 0.67)$

Figure 5.5: Power of detecting a significant treatment effect overall according to sample size for different magnitudes of treatment effect (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C. This is due to the amount of available data on early transitions when β_{12} was estimated individually, compared to models B and C where constraints were imposed on early transitions.

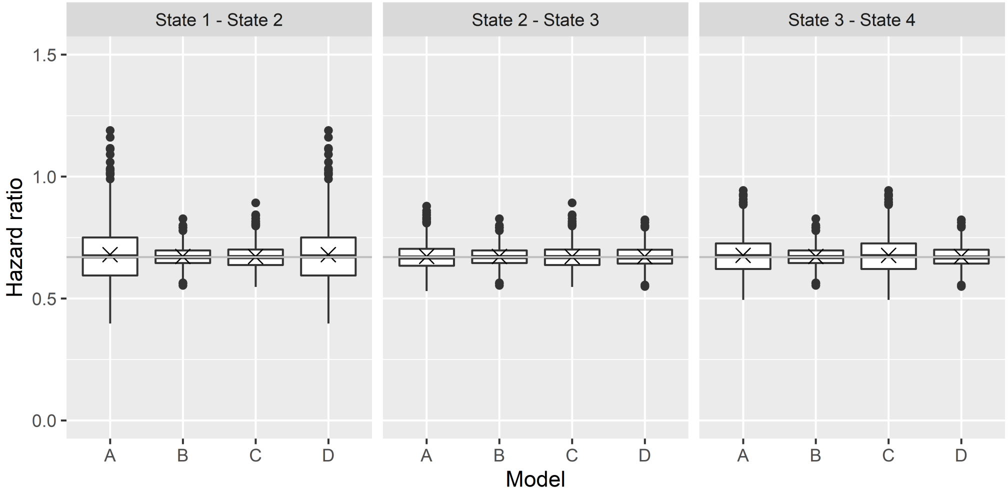
When there were unequal treatment effects, bias in the estimated treatment ef-



(d) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.9)$

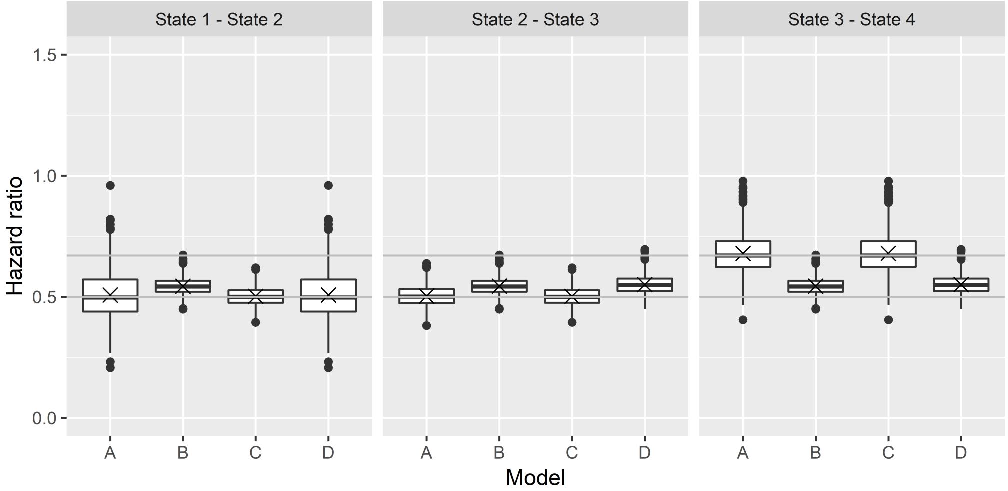
Figure 5.5: Power of detecting a significant treatment effect overall according to sample size for different magnitudes of treatment effect (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

fects may occur if the model is mis-specified. An example is shown in Figure 5.6 when there were high treatment effects on the early transitions, and a moderate treatment effect on the final transition, i.e. $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$. The unconstrained model A demonstrated unbiased treatment effect estimates on all transitions. Model C also led to unbiased treatment effects because appropriate constraints were imposed, with $\beta_{12} = \beta_{23} \neq \beta_{34}$. Bias occurred, as expected, in models B and D where β_{34} is constrained to be equal to one or more of the earlier treatment effects. Specifically, in the fully constrained model, where $\beta_{12} = \beta_{23} = \beta_{34}$, the early treatment effect estimates, $e^{\beta_{12}}$ and $e^{\beta_{23}}$ were attenuated towards the null, whereas $e^{\beta_{34}}$ is estimated to be larger in magnitude than its true value. In Model D where $\beta_{12} \neq \beta_{23} = \beta_{34}$, the estimated treatment effect on the first transition, $e^{\beta_{12}}$ was unbiased as expected because there were no constraints imposed, and similar variability in the estimates were observed compared to model A. However, the treatment effects $e^{\beta_{23}}$ and $e^{\beta 34}$ had similar bias as model B, with $e^{\beta_{23}}$ attenuated towards the null, and $e^{\beta_{34}}$ larger in magnitude. Note that whilst conclusions about any bias in the estimated treatment effects can be made by examining plots of the estimates, it may be more useful in future to examine the bias itself particularly when there are different treatment effects on each transition.



(a) Base Case, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$

Figure 5.6: Distribution of point estimates of the treatment effect (hazard ratio) when $N = 2000$



(b) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$

Figure 5.6: Distribution of point estimates of the treatment effect (hazard ratio) when $N = 2000$ (Maximum length of follow-up= 60 days, Assessment frequency=Daily, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

5.4.3 Coverage

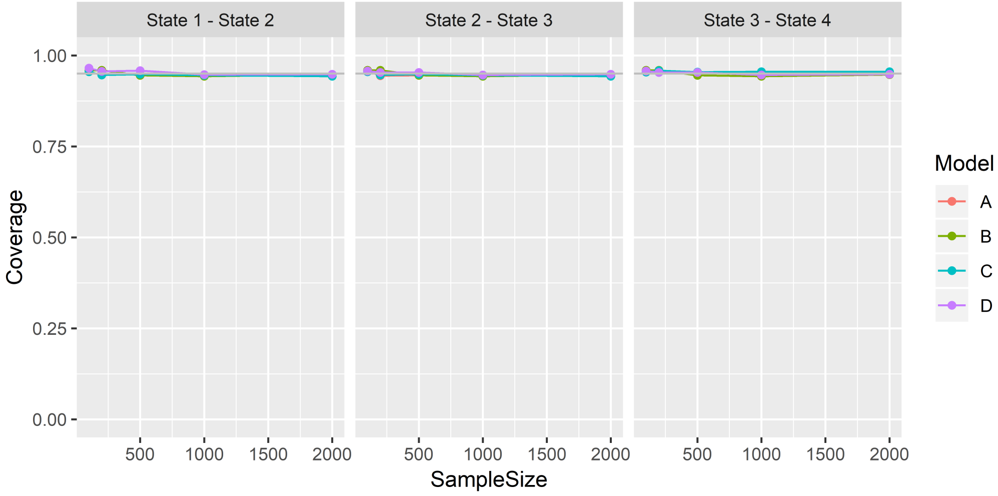
Coverage of the 95% confidence intervals for each treatment effect on each transition have been examined and were adequate in all scenarios, except those where treatment effects differed, and in some cases when the baseline transition intensity differed (Appendix C.5). Figure 5.7 illustrates the base case, where all treatment effects were equal to 0.67; the estimated coverage was broadly consistent with 95%

for all sample sizes and all MSM models.

Poor coverage occurred when there were unequal treatment effects, illustrated in Figure 5.7 using the same example of high treatment effects on the early transitions, and a moderate treatment effect on the final transition. The unconstrained model A, and model C, which used appropriate constraints on the treatment effects for the early transitions led to adequate coverage for all sample sizes. Fitting model B, where the treatment effects were all constrained to be equal, to data generated under model C led to decreased coverage to around 75% for $e^{\beta_{12}}$ and $e^{\beta_{23}}$, and to less than 10% for $e^{\beta_{34}}$. Similarly, fitting model D demonstrated poor coverage for treatment effects $e^{\beta_{23}}$ at approximately 75% and 15% for $e^{\beta_{34}}$, whereas the coverage for $e^{\beta_{12}}$ was adequate. This was because the treatment effect was unconstrained on the transition from State 1 to State 2, but the treatment effect on the transitions from State 2 to State 3 was constrained to be equal to the treatment effect on the transition from State 3 to State 4. When poor coverage occurred, the problem was exacerbated as the sample size increased, which is expected because the confidence intervals are more precise. The confidence intervals are therefore less likely to include the true treatment effect when the point estimate of the treatment effect is biased.

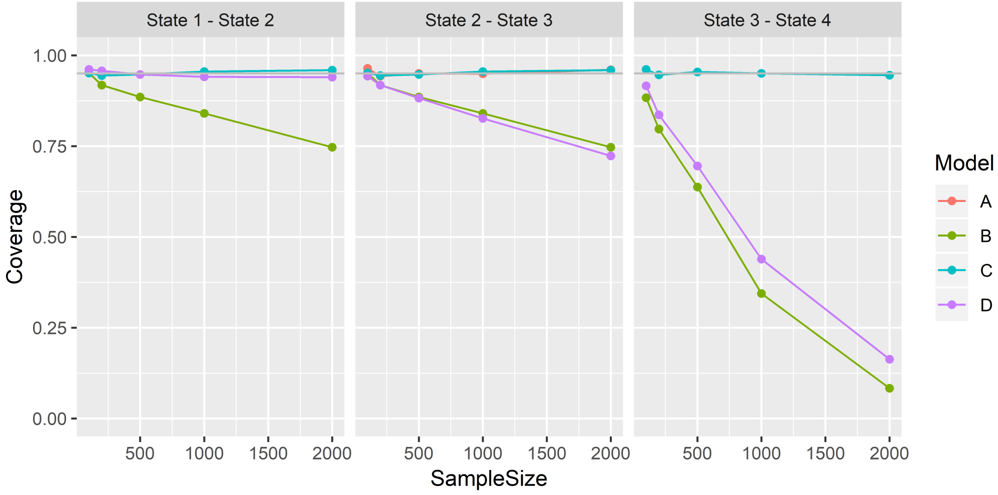
Appendix C.5 includes plots for the coverage of the estimated treatment effects under different baseline transition intensities. On the whole, 95% coverage was achieved, however there were some cases for small sample sizes where this was not the case, which is due to models not converging. In particular, when $\boldsymbol{q}_0 = (0.05, 0.01, 0.01)$ and the sample size was equal to 100, Model A and Model C had low coverage for $e^{\beta_{34}}$, which could be due to a smaller number of entries to the third state combined with a low number of transitions between state 3 and state 4. Low coverage was also observed for models A and D when $\boldsymbol{q}_0 = (0.01, 0.05, 0.01)$ and the sample size was smaller than 500, which could be due to a low number of transitions between state 1 and state 2. However the coverage was adequate for later transitions in all models because there were more data available from the 70% and 15% participants who started in states 2 and 3 respectively. When the baseline transition intensities were $\boldsymbol{q}_0 = (0.05, 0.01, 0.01)$ and the sample size was equal to

100 or 200, coverage was low for $e^{\beta_{12}}$ under models A and D where the treatment effects are estimated individually (i.e. without a constraint on β_{12}). Similarly, coverage was lower for $e^{\beta_{34}}$ under models A and C where there was no constraint imposed on β_{34}.



(a) Base Case

Figure 5.7: Coverage for nominal 95% confidence intervals



(b) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$

Figure 5.7: Coverage for nominal 95% confidence intervals

5.5 Discussion

Summary of results

RCTs of strategies for prevention of diseases and medical conditions often involve repeated assessments of the severity of disease at multiple time-points. The potential estimands from different models that could be applied to data of this structure include odds ratios for a binary endpoint, hazard ratios for a TTE endpoint, and transition-specific hazard ratios obtained from MSM. It has been demonstrated in the literature and through secondary analysis of two PU trial datasets in Chapter 4 that MSM can provide a deeper understanding of the natural history of a disease and how treatment acts at each stage of the disease pathway [116, 124, 125].

In this chapter, a comprehensive simulation study was conducted and showed that, depending on the estimand of interest and underlying natural history of the disease, analysis using MSM has the potential to have a substantial impact on power, or equivalently a reduction in sample size compared to logistic regression for a binary endpoint or a Cox PH regression model for a TTE endpoint. Greatest improvements in efficiency were observed when early changes could be observed due to frequent assessments, high early transition rates or larger effects on early transitions. Where the design featured long intervals between assessments, slow transition through early states relative to later states or low treatment effect on early states relative to the treatment effect on later states, there was little to be be gained from MSM in terms of trial efficiency, although estimation of the disease development over time may be of interest.

Whilst greatest efficiencies were shown to arise when treatment effects for different transitions were equal, fitting constrained models when the true process has a different form could result in substantial bias and poor coverage.

Implications on clinical settings

The results of the simulation study suggest that in the motivating example for a PU prevention trial, the length of follow-up could be halved from 60 to 30 days or assessments conducted every 2 or 3 days (in line with current practice) to provide similar levels of power as would be obtained by Cox or binary logistic regression models applied to daily measurements for 60 days provided that the treatment effect acts across the disease progression pathway. This has the potential to reduce trial resource use by using fewer patients, with savings in assessor time and data management. In many scenarios, fewer patients need to be recruited overall and fewer are unnecessarily exposed to inferior treatments. Moreover, evidence of treatment effectiveness (or not) will emerge more quickly leading to quicker changes in practice for subsequent patients. However, this should be considered in conjunction with the relevant clinical research question (estimand) since the overall significance level for MSM reflects treatment effects across all transitions. For example if primary interest lies in preventing Severe disease then the commonly used methods may be sufficient, and have the advantage that the resulting significance level is directly related to a single treatment effect. If, however interest lies in assessing whether the treatment can reduce transitions at any stage along the pathway, then MSM may lead to more efficient designs at lower cost.

This simulation study provides a comprehensive set of results under a range of scenarios and compared MSM with simpler models (logistic regression and the Cox PH model), in addition to reviewing the impact of applying constraints to treatment effects within MSM. Constraining treatment effects to be equivalent within the MSM did provide additional power but should be used with caution as they may not be a realistic representation if the true treatment effects differ between disease stages, as demonstrated through the observed bias of the treatment effect estimates. The decision over whether to apply constraints should be informed by analysis of pre-existing datasets in conjunction with discussion with clinical experts to inform clinical plausibility of such constraints. Sensitivity analyses should be conducted to assess the robustness of the models to different constraints.

If researchers or funders have a preference for the simpler models, the results also show important implications of the length of follow-up for these models, with a 30 day follow-up period being optimal under the parameters of the simulation study.

Furthermore, if the baseline risk of developing a Category 2+ PU is low, then the binary and Cox regression models lead to a substantial loss of power. However, if a Category 1 PU is chosen as the endpoint of interest to increase power or reduce sample sizes, this could be at risk of misclassification (see Chapter 6 for further details of misclassification).

Some statistical considerations

It is important to try to understand how the additional power from MSM arises, given that the number of Category 2+ PUs observed was the same for all analyses. More data on early skin changes is included in MSM and this, together with the structure of the model, which links the different transitions together, is the source of the additional information. If the MSM is not consistent with the observed natural history of the disease of interest, then either the predicted increase in power will not manifest, or spurious increases in power will result. Therefore, it is imperative that a good model is adopted and (in line with good statistical practice) the fit of the model is checked carefully.

This simulation study makes a number of assumptions, including that censoring patterns are independent of skin status or patient condition, and that the MSM allows progression only. In another setting it may be important to allow transitions between states in both directions and this should be considered for further research but is beyond the scope of this thesis. Furthermore, this simulation study was conducted assuming that the disease state was at a patient level, specifically, the “worst observed skin state”, however in practice the underlying data are available for multiple skin sites. As demonstrated in Section 4.4.1, there was no evidence that accounting for individual skin sites within patients led to a more adequate model, but this may differ for other disease areas. As previously described, MSM methods exist for correlated disease processes such as psoriatic arthritis [111], but the models are more complex and software is limited when there is interval censoring. As such, conducting a simulation study that is as comprehensive as this would be computationally intensive, and the estimand of interest would need careful consideration by the clinical team.

A further assumption in this simulation study was that the states observed were always correct, however as identified in the literature, the state may be at risk of misclassification. In PU prevention trials, skin assessment is based on the appearance of the skin, which requires expert knowledge particularly for the earlier stages of PU development. Category 2+ PU is often used as the endpoint of interest because it is considered to be less prone to error [31]. Although classification may be less reliable, some researchers analyse Category 1 PUs because it is clinically important, and is a strong prognostic marker for Category 2+ PU [42]. The higher incidence of Category 1 PUs may also influence the decision to use it as a primary endpoint because a smaller sample size may be required compared to using a Category 2 PU. Nonetheless, the impact of misclassified states in the estimation of treatment effects and power of a trial is explored in Chapter 6.

Overall summary

In summary, this simulation study demonstrated that logistic regression and Cox PH regression may be inefficient in terms of sample size and frequency of assessments compared to MSM for analysing trials where panel data collected for a progressive disease with an ordinal outcome are available. This first simulation study has also demonstrated that MSM have the potential to improve trial design through increased power subject to further investigation of dataset characteristics that may arise in a realistic setting. However, any gains may disappear, and bias and/or poor coverage will result if models are misclassified.

Chapter 6

Misclassification

6.1 Introduction

Up to this point the observed state in a MSM has been assumed to be true and not subject to misclassification. However, given the subjective nature of PU classification identified in the literature review, misclassification is entirely possible. This is particularly the case if assessors are not subject experts.

When designing clinical trials an important decision is the choice of assessment process, including measurement instrument, mode of administration and assessor. For PUs the gold standard measurement instrument is the NPUAP/EPUAP/PPPIA guidelines [32]. The literature review in Chapter 2.3 identified that there are concerns of misclassified PU assessments and therefore the gold standard could be considered the use of expert dermatologists for assessment at regular intervals. One question of interest in the design of trials is whether less expert assessors could be used for assessment. For example, in PU research, routine care staff could assess skin sites more frequently than experts, but each measurement may be subject to some error, which is a major concern highlighted in the literature review. In order to investigate this issue, it is important to understand the potential effect of misclassification on accuracy and precision of treatment effects. This is the subject of this chapter.

Throughout the chapter, measurement error is used to describe deviations from the latent data for continuous outcomes, misclassification is used for dichotomous or categorical outcomes and under-ascertainment/over-ascertainment is used for events.

6.1.1 Measurement error of continuous outcomes

Mis-measurement can occur for different types of variables and it is important to have an understanding of the potential implications. Therefore, for completeness, the impact of mis-measured continuous variables is briefly discussed here. Measurement error of continuous variables can be classified as classical error where the observed data is equal to the latent data plus a random component with zero mean and constant variance, systematic error where the observed data is systematically different from the latent data, or differential where the error is dependent on the outcome, conditional on the true value of the covariate. A fourth type of measurement error described is the Berkson error, which is the opposite of the classical error, whereby the latent data is equal to the observed data plus a random component with zero mean and constant variance [153].

Nab *et al* [154] presented an illustrative example for the measurement of haemoglobin where the equipment used in the trial appeared to give lower values of haemoglobin compared to certified measurements with mean (standard deviation (sd)) values reported as 135 $(0.96) g/L$ compared to 137 $(3.2) g/L$ respectively. The authors conducted a simulation study to explore the potential impact on the trial results under different types of measurement error. In the first instance they explored classical measurement error whereby the measurement introduces an additional component of variation in the outcome, independent of treatment allocation and true level of haemoglobin. The result, also demonstrated algebraically, is that the treatment effect estimate is unbiased, however there will be an increased type II error (reduced power) due to a larger variance in the observed data compared to the latent (true) values.

Nab *et al* [154] also consider systematic measurement error in which the values obtained from one method are systematically different from those obtained from another method (i.e. there is a location shift). This could be additive, whereby the observed data are always a constant, c, further away from the latent data, independent of their values. In this case, the arithmetic difference between treatment groups is unbiased because the constant value will be cancelled out in the calculation. However, when one method provides values that are multiplicatively different to the latent values, for example, the observed values may be c times higher than the latent value, then the arithmetic difference observed is also increased multiplicatively by c, thus leading to a biased result, unless the true difference is 0. It may be appropriate to use the log transformation on the measurements, which will lead to an unbiased estimate of the treatment effect but Nab *et al* reported that there will be an increase or decrease in type II error depending on the value of c in the case of multiplicative measurement error, because the variance will be affected, however type I error will be unaffected [154].

So far, measurement error is assumed to be similar across treatment groups. When the measurement error differs according to treatment groups (differential mismeasurement), the estimate of the treatment difference will be biased with the direction and magnitude of bias dependent on the nature of the measurement error. For example, if the endpoint is subjective in some way and the patient or assessor knows the treatment allocation, there may be a tendency to over or under value the outcome in one group. In line with the assessment of systematic measurement error, Nab *et al* [154] reported that there will be an increase or decrease in type II error depending on the nature of the differential measurement error, because the variance will be affected. Nab *et al* proposed alternative estimators for models with mismeasured continuous outcomes.

The Cochrane Risk of Bias (RoB) guidance notes that measurement error of continuous outcomes is often assumed to be additive, therefore whilst it is important to consider the accuracy in the choice of a continuous outcome measures, there will usually be a low risk of bias in measurement of the treatment effect providing the error is not multiplicative, or differential between treatment groups.

6.1.2 Misclassification of dichotomous or categorical outcomes

Misclassification of a dichotomous or categorical outcome may occur if the method of outcome measurement is not the gold standard, or if the assessment requires some level of judgement. For diagnostic tests with a dichotomous outcome, misclassification is usually quantified in terms of sensitivity and specificity. Sensitivity is defined as the proportion of true positives that are correctly identified, and specificity is the proportion of true negatives that are correctly identified by the test [155].

Some trials use methods to adjudicate outcomes that may be subject to observer bias. This is dependent on the trial context and logistical considerations but could be done at the site or through central review. For example, Godolphin *et al* conducted a simulation study to investigate the role of central adjudication in stroke RCTs at risk of misclassification of binary or ordinal outcomes [156]. The results showed that if as little as 2.1% of participants were misclassified differentially this led to a different trial result when the outcome was binary. They also found that, in trials with an ordinal outcome, misclassification between 1.9% and 27.8% could affect the trial result, with larger trials being more sensitive to misclassification. The authors suggest this could be due to larger trials being able to detect smaller differences and with greater precision. In comparison, in the assessment of non-differential misclassification, the authors found that the level of misclassification that could affect trial results increased with the event rate and the trial sample size. The recommendation from this paper was that central adjudication is important for stroke trials without sufficient blinding for outcome assessment but that it may not be necessary for trials with adequate blinded outcome assessment. Kahan *et al* compared different approaches for adjudication of outcomes in clinical trials and conducted a simulation study to investigate the number of assessors required, whether on site or central assessment should be conducted, whether all outcomes should be adjudicated or only the events of interest, and finally whether central assessment with multiple assessors should be conducted independently or via consensus [157]. The conclusions of the simulation study were that whilst outcome adjudication is important for trials with misclassified outcomes, the decision of which approach to take should be made in the context of each clinical trial. Overall, the effect of misclassification of outcomes is dependent on multiple factors including expected incidence of the event and extent of misclassification, and should therefore be considered at the design stage of a trial [156, 157].

6.1.3 Over-ascertainment or Under-ascertainment of events

In addition to measurement error of continuous outcomes, and misclassification of dichotomous or categorical outcomes, analyses of TTE outcomes may also be affected by misclassification through over-ascertainment or under-ascertainment of events. For example, the diagnosis of progression of disease in some cancer clinical trials is assessed using RECIST [158] criteria based on a radiological scan of the patient. Diagnostic scans are rarely perfectly sensitive and specific, which means there may be differences in the diagnosis of progression. Hróbjartsson *et al* [159] conducted a systematic review to quantify the effect of observer bias in RCTs with binary outcomes. The trials included in the review utilised both blind and unblind assessors and the review observed that the estimated treatment effects (hazard ratios) from the analysis of the unblinded outcome assessors data were, on average, 36% larger than that of the analysis of data recorded by their blinded counterpart.

Simulation studies have been conducted to assess the impact of measurement error in trials with a TTE endpoint in terms of bias [160, 161]. A consistent conclusion was that non-differential measurement error led to attenuated treatment effects. On the other hand, differential measurement error did lead to bias in the treatment effect with a likely increase in it's magnitude [161]. There was less attenuation in the treatment effect for scenarios with longer assessment intervals with a possible explanation that the event was less likely to be diagnosed early if there was a tendency to over-report [160].

6.1.4 Measurement error of covariates

So far, the measurement error described has focused on outcomes, but it may also occur in the assessment of covariates. Brakenhoff *et al* conducted a systematic review to investigate the reporting of measurement error in exposure and confounder variables, and any methods used to account for it in the analyses [153]. The review was of research published in high-impact medical and epidemiological journals in 2016. The key findings from this review were that, of the 565 reviewed texts, 247 mentioned measurement error and of these, only 18 investigated or corrected for the error, leading to the conclusion that the potential impact was often ignored and misunderstood.

The review outlined that for classical error, even if the exposure variable is measured without error, any error in one or more of the confounders may lead to bias in the estimated relation between exposure and outcome, although the direction and extent of this bias is unpredictable. For systematic and differential error in models estimating the effect of a covariate on outcome, the potential bias can occur in either direction. Berkson error on the other hand, rarely leads to bias in the estimates of the effect of a covariate, but may reduce the precision. Similarly, for categorical variables, the direction and magnitude of bias in the estimated effect on outcome is difficult to predict and will be context specific.

For the remainder of this chapter, the focus is on the potential mismeasurement of outcomes with acknowledgement that it is critical for trials to ensure that assessment of bias through measurement error of both outcomes and covariates are considered prior to analysis. For RCTs the covariates are often measured at baseline and, for the primary analysis, are likely to be the allocated treatment and any randomisation factors, which could be chosen such that they are at very low risk of measurement error. Differential measurement error of baseline variables is unlikely to be an issue in RCTs because the baseline assessments are usually conducted prior to randomisation.

6.1.5 Misclassification of outcomes in multi-state models

Misclassification has been incorporated into MSM analysis in various clinical settings. Cook and Lawless considered the reasons for unexpected improvements in a chronic disease [162]. Depending on the context, they suggested that it may be due to random fluctuations in the condition, misclassification of the discrete disease states, or errors in the measurement of an underlying continuous score. They highlighted that in some settings it is common to confirm movement to a different state through repeated observations of that state, which was an approach adopted for the confirmation of a Category 1 PU in some of the papers identified in the literature review (Chapter 2.3).

Alternatively, the probability of misclassification may be modelled jointly with the MSM called a Hidden Markov Model (HMM) which was the approach taken by Van den Hout and Matthews when they analysed cognitive ability data [163]. In this example a 3-state illness-death model was used whereby the first 2 transient states ("not cognitively impaired" and "cognitively impaired") were at risk of misclassification, but the final absorbing state of death was measured without error. The authors used a piecewise constant hazards model to analyse the data and noted that the methods could allow regression from state 2 but not the absorbing state. A further example of a HMM was given by Jackson *et al* who analysed data to assess disease progression and prognosis for patients from 6 months after either a single lung, double lung or heart-lung transplantation [118]. These patients were considered at risk of a chronic condition characterised by declining lung function called Bronchiolitis obliterans syndrome (BOS), which is diagnosed using the forced expiratory volume in 1 second (FEV$_1$) every 3 to 6 months [118]. The assessment of BOS may be subject to misclassification because the assessment of FEV$_1$ is sensitive to factors such as infection affecting lung function and resulting in short term fluctuation. One approach was to use central assessment to classify the patient's BOS status, but Jackson *et al* proposed a hidden Markov model to simultaneously estimate the transition rates of an illness-death model and the probability of state misclassification [118]. This method was extended to analyse screening data for abdominal aortic aneurysms where patients could be in 1 of 4 disease states according to their risk status, which was best predicted by aortic diameter [164]. The method published by Jackson *et al* allowed for any number of transitions, and misclassification between any pair of states, whilst allowing for different sets of independent variables for the transition rates and misclassification probabilities [164].

The methods proposed by Jackson were illustrated for progression models with misclassification where observed regression was assumed to be an artefact of misclassification however in some situations there both regression and misclassification is clinically plausible. HMM are at risk of identifiability issues when different misclassification and transition matrix combinations lead to the same likelihood value (see Section 6.4.1). This may be more likely to occur as the models increase in complexity such as when regression is modelled as well as progression, or when fewer data are available for model estimation. However, an example of when a HMM for a process with regression has led to useful insights is a publication by Gangnon *et al* [165]. The authors investigated the impact of misclassification of age-related macular degeneration (AMD) on the baseline intensity and covariate effects on the disease process, which consisted of 5 AMD states, and death as the absorbing state. The process included 12 transitions, with progression and regression permitted between adjacent AMD states (with the exception that regression from the final AMD state was not permitted), and progression to the absorbing state, from each of the AMD states. The authors utilised MSM with misclassification in continuous time and identified that ignoring misclassification tended to lead to attenuated covariate effects on some transitions, and that regression of AMD disease was largely explained by misclassified states. Furthermore, the authors concluded that in the AMD setting there is a need for ongoing assessment of the data, which are collected as part of a 20 year cohort, to attempt to reduce the extent of misclassification in the dataset. The authors did not discuss the risk of identifiability or strategies to overcome this, but may have benefited from a large dataset of $4,379$ patients with a total of $12,640$ assessments.

HMM benefit from informative initial values of the misclassification probabilities to reduce the risk of non-identifiability, and the reliability of the results can be assessed through sensitivity analyses with different initial values of the misclassification probabilities. One example of this is for a discrete time MSM accounting for misclassification when the Markov assumption was not considered appropriate [166–168]. In this example by Bacchetti *et al* 5 state MSM was used to analyse biopsy measured liver fibrosis data where where there was interest in whether patient prognosis was dependent on their disease history [166–168]. The models incorporated a time-dependent covariate for the MSM to denote the length of time spent in the current state. In addition, the models estimated the probability of misclassifying the transient states and the absorbing state, including both over and under-reporting. A sensitivity analysis provided reassurance that the model was insensitive to the initial misclassification matrix specification but the authors recommend obtaining more reliable estimates of the misclassification relevant to the study, or using a measure with lower levels or no misclassification, which will also help with computation burden. The method used by Bacchetti *et al* was implemented using the R package *mspath*.

Overall, HMM are a possible approach to analysing misclassified data in the MSM setting, but careful consideration must be given to the MSM structure, likely misclassification patterns and initial misclassification probabilities. The remainder of this chapter will explore the extent of misclassification in the PU setting with an application of HMM to the PRESSURE dataset.

6.2 Pressure ulcer misclassification

As identified in the literature review of PU prevention trials in Chapter 2.3, inter-rater reliability of skin status is a common concern because the assessment is made based on the appearance of the skin. There are 2 illustrative datasets available to assess the inter-rater reliability of PU assessments. These datasets are described here and summarised in Table 6.1. Note that the number of paired skin site level assessments are more than the number of participants because each participant had multiple skin sites assessed by each rater.

Table 6.1: Summary of PU inter-rater reliability datasets

Dataset	**Gold standard**	**Comparator**	**Timing**	**Number of paired skin site assessments**	**Number of patients**
PRESSURE	CRN Co-ordinator	CRN	Pre-trial	107	16
			New CRNs	233	35
			Repeat (CRNs in trial ≥ 1 year)	134	20
	CRN	WN	Pre-trial	$2,396$	109
			During trial	$2,606$	331
PURAF	TVN	WN	During study	$2,262$	230

CRN (Clinical Research Nurse); WN (Ward Nurse); TVN (Tissue Viability Nurse)

1. PRESSURE: Recall from Chapter 4 that this is an RCT comparing 2 types of mattresses in acute and elective hospital patients for PU prevention. The trial originally planned to use data collected by non-specialist ward nurses (WN) but also collected data from clinical research nurses (CRNs) to assess the reliability of the WN assessments. To ensure the CRN assessments were reliable, an inter-rater reliability study was conducted comparing assessments between the CRN co-ordinator (gold standard) and the individual CRN; data were collected before the trial started, when new CRNs were appointed and when the trial was first in place at each centre and when the CRNs had been in post for at least a year. Inter-rater reliability data between CRNs and WNs, who are expected to be less experienced than the CRN in terms of skin assessments, were also collected before and during the trial. Assessment of PU classification reliability has been published previously using the PRESSURE dataset [31].

2. PURAF: As part of the development of a new PU Risk Assessment Instrument, the PURPOSE-T, a clinical evaluation was conducted whereby 230 participants with 2262 paired skin sites were assessed by a WN and a member of the tissue viability team (equivalent to the gold standard) independently at a single time point. This assessment included a detailed skin assessment.

PRESSURE

Overall, there was a total of 474 paired skin site assessments between the CRN-co-ordinator and CRNs across the duration of the trial, of which there was perfect agreement in 449 (94.7%). The detail of these is presented in Table 6.2; it is noteworthy that discrepancies were only by 1 category on the PU assessment scale. The

Table 6.2: Cross tabulation of PRESSURE trial skin assessments by CRN co-ordinator and CRN

	CRN assessment (pre-trial)					
CRN co-ordinator (Gold standard)	**State**	**1**	**2**	**3**	**4**	**Total**
	1	47 (100%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	47 (100%)
	2	1 (3.2%)	30 (96.8%)	0 (0.0%)	0 (0.0%)	31 (100%)
	3	0 (0.0%)	1 (5.3%)	18 (94.7%)	0 (0.0%)	19 (100%)
	4	0 (0.0%)	0 (0.0%)	0 (0.0%)	10 (100%)	10 (100%)
	Total	48	31	18	10	107
	New CRN assessment					
CRN co-ordinator (Gold standard)	**State**	**1**	**2**	**3**	**4**	**Total**
	1	129 (97.7%)	3 (2.3%)	0 (0.0%)	0 (0.0%)	132 (100%)
	2	2 (3.1%)	59 (92.2%)	3 (4.7%)	0 (0.0%)	64 (100%)
	3	0 (0.0%)	5 (18.5%)	22 (81.5%)	0 (0.0%)	27 (100%)
	4	0 (0.0%)	0 (0.0%)	0 (0.0%)	10 (100%)	10 (100%)
	Total	131	67	25	10	233
	CRN assessment (repeated)					
CRN co-ordinator (Gold standard)	**State**	**1**	**2**	**3**	**4**	**Total**
	1	73 (92.4%)	6 (7.6%)	0 (0.0%)	0 (0.0%)	79 (100%)
	2	3 (7.0%)	40 (93.0%)	0 (0.0%)	0 (0.0%)	43 (100%)
	3	0 (0.0%)	1 (14.3%)	6 (85.7%)	0 (0.0%)	7 (100%)
	4	0 (0.0%)	0 (0.0%)	0 (0.0%)	5 (100%)	5 (100%)
	Total	76	47	6	5	134

identification of a Category 2+ PU was always observed without error, however only 25 were observed in the comparison between CRN co-ordinator and CRNs across the trial. Despite this small sample, it is reasonable to assume in this setting that the roles of CRN co-ordinator and CRN can both be considered the gold standard, which allows assessment of the reliability of WN assessments. The results across the trial are displayed in Table 6.3 with a total of $5,002$ complete paired skin site assessments. Misclassification is clearly present with a total of $3,924$ (78.4%) in perfect agreement. Disagreements were more extreme than in the comparison between CRN and the CRN co-ordinator, with the assessment of a Category 2 being both under and over reported by the WN. A total of 35 (15.1%) of the 232 Category 2+ PUs observed by the CRN were under-reported as Altered skin by the WN. Conversely, 29 (13.1%) of the Category 2+ PUs reported by the WN were actually Altered skin according to the gold standard CRN assessment. There were differences between the pre-trial and mid-trial assessments, with for example, 26.1% of the true Altered skin assessed as healthy by the WN pre-trial, compared to 63.2% mid-trial. It is important to note that the data reported mid-trial in Table 6.3 are complete data only; there were an additional $1,144$ skin site assessments that were not completed by the WN, 163 unavailable from the CRN and 175 unavailable from both. The dataset provided pre-trial in Table 6.3 was not affected by missing data and therefore may be a more reliable reflection of the misclassification by the WN compared to the CRN.

PURAF

The second dataset from the PURAF study shows similar patterns of misclassification to that of the CRN and WN comparisons in PRESSURE (Table 6.4). That is, there was perfect agreement in $1,766$ (78.1%) paired skin site observations between the TVN and WN. Both under and over-reporting of Category 2+ PUs were observed, with 19 (29.2%) Category 2+ PUs reported as Altered skin by the WN and 7 (11.7%) of the Category 2+ PUs reported by WN were actually Altered skin according to the gold standard assessment [40].

Table 6.3: Cross tabulation of PRESSURE skin assessments between the CRN co-ordinator and WN

	WN assessment (pre-trial)					
	State	**1**	**2**	**3**	**4**	**Total**
CRN co-ordinator (Gold standard)	**1**	1,239 (91.9%)	92 (6.8%)	10 (0.7%)	7 (0.5%)	1,348 (100%)
	2	187 (26.1%)	442 (61.7%)	65 (9.1%)	22 (3.1%)	716 (100%)
	3	11 (7.5%)	47 (32.2%)	82 (56.2%)	6 (4.1%)	146 (100%)
	4	7 (3.8%)	27 (14.5%)	8 (4.3%)	144 (77.4%)	186 (100%)
	Total	1,444	608	165	179	2,396
	WN assessment (mid-trial)					
	State	**1**	**2**	**3**	**4**	**Total**
CRN co-ordinator (Gold standard)	**1**	1,770 (93.7%)	107 (5.7%)	13 (0.7%)	0 (0.0%)	1,890 (100%)
	2	343 (63.2%)	177 (32.6%)	16 (2.9%)	7 (1.3%)	543 (100%)
	3	42 (33.1%)	41 (32.3%)	39 (30.7%)	5 (3.9%)	127 (100%)
	4	6 (13.0%)	8 (17.4%)	1 (2.2%)	31 (67.4%)	46 (100%)
	Total	2,161	333	69	43	2,606

Table 6.4: Cross tabulation of PURAF Skin assessments between the TVN and WN

	WN assessment					
	State	**1**	**2**	**3**	**4**	**Total**
TVN (Gold standard)	**1**	1,358 (88.0%)	180 (11.7%)	2 (0.1%)	4 (0.3%)	1,544 (100%)
	2	261 (41.1%)	356 (56.1%)	11 (1.7%)	7 (1.1%)	635 (100%)
	3	4 (22.2%)	3 (16.7%)	7 (38.9%)	4 (22.2%)	18 (100%)
	4	0 (0.0%)	19 (29.2%)	1 (1.5%)	45 (69.2%)	65 (100%)
	Total	1,623	558	21	60	2,262

Summary

For both the PRESSURE and PURAF datasets, misclassification by the WN exists; the disagreements differed by more than one category and there was some uncertainty around the diagnosis of a Category 2+ PU. The results are consistent with the concerns identified in the literature review and may be applicable to other clinical settings where researchers may be reliant on less experienced staff, or a less reliable outcome measure, for example due to costs or availability of resources. It is therefore important to understand the impact of potential misclassification in terms of bias and loss of power and to use methods to account for misclassification in a

MSM setting.

It is clear that misclassification has been incorporated into MSM analysis in a variety of settings, but a comprehensive assessment of the impact on power and sample size of a trial designed using an MSM where the outcomes are at risk of misclassification is required as an extension to the simulation study conducted in Chapter 5.

6.3 Aim

The aim of this chapter is to explore how misclassification can be incorporated into the analysis of PU trial data in the MSM setting using hidden Markov models (HMM).

Objectives

1. Apply 4 state progression HMM to PRESSURE WN assessments.
2. Assess the sensitivity of analysis to different starting values of misclassification probabilities.
3. Apply 4 state progression MSM to PRESSURE WN assessments

6.4 Methods

6.4.1 Notation

This section introduces the notation for HMM using the methods described by Jackson *et al* [164]. Let $\boldsymbol{Y}$ denote the true disease process as introduced in Section 4.3.1. Misclassification of a state occurs when the latent (true) state r is incorrectly observed as state s where $r \neq s$. Let $\boldsymbol{Y}^*$ denote the observed process such that $\boldsymbol{Y}^* = \{Y_t^* \mid t \in (0, \infty)\}$, $Y_t^* \in S^* = \{1, 2, ...D\}$. Note that the state space for the observed process is assumed to be the same as the state space for the latent process, that is $S^* = S$. This reflects that there may be errors in correctly classifying the true disease state. The probability of misclassification needs to be modelled jointly with the underlying disease process for unbiased estimation of the model parameters including treatment effects. The probability of misclassification at time t is defined as

$$e_{rs} = P(Y_t^* = s|Y_t = r) \tag{6.1}$$

These misclassification probabilities correspond to the rs^{th} entry of a $D \times D$ misclassification matrix, $\boldsymbol{E}$ given by

$$\boldsymbol{E} = \begin{pmatrix} e_{11} & e_{12} & ... & e_{1D} \\ e_{21} & e_{22} & ... & e_{2D} \\ ... & ... & ... & ... \\ e_{D1} & e_{D2} & ... & e_{DD} \end{pmatrix}, \tag{6.2}$$

where each row sums to 1, and $e_{rr} = 1$ if there is no misclassification of state r. Suppose that individual i is observed at W time points. Dropping the i for simplicity, the observed disease states are recorded as $\boldsymbol{y^*} = (y_1^*, ..., y_w^*)$. The contribution of individual i to the likelihood function is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}|\boldsymbol{y^*}, \boldsymbol{x}) &= p(Y_1^* = y_1^*, ...Y_W^* = y_W^*) \\ &= \sum_{\boldsymbol{y} \in \Omega_W} p(Y_1^* = y_1^*, ..., Y_W^* = y_W^*|Y_1 = y_1, ..., Y_W = y_W)p(Y_1 = y_1, ..., Y_W = y_W) \end{aligned} \tag{6.3}$$

where θ is the vector of model parameters, $\boldsymbol{x}$ is the vector of p covariates and Ω_W is the set of possible paths of latent states at times $t_1, ..., t_w$. Note that θ and $\boldsymbol{x}$ have been suppressed on the right hand side. It is assumed that for every pair of observed states, Y_v^* and Y_w^*, $v \neq w$, the misclassification at time t_w is independent of both the misclassification and the latent states at other times. This can be expressed in notation through

$$P(Y_v^* = y_v^*, Y_w^* = y_w^* | Y_v = y_v, Y_w = y_w) = P(Y_v^* = y_v^* | Y_v = y_v) P(Y_w^* = y_w^* | Y_w = y_w) \tag{6.4}$$

Assuming the Markov property, the individual's contribution to the likelihood function can therefore be written as

$$L_i(\boldsymbol{\theta} | \boldsymbol{y^*}, \boldsymbol{x}) = \sum_{y_1} \mathrm{P(Y_1^*|Y_1)P(Y_1)} \sum_{\mathrm{y_2}} \mathrm{P(Y_2^*|Y_2)P(Y_2|Y_1)} ... \sum_{\mathrm{y_W}} \mathrm{P(Y_W^*|Y_W)P(Y_W|Y_{W-1})} \tag{6.5}$$

where $P(Y_t^*|Y_t)$ is the misclassification probability, and $P(Y_t|Y_{t-1})$ is the transition probability of the latent disease process. The full likelihood function is then given as described in Section 4.3.1 by

$$L = \prod_{i=1}^{N} L_i(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x}) \tag{6.6}$$

This model may be estimated by maximising the log-likelihood function using the *msm* package in R. The misclassification probabilities are often modelled using the logit link function, and a dependence on time or covariates may be incorporated. For example, Bhatt *et al* incorporated age as a time-dependent covariate for estimating the misclassification probability of a cancer screening program [169]. The incorporation of covariates in the estimation of the misclassification probabilities may also go some way to accounting for differential bias in a trial if suspected. The distribution of the latent first state may be modelled through multinomial logistic regression if it is unknown, however, it is unlikely that the first state is unknown in a clinical trial because disease status would be assessed as part of the eligibility criteria. In this instance, a vector of known probabilities for each state may be specified, or an indicator variable could be used, which allows the specification of assessments, which have been measured with no error. If none of these options are specified, the *msm* package in R will assume that all participants start in the same state.

Identifiability

HMM are at risk of identifiability issues when different misclassification and transition matrix combinations lead to the same likelihood value, so that the model parameters are not estimated correctly or the maximum likelihood estimation procedure does not converge. This may be more likely to occur as the models increase in complexity or when fewer data are available for model estimation. Therefore it is important to specify any known features of the data in the model, such as which misclassifications are not possible. Alternatively, the misclassification probabilities can be set to pre-specified values [102] informed by previous research such as an inter-rater reliability study or through some form of elicitation from experts in the clinical application area.

6.4.2 Analysis of PRESSURE trial dataset

To illustrate the application of a HMM the PRESSURE trial dataset was re-analysed using the WN data collected daily.

$$\mathbf{E} = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.6 & 0 \\ 0 & 0.1 & 0.1 & 0.8 \end{pmatrix}. \tag{6.7}$$

In the first instance, the starting values of the misclassification matrix were according to (6.7), which was informed by the data reported in Table 6.3.

$$\mathbf{E} = \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.6 & 0 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix}. \tag{6.8}$$

Second, a HMM will be fitted using a more flexible misclassification matrix according to (6.8). Note that these starting values allow flexibility to estimate where the true misclassifications occurred, and may me useful in situations where there is greater

Table 6.5: Observed state transitions for the PRESSURE trial WN daily assessments

From state ↓	**To state →**			
	1	**2**	**3**	**4**
1	3570	378	36	32
2	777	1755	75	72
3	88	163	256	29

uncertainty about the true misclassification.

Note that throughout the reanalysis of these datasets the observation of the Severe disease state by the WN is assumed to be accurate but the Severe disease state may be reported as Mild or Pre-Clinical disease as informed by Table 6.3. That is, the assessment of the severe disease state is assumed to be 100% specific, but sensitivity is $< 100\%$. Furthermore, the first observation for each patient is assumed to be correct because accurate assessment of the skin was required to determine eligibility.

To assess the sensitivity of results to using misclassified data, a MSM was applied to the WN data assuming they were correct.

6.5 Results

The observed state transitions for the PRESSURE trial WN daily assessments are presented in Table 6.5. Note that there was both forward and backward movement between the Healthy, Mild and Pre-clinical disease states, which is not in line with the assumed disease process where backwards transitions are not permitted. However, in reality it may be reasonable for earlier skin changes to be observed to improve.

In the first instance, the HMM applied to the PRESSURE trial WN data did not converge because the structure of the misclassification matrix were not compatible with the observed data. To enable the model to fit, the starting values for the misclassification matrix were relaxed as in 6.8 in order to allow any combination of

Table 6.6: Observed state transitions for the PRESSURE trial WN daily assessments: most severe state carried forwards

From state ↓	**To state →**			
	1	**2**	**3**	**4**
1	1472	150	11	15
2	0	3699	68	63
3	0	0	1698	55

observed and true states. However, whilst the model converged, the Hessian was not positive definite. Various solutions were attempted such as applying a scaling factor to normalise the likelihood as guided by associated documentation for the *msm* package [102], and a range of different starting values of the misclassification matrix were used, but the the Hessian continued to not be positive definite. This could be due to the relatively small number of observed state occupancies from Mild disease, which accounted for just 7.4% (536/7231) of all observed state occupancies.

When the observed data were assumed to be correct, and the most severe state was carried forward, the observed state occupancies were as presented in Table 6.6. The estimates of the transition intensities in the MSM applied to observed WN data (where the most severe state was assumed to be true) led to different conclusions of the treatment effect compared to the analysis conducted in Chapter 4(Table 6.7). For the transitions 1 → 2 and 2 → 3, the treatment effects were similar in magnitude with slightly wider confidence intervals. However, for the transition from 3 to 4, the treatment effect was attenuated towards a HR of 1 with an estimate of 1.10 $(0.78, 1.55)$ compared to 0.76 $(0.55, 1.05)$ estimated from the MSM applied to the gold standard data in Chapter 4. This attenuation is in line with expectations if there was non-differential measurement error [160, 161]. In this situation there would be a higher number of category 2+ PUs reported by the ward nurses. If there was non-differential measurement error, then the relative difference would be diluted by the additional observations of the absorbing state.

Table 6.7: Results of MSM applied to observed PRESSURE trial data (ward nurse assessments)

Transition	**Baseline transition intensity (95% CI)**	**HR ((95% CI)**
$1 \to 2$	$0.055\ (0.047, 0.064)$	$1.16\ (0.86, 1.56)$
$2 \to 3$	$0.017\ (0.015, 0.020)$	$0.99\ (0.72, 1.35)$
$3 \to 4$	$0.031\ (0.026, 0.036)$	$1.10\ (0.78, 1.55)$

6.6 Discussion

In the PU setting, inter-rater reliability studies have helped to provide reassurance that misclassification is independent of treatment allocation (Chapter 2.3). In this chapter, the cross tabulations of PU assessment suggest misclassification is likely in the trial datasets, however there were challenges with fitting HMM to the data.

If the analysis model is for a binary or TTE endpoint, literature on misclassified dichotomous or categorical outcomes and over or under ascertained events shows that analysis results could be biased. Therefore it would be wise to continue assessing misclassification even when the outcome is binary or TTE, and carefully consider the impact of findings in the analysis. If the analysis is a MSM, the findings of an inter-rater reliability study can be used to inform both whether a HMM is required, and if so, the initial values, possible covariates and structure of the misclassification matrix.

HMM were applied to the PRESSURE trial dataset of WN assessments with starting values of the misclassification parameters initially informed by skin site level summaries in Section 6.2. Note that these starting values were likely to be a worst case scenario for the PU setting because discrepancies due to incorrectly swapping left and right sides will have been absorbed when data were aggregated to patient level. However, the application of HMM to the case study dataset demonstrated difficulties in fitting the model for a range of starting values for the misclassification probabilities.

In dermatology there is heterogeneity between patients in skin types, pigmentation and morphology, which makes accurate assessment more difficult [23, 66, 67, 93, 94]. Allowing misclassification probabilities to be conditional on skin type, modelled using logistic or multinomial regression, could accommodate this heterogeneity in the appearance of skin. However, this results in a much more complicated model with many more parameters, so that such analyses may only be feasible in very large studies. Furthermore, whilst the motivating datasets arose from the acute hospital setting, community care is moving towards remote clinical assessments for both routine care and research, partly in response to the COVID-19 pandemic [170]. Virtual assessments of conditions requiring a visual assessment may lead to increasingly misclassified disease states either due to subjective assessments being more difficult if image quality is not adequate, or because of technology availability.

Applying an MSM to the misclassified data showed that the treatment effects on the later transition were attenuated towards the null as expected. Therefore, if misclassified outcomes are likely, a modelling strategy needs to encompass the misclassification so that unbiased treatment effects can be estimated. Therefore, despite the challenges encountered with the example dataset in this chapter, it is of interest to understand when and how misclassification can be accounted for in a MSM analysis. For example, whether less experienced staff could be used to conduct more frequent assessments for longer to provide a similar level of power that would be obtained by using a gold standard assessor. It is therefore important to understand the impact on power, bias and coverage of using HMM compared to MSM in the design of trials when there are misclassified outcomes, under different assessment schedule and patterns of misclassification.

Chapter 7

Impact of misclassification on power, bias and coverage

7.1 Introduction

The previous chapter described the use of HMM to analyse data subject to misclassification. The application of HMM to existing data was problematic with issues of non-convergence and non-identifiability. The inter-rater reliability for the illustrative datasets in Chapter 6 showed that there were different patterns of misclassification of PU disease state depending on the expertise of the assessor. More experienced assessors tended to accurately assess the absorbing state with only minor misclassification of transient states with adjacent states. Less experienced assessors were less accurate in their assessments with misclassification of the absorbing state also observed. The summaries of inter-rater reliability in the PRESSURE and PURAF studies were used to inform the design of the simulation study, described in this chapter, to assess the impact of misclassification on power, bias and coverage in the MSM setting. The simulation study builds on the results from Chapter 5, encompassing recommendations from the use of MSM under the gold standard method of assessments to explore the potential impact of misclassification. That is, maximum lengths of follow-up of 30 or 60 days with assessments conducted every 1, 2, or 3 days.

7.2 Aim

The aim of the simulation study is to assess bias and coverage of estimated hazard ratios and power of hypothesis tests for a given sample size when; 1) misclassification is ignored in the analysis of data arising in trials designed using MSM as the primary analysis method, and 2) when misclassification is incorporated into the primary analysis through the use of hidden Markov models (HMM).

Objectives

1. Define scenarios to be evaluated.
2. Apply 4 state progression MSM to simulated latent and misclassified datasets and apply 4 state progression HMM to simulated misclassified datasets.
3. Report the power, bias and coverage for each method.
4. Provide advice on the implication for trials in terms of, for example, relative efficiency of choice of assessor and frequency of assessment.

7.3 Methods

7.3.1 Simulation study plan

The ADEMP general framework is used to outline the plan of the simulation study [148]. The aim is described in Section 7.2.

Data Generating Mechanism

The following were fixed for each scenario and are in line with the Base Case from the previous simulation study: number of patients, $N = 1,000$ (chosen because it achieved over 90% power with the unconstrained model in the base case, compared to 500 participants for which approximately 70% power was achieved); number of simulations $= 1,000$; control group transition intensities, $\boldsymbol{q}_0(t) = (0.05, 0.05, 0.03)$;

moderate treatment effect on each transition, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$. The parameters that vary are based on the results of the first simulation study. In particular, there are 2 options for length of follow-up (30 and 60 days) and 3 options for the assessment interval (1 day, 2 days and 3 days). Each of these were were shown to provide adequate power in Chapter 5, but it is of interest to compare say, misclassified daily outcomes with gold standard outcomes collected less frequently. In all scenarios, the proportions of patients in states 1 (Healthy), 2 (Pre-clinical) and 3 (Mild disease) at baseline ($t = 0$) were 15%, 70% and 15% respectively and patients were allocated in a 1:1 ratio to one of two treatment groups (intervention and control).

The latent process was simulated using the same method as described in Chapter 5, with an additional transition from the time of entry into the absorbing state to the end of the follow up period. This was included to allow the misclassification of the latent absorbing state beyond the time that the latent absorbing state is truly entered. Throughout patients were censored at a rate of 5% per day. The observed process was generated by applying misclassification probabilities to the latent process, informed by the illustrative data in Section 6.2. There were 4 patterns of misclassification for transitions to the absorbing state

1. The sensitivity and specificity for observing the absorbing state are both equal to 1, so that $e_{44} = P(Y_t^* = 4|Y_t = 4) = 1$ and $P(Y_t^* \neq 4|Y_t \neq 4) = 1$. This means that all absorbing states were correctly observed and no transient states are incorrectly recorded as an absorbing state.

2. The sensitivity for observing the absorbing state is **not** equal to 1, $e_{44} = P(Y_t^* = 4|Y_t = 4) \neq 1$, but the specificity remains equal to 1, $P(Y_t^* \neq 4|Y_t \neq 4) = 1$. This means that some absorbing states were incorrectly observed as a transient state but that no transient states were recorded as an absorbing state.

3. The sensitivity for observing the absorbing state is equal to 1, $e_{44} = P(Y_t^* = 4|Y_t = 4) = 1$, but the specificity is **not** equal to 1, $P(Y_t^* \neq 4|Y_t \neq 4) \neq 1$.

This means that all absorbing states were correctly observed but that some transient states were incorrectly recorded as an absorbing state.

4. The sensitivity for observing the absorbing state is **not** equal to 1, $P(Y_t^* = 4|Y_t = 4) \neq 1$ and, the specificity is **not** equal to 1, $P(Y_t^* \neq 4|Y_t \neq 4) \neq 1$. Here some absorbing states were incorrectly observed as a transient state and some transient states were incorrectly recorded as an absorbing state.

Therefore, with 4 scenarios for the misclassification of the absorbing state, 3 assessment frequencies, and 2 follow-up lengths, there were 4x3x2 = 24 scenarios for Part *I* of the simulation study (see Table 7.1 for a summary).

The simulation study was developed in 3 parts related to the possible misclassification patterns;

- Part *I* consisted of misclassification between all transient states, but any misclassification of the absorbing state was only with the adjacent state.
- Part *II* consisted of misclassification only with adjacent states for all states.
- Part *III* consisted of misclassification between transient states and allowed misclassification of the absorbing state with a non-adjacent state in addition to the adjacent state.

Part *III* most closely reflects the misclassification observed in the WN assessments in the PU case studies (Chapter 6), whereas Part *II* reflects scenarios where the misclassification is less extreme. This situation is closer to the misclassification levels observed for CRN relative to the CRN coordinator in the PRESSURE trial, where the maximum difference was one state apart. Part *I* is a general case where there may be greater uncertainty in the transient states, but the absorbing state is either measured without error or may be misclassified as the adjacent state only. Note that for Part *II* and Part *III*, the assessment interval and length of follow-up remain fixed at 1 day and 60 days respectively, which means there are just 4 scenarios within each of these parts of the simulation study. Full details of the simulation parameters for each scenario are provided in Tables 7.1, 7.2 and 7.3 corresponding to Part *I*, Part *II*, and Part *III* respectively.

Throughout these scenarios, the misclassification is assumed to be the same for each treatment group for simplicity, although the programs could be updated to explore the impact of imbalanced misclassification between treatment groups.

7.3.2 Estimand and target

In order for the simulation study to evaluate efficiency of the analysis models, the estimand is defined as the estimated coefficients for treatment. To assess power and type 1 error rate, the target is the null hypothesis as outlined in 5.3.1.

7.3.3 Methods to be evaluated

The analysis methods evaluated for each scenario are;

- Model A - MSM applied to the latent process.
- Model B - HMM fitted to the observed data.
- Model C - MSM fitted to the misclassified observed data.

7.3.4 Performance

Mean coverage and power were compared between analyses with Hochberg's method for multiple hypothesis tests [151] used to assess statistical significance of treatment effects on each transition, in line with the testing procedure used in Chapter 5. Bias was reported graphically through box plots of estimated hazard ratios.

A formal sample size calculation for the number of simulations was not conducted, but a total of $1,000$ simulations were completed for each scenario. The same datasets were used to compare statistical methods but different datasets were generated for each scenario being considered.

Table 7.1: Misclassification simulation parameters, Part I (Misclassification of all transient states, misclassification of the absorbing state with the adjacent state at most)

Scenario	**Assessment frequency**	**Length of follow-up**	**Misclassification probabilities**	**Description**
1*	Daily	60 days	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.6 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	No misclassification of absorbing state
2	Every 2 days			
3	Every 3 days			
4	Daily	30 days		
5	Every 2 days			
6	Every 3 days			
7	Daily	60 days	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.6 & 0 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$	Under-reporting of absorbing state
8	Every 2 days			
9	Every 3 days			
10	Daily	30 days		
11	Every 2 days			
12	Every 3 days			
13	Daily	60 days	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	Over-reporting of absorbing state
14	Every 2 days			
15	Every 3 days			
16	Daily	30 days		
17	Every 2 days			
18	Every 3 days			
19	Daily	60 days	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$	Both under- and over-reporting of absorbing state
20	Every 2 days			
21	Every 3 days			
22	Daily	30 days		
23	Every 2 days			
24	Every 3 days			

* denotes the base case

Table 7.2: Misclassification simulation parameters, Part II (Misclassification of all states with the adjacent state at most)

Scenario	**Assessment frequency**	**Length of follow-up**	**Misclassification probabilities**	**Description**
25	Daily	60 days	$\begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	No misclassification of absorbing state
26			$\begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$	Under-reporting of absorbing state
27			$\begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0 & 0.3 & 0.6 & 0.1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	Over-reporting of absorbing state
28			$\begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0 & 0.3 & 0.6 & 0.1 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$	Both under-and over-reporting of absorbing state

Table 7.3: Misclassification simulation parameters, Part III (similar to PU case studies)

Scenario	**Assessment frequency**	**Length of follow-up**	**Misclassification probabilities**	**Description**
29	Daily	60 days	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.1 & 0.3 & 0.6 & 0 \\ 0 & 0.1 & 0.1 & 0.8 \end{pmatrix}$	Under-reporting of absorbing state
30			$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.3 & 0.5 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	Over-reporting of absorbing state
31			$\begin{pmatrix} 0.8 & 0.1 & 0 & 0 \\ 0.3 & 0.5 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0.1 & 0.1 & 0.8 \end{pmatrix}$	Both under-and over-reporting of absorbing state

7.4 Results

Part I: Misclassification of transient states only (Scenarios 1 to 6)

The power and coverage observed for Part I are presented in Table 7.4. Note that throughout the results in this chapter, the Monte Carlo Standard Error for the estimates of power were considered sufficiently small at < 0.016 and are provided in Appendix D.1. Examination of the base case demonstrates that estimated treatment effects (hazard ratios) were unbiased when obtained from the appropriate HMM, Model B (Figure 7.1). In contrast, the model ignoring misclassification, Model C, which was applied to data whereby the most severe assessment observed was carried forward, led to attenuated treatment effects on the earlier transitions, but a slightly larger treatment effect on the transition to the absorbing state. Furthermore, under the latent process, where the data were observed without error and analysed appropriately (Model A), the power was 97.7%, which reduced to 92.2%

when misclassification was present and appropriately analysed (Model B), and reduced further to 85.2% under Model C. For comparison, the power obtained from Model B and Model C is 94.4% and 87.2% respectively, relative to the power achieved under Model A. The coverage of the 95% confidence intervals for each hazard ratio ranged from 95.4% to 96.0% for Model A and Model B, apart from the transition from state 1 to state 2 in model B, which had slightly higher coverage at 97.8%. Model C resulted in inadequate coverage, with 49.9% for $e^{\beta_{12}}$; 3.3% for $e^{\beta_{23}}$ and 89.5% for $e^{\beta_{34}}$. These results were generally consistent throughout this pattern of misclassification for scenarios 1 to 6, under different lengths of follow-up and assessment frequencies. For example, in scenario 5 where patients were assessed every 2 days for up to 30 days, the estimated treatment effects were unbiased for Model A and Model B (Figure 7.2), however there was increased variability in the point estimates obtained for $e^{\beta_{12}}$ from Model B. Model C led to biased treatment effect estimates in the same directions as for the base case. The power for Model A was estimated as 93.4%, compared to 82.0% for model B and 75.1% for model C. The coverage ranged from 94.4% to 96.4% for Model A and Model B, apart from the transition from state 1 to state 2 in model B, which had slightly higher coverage at 98.5%. Model C continues to be inadequate in terms of coverage, however there are slight improvements with 68.0% for $e^{\beta_{12}}$; 18.4% for $e^{\beta_{23}}$ and 90.6% for $e^{\beta_{34}}$. These improvements in coverage for model C may be because the longer intervals between assessments led to a simpler likelihood function because the set of possible paths for the latent process is smaller. Within scenarios 1 to 6, the misclassification probabilities were estimated without bias indicating that this level of misclassification did not lead to identifiability issues, Figure 7.3 is an example with the estimated misclassification probabilities under the base case, and the remaining scenarios are available in Appendix 7.3.1.

Table 7.4: Part I (Misclassification of all transient states, misclassification of the absorbing state with the adjacent state at most): Power and coverage

Scenario	Power (%)			Coverage (95% CI)								
	Model A	Model B	Model C	Model A			Model B			Model C		
				$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$
No misclassification of absorbing state												
1	97.7%	92.2%	85.2%	95.8%	95.4%	95.5%	97.8%	95.9%	96.0%	49.9%	3.3%	89.5%
2	95.8%	88.2%	84.0%	93.1%	95.0%	95.5%	97.4%	95.1%	95.2%	64.4%	16.1%	90.0%
3	96.1%	84.7%	83.8%	94.8%	95.5%	95.5%	99.3%	96.7%	95.0%	73.5%	34.2%	90.3%
4	94.2%	86.9%	74.0%	96.6%	95.0%	95.4%	97.3%	94.8%	95.5%	54.9%	3.6%	89.9%
5	93.4%	82.0%	75.1%	96.3%	94.4%	94.5%	98.5%	96.4%	95.0%	68.0%	18.4%	90.6%
6	93.7%	78.1%	75.0%	95.0%	95.7%	95.3%	99.7%	94.8%	95.1%	70.1%	32.9%	90.8%
Under-reporting of absorbing state												
7	96.0%	91.6%	75.6%	95.6%	95.9%	95.7%	97.6%	96.4%	95.6%	53.1%	4.0%	91.1%
8	95.6%	86.3%	75.7%	95.4%	96.3%	95.7%	98.8%	95.6%	95.5%	61.3%	17.5%	90.9%
9	95.9%	85.7%	75.2%	94.3%	96.3%	94.0%	99.4%	96.2%	94.8%	75.7%	33.9%	90.6%
10	94.3%	87.6%	66.1%	93.8%	95.1%	94.6%	96.8%	94.8%	94.2%	52.8%	2.6%	90.3%
11	94.1%	81.9%	64.9%	93.9%	95.3%	94.4%	98.8%	96.4%	94.7%	67.9%	18.8%	90.6%
12	93.3%	72.9%	61.8%	95.2%	95.2%	95.7%	99.4%	94.5%	96.2%	75.3%	33.4%	93.5%
Over-reporting of absorbing state												
13	96.5%	90.2%	76.1%	96.2%	94.6%	95.1%	96.9%	93.0%	93.4%	52.0%	4.2%	86.2%
14	95.8%	76.6%	69.7%	95.5%	94.7%	94.7%	95.7%	86.9%	91.9%	62.1%	17.4%	86.5%
15	96.5%	93.4%	69.3%	94.1%	93.5%	96.0%	97.0%	56.9%	75.6%	71.5%	31.4%	89.2%
16	94.4%	81.0%	68.2%	94.0%	95.0%	95.6%	98.5%	92.9%	95.8%	50.7%	3.9%	82.2%
17	94.8%	80.2%	61.8%	94.3%	94.2%	95.2%	93.7%	85.8%	89.2%	66.7%	19.9%	85.9%
18	93.2%	89.8%	64.1%	94.2%	95.2%	94.9%	95.6%	55.5%	73.0%	74.2%	36.8%	88.1%
Both under- and over-reporting of absorbing state												
19	96.6%	88.8%	72.0%	95.3%	95.0%	95.2%	95.3%	92.9%	94.8%	50.5%	3.3%	83.7%
20	96.8%	91.6%	65.5%	96.0%	95.5%	95.2%	93.1%	70.3%	9.7%	65.3%	18.9%	83.7%
21	96.3%	55.1%	64.0%	95.4%	95.8%	96.0%	98.7%	43.9%	81.9%	73.8%	33.3%	86.9%
22	94.9%	84.0%	60.5%	94.2%	95.2%	95.4%	97.5%	93.5%	94.9%	52.9%	3.9%	82.4%
23	92.2%	84.3%	54.3%	95.1%	94.0%	94.1%	96.5%	77.5%	82.9%	63.5%	17.8%	79.7%
24	92.8%	35.3%	54.0%	96.2%	94.8%	95.8%	99.7%	72.3%	91.4%	71.8%	33.9%	85.1%

Part I: Misclassification of transient states and under-reporting of the absorbing state (Scenarios 7 to 12)

Under the scenario where the assessments were conducted daily for up to 60 days (Scenario 7), the estimated treatment effect estimates followed a similar pattern to that observed where there was no misclassification of the absorbing state, across all models (Figure 7.4). The power obtained from model A and model B was 96.0% and 91.6% respectively. This is similar to the reduction observed when there was no misclassification of the absorbing state. In contrast, the power under model C, which

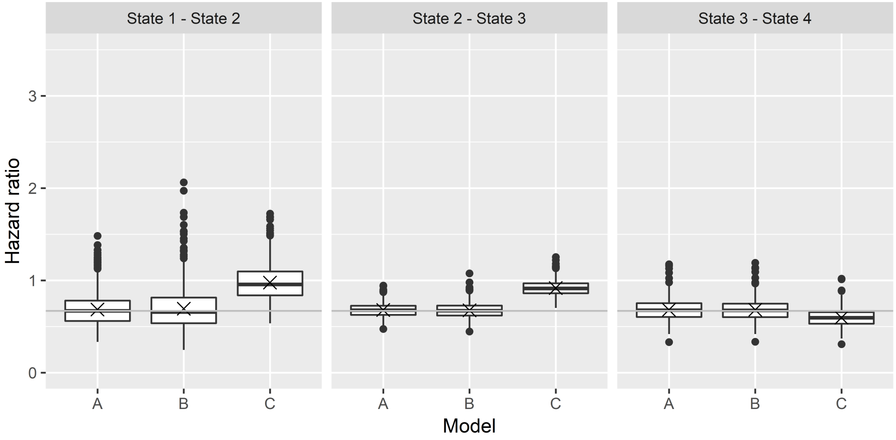


Figure 7.1: Scenario 1, Base Case: Box plot of point estimates for hazard ratios (No misclassification of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

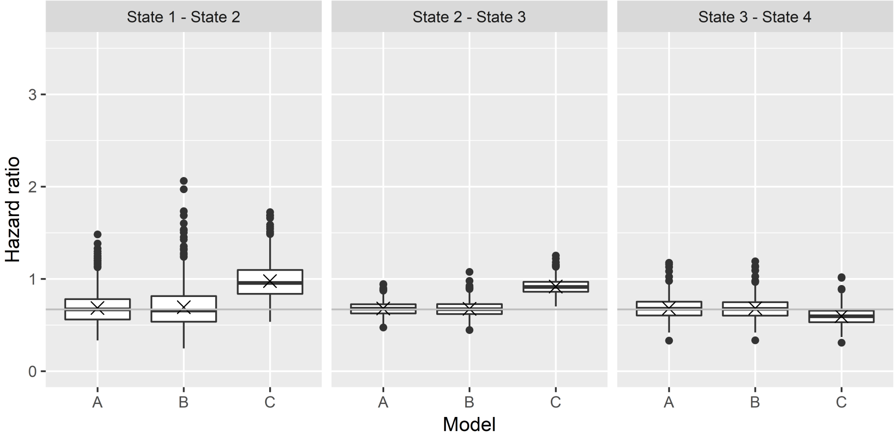


Figure 7.2: Scenario 5: Box plot of point estimates for hazard ratios (No misclassification of absorbing state, Assessment frequency=every 2 days, length of follow-up=30 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

ignored misclassification, was 75.6%. Thus whilst under-reporting of the absorbing state did not have an effect on power when the appropriate HMM was used, ignoring misclassification could lead to a substantial loss of power. The coverage of the 95% confidence intervals for each transition hazard ratio was adequate for Model A and Model B with the exception of the transition between state 1 and state 2 (Table 7.4), which was equal to 97.6%. The coverage for Model C was also similar to that reported for the scenario with no misclassification of the absorbing state, with 53.1% for $e^{\beta_{12}}$; 4.0% for $e^{\beta_{23}}$ and 91.1% for $e^{\beta_{34}}$. These results are overall consistent with the scenarios where there was no misclassification of the absorbing state across different lengths of follow-up and assessment frequency. The main difference between results

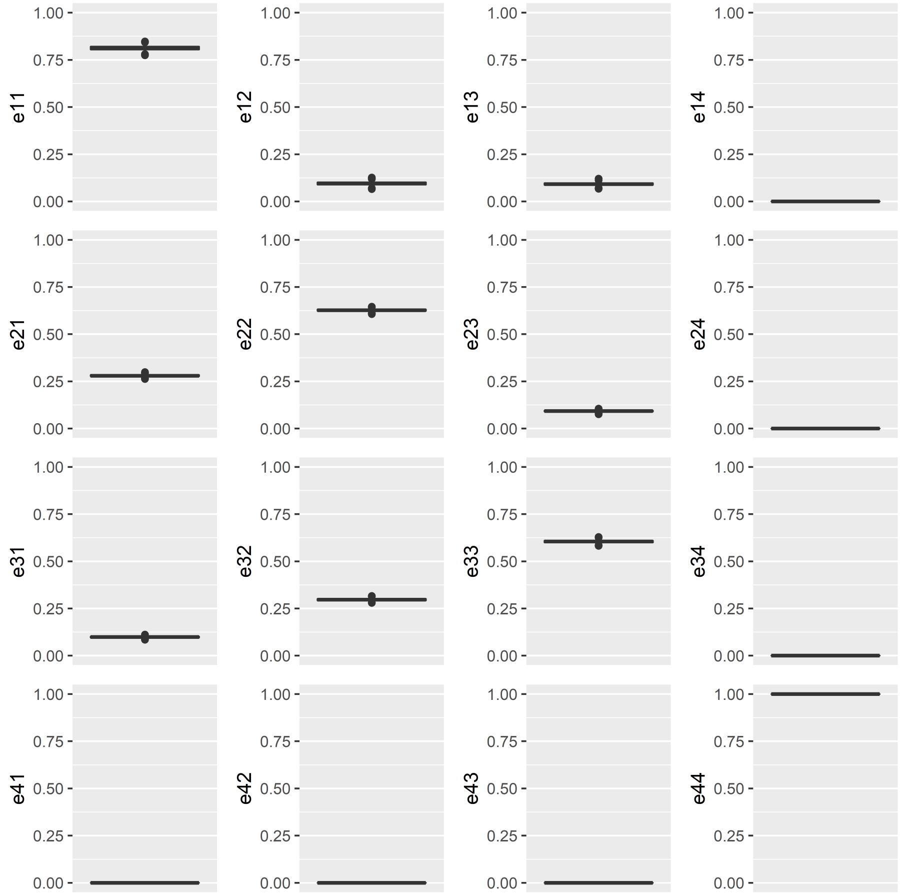


Figure 7.3: Scenario 1, Base Case: Box plot of misclassification probability estimates obtained from model B (No misclassification of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

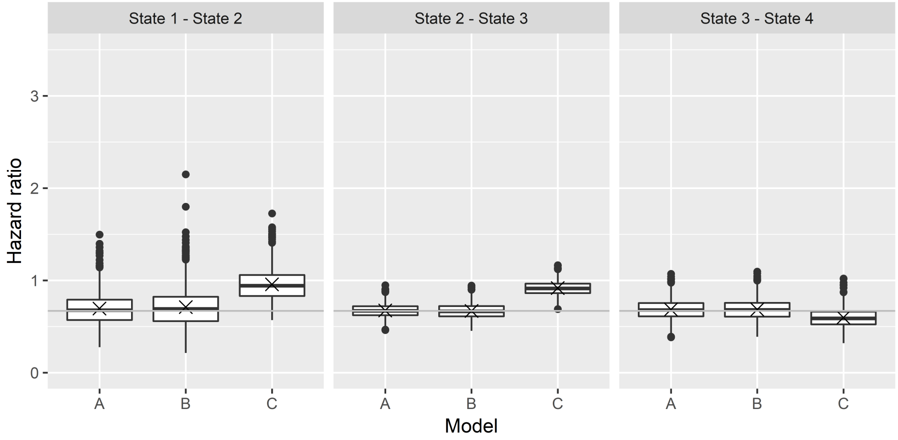


Figure 7.4: Scenario 7: Box plot of point estimates for hazard ratios (Under-reporting of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

from models A and B for scenarios 7 to 12, is that whilst the misclassification probabilities were estimated without bias, there was increased variability in the estimated misclassification probabilities for the absorbing state (for an example, see figure 7.5, although this did not appear to impact on the bias of the estimated treatment effects or the power. However, for smaller datasets, the models may encounter identifiability issues when estimating misclassification probabilities.

Part I: Misclassification of transient states and over-reporting of the absorbing state (Scenarios 13 to 18)

Under the scenario, where the assessments were conducted daily for up to 60 days (Scenario 13), the estimated treatment effect estimates were similar in terms of bias to the previous scenarios (Figure 7.6). The power obtained from model A remained high, at 96.5% whereas model B achieved 90.2% power and the power under model C was 76.1%, which were all in line with the reduction in power observed when there was under-reporting of the absorbing state. The coverage of the 95% confidence intervals for each transition was adequate for Model A as expected, but under Model B coverage was slightly reduced coverage for the transitions between state 2 and state 3, and between state 3 and state 4 at 93.0% and 93.0% respectively. The coverage for Model C was similar to that reported for the previous scenarios, with 52.0% for $e^{\beta_{12}}$; 4.2% for $e^{\beta_{23}}$ and 86.2% for $e^{\beta_{34}}$.

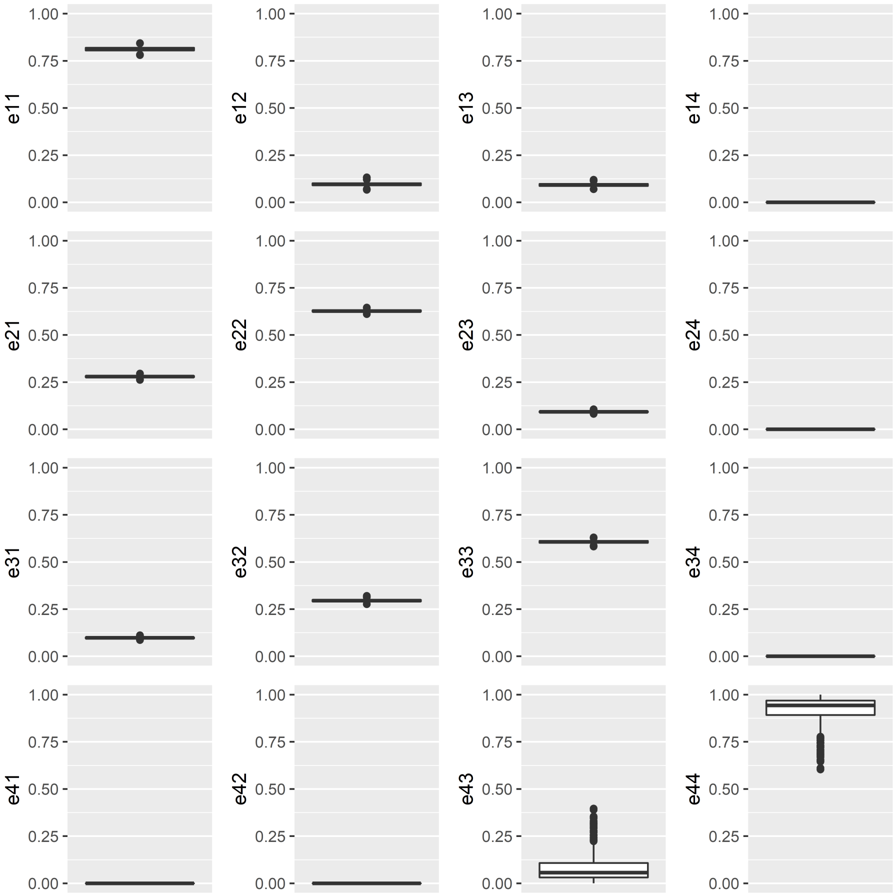


Figure 7.5: Scenario 7: Box plot of misclassification probability estimates obtained from model B (Under-reporting of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

The mean power for model B was affected by changes in the assessment frequency. When the length of follow-up was 60 days, daily assessments led to mean power of 90.2% in line with previous scenarios. Assessments conducted every 2 days led to a reduction in power to 76.6%, however assessments every 3 days led to increased power of 93.4%. When the length of follow-up was 30 days the mean power under a HMM was approximately 80% when the assessments were daily or every 2 days, but when the assessments were every 3 days the power increased to 89.8%.

In this set of scenarios, the distribution of the estimated treatment effects on each transition from model B were also affected as the assessment interval increased and/or the length of follow-up decreased. The increased power when observations were every 3 days was likely to be an artefact of biased treatment effects. For example, in scenario 15 when the assessments were conducted every 3 days for 60 days, the estimates of $e^{\beta_{12}}$ were attenuated towards the null, whereas the estimates of $e^{\beta_{23}}$ and $e^{\beta_{34}}$ were biased away from the null (Figure 7.7). Meanwhile, in scenarios where assessments were conducted daily or every 2 days, the estimates of $e^{\beta_{23}}$ and $e^{\beta_{34}}$ were unbiased in all scenarios. Similarly $e^{\beta_{12}}$ was estimated without bias with the exception of scenario 16 (daily assessments for 30 days) which showed attenuation towards the null.

Throughout scenarios 13 to 18 the mean coverage was inadequate for both $e^{\beta_{23}}$ and $e^{\beta_{34}}$ when assessments were conducted less frequently than daily. There was some loss of coverage for $e^{\beta_{23}}$ when assessments were conducted daily, and there was increased coverage for $e^{\beta_{12}}$ when assessments were conducted daily for 30 days. The misclassification probabilities were, on average, estimated without bias throughout scenarios 13 to 18, which suggests that identifiability was not a concern for Model B.

Overall, over-reporting of the absorbing state had the potential to lead to biased treatment effects with inadequate coverage if the assessment interval was less frequent than daily or if the length of follow-up was shorter than 60 days.

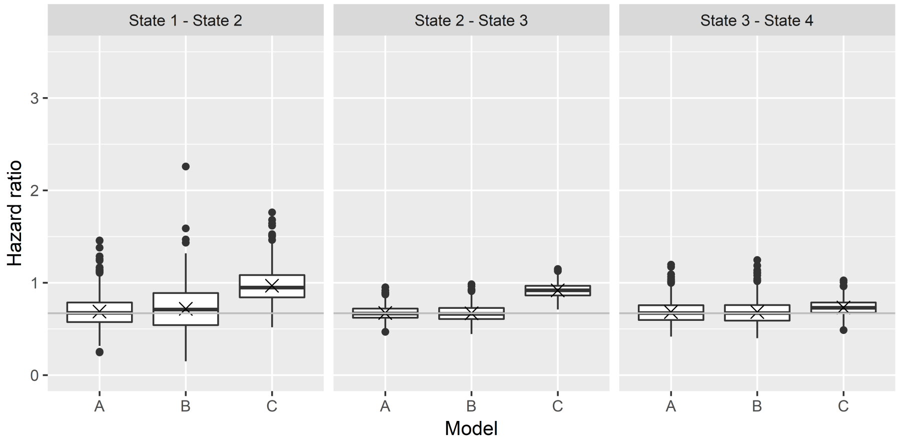


Figure 7.6: Scenario 13: Box plot of point estimates for hazard ratios (Over-reporting of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

Part I: Misclassification of transient states and both under- and over-reporting of the absorbing state (Scenarios 19 to 24)

When assessments were conducted daily for up to 60 days (Scenario 19), the power obtained from model A remained consistent, at 96.6%, model B achieved 88.8%

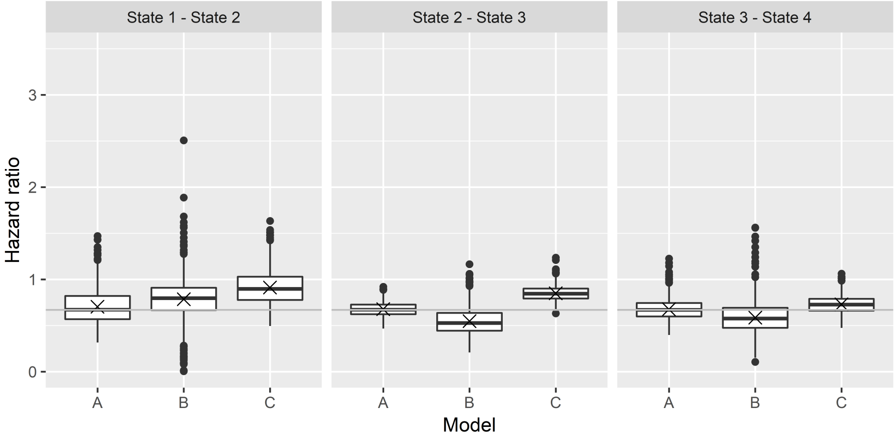


Figure 7.7: Scenario 15: Box plot of point estimates for hazard ratios (Over-reporting of absorbing state, Assessment frequency=every 3 days, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

power and Model C had 72.0% power (Table 7.4). For Scenario 19, the coverage of the 95% confidence intervals for each transition hazard ratio was adequate for Model A, but under Model B the coverage was observed to be 95.3%, 92.9% and 94.8% for $e^{\beta_{12}}$, $e^{\beta_{23}}$ and $e^{\beta_{34}}$ respectively. The coverage for Model C was similar to previous scenarios.

Throughout Scenarios 19 to 24, the estimated treatment effects were similar to that observed when there was only over-reporting of the absorbing state with the exception of scenarios 21 and 24 when assessments were conducted every 3 days for 60 an 30 days respectively. In both scenarios there was greater attenuation towards the null for $e^{\beta_{12}}$ (Figures 7.8 and 7.9). For $e^{\beta_{23}}$ and $e^{\beta_{34}}$ there was greater variability in the point estimates and, compared to scenarios 15 and 18 they were not biased away from the null, but there was in fact evidence of attenuation towards the null particularly for scenario 24. The power was adversely affected in these two scenarios at 55.1% for scenario 21 and 35.3% for scenario 24. The coverage was also poor, particularly for $e^{\beta_{23}}$ at 43.9% for scenario 21 and 72.3% for scenario 24.

The misclassification probabilities were, on average, estimated without bias throughout scenarios 19 to 24, which suggests that identifiability was not a concern for Model B. Therefore, the impact of both under- and over-reporting of the absorbing state was similar to that observed when there was only over-reporting of the absorbing state. That is, the treatment effects are at risk of bias with inadequate coverage if the assessment interval was less frequent than daily or if the length of follow-up was shorter than 60 days.

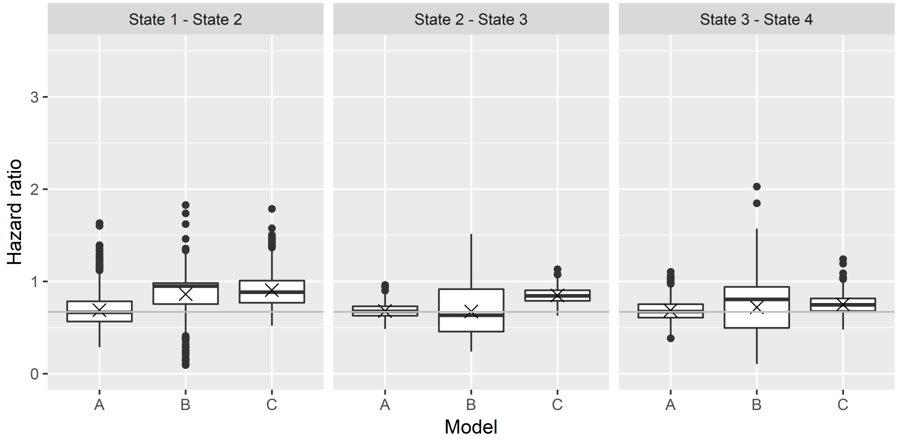


Figure 7.8: Scenario 21: Box plot of point estimates for hazard ratios (Under-and over-reporting of absorbing state, Assessment frequency=every 3 days, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

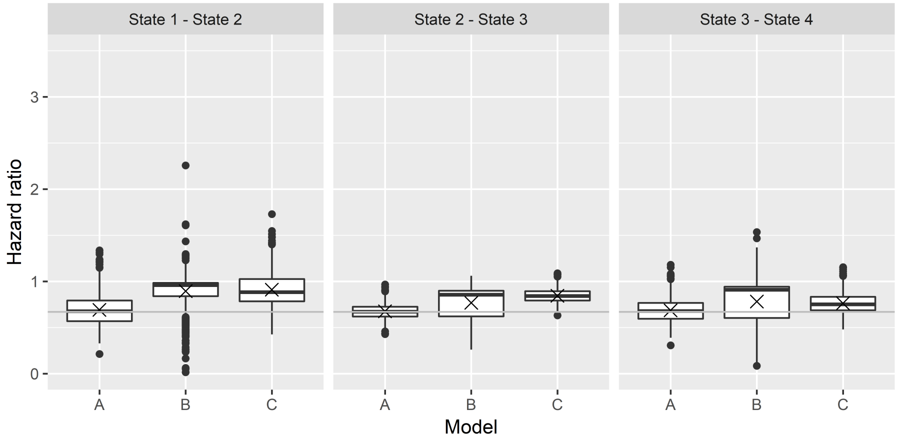


Figure 7.9: Scenario 24: Box plot of point estimates for hazard ratios (Over-reporting of absorbing state, Assessment frequency=every 3 days, length of follow-up=30 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

Part II: Misclassification of adjacent states only (Scenarios 25 to 28)

Scenarios 25 to 28 explored the case where misclassification only occurred between adjacent states. The results of this part of the simulation study were, on the whole, consistent with Part I. The behaviour of the values of the estimated treatment effects drew similar conclusions and the effects on power and coverage were similar as shown in Table 7.5. It is important to note that in scenario 26 where there was only under-reporting of the absorbing state, there was increased variability in the estimated misclassification probability e_{43} (Figure 7.10) which may be an indication that identifiability was an issue in some of the simulated datasets.

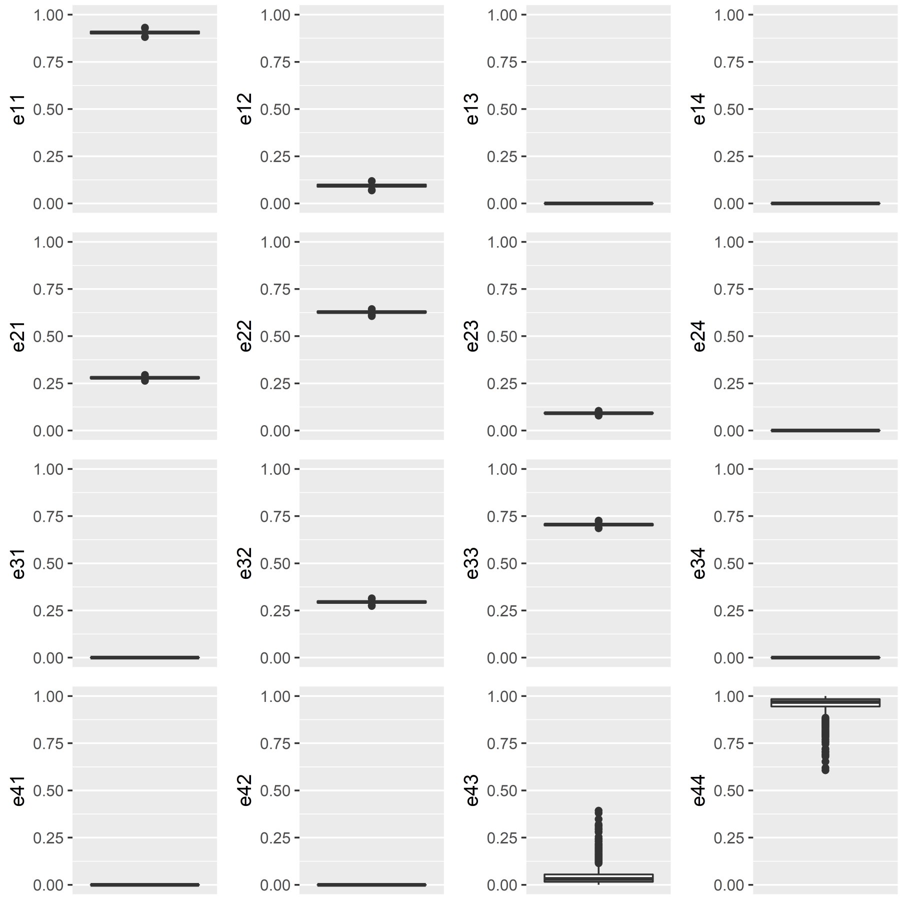


Figure 7.10: Scenario 26: Box plot of misclassification probability estimates obtained from model B (Under-reporting of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

Table 7.5: Part II (Misclassification of all states with the adjacent state at most): Power and coverage

Scenario	**Power** (%)			**Coverage** (95% **CI**)								
	Model A	**Model B**	**Model C**	**Model A**			**Model B**			**Model C**		
				$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$
No misclassification of absorbing state												
25	96.7%	92.7%	86.5%	95.1%	95.0%	95.3%	96.8%	95.5%	95.3%	69.8%	16.7%	91.7%
Under-reporting of absorbing state												
26	96.5%	92.6%	78.3%	94.9%	95.6%	95.1%	97.3%	96.1%	95.5%	66.3%	16.8%	91.1%
Over-reporting of absorbing state												
27	97.3%	91.7%	77.2%	95.1%	96.2%	95.4%	96.6%	93.9%	94.4%	68.2%	16.7%	83.7%
Both under- and over-reporting of absorbing state												
28	96.8%	86.3%	71.3%	95.3%	94.3%	94.7%	96.4%	92.5%	94.2%	68.0%	16.0%	79.5%

Part III: Pressure ulcer setting (Scenarios 29 to 31)

Scenarios 29 to 31 were for the case where misclassification of the absorbing state could have been to more than one state apart. The results of this part of the simulation study were similar the results in Part I and Part II. There was some loss of coverage for the HMM (Model B) when the absorbing state could be over-reported. The mean power ranged from 88.6% to 91.4% (Table 7.6). Furthermore, the treatment effect estimates obtained from Model B remained unbiased in each of these scenarios, with scenario 31 demonstrated as an example in Figure 7.11. Similarly, the misclassification probabilities were, on average, estimated without bias throughout scenarios 29 to 31, which suggests that identifiability was not a concern for Model B in these cases.

Table 7.6: Part III (similar to PU case studies): Power and coverage

Scenario	**Power** (%)			**Coverage** (95% **CI)**								
	Model A	**Model B**	**Model C**	**Model A**			**Model B**			**Model C**		
				$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$	$e^{\beta_{12}}$	$e^{\beta_{23}}$	$e^{\beta_{34}}$
Under-reporting of absorbing state												
29	96.8%	91.4%	77.9%	95.1%	96.0%	95.4%	97.7%	94.4%	94.7%	63.8%	14.6%	92.0%
Over-reporting of absorbing state												
30	97.2%	90.3%	35.1%	95.6%	95.5%	96.1%	94.3%	93.8%	92.3%	67.5%	11.7%	6.6%
Both under- and over-reporting of absorbing state												
31	96.5%	88.6%	33.7%	94.7%	93.9%	95.5%	93.3%	91.5%	92.9%	65.9%	10.7%	5.2%

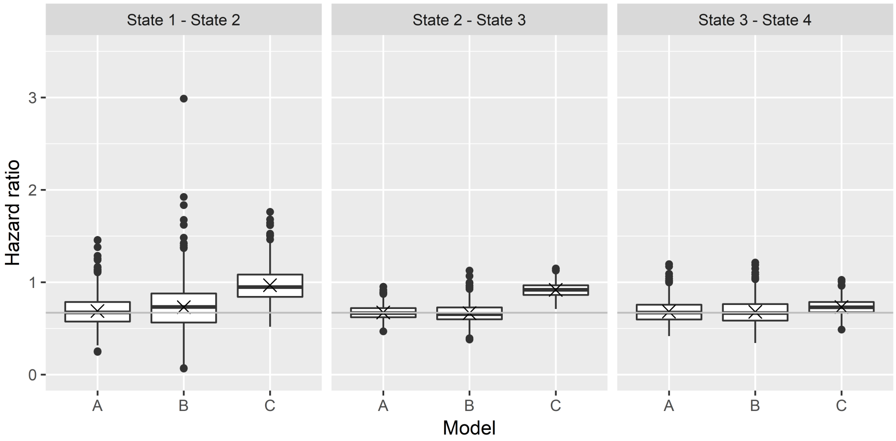


Figure 7.11: Scenario 31: Box plot of point estimates for hazard ratios (Under- and over-reporting of absorbing state, Assessment frequency=daily, length of follow-up=60 days, $N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

7.5 Discussion

The impact of misclassified outcomes on power, bias and coverage was evaluated in a simulation study comparing both HMM and MSM applied to observed data with MSM applied to latent data. The assessment interval and length of follow-up affected both the power, coverage and bias of the treatment effect estimates, with reduced length of follow-up and longer assessment intervals leading to less reliable model conclusions when data were misclassified. The motivating problem for this simulation study was informed by data provided by experts and non-experts in the PU setting discussed in Chapter 6. However, the findings of this chapter are relevant to any setting where the disease process may be measured with error.

Overall, the simulation study showed that when there was only under-reporting or no misclassification of the absorbing state, the power obtained from using a HMM was at least 70% of that achieved from an MSM fitted to data from the latent process, irrespective of varied follow-up length and assessment frequency. At least 90% power was achieved from applying MSM to the true data of 1000 participants when the maximum length of follow-up ranged from 30 to 60 days, and when the assessment intervals were up to 3 days long. Meanwhile, to achieve 90% power with 1000 patients and misclassified data, assessments should be conducted daily with a maximum follow-up of 60 days.

When there was over-reporting of the absorbing state, the power, bias and coverage were adequate when assessments were conducted daily for up to 60 days. However when the length of follow-up was reduced to 30 days, or when the assessment intervals was increased to every 2 or 3 days, there was a more substantial impact on power with HMM providing as low as 38.0% of the power available from using a MSM applied to the true data. Coverage was also worse for the HMM when there was over-reporting of the absorbing state. In almost all scenarios Model C led to reduced power compared to HMM, and poor coverage.

The results of the simulation study indicated that the use of a HMM will reduce the power compared to a MSM fitted to the latent disease process (i.e. the case where states are observed without error), and the decision on whether to conduct a trial using an outcome measure at risk of misclassification should be based on a variety of factors including: the extent of misclassification, availability of accurate measures of disease (e.g. by expert assessors), and the financial cost of using an accurate measure of the disease compared to a less reliable method.

Whilst HMM have the potential to account for misclassified states, it is always preferable to consider the quality of the outcome assessment at the planning stage of the trial. If the absorbing state was only at risk of being under-reported, then the simulation study indicated that HMM were adequate to account for this misclassification. However, if there was any over-reporting of the absorbing state, the results of the simulation study suggest that the HMM may lead to unreliable estimates of the treatment effects and low power and potential identifiability issues. Therefore, it may be appropriate to consider adjudication of the absorbing state, if it is observed, to provide confidence that it is recorded without error, which is in line with recommendations for binary and TTE endpoints [156, 157, 160, 161]. In future simulation studies it may be useful to examine the estimated misclassification probabilities on a logistic scale, rather than a log-logistic scale so that any skewness of estimates can be assessed more easily.

In the context of PU prevention trials gold standard assessments should continue to be used. Despite the simulation study showing that in the PU example, the relative power of applying a HMM to misclassified data was at least 90% of the relative power achieved from using an MSM on the true data, the absorbing state in PU prevention trials may be misclassified which increases the risk of bias and poor coverage. Throughout the simulation study, there were some situations where the HMM converged but the Hessian was not positive definite, although the majority of models could be fitted with no problems. However, there were challenges in fitting the models to daily data provided by ward nurses in the re-analysis of the PRESSURE trial (Chapter 6) which suggests that further work is required to examine when HMM may run into model convergence issues.

Given that identifiability of HMMs is an acknowledged problem in the literature, it is important to assess the sensitivity of the results to different starting values for the optimisation algorithms. If problems with convergence arise then they may be addressed by using initial values that are informed by inter-rater reliability studies or by constraining parameters to plausible ranges. Yi et al [171] propose alternative inference methods for situations, which may be at risk of model misspecification or when model estimation becomes computationally intensive due to complex models or large samples. The methods proposed are based on pairwise likelihood function formulation where a composite likelihood was derived from marginal log-likelihoods [172]. They also utilise an Expectation Maximisation (EM) algorithm, which Cook and Lawless advise is beneficial when the number of processes or assessment times is large [140].

The models examined in the simulation study were based on the Markov assumption, however as discussed in Chapter 5 this may not be valid for studies with prolonged follow-up. If the Markov assumption is not valid, a semi-Markov model can be used if the exact transition times are known. However, these models are not appropriate for panel data because the length of time is unknown. Kang and Lagakos developed methods for a semi-Markov process and misclassification for panel data in continuous time by specifying a minimum length of time spent in each state to limit the number of possible pathways [173].

Another limitation of the simulation study was the assumption of a common misclassification matrix for each treatment group and in some situations it may be important to assess the impact of differential misclassification. In the PU setting, inter-rater reliability studies have been used to provide reassurance that misclassification is independent of treatment allocation [106]. Utilising assessors who are blind to treatment allocation would help to overcome the possibility of differential misclassification, however in some settings blinded assessments are not possible. For example, the interventions in both the PRESSURE and PRESSURE2 trials were mattresses with different modes of action. To conduct a blinded assessment, the participants would need to be moved off the mattress, which is extremely burdensome to both the participant and ward staff. Therefore, inter-rater reliability studies or validation of the outcome are a pragmatic solution to check for differential misclassification. The findings can be used to inform both whether a HMM is required, and if so, the initial values, possible covariates and structure of the misclassification matrix.

One of the reasons for considering the use of less experienced staff in the assessment of PU prevention trials was that they are involved in the patients' day to day care, which means they can record data more frequently, thereby reducing the potential for missing data and reducing the burden on trial patients by only requiring a single assessment as part of their routine care. Furthermore, if research staff are used, they may record that a skin site has not been assessed because a bandage or dressing is in situ, but this may be indicative of an existing PU and would be observed by those changing the dressings. Therefore, with the recommendation that gold standard assessors should be used in PU prevention trials, it is important to recognise the composite nature of the state definition in the motivating example, and how missing data may arise and consequently affect the analysis.

The definition of disease state in the PU trial case studies ignored missing data at both the patient and skin site level. Missing data may occur at a patient level, perhaps because of patient related reasons such as being too unwell to be assessed or for logistical reasons such as there being no assessor available. Missing data may also occur at a component or skin site level, perhaps because the patient cannot be turned over, or because a dressing is in situ. Therefore skin classifications that are not recorded may be associated with the PU stage itself. The method of state definition defined in Section 4.4.1 ignored missing data and used all *available* skin sites to obtain the patient level state; this fails to account for reasons for missing data and the quantity of missing data. Failure to explicitly model this missing data mechanism may result in biased estimates of the rate of PU onset and change. Therefore the potential impact of missing data will be examined in the next chapter.

Chapter 8

Missing data

8.1 Introduction

The simulation studies in Chapters 5 and 6 both assumed that data were observed according to some pre-specified visit schedule; in Chapter 5 data were assumed to be accurate and complete up to the point of censoring, and in Chapter 6 data were complete up to the point of censoring, but were subject to misclassification. There were substantial levels of missing outcome data in the motivating datasets, which are discussed later in Section 8.3.1, however missing outcomes were ignored in the original application of MSM to these data in Chapter 4.

Missing outcomes are ubiquitous in medical research, and depending on the data collection schedule may occur in different patterns. For example, in longitudinal data missing outcomes may occur in a monotonic pattern such that if one observation is missing for an individual then all subsequent observations for that individual are missing. This may occur in a trial if an individual withdraws from further data collection, or is lost to follow-up. Alternatively, data may be missing in a sporadic pattern including intermittently missing. Longitudinal data are susceptible to both sporadic and monotonic missing data patterns, particularly if the follow-up period is long and the visit schedule is burdensome to patients [174].

Approaches to handling missing outcome data in medical research have evolved over time and the choice of method should depend on the analysis model and the missing data mechanism itself. Rubin [175] defined the missing data mechanisms described below, which have been widely adopted and are used throughout this Chapter.

First, some additional notation to be used in this chapter is introduced. Let $\boldsymbol{R}$ be a vector of indicator variables that takes the value 1 if the outcome is observed for an individual and 0 otherwise. Let $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ denote the observed and missing data respectively. Then the probability of observing the outcome is given by

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \tag{8.1}$$

Missing completely at random (MCAR)

Using the framework of Rubin [175], a planned measurement is defined to be "missing completely" at random (MCAR) if the reasons for the data being missing are unrelated both to the outcome itself, and to any patient characteristics or other data that are observed in the dataset. In the PU trial setting, data would be MCAR if the assessment did not take place because the researcher was unwell. Formally, this means that

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) = P(\boldsymbol{R}) \tag{8.2}$$

Missing at random (MAR)

Planned measurements are "missing at random" (MAR) if the probability of the outcome being missing is dependent on other observed variable(s). An example of this in the PU setting could be that patients who are completely immobile may be more likely to have missing data if more than one researcher is required to turn the patient for assessment. In this case, it could be reasonable to assume that the missing data are a random subset of the data for all patients with a similar mobility. Formally, this means that

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) = P(\boldsymbol{R}|\boldsymbol{Y}_{obs}) \tag{8.3}$$

Note that this is a conditional statement, so that if the missing data mechanism is MAR conditional on mobility status, resulting analyses will only be unbiased if mobility status is adjusted for in the analysis.

Missing not at random (MNAR)

Data are "missing not at random" (MNAR) if the probability of the outcome data being missing is dependent on unobserved data, that is if the missing outcome is dependent on its value independently of other data, for example if the unobserved outcomes are systematically different to the observed outcomes. In PU trials this might occur if people with dressings that cannot be removed are more likely to have PUs. Formally, this means that

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) = P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \tag{8.4}$$

Recall from Section 4 that Grüger *et al* [129] defined four observation schedules and examined whether they were informative or not in a simulation study. For RCTs it is common for observation schemes to be fixed in advance, however any deviation from the pre-specified schedule is only ignorable if the deviation independent of the current disease state itself.

The method of handling missing data largely depends on the missing data mechanism. If we can assume assessments are MCAR, then we can continue to use continuous time MSM for panel data as described so far in the thesis [130].

Farewell and Tom [176] described multiple methods for analysing longitudinal data with an informative observation scheme in the context of MSM. The first example was motivated by a study where the outcome was the occurrence of a serious coronary heart disease event, which could be fatal or non-fatal. The fatal event is recorded for the full cohort through the use of registry data, but incidence of non-fatal events may be missing if the subject was lost to follow-up. The MSM structure included a specific state for subjects who are lost to follow-up, and an unobservable state for a non-fatal event that occurs after they have been lost to follow-up. In order for the model to be fitted, the authors made assumptions about the transition rates from the unobservable state that were consistent with the observable states.

These were that the risk of the fatal event was the same for patients who remain in follow up, and those who were lost to follow up after a non-fatal event (i.e. MAR). A further assumption was that the hazard ratio for a fatal event relative to a non fatal event for patients who were lost to follow up was proportional to the hazard ratio for patients who were not lost to follow-up. This particular approach would be worth considering for settings where the entry to the fatal state is always observed.

The second example described by Farewell and Tom [176] is for analysis of PsA clinic data where subjects were assessed as part of their disease monitoring protocol every 6 months but the health assessment questionnaire (HAQ), which quantifies physical functional disability was collected annually. The analysis assessed the relationship between the outcome, physical functional disability, and other independent variables including rapidly changeable time-dependent variables such as the number of permanently damaged joints at the 6 monthly assessments. One approach could have been to analyse the HAQ as assessed annually, however this would ignore the changing values of the time-dependent variables in between assessments. Similarly, analysing the data at 6 monthly intervals led to a missing data problem for the annual HAQ score. A MSM for the outcome and the time-dependent variable was proposed, which increased the number of states from 3 to 9, where each state represented the combination of the last known state based on the HAQ and the level of disease activity. A detailed explanation of this approach is provided in [141], however, this example is only relevant to studies that have different assessment schedules for outcome and time-dependent covariates.

The third example described by Farewell and Tom [176] introduced the idea of a HMM structure to allow fitting of an MSM with missing or partially observed states. This example was in the PsA disease setting again, but modelled the disease activity to identify variables associated with remission. Within a three transient state MSM, the authors used a HMM to account for partial data on the joints in each hand in the state definition. That is, the observed states were based on available data only and were at risk of misclassification if data were incomplete for some of the joints on the hand. The use of a HMM enabled the authors to jointly model the MSM and diagnostic uncertainty caused by missing component data of a composite outcome. The authors also compared the estimates to different definitions of remission concluding that a variety of models could be used to assess the sensitivity of results to missing data.

The final example in Farewell and Tom [176] extended the PsA example to a multi-level analysis where the states of 14 joint locations were modelled, with patient-specific random effects to account for the clustering of joints within patients. In this example, explicitly analysing the states of the individual joint locations avoids the need to consider how partial data in a composite outcome are handled. However, the coding of the maximum likelihood estimation algorithm is more complex and readily available software packages do not accommodate both random effects and data that are interval censored.

Efthimiou et al [134] used an MSM for missing data in a trial where data were collected longitudinally and the patients could be in one of two states (non-response or response). The analysis dataset had no missing intermittent values but there were monotonic missing data once a patient was lost to follow-up. Because interest centred on inferences about treatment effects for patients who were lost to follow up, two additional states were added to the original two state model, "unobserved non-response" and "unobserved response". A variety of models with different permitted transitions were explored to assess sensitivity of the treatment effect estimates under different missingness mechanism assumptions including MCAR, MAR, MNAR, and using single imputation strategies such as Last Observation Carried Forward, and, all missing data were non-responders. The models were fitted within a Bayesian framework but the authors advised that a frequentist approach is possible, through the *msm* package in *R* for example.

Farewell, Su and Jackson [177] used a partially hidden Markov MSM to analyse psoriatic arthritis (PsA) data, where the disease state, minimal disease activity, is determined based on 7 components. At some assessments there were missing components so that the overall state was unknown. The authors assumed that these missing components were MAR because the reasons were most likely to be logistical due to the specific visit schedule for this disease. The likelihood was taken over all possible state pathways between observed components. If the data were assumed MNAR then the authors proposed inclusion of a missing data indicator for each component of the outcome, however this could lead to a complicated model as the number of states and components increase.

Heckman [178, 179] introduced a selection model as a method for handling selection bias in the data collection and subsequent analysis. The selection model jointly models the probability of observing the outcome and the outcome conditional on a set of independent variables. Cole *et al* [180] later extended this by developing a selection model for discrete time MSM with an application to a breast cancer trial. In this model, the disease states represented levels of a categorised quality of life scale, but patients with better quality of life may be more likely to provide a quality of life assessment and therefore any missing observations could be indicative of lower quality of life. The selection model by Cole *et al* explicitly modelled the MSM and the probability of the observation being missing conditional on the latent disease state. In 2010, van den Hout and Matthews [181] also proposed a selection model to jointly model the disease process and the probability of observing each state, to handle non-ignorable missing values when estimating stroke free and total life expectancy. The method proposed by van den Hout and Matthews extends Cole *et al* to continuous time models. These are relevant to this thesis and will be outlined in more detail in the next section.

Joint models of the disease process and the observation scheme have been widely used in the MSM setting. Chen, Yi and Cook [182, 183] published work in this area for both discrete and continuous time processes using an EM algorithm to fit their models. The authors noted that identifiability is a concern but discussed conditions for when model parameters are identifiable for both discrete and continuous time processes. Their discussion is shown for an observation scheme where data were missing in between observations of the same state. For example, they demonstrated the general case when the true complete data for individual i were $\boldsymbol{y} = (y_1, ...y_{j-2}, y, y, y, y_{j+2}, ..., y_W)$ but the observed data were

$\boldsymbol{y^*} = (y_1, ...y_{j-2}, y, m, y, y_{j+2}, ..., y_W)$ where m denotes a missing observation. The authors used the fact that under a progression model, m can only take the value y to demonstrate that the parameters of the missing data model were identifiable.

The methods used by van den Hout and Matthews [181] and Chen [182, 183] where the probability of observing the data were modelled is relevant to settings where patients are scheduled to attend at regular assessments such as in a clinical trial. In other settings it might be relevant to model the observation times themselves [183]. For example, Lange *et al* developed a method to jointly model informative observation times and misclassified data [184]. The method can be implemented in *R* through the *cthmm* package but there are acknowledged limitations including potential for identifiability problems and possibility of model misspecification. A further example of this type of model was published by Gasparini *et al* [185] who jointly modelled the longitudinal process and observation process of a routine data source. In this example the authors expected the observation times to be correlated with the underlying disease severity such that patients with more severe disease would have more assessments. The model for the observation scheme analysed the time to each observation rather than the probability of whether an assessment took place. This is particularly useful for observational data rather than a trial where the assessment times are pre-specified and are expected to be the same for all patients.

The PU case studies described throughout this thesis were subject to missing data both at a patient level and a skin site level, and the missing data mechanism may be informative. For example, a skin site assessment may be incomplete because a bandage or dressing was in situ, but this may be indicative of an existing PU. The definition of patient level state first described in Chapter 4 assumed missing data were ignorable but it is important to reconsider the state definition in including how missing data may arise and consequently affect the analysis. Based on the discussion of incomplete observation schemes in the MSM literature, a selection model will be explored to jointly model the MSM and the missing data mechanism in line with van den Hout and Matthews [181]. Misclassification of the state based on partial data will also be considered using a HMM in line with Farewell and Tom [176].

8.2 Aim

The aim of this chapter is to assess the sensitivity of analysis results to different definitions of the missing data mechanism in the PRESSURE2 trial.

8.2.1 Objectives

- Examine reasons for missing data in PRESSURE2.
- Propose candidate definitions for the missing data mechanism in PRESSURE2.
- Demonstrate how a HMM can be equivalent to a selection model.
- Analyse PRESSURE2 using HMM to jointly model the disease process and the missing data mechanism.

8.3 Methods

8.3.1 Missing data in PRESSURE2

In order to determine which methods are most appropriate, the reasons for missing data in PRESSURE2 were explored. Recall that in the PRESSURE2 trial, the assessment schedule consisted of twice weekly assessments for the first 30 days and once weekly thereafter until the participant completed the treatment phase. At each assessment 14 pre-specified skin sites should have been assessed by the research nurse and the skin state recorded. So far, missing assessments at either the patient level or skin site level have been ignored. The remainder of this chapter investigates the reasons for missing data in PRESSURE2 to inform criteria for non-ignorable missing measurements, and the sensitivity of analysis models under these criteria.

Missing data at both the skin site level and patient level were anticipated at the start of the PRESSURE2 trial, which led to the reasons for missing data at each level being collected. In preparation, a number of reasons for missing data were pre-specified on the data collection forms, in addition to allowing free text fields. The first step in establishing the most realistic missing data mechanism was

Table 8.1: Possible reasons for not recording PU category

Probable association with the latent PU state	**Reasons**
Very likely to be associated with adverse PU state	Dressing in situ, Incontinence Associated Dermatitis or moisture lesions, device-related ulcer, blisters, scuffs, excoriated
Possibly associated with adverse PU state	Bandage in situ, unable to assess due to medical device in situ, unable to assess because participant is unwell, unable to move participant, infection control measures, staff safety concerns
Very unlikely to be associated with adverse PU state	Cast in situ, other chronic wound, surgical wound or bruising, traumatic wound or bruising, dermatological skin condition, participant has been discharged
Unknown association	Unable to assess, unable to assess because the participant refused, missed by research nurse, participant unavailable, participant transferred to another inpatient facility, participant withdrawn from trial, participant died, family member unavailable or refused, Not appropriate to assess, lack of staff capacity, hospital transfer, other reason, reason unknown

to ascertain whether the reasons for missing data could be associated with the latent PU state itself. Using subject-specific expertise, my supervisor Professor Jane Nixon categorised each potential reason according to the probable association with the true disease state based on clinical knowledge, blind to treatment allocation (Table 8.1).

Following these categorisations, the frequency of reasons for missing skin site level assessments were presented by skin site in Figure 8.1. Whilst there was a large quantity of unrecorded reasons for missing, this plot shows patterns for different skin sites. In particular, compared to other skin sites, the sacrum and buttocks were most likely to be missing for reasons thought to be very likely associated with the latent PU state. The heels and ankles also have similar patterns to each other, with reasons for missing data thought to be at least possibly associated with the latent PU state.

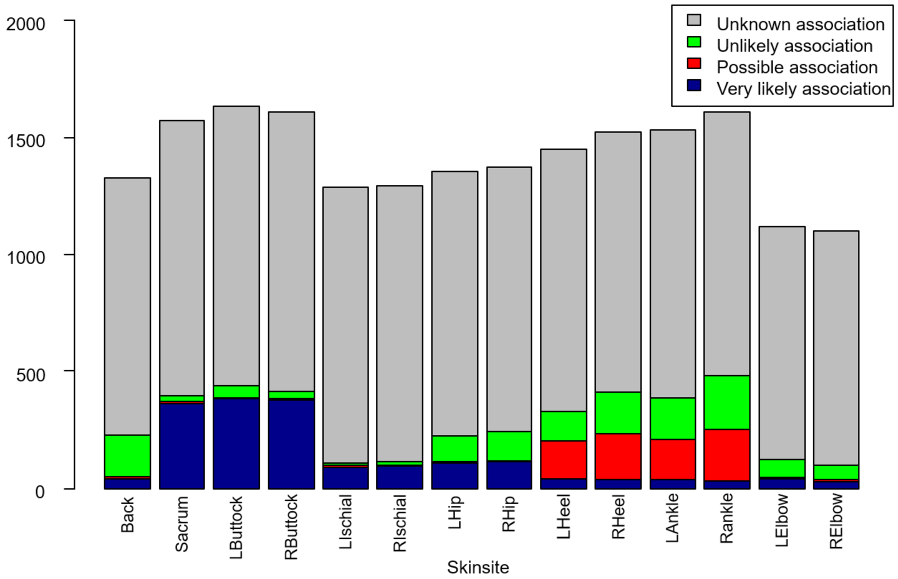


Figure 8.1: Frequency of likely association of reasons for missing PU state data with true PU state at skin site level

The specific reasons, where available, are summarised in Figure 8.2. From this plot, the most common reason for the PU state not being recorded on the sacrum and buttocks was incontinence associated dermatitis (IAD). The reasons for missing data at the heels and ankles were either a bandage or cast being in situ. It is likely that if the heel has a bandage or cast in situ then the ankle on the same side will also have a bandage or cast in situ. In the original examination of association between missing data reason and the underlying PU state it was thought that a dressing in situ would be highly likely to be related to the underlying state. However, the skin sites most frequently reported to have a dressing were the hips were also reported to have surgical wound/bruising suggesting that dressings at these skin sites may in fact have been for a surgical wound, which is very unlikely to be associated with the latent PU state.

The total number of observed PU skin assessments for each skin site are summarised in Figure 8.3. Here, similar patterns are identified; the sacrum, buttocks, heels, ankles and elbows are all less likely than the back, ischial tuberosities and hips to have a healthy PU state recorded. The sacrum, buttocks and heels have the greatest frequencies of a Category 1 PU and Category 2 PU recorded (also see Figure 8.4) compared to the other pre-specified skin sites .

Recall that the analysis of patient level binary and TTE event outcomes in Chap-

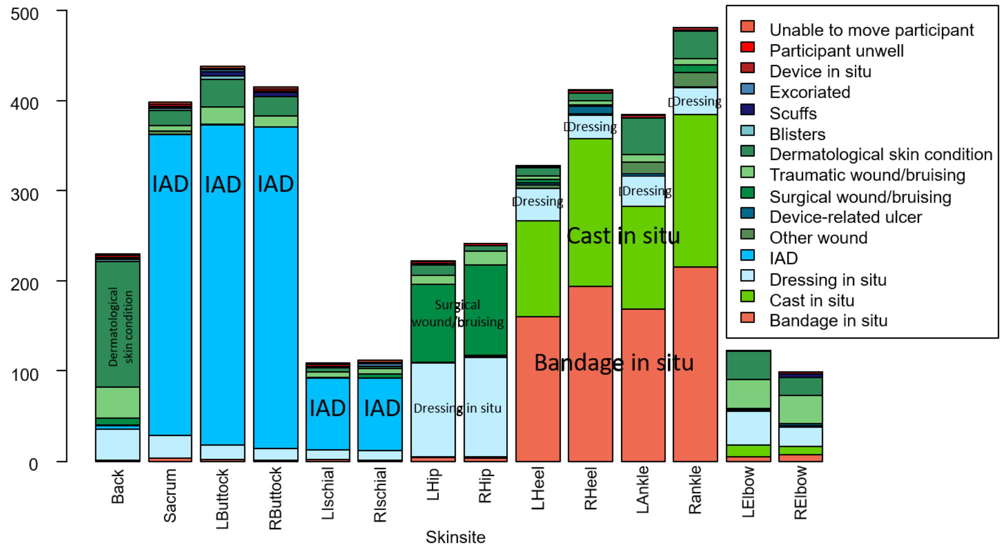


Figure 8.2: Frequency of detailed reasons for missing PU state data at skin site level

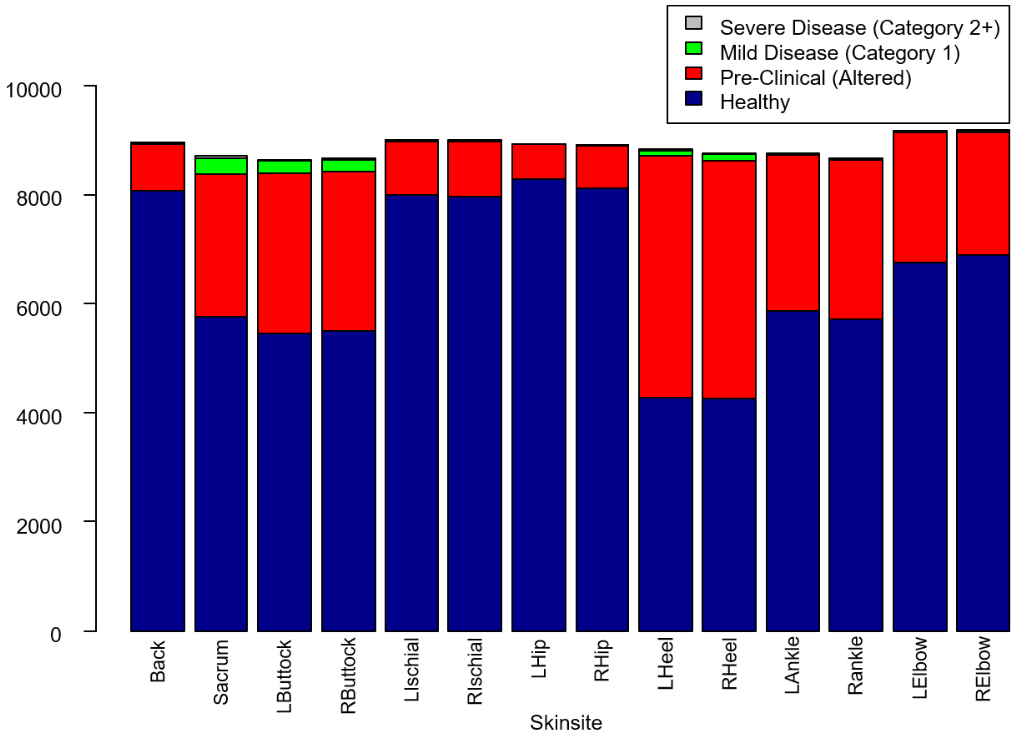


Figure 8.3: Frequency of observed states at skin site level in the PRESSURE2 trial

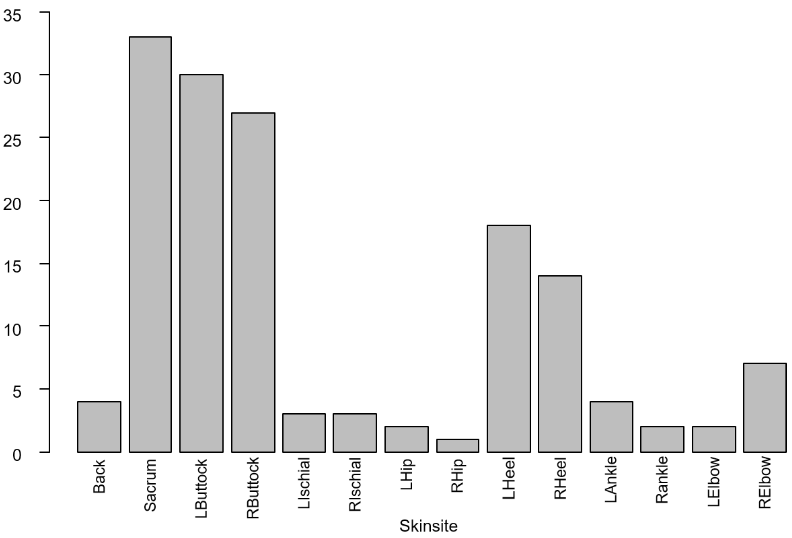


Figure 8.4: Frequency of Severe disease (Category 2+ PU) states at skin site level in the PRESSURE2 trial

ter 3 concluded that it was appropriate to analyse aggregate patient data rather than skin site level data because approximately 90% of the variability in outcomes was at the patient level. However, the patterns observed in the reasons for missing data and the non-missing skin assessments have identified 5 key skin sites/components (out of 14), which if missing, are likely to be non-ignorable. Therefore, it is appropriate to consider a selection model of the joint probability of the MSM and the missing data. In the presence of partial data, these findings also inform the criteria for the definition of patient-level missing data.

8.3.2 Defining missing data

Before using a selection model to analyse the PRESSURE2 data, the disease states need to be defined in the presence of missing data. The patient level state definition was described in Chapter 4. Recall that the state for participant i at time t_{iw}, denoted by $Y(t_{iw})$ is defined by taking the most severe state of observed skin sites. The original approach to defining $Y(t_{iw})$ was to ignore missing data and define the state based on available data. That is,

$$Y(t_{iw}) = \begin{cases} missing, & \text{if } d_{iw} = 0 \\ \max_k\left(X_k(t_{iw})\right), & \text{otherwise} \end{cases} \tag{8.5a}$$

where $X_k(t_{iw})$ denotes the classification of skin site k and d_{iw} denotes the number of components that are observed at time t_{iw} for participant i, $d_{iw} \leq K = 14$. However, this approach assumes that missing skin sites are healthy. In the PRESSURE2 data it is clear that missing data may be associated with an adverse PU classification. One alternative option may be to take a zero tolerance approach to missing data whereby if any components are missing at time t_{iw} for participant i, then $Y(t_{iw})$ is considered missing. Formally, this is given by

$$Y(t_{iw}) = \begin{cases} missing, & \text{if } d_{iw} \neq K \\ \max_k \left(X_k(t_{iw})\right), & \text{otherwise} \end{cases} \tag{8.5b}$$

Whilst this approach does not ignore missing data, it is inefficient in that it will exclude relevant information. For example, if there was a small number of missing components, which do not usually contribute to the overall state, and for which the reason for missing assessment is unlikely to be associated with the true state, then it may be acceptable to ignore these sites when calculating the composite outcome. One option could be to pre-specify a subset of K^* 'key' components, and let d^*_{iw} denote the number of 'key' components that are observed for patient i at time t_{iw}, where $d^*_{iw} \leq K^*$. The definition of $Y(t_{iw})$ can then be given by

$$Y(t_{iw}) = \begin{cases} missing, & \text{if } d^*_{iw} \neq K^* \\ \max_k \left(X_k(t_{iw})\right), & \text{otherwise} \end{cases} \tag{8.5c}$$

The remainder of this chapter will explore the sensitivity of analyses to the three ways of defining the composite disease state defined by 8.5a, 8.5b and 8.5c and will be referred to as Definition A, Definition B and Definition C herein.

8.3.3 Selection model

This section describes the approach taken by van den Hout and Matthews [181] for jointly modelling the continuous time disease process and the probability of observing each state.

Let $\boldsymbol{Y}$ denote the disease process where at time t, $t \geq 0$, individual i is in state $Y_t \in S = \{1, 2, ..D\}$ (note that the index for patient i is suppressed hereafter). The process allows progression only so that participants cannot recover from transient states. Additionally, there is an indicator variable, R_t, which takes the value 1 if the state is observed at time t and 0 otherwise. The conditional probability that $Y_t = y$ is observed is defined as $p_y(t) = P(R_t = 1|Y_t = y, \boldsymbol{x}(t))$ and $\boldsymbol{x}(t)$ is a vector of (possibly time-varying) covariates. It is assumed that these probabilities can be modelled using logistic regression, so that the model for the bivariate distribution of Y_{t_w} and R_{t_w} for an observed time interval $(t_{w-1}, t_w], w \geq 2$ is given by

$$P\{Y_{t_w} = y, R_{t_w} = \nu|Y_{t_{w-1}}, \boldsymbol{x}(t_{w-1})\} = P\{Y_{t_w} = y|Y_{t_{w-1}}, \boldsymbol{x}(t_{w-1})\}p_y(t_w)^{\nu}\{1 - p_y(t_w)\}^{1-\nu} \tag{8.6}$$

where $\nu = 1$ if the state is observed, and $\nu = 0$ if the state is missing. That is, the joint probability of the outcome and the missing status indicator, conditional on the previous outcome and covariates, is the product of the conditional probability of the outcome, and the probability of the missing status. Note that, if the absorbing state is always observed if it is entered, then if $Y_t = D$ it follows that $p_y(t) = 1 \quad \forall \quad t$. In line with the assumption made in Chapter 6 that the true state is known at baseline to ensure eligibility for the clinical trial, it is assumed that $R_0 = 1$.

The model parameters denoted by the vector $\boldsymbol{\theta}$, are estimated by maximising the likelihood function. For an individual, i with W assessment times $\boldsymbol{t} = (t_1, ..., t_W)$, complete (but possibly partially observed) data $\boldsymbol{y^c} = (y_1, ..., y_W)$, and observation indicators $\boldsymbol{r} = (\nu_1, ..., \nu_W)$, the contribution to the likelihood function is given by

$$L_i^c(\boldsymbol{\theta}|\boldsymbol{y^c}, \boldsymbol{r}, \boldsymbol{x}) = P(Y_{t_1} = y_1, Y_{t_2} = y_2, ..., Y_{t_W} = y_W, R_{t_1} = \nu_1, R_{t_2} = \nu_2, ..., R_{t_W} = \nu_W) \tag{8.7}$$

where superscript c is used to denote 'complete'. Under the Markov assumption that transitions depend on current disease stage and not the disease history of the patient and time homogeneity, we have

$$L_i^c(\boldsymbol{\theta}|\boldsymbol{y^c}, \boldsymbol{r}, \boldsymbol{x}) = P(Y_{t_1} = y_1, R_{t_1} = \nu_1) \prod_{w=2}^{W} P(Y_{t_w} = y_w, R_{t_w} = \nu_w|Y_{t_{w-1}} = y_{w-1}) \tag{8.8}$$

If the observed state at time $t_w, w \in \{2, ..., W\}$ is equal to $1, ..., D-1$, then using 8.6

$$\begin{aligned} L^c_{iw} &= P(Y_w = y_w, R_w = \nu_w | Y_{w-1} = y_{w-1}) \\ &= P(Y_w = y_w | Y_{w-1} = y_{w-1}) p_{y_w}(t_w)^{\nu_w} \{1 - p_{y_w}(t_w)\}^{1-\nu_w} \end{aligned} \tag{8.9}$$

Under the assumption that, if it occurs, the absorbing state is always observed and the exact time is known, then if it is observed at time t_w the contribution to the individual's likelihood function is given by

$$L^c_{iw} = \sum_{s=1}^{D-1} P(Y_w = s \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}) q_{sD}(t_{W-1} \mid \boldsymbol{\theta}, \boldsymbol{x}) \tag{8.10}$$

and if they are right censored at time t_w then we assumed the individual is alive but with an unknown state such that

$$L^c_{iw} = \sum_{s \in C} P(Y_w = s \mid Y_{w-1} = y_{w-1}, \boldsymbol{\theta}, \boldsymbol{x}) \tag{8.11}$$

where C denotes the set of possible states at the time of censoring t_w [102, 130].

Therefore, for individual i with assessment times $\boldsymbol{t} = (t_1, ..., t_W)$, complete data $\mathbf{y^c} = (y_1, ..., y_W)$, and observation indicators $\boldsymbol{r} = (\nu_1, ..., \nu_W)$, the contribution to the likelihood function is given by

$$L^c_i(\boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{x}) = P(Y_1 = y_1) \prod_{w=2}^{W} L^c_{iw} \tag{8.12}$$

If there are missing measurements, the likelihood function contribution for individual i is derived by summing over all possible missing states

$$L_i(\boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{x}) = P(Y_1 = y_1) \sum_{\boldsymbol{y}^{\mathbf{c}} \in \Omega(\mathbf{y})} L^c_i(\boldsymbol{y}^c) \tag{8.13}$$

where $\Omega(\boldsymbol{y})$ is the set with all possible paths of latent states.

8.3.4 Hidden Markov model

In order to explore estimation of the selection model described in Section 8.3.3, I simulated a dataset with a similar structure to that used in the original article of van den Hout and Matthews [181]. This was a 3 state illness-death model with a maximum of 9 daily assessments. Dr Ardo van den Hout kindly provided the *R* code used to implement the proposed method and I updated it to suit my simulated dataset. However, implementation of the 'manual' code led to longer computational time compared to other programs used throughout this thesis where an *R* package (*msm*) has been available. Furthermore, the available code relies on the user manually updating the code throughout to align with their dataset. This could become quite burdensome as the number of states and number of missing assessments increase. As an alternative, I demonstrate here how the selection model can be re-formulated as a HMM as in Chapter 6 and implemented using the *msm* package in *R*.

For analysis of data with missing state outcomes, assume that the latent process $\boldsymbol{Y}$ has a state space $S = \{1, 2, ..D\}$, and that the observed process, $\boldsymbol{Y}^*$ has a state space $S^* = \{1, 2, ..., D, m\}$, where m is used to denote a missing state. Recall from Chapter 6 that the probability of misclassification at time t is defined as

$$e_{rs} = P(Y_t^* = s|Y_t = r). \tag{8.14}$$

where $r \neq s$. Now assume that misclassification of a state only occurs through missing data. That is, $e_{rr} = 1 - e_{rm}$ where $e_{rm} = P(Y_t^* = m|Y_t = r)$ is the probability that the assessment is not completed conditional on the latent state $Y_t = r$. It follows from Section 8.3.3 that

$$\begin{aligned} e_{rm} &= P(Y_t^* = m|Y_t = r) \\ &= P(R(t) = 0|Y_t = r) \\ &= 1 - p_r(t) \end{aligned} \tag{8.15}$$

where $R(t)$ is the indicator of whether the state was observed at time t and $p_r(t)$

is the probability that the state was observed (not missing) at time t conditional on $Y_t = r$. Recall that in a HMM the misclassification probabilities are jointly modelled with the MSM process, and the misclassification probabilities are most commonly estimated using a logit link function. This is analogous to the approach taken in the selection model, which jointly models the MSM and the probability of observing the data.

Under the HMM, suppose individual i has W assessment times $\mathbf{t} = (t_1, ..., t_W)$ and 'observed' states $\mathbf{y}^* = (y_1^*, ..., y_W^*)$, then the contribution of individual i to the likelihood function for the HMM is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}|\boldsymbol{y^*}, \boldsymbol{x}) &= p(Y_1^* = y_1^*, ...Y_w^* = y_W^*) \\ &= \sum_{\boldsymbol{y} \in \Omega(\boldsymbol{y})} p(Y_1^* = y_1^*, ..., Y_w^* = y_W^*|Y_1 = y_1, ..., Y_W = y_W)p(Y_1 = y_1, ..., Y_W = y_W) \end{aligned} \tag{8.16}$$

Where $\Omega(\boldsymbol{y})$ is the set of all possible paths of the latent states. It is assumed that the misclassification, or missingness in this case, at time t_w is independent of both the misclassification (missingness) and the latent states at other times. This assumption allows the following

$$P(Y_w^* = y_w^*, Y_{w+1}^* = y_{w+1}^*|Y_w = y_w, Y_{w+1} = y_{w+1}) = P(Y_w^* = y_w^*|Y_w = y_w)P(Y_{w+1}^* = y_{w+1}^*|Y_{w+1} = y_{w+1}) \tag{8.17}$$

Assuming the Markov property, the individual likelihood function can therefore be written as

$$L_i(\boldsymbol{\theta}|\boldsymbol{y^*}, \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \Omega(\boldsymbol{y})} p(Y_1^*|Y_1)...p(Y_W^*|Y_W)p(Y_1)p(Y_2|Y_1)...p(Y_W|Y_{W-1}) \tag{8.18}$$

where $P(Y_t^*|Y_t)$ is the probability that the state is 'misclassified' as missing and $P(Y_t|Y_{t-1})$ is the transition probability of the latent process. Under the assumption that the first state is never missing, that is

$$P(Y_1^* = y_1^*|Y_1 = y_1) = \begin{cases} 1, & \text{if} \quad y_1^* = y_1, y_1^* \neq m \\ 0, & \text{otherwise} \end{cases} \tag{8.19}$$

the likelihood function can be written as

$$L_i(\boldsymbol{\theta}|\boldsymbol{y}^*, \boldsymbol{x}) = p(Y_1 = y_1) \sum_{\boldsymbol{y} \in \Omega(\boldsymbol{y})} p(Y_2|Y_1)p(Y_2^*|Y_2)...p(Y_W|Y_{W-1})p(Y_W^*|Y_W) \qquad (8.20)$$

Note that at time $t_W, W \in \{2, ..., w\}$,

$$\begin{aligned} L_{iw} &= P(Y_w|Y_{w-1})P(Y_w^*|Y_w) \\ &= P(Y_w|Y_{w-1})e_{y_w y_w}^{\nu} e_{y_w y_w^*}^{(1-\nu)} \\ &= P(Y_w|Y_{w-1})p_y(t_w)^{\nu}(1 - p_y(t_w))^{(1-\nu)} \end{aligned} \qquad (8.21)$$

where ν is the indicator, which takes the value 1 if the state has been observed, and 0 if the state is missing. Therefore, the likelihood function in the HMM framework is equivalent to the likelihood function for the selection model described in Section 8.3.3, as given in 8.13 and 8.22

$$L_i(\theta|\mathbf{y}^*, \mathbf{x}) = p(Y_1 = y_1) \sum_{\mathbf{y} \in \Omega(\mathbf{y})} L_{iw}. \qquad (8.22)$$

Therefore, a selection model could be specified in a HMM framework and estimated using the *msm* package in *R*.

8.3.5 Analysis method for PRESSURE2

HMM are used throughout this section to jointly model the MSM and probability of missing data. In order to assess the sensitivity of the analyses to the missing data definitions, the misclassification matrix for the model will be specified in line with 8.23 where e_{rm} is defined in Section 8.3.4.

$$E = \begin{pmatrix} 1-e_{1m} & 0 & 0 & 0 & e_{1m} \\ 0 & 1-e_{2m} & 0 & 0 & e_{2m} \\ 0 & 0 & 1-e_{3m} & 0 & e_{3m} \\ 0 & 0 & 0 & 1-e_{4m} & e_{4m} \end{pmatrix} \tag{8.23}$$

For definitions A and C where partial data may be used to define $Y(t_{iw})$, the state may be misclassified. For example, suppose the maximum of the available components lead to state 2 being 'observed', but that one of the unobserved components is classified as state 3, then the overall state has been misclassified. Only under-reporting is possible in these cases because the state is defined by taking the maximum from a set of components. An alternative misclassification matrix for the selection model is therefore specified in line with 8.24 whereby the model accounts for potential under-reporting of states based on the available data and is a similar approach taken by Farewell and Tom [176] in their analysis of PsA data using partial information on the number of active joints. Note that misclassification is assumed to be not possible under definition B where the overall state is only derived if the component data are complete.

$$E_{partial} = \begin{pmatrix} 1-e_{1m} & 0 & 0 & 0 & e_{1m} \\ e_{21} & 1-\sum_{s\in\{1,m\}} e_{2s} & 0 & 0 & e_{2m} \\ e_{31} & e_{32} & 1-\sum_{s\in\{1,2,m\}} e_{3s} & 0 & e_{3m} \\ e_{41} & e_{42} & e_{43} & 1-\sum_{s\in\{1,2,3,m\}} e_{4s} & e_{4m} \end{pmatrix} \tag{8.24}$$

A summary of the definitions and associated misclassification mechanisms used in the selection model have been summarised in Table 8.2. The model IDs correspond to the definitions of the state, with $A1$, B and $C1$ corresponding to the models

Table 8.2: Missing data definitions and corresponding misclassification matrices to be investigated

Definition of $Y(t_{iw})$	**Misclassification matrix**	**Model ID**
State derived using all available data, missing if no components available		
$Y(t_{iw}) = \begin{cases} missing, & \text{if } d_{jl} = 0 \\ \max_k\left(X_k(t_{iw})\right), & \text{otherwise} \end{cases}$	E	**A1**
	$E_{Partial}$	**A2**
State derived if no components are missing		
$Y(t_{iw}) = \begin{cases} missing, & \text{if } d_{jl} \neq K \\ \max_k\left(X_k(t_{iw})\right), & \text{otherwise} \end{cases}$	E	**B**
State derived using all available data, missing if any key components are missing		
$Y(t_{iw}) = \begin{cases} missing, & \text{if } d^*_{jl} \neq K^* \\ \max_k\left(X_k(t_{iw})\right), & \text{otherwise} \end{cases}$	E	**C1**
	$E_{Partial}$	**C2**

where misclassification is assumed to occur only when missing states are recorded. Models $A2$ and $C2$ are the models for definition A and C when under-reporting of states can also occur because the states are defined using partial component data. Note also that no covariate has been included in the models for the misclassification probabilities, although this is a straightforward extension of the model.

8.4 Results

In the PU example, the components considered mandatory for definition of the overall state are based on the data exploration described in Section 8.3.1 and clinical opinion. Therefore there will be 5 ‘key’ components, or skin sites, namely the sacrum, buttocks (×2), and heels (×2). Applying the 3 proposed definitions yields different levels of missing data, demonstrated in Table 8.3. Definition A, which ignores missing component data and derives the state based on all available data has the smallest number of missing observations, with 951 (11.2%) occasions where a visit was made, but no skin site data were available. Definitions B and C have similar levels of missing data to each other, with Definition B being the most stringent leading to $2,764$ (32.5%) missing observations, and Definition C resulting in $2,158$ (25.4%) observations. There were $2,764$ occasions when there was at least 1 component missing and of these, the majority (78.1%) included at least 1 key component.

The estimated transition hazard ratios with corresponding 95% confidence intervals for the effect of treatment are consistent across all definitions for all transitions as shown in Figure 8.5. The confidence intervals all include 1 for the transition from $1 \rightarrow 2$. For the transition from $2 \rightarrow 3$, whilst the point estimates and confidence intervals for the hazard ratio are similar, under Definition A and Definition C (Model $C2$) the confidence intervals include 1, whereas the confidence intervals for the hazard ratio under Definition B and Definition C (Model $C1$) lie to the left of 1 indicating that the alternating pressure mattress confers a significant reduction in the transition to a Category 1 PU under these definitions and models. All of the definitions led to similar results for the transition from state 3 to state 4 with similar point estimates of the hazard ratio and all of the confidence intervals lie to the left of 1. Under a progression only MSM with no misclassification the likelihood function for the selection model is equivalent to the likelihood function for the MSM if the observed states before and after the missing states are the same because the missing state is known (i.e. there is only one possible pathway). Inspection of these 'start' and 'end' states for the sets of possible pathways in the selection model are shown in Table 8.4, for example, there were 36 cases of missing data sandwiched between two State 1 observations and 11 between State 1 and State 2. Overall a large number of missing observations were between the same 'start' and 'end' states for all definitions of the missing data mechanism; specifically under definition A, 643 (67.6%) missing observations lie between the same states, compared with $1,151$ (41.6%) under definition B and $1,033$ (47.9%) under definition C. This means that there is more information contributing to the likelihood than originally anticipated, and may explain why there is little difference in the width of the estimated confidence intervals. There was a large number of missing observations for which the 'end' state is unknown and will be discussed later in relation to the impact on identifiability.

Table 8.3: Missing data definitions and corresponding misclassification matrices to be investigated

Missing data definition	**From state ↓**	**To state →**				
Definition *A*		**1**	**2**	**3**	**4**	**Missing**
	1	561	130	10	5	54
	2	0	4,987	133	66	519
	3	0	0	1,099	38	128
	4	0	0	0	0	0
	Missing	34	380	105	18	250
Definition *B*		**1**	**2**	**3**	**4**	**Missing**
	1	475	94	8	5	112
	2	0	3,417	82	49	1,000
	3	0	0	761	29	203
	4	0	0	0	0	0
	Missing	44	569	154	44	1,449
Definition *C*		**1**	**2**	**3**	**4**	**Missing**
	1	491	106	9	5	95
	2	0	3,941	98	54	855
	3	0	0	844	33	192
	4	0	0	0	0	0
	Missing	42	533	146	35	1,016

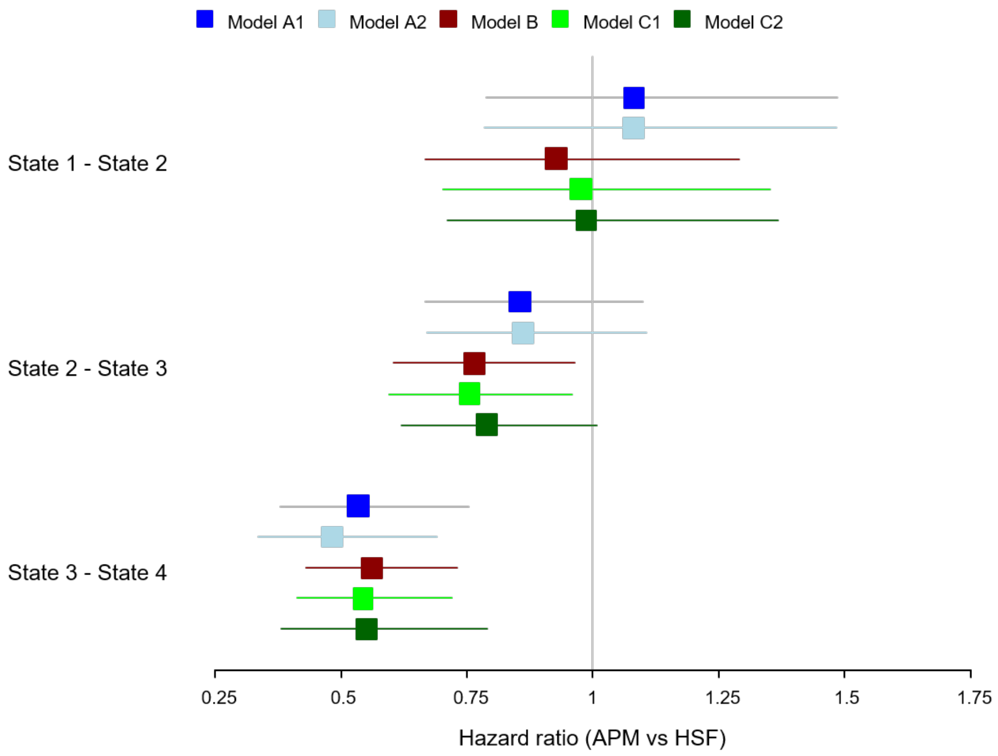


Figure 8.5: Forest plot of estimated treatment effects

Table 8.4: Number of missing observations between observed states

Missing data definition	**Start state ↓**	**End state →**				
Definition A		**1**	**2**	**3**	**4**	**Missing**
	1	36	11	1	2	12
	2	0	491	24	23	171
	3	0	0	116	5	59
Definition B		**1**	**2**	**3**	**4**	**Missing**
	1	38	55	7	9	134
	2	0	927	84	76	1,080
	3	0	0	186	12	156
Definition C		**1**	**2**	**3**	**4**	**Missing**
	1	45	37	3	9	94
	2	0	824	69	55	719
	3	0	0	164	8	131

The estimated misclassification probabilities with 95% confidence intervals are presented in Figure 8.6. From this plot it is clear that there are some spurious results, particularly when the latent state is State 4. This could arise due to the small amount of data available for the final state. Appendix E includes an extended plot of the probabilities of misclassifying a state based on partial data, i.e. e_{12}, e_{13}, e_{23}. These results indicated that the probability of misclassifying a state based on partial data is likely to be very small (Models $A2$ and $C2$). A pragmatic decision could therefore be that these misclassification probabilities do not need to be explicitly modelled. Figure 8.6 provides a closer look at the estimated probability of missing measurements, conditional on the latent state for Model $A1$, Model B, and Model $C1$. For model $A1$ where missing skin sites are ignored, the probabilities of missing data given the true state are generally small, with less than 10% probability of the transient states being missing conditional on the true state. These results are expected given that Definition A leads to lower levels of missing data in the dataset. In comparison, models B and $C1$ led to higher estimated probabilities of missing data. In particular, for each definition, the estimated probability of the state being missing, given that the latent state is Healthy, is approximately 15%, whereas the probability of missing data increases to between 20% and 30% when the latent state is 2 (Altered) or 3 (Category 1) with Definition B leading to a higher probability. The estimated probability of missing data conditional on the latent state being absorbing is 40% for Model A and 80% for models B and $C1$. Table 8.4 shows the number of missing observations between observed states. Under definition A there were 242 missing observations for which the 'end' state was unknown, $1,370$ such observations under definition B and 944 under definition C. These missing observations could have occurred because there were repeatedly incomplete observations due to say, the patient being too unwell, before they reached the end of their treatment phase. If the absorbing state was always known to be observed, for example if the absorbing state was death, then this state would never be missing, however in the PU setting the absorbing state may be missed. In future research, as with the conclusions from Chapter 6, an assessment of patient PU status at the end of

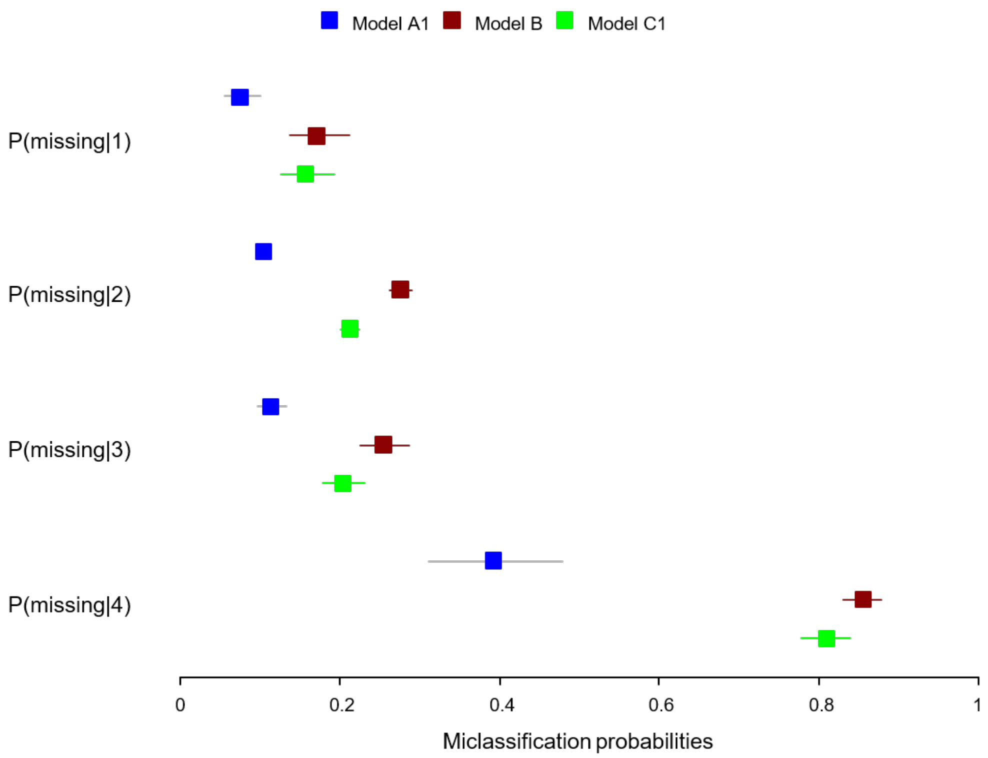


Figure 8.6: Forest plot of estimated misclassification probabilities excluding Models $A2$ and $C2$

their trial participation would be needed to confirm whether the absorbing state was entered and therefore reduce the risk of non-identifiability.

8.5 Discussion

The available methodology for dealing with missing data in the context of MSM was summarised in this chapter. I have shown how a selection model could be constructed using HMM applied to the PRESSURE2 dataset and assessed the sensitivity of analyses to different data mechanisms. The dataset included detailed reasons for missing data, which led to a conclusion that the mechanism was MNAR. The patterns of missing data and outcomes at the skinsite component level led to three potential definitions to determine when the overall state was missing in the dataset. These definitions were based on the available component, or skin site, level data considering both the quantity and any specific 'key' components that were missing. Definition A assumed that the missing data were ignorable, i.e. that the missing components were no worse in severity than those observed. Definition

B assumed that any missing data were non-ignorable, and definition C assumed that only missing 'key' components were non-ignorable. A HMM was then used to jointly model the MSM and the probability of data being missing under the 3 proposed definitions. These analyses demonstrated that the results in PRESSURE2 first presented in Chapter 4 were not sensitive to the missing data model used. However, this result may not be generalisable to other datasets, so careful consideration of the definition should be given during the trial design. In some situations, such as when the state definitions represent a composite outcome from a range of quality of life markers, collecting the reasons for missing data would be challenging. In these instances, a range of possible scenarios should be discussed with clinicians and patient representatives to establish likely missing data mechanism.

One of the challenges in the analyses applied to this dataset was the identifiability of misclassification probabilities associated with the absorbing state. In the example discussed by Van den Hout and Matthews [181] the absorbing state was always observed if it occurred whereas in the PU example, the absorbing state may not be observed even if it occurs. This causes a major issue for convergence of the maximum likelihood optimisation algorithm, and may well be apparent in other applications where the absorbing state can be misclassified. A solution in future trials would be to include an exit assessment conducted by an expert to confirm whether the absorbing state has been entered. A further issue of model fit occurred when the model included misclassification of states when defined using partial data. Specifically, wide confidence intervals were estimated for the misclassification probabilities associated with the absorbing state, which could be due to a small amount of data available to estimate these parameters. In the PU setting, the results suggested that states defined using partial data were at low risk of being misclassified and therefore it could be ignored, however in another setting sensitivity to this should be assessed.

Although the analysis presented in Chapter 4 indicated that analysing the PU datasets at the skin site level did not improve model fit, in some disease settings it might be appropriate to account for correlated disease processes in the analysis. Zhang et al [186] extended the selection model used in this chapter and published by Van den Hout [181] to jointly model interval censored data with within-unit clustering and MNAR data using a Monte Carlo EM algorithm. They applied the method to a clinical trial dataset and also conducted a simulation study, which concluded that the proposed methods have good operating characteristics, and that inappropriate models can lead to biased treatment effect estimates. For correlated data within an MSM this method could be used to address missing data problems.

This chapter has focused on analysing data with missing outcomes, but we have not attempted to accommodate missing covariates since they are rarely a major problem for RCTs. However, time-varying covariates may be important and are more likely to be subject to missingness. This topic has been discussed in the context of MSM by Lou et al [187] who proposed using an EM algorithm for an MSM with missing dichotomous covariates. Other studies have used multiple imputation of missing data. For example, Eleuteri *et al* [188] used an MSM (specifically a competing risks model) to analyse data on patients treated for uveal melanoma; in this analysis there were missing covariates including continuous variables. Multiple imputation using the Alternating Conditional Expectations algorithm, coupled with an approximate Bayesian bootstrap was used to accommodate missing data during estimation of model parameters [188].

The focus of this chapter was to summarise available methodology for dealing with missing data in the MSM context and to analyse the PRESSURE2 dataset in detail with respect to missing composite outcomes. Future work could include a simulation study investigating this the impact of ignoring missing data on the power of a trial designed to by analysed by a MSM, including different missing data patterns, as well as the impact of both misclassification and missing data.

Chapter 9

Discussion

9.1 Summary of key findings

Current methods of design and analysis

This thesis included a review of the literature surrounding the motivating problem of PU prevention trials. The findings were that longitudinal discrete outcome data are commonly aggregated to a binary or TTE outcome, which is an inefficient use of the data provided by participants. In Chapter 3 I applied logistic regression and Cox PH regression to 2 published trial datasets that were used to illustrate methods throughout the thesis and motivate methodological choices for simulation studies. These analyses were conducted at the patient level and skin site level and demonstrated that over 90% of the total variance was due to between patient variability. Therefore, the analyses throughout the remaining chapters were conducted at the patient level. In the PRESSURE case study, the effect of the intervention was not shown to be statistically significant for the binary or time to event analyses. In the PRESSURE2 case study, both the binary and TTE analyses concluded a statistically significant treatment effect suggesting that the intervention provided a benefit to patients in terms of the onset of Category 2+ PU. For both case studies, inspection of the KM-plots suggested that the proportional hazards assumption was not valid.

MSM analysis of PU data

Recognising the limitations of aggregation of longitudinal measurements to a single outcome measure, in Chapter 4 I introduced MSM and applied a 4 state progression MSM to the 2 case study datasets demonstrating that they provide a deeper insight into the natural history of the disease and the disease stage at which treatments may have a greater benefit. For example, for the PRESSURE2 trial, the treatment effect was not statistically significant for the early transitions ($1 \rightarrow 2$ and $2 \rightarrow 3$), but a treatment benefit was observed on the final transition, $3 \rightarrow 4$. This finding was consistent with the Kaplan-Meier estimates from Chapter 3 that suggested a delayed treatment effect. For both datasets the fit of the estimated 4 state progression model was shown to be adequate upon inspection of the observed and model-fitted prevalence plots.

Impact on power and sample size for disease prevention trials designed using MSM

I defined a hypothesis test for multiple effect estimates in the MSM setting in Chapter 5 and conducted a simulation study to assess the impact on power and sample size of analysing longitudinal assessments of outcomes using MSM compared to logistic and Cox PH regression models of aggregated outcomes. In some scenarios there was increased power, or reduced sample size, but the baseline transition intensities and treatment effects were influential in these conclusions. That is, greatest improvements in efficiency were observed when early changes could be observed due to frequent assessments, high early transition rates or larger effects on early transitions. Where the design featured a short follow-up period, long intervals between assessments, slow transition through early states relative to later states or low treatment effect on early states relative to the treatment effect on later states, there was little to be be gained from MSM in trial efficiency.

Impact of misclassified outcomes on power, bias and coverage

Because PU categories can be misclassified, in Chapter 6 I introduced HMM and applied them to ward nurse assessments provided in the PRESSURE dataset. From a range of starting values, the models were found to converge but the Hessian was not positive definite. A MSM was applied to the observed ward nurse data where the most severe observation was carried forward, and the analysis was shown to be sensitive to the misclassification with an attenuated treatment effect on the final transition. To further explore the impact of misclassification on power, bias and coverage, in Chapter 7 I designed and implemented a wide-ranging simulation study. Inter-rater reliability studies conducted alongside PRESSURE and the PURAF diagnostic accuracy study were used to inform various scenarios for the likely misclassification patterns. The results showed that HMM could lead to unbiased results with little loss of power for plausible trial scenarios, for example, when assessments were conducted daily the power achieved by a HMM applied to misclassified data was at least 90% of that achieved by a MSM on the latent data. Ignoring misclassification or model misspecification in the analysis was shown to lead to biased results and poor coverage. When misclassification of the absorbing state was possible there were issues with bias and poor coverage which was supported by the reanalysis of the PRESSURE dataset. Therefore, for PU prevention trials, the gold standard assessment should continue to be used until a more robust modelling strategy has been developed to deal with situations with misclassified absorbing states. If misclassified assessments are unavoidable in PU trials then the assessments should be conducted daily for up to 60 days to minimise the risk of bias and to maximise power.

Missing data in PU trials

Missing assessments are common in longitudinal data. In Chapter 8, using methods developed by Cole [180] and van den Hout and Matthews [181], I demonstrated that selection models can be specified within the HMM framework in order to conduct sensitivity analyses to different MNAR assumptions. This was applied to data from the PRESSURE2 trial. The results demonstrated that in this particular example, the results were not sensitive to different assumptions about the missing data mechanisms. A similar analysis approach could be taken in other clinical settings and may have different conclusion depending on the amount of missing data and the mechanisms driving its absence.

The following sections further summarise implications for practice for key stakeholders involved in clinical trial design; patients, clinical researchers and trial statisticians. This is followed by a section on the limitations of the research together with suggestions for future research and a final conclusion.

9.2 Implications for practice

9.2.1 Patients

The results of this thesis have been communicated to the Pressure Ulcer Research Service User Network (PURSUN). Patients were shown the additional information that can be extracted from estimated MSM during analysis of trial results, as well as demonstration of when treatments may be effective and that by using MSM trials may be less burdensome for patients through reduced assessment frequency and length of follow-up. Furthermore the members of PURSUN thought it was important for funders to save money and for researchers to have more efficient trial assessment schedules. The over-arching feedback was that, for patients, prevention is always better than treatment because PUs can have a detrimental impact on mental health as well as physical health. The group therefore particularly welcomed methods that may help to identify interventions that prevent early stages of pressure damage.

9.2.2 Key stakeholders

In addition to patients, the findings of this thesis have important implications for key stakeholders such as policy makers, health economists and commissioners. In particular, decision makers often require economic evaluations alongside clinical effectiveness to guide their recommendations [142]. The models proposed in this thesis have the potential to align the primary analysis model with multi-state models that are often used in health economics analyses. Discussion of the appropriate clinical effectiveness model will help ensure that an appropriate clinical model is also used for the cost effectiveness analysis. This will consequently strengthen the the interpretation and conclusions of the analyses together.

9.2.3 Clinical researchers

The literature review highlighted that there were two main outcomes of interest to researchers of PU prevention interventions. These were Category 1 PUs or Category 2 PUs. Both have their merits with Category 1 PUs being more common and therefore leading to potentially smaller trial sample sizes, however there was a common concern in the literature that Category 1 PUs are difficult to assess. Some researchers may therefore use a Category 2 as the primary endpoint of a trial because it is less likely to be misclassified, accepting that it will lead to larger trial sample sizes in order to detect a statistically significant treatment effect. The results of this thesis demonstrated that MSM can analyse the occurrence of more than one clinically important endpoint in addition to modelling misclassification of disease state, which is appealing to researchers where there can be diagnostic uncertainty. Reanalysis of the case study datasets showed that the treatment effects seemed to be strongest in patients who already developed a category 1 PU. One conclusion might have been to restrict future clinical trials to patients with a pre-existing category 1 PU in order to minimise the sample size required to detect a difference in the incidence of Category 2 PUs. However there are a number of limitations here. First, the proportion of patients presenting with a Category 1 PU is small and therefore the recruitment rate may be slower. Second, the trial results would be relevant only to those patients with a pre-existing Category 1 PU and are not relevant to the whole patient population at risk of developing a PU. Clinical trials should be designed to answer a clinical question and strategies to reduce the sample size (such as restricting the patient population to maximise the incidence of an event) should ensure that the trial results remain relevant to the clinical community.

The ability to accommodate misclassification using a HMM means that researchers have more flexibility in their choice of assessor, which could lead to cost savings and more convenient assessments to align with standard care. It is important to have an idea of the likely misclassification structure, which could be determined through an inter-rater reliability study using gold standard assessors, or through an elicitation exercise with clinical experts. If there is likely misclassification, in particular over-reporting, of the absorbing state, researchers should use the gold standard assessment where possible. If this is not possible then strategies to minimise the risk of over-reporting of the absorbing state such as endpoint adjudication should be used.

This thesis has demonstrated that non-ignorable missing data can be accommodated in the analysis, but researchers should continue to minimise levels of missing data where possible. Missing data is almost inevitable in any clinical trial though, particularly when outcomes are measured repeatedly through time. Where it occurs it is important to collect the reason and assess the likely association with the latent disease state to inform the appropriate analysis model.

9.2.4 Trial statisticians

The findings in this thesis have highlighted the importance of some of the trial design decisions from a statistical perspective. It demonstrated that deviating from common and widely understood methods of analysis can lead to improvements in power or reductions in sample sizes when using a method that uses more of the data collected and suits the underlying data structure. Regardless of the method of analysis, the length of follow-up, assessment frequency, mitigation of assessment bias and underlying assumptions of the model need to be carefully considered at the design stage to ensure adequate power.

For PU trials such as those described in the case studies, if an MSM was used, under some plausible scenarios length of follow-up could be reduced from 60 to 30 days and assessment can continue to be conducted every 2 or 3 days. However, if the baseline transition intensities were low on the early transitions, a Cox PH model comparing the time to development of a Category 2+ PU may be more appropriate. If the treatment effects were greater or similar for the early transitions relative to the final transition then an MSM should be used. Note that the case studies were for interventions where the outcome was a PU on any skin site, however for trials where the intervention is for a specific skin site, more work would be required to understand the natural history of the disease process before designing a clinical trial using an MSM. In the absence of this evidence it may be more appropriate to design a trial using more common methods such as a Cox PH or logistic regression with a pre-planned MSM as a key secondary analysis.

There are potential barriers to using MSM. If statisticians are interested in exploring the potential efficiencies of using a MSM for trial analysis, simulation studies are likely to be necessary. Sample size calculators able to accommodate the range of models required are are not widely accessible. However, not all statisticians and researchers are experienced in designing and conducting simulations in this context. The code used to conduct the simulations in Chapter 5 was made available on request through publication of the results [189]. The code used in Chapter 7 will also be made available in a similar way upon publication. In September 2021, Jackson developed the *simmulti.msm* function to simulate panel data from a MSM and HMM within the *msm* package in *R*. There are other examples of software available to simulate data from an MSM including the *simMSM* package [190] which was used by Le Rademacher *et al* to develop an RShiny package to illustrate the impact on the power of using an MSM with up to 5 states in a 2 arm trial with equal allocation assuming all participants start in State 1. This package is useful as an introduction to exploring the design of clinical trials using MSM, however the power is reported for individual transitions, with no adjustment for multiplicity.

Misclassification of outcomes can lead to biased estimates of model parameters if not properly accounted for in the design or analysis. Misclassification of PU categories is well documented in the literature, especially for early skin changes. A number of methods for mitigating this have been suggested, such as selecting an outcome that is at low risk of misclassification, using adjudication of the outcomes or incorporating the uncertainty in the model. The analysis in this thesis demonstrates that, if misclassification is possible, HMM will give unbiased parameter estimates with little loss of power. Ignoring misclassification was shown to lead to biased results and poor coverage. In the situation where the absorbing state is at risk of misclassification, adjudication of that state should be strongly recommended at the design stage to minimise the risk of biased treatment effect estimates, ensure adequate coverage and reduce the risk of identifiability issues.

If missing data are assumed to be MCAR then MSM can be fitted by ignoring the missing data. If the data are assumed to be MAR then the likelihood can be taken over all possible pathways of the disease process [177]. If the data are assumed to be MNAR, selection models are recommended where the MSM and observation process are jointly modelled. In Chapter 8 HMM were shown to be equivalent to selection models with lower computational burden. Any assumptions about the missing data mechanism should be assessed through sensitivity analyses. For PU prevention trials where multiple skin sites are assessed, it is important to first define missingness based on fully or partially complete assessments. In the context of PRESSURE2, the frequency of missing cases could be substantially reduced by restricting the definition to the 5 key skin sites where the majority (79.8%) of Category 2+ PUs occurred, without affecting the model parameter estimates. In general, composite outcomes should be defined based on both empirical analysis of the components and detailed discussions with subject specialists. In the analysis of PRESSURE2, there was little difference in results irrespective of the missing data models used. This provided confidence in the analysis results and should be considered in any setting where outcome data may be partially complete.

9.3 Limitations and areas for further research

The literature review examined key features of trial design and analysis but did not extract planned power. This may be useful to extract in further research to highlight the limitations of previous clinical trials of PU prevention interventions to the relevant clinical audience.

A thorough examination of continuous time MSM for the design of disease prevention clinical trials has been conducted, with a comparison of MSM to models of a single binary or TTE outcome. However, continuous time MSM were not the only option for analysis of this data type and other methods could be explored for this setting in further research. For example, a discrete time MSM may be appropriate for trials given the regular and common assessment times, or a generalised linear mixed model of a binary or ordinal outcome could also be fitted. Note that the analysis methods were informed by a review of pressure ulcer trials but further research could examine methods used in other clinical areas with similar issues in discrete longitudinal data, such as low event rates, multiple outcome states and misclassification. The MSM assessed in this thesis was based on a 4-state progression only model, which may not be appropriate for other disease settings. In cases where the disease state can improve (regression) the models become more complicated and there are consequently additional challenges to consider. First, for clinical trials consideration must be given a priori to whether the treatment affects both progression and regression and initial guesses must be made for the sample size calculation. Note that, in some contexts, it may be important to use a model that allows backward transitions to account appropriately for the natural history of the disease even if the effect of treatment on backwards transitions is not of interest. The simulation study results in Chapter 5 showed that the efficiency of MSM compared to a Cox model were sensitive to the baseline transition intensity rates and treatment effects. Furthermore, if an outcome at risk of misclassification is used, there is an increased risk of non-identifiability because the observed progression and regression cannot easily be distinguished from misclassification.

Additional work is required to further understand the place of HMM for analysis of misclassified outcome data including exploring a range of starting values for misclassification probabilities to assess the risk of non-identifiability and convergence issues. This should be designed with a clinical setting in mind. Furthermore, it is important to investigate the impact of differential misclassification. There is evidence in the literature to suggest that differential measurement error for time to event outcomes can lead to biased treatment effects which may lead to treatments being incorrectly concluded as effective.

Lindsey [191] explained that in longitudinal data, if there are variables that change over time, treatment groups can remain comparable but the time-varying covariates are no longer randomised. This means that patient outcomes are conditional on their history and the analysis should take this into account. The Markov assumption made throughout this thesis is such that the process depends only on the current state, and not on the history up until that point. The validity of this assumption should be assessed for each individual trial and clinical area. Further simulation studies could also be conducted to quantify the impact of deviations from the Markov assumption, for example trials with a longer follow-up period, or trials where important covariates have not been collected. Lindsey [191] advises that caution should also be taken when making causal conclusions from the results of a MSM analysis that could influence treatment recommendations. For example, in the PRESSURE2 trial, the largest treatment effect was observed on the final transition, and so a recommendation might have been to implement the intervention only in participants who are in state 3. However, the treatment effect estimates are based on an ITT analysis, and did not account for treatment compliance or other time varying covariates. As such, the individual treatment effect estimates should not necessarily be considered causal, but rather an estimate of the effect of an intervention policy. To extract treatment effects for a future trial in the population of patients who develop Category 1 PU, a model that includes time-dependent confounding variables could be fitted to the existing data. The adjusted treatment effect estimate might then be considered closer to the causal effect in the future trial. In the PU example, participants may start in any of the transient states and treatment is allocated on a 1:1 basis within each starting state. Therefore, exploratory analyses could be conducted in subgroups of participants who start in each of the transient states to estimate the effect of starting treatment in each state.

Chapter 8 examined the use of a HMM to jointly model the Markov disease process and the missing data mechanism. Although treatment effect estimates and analysis conclusions were consistent in the PU example, a simulation study is required to assess the impact on sample size and power with other missing data mechanisms and quantities of missing data.

The overall state and missing data mechanism definitions were informed by the distribution of missing data among individual components. This was appropriate for the motivating example, where variability in PU development was explained more so at the participant level rather than the component level. However, there are other clinical settings, for example in psoriatic arthritis, where change in disease state at the component level is relevant. There is no published work on the design of clinical trials when the data have this complex structure. The decision for the analysis approach depends on the discussions for the appropriate estimand. However, it is difficult to conceive an appropriate estimand for a trial based on such complex data and it may be considered more appropriate to evaluate an intervention at a patient level, with exploratory analyses conducted at the component level.

Throughout the simulations, it was assumed in the design that the treatment effects on each transition were independent of each other, however further research could examine the correlation between them and consequential impact on power. This could involve simulating event times from a multivariate distribution, with careful consideration of the off diagonals of the covariance matrix. For exactly observed observation times, Wu and Cook [146] proposed formulae for trial sample size calculations, however a greater understanding of how the correlation between treatment effects could be incorporated might be useful. This could be extended further to inform the design of earlier phase trials of PU prevention strategies, such as phase II trials to increase the chance of taking forward the most beneficial treatments to Phase III. Early phase trials commonly use surrogate endpoints for efficacy for which operational criteria were proposed by Prentice [192]. A candidate surrogate endpoint in a setting where an MSM might be used could be an earlier transient state. For example, in the PU setting, altered skin or a Category 1 PU might be sensible candidates for a surrogate endpoint for Category 2 PUs. However, Chapter

6 reported that misclassification was higher for early skin changes, which would lead to challenges in determining appropriate endpoints for earlier phase trials in the PU setting. Hidden Markov models can be used to account for misclassification of outcomes, however there were issues in model performance when the final state was at risk of misclassification, which would need to be explored further before recommendations could be made for a Phase II endpoint.

Whilst the code for each simulation study will be made available on request following publication, further work could include generalising the code to k states. This code could then be placed on a platform such as *GitHub*, or implemented in an RShiny application to enable use by those who are less familiar with the *R* programming language.

9.4 Conclusion

This thesis was motivated by PU prevention trials that are challenging to conduct due to large sample size requirements arising from low PU incidence. I have demonstrated how MSM can be used in the PU setting, including consideration of misclassification of outcomes by less expert assessors, and missing outcome data. A hypothesis test based on multiple effect estimates in the MSM setting was proposed and used in a comprehensive simulation study to explore the impact on power and sample size of using MSM as the primary analysis method, compared to methods based on a single endpoint. Scenarios were based on gold standard assessments for a range of assessment schedules, baseline transition intensities, and treatment effects. A further simulation study explored the impact on power and bias of misclassified assessments. New candidate definitions of PU state in the presence of missing data were proposed, and selection models were shown to be easily implemented under the HMM framework. Overall, this thesis has demonstrated how sophisticated methods can be used to improve the efficiency of disease prevention trials where participants pass through a series of discrete health states. Further work is required to develop robust modelling strategies for misclassified data, and to further explore the impact of missing data on power and sample size requirements.

Bibliography

[1] Sarah Brown, Isabelle L Smith, Julia M Brown, Claire Hulme, Elizabeth McGinnis, Nikki Stubbs, E Andrea Nelson, Delia Muir, Claudia Rutherford, Kay Walker, et al. Pressure relieving support surfaces: a randomised evaluation 2 (pressure 2): study protocol for a randomised controlled trial. *Trials*, 17(1):1–12, 2016.

[2] National Institute for Health Research. Best research for best health: The next chapter, 2021.

[3] Shaun Treweek, Doug G Altman, Peter Bower, Marion Campbell, Iain Chalmers, Seonaidh Cotton, Peter Craig, David Crosby, Peter Davidson, Declan Devane, et al. Making randomised trials more efficient: report of the first meeting to discuss the trial forge platform. *Trials*, 16(1):1–10, 2015.

[4] Patricia Healy, Sandra Galvin, Paula R Williamson, Shaun Treweek, Caroline Whiting, Beccy Maeso, Christopher Bray, Peter Brocklehurst, Mary Clarke Moloney, Abdel Douiri, et al. Identifying trial recruitment uncertainties using a james lind alliance priority setting partnership–the priority (prioritising recruitment in randomised trials) study. *Trials*, 19(1):1–12, 2018.

[5] Stephen J Walters, Inês Bonacho dos Anjos Henriques-Cadby, Oscar Bortolami, Laura Flight, Daniel Hind, Richard M Jacques, Christopher Knox, Ben Nadin, Joanne Rothwell, Michael Surtees, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the united kingdom health technology assessment programme. *BMJ open*, 7(3):e015276, 2017.

[6] Amanda Jane Blatch-Jones, Wei Pek, Emma Kirkpatrick, and Martin Ashton-Key. Role of feasibility and pilot studies in randomised controlled trials: a cross-sectional study. *BMJ open*, 8(9):e022233, 2018.

[7] Esther Herbert, Steven A Julious, and Steve Goodacre. Progression criteria in trials with an internal pilot: an audit of publicly funded randomised controlled trials. *Trials*, 20(1):1–9, 2019.

[8] Kerry NL Avery, Paula R Williamson, Carrol Gamble, Elaine O'Connell Francischetto, Chris Metcalfe, Peter Davidson, Hywel Williams, and Jane M Blazeby. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ open*, 7(2):e013537, 2017.

[9] Jenny L Donovan, Leila Rooshenas, Marcus Jepson, Daisy Elliott, Julia Wade, Kerry Avery, Nicola Mills, Caroline Wilson, Sangeetha Paramasivan, and Jane M Blazeby. Optimising recruitment and informed consent in randomised controlled trials: the development and implementation of the quintet recruitment intervention (qri). *Trials*, 17(1):1–11, 2016.

[10] Kimberly A Mc Cord, Rustam Al-Shahi Salman, Shaun Treweek, Heidi Gardner, Daniel Strech, William Whiteley, John PA Ioannidis, and Lars G Hemkens. Routinely collected data for randomized trials: promises, barriers, and implications. *Trials*, 19(1):1–9, 2018.

[11] Paula R Williamson, Douglas G Altman, Heather Bagley, Karen L Barnes, Jane M Blazeby, Sara T Brookes, Mike Clarke, Elizabeth Gargon, Sarah Gorst, Nicola Harman, et al. The comet handbook: version 1.0. *Trials*, 18(3):1–50, 2017.

[12] Matthew L Costa, Xavier L Griffin, Juul Achten, David Metcalfe, Andrew Judge, Rafael Pinedo-Villanueva, and Nicholas Parsons. World hip trauma evaluation (white): framework for embedded comprehensive cohort studies. *BMJ open*, 6(10):e011679, 2016.

[13] KL Haywood, XL Griffin, J Achten, and ML Costa. Developing a core outcome set for hip fracture trials. *The bone & joint journal*, 96(8):1016–1023, 2014.

[14] Linda Kwakkenbos, Edmund Juszczak, Lars G Hemkens, Margaret Sampson, Ole Fröbert, Clare Relton, Chris Gale, Merrick Zwarenstein, Sinéad M Langan, David Moher, et al. Protocol for the development of a consort extension for rcts using cohorts and routinely collected health data. *Research integrity and peer review*, 3(1):1–9, 2018.

[15] Clare Relton, David Torgerson, Alicia O'Cathain, and Jon Nicholl. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. *Bmj*, 340, 2010.

[16] Philip Pallmann, Alun W Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V Hampson, Jane Holmes, Adrian P Mander, Lang'o Odondi, Matthew R Sydes, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):1–15, 2018.

[17] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[18] Jay JH Park, Ellie Siden, Michael J Zoratti, Louis Dron, Ofir Harari, Joel Singer, Richard T Lester, Kristian Thorlund, and Edward J Mills. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*, 20(1):1–10, 2019.

[19] Erin L Ashbeck and Melanie L Bell. Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data. *BMC medical research methodology*, 16(1):1–8, 2016.

[20] Mary L Miller, Denise J Roe, Chengcheng Hu, and Melanie L Bell. Power difference in a χ 2 test vs generalized linear mixed model in the presence of missing data–a simulation study. *BMC medical research methodology*, 20(1):1–12, 2020.

[21] Jane Nixon, Gillian Cranny, Cynthia Iglesias, E Andrea Nelson, Kim Hawkins, Angela Phillips, David Torgerson, Su Mason, and Nicky Cullum. Randomised, controlled trial of alternating pressure mattresses compared with alternating pressure overlays for the prevention of pressure ulcers: Pressure (pressure relieving support surfaces) trial. *Bmj*, 332(7555):1413, 2006.

[22] Jane Nixon, Isabelle L Smith, Sarah Brown, Elizabeth McGinnis, Armando Vargas-Palacios, E Andrea Nelson, Susanne Coleman, Howard Collier, Catherine Fernandez, Rachael Gilberts, et al. Pressure relieving support surfaces for pressure ulcer prevention (pressure 2): clinical and health economic results of a randomised controlled trial. *EClinicalMedicine*, 14:42–52, 2019.

[23] J Nixon, D McElvenny, S Mason, J Brown, and S Bond. A sequential randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores. *International Journal of Nursing Studies*, 35(4):193–203, 1998.

[24] John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.

[25] GOV.UK. Essence of care, 2010.

[26] Nicky Cullum, Hannah Louise Buckley, Joanne Dumville, Jill Hall, Karen Lamb, Mary Teresa Madden, Richard Morley, Susan Mary O'Meara, Pedro Rafael Saramago Goncalves, Marta O Soares, et al. Wounds research for patient benefit: a 5 year programme of research. *Health technology assessment*, pages 1–303, 2016.

[27] NHS. Safety thermometer.

[28] National Patient Safety Agency. Seven steps to patient safety: The full reference guide, 2004.

[29] Isabelle L Smith, Jane Nixon, Sarah Brown, Lyn Wilson, and Susanne Coleman. Pressure ulcer and wounds reporting in nhs hospitals in england part 1: audit of monitoring systems. *Journal of tissue viability*, 25(1):3–15, 2016.

[30] Susanne Coleman, Isabelle L Smith, Jane Nixon, Lyn Wilson, and Sarah Brown. Pressure ulcer and wounds reporting in nhs hospitals in england part 2: survey of monitoring systems. *Journal of tissue viability*, 25(1):16–25, 2016.

[31] Jane Nixon, Helen Thorpe, Helen Barrow, Angela Phillips, E Andrea Nelson, Susan A Mason, and Nicky Cullum. Reliability of pressure ulcer classification and diagnosis. *Journal of advanced nursing*, 50(6):613–623, 2005.

[32] EPUAP, NPUAP, and PPPIA. National pressure ulcer advisory panel (npuap), european pressure ulcer advisory panel (epuap) and pan pacific pressure injury alliance (pppia). *Prevention and Treatment of Pressure Ulcers: Quick Reference Guide*, pages 1–72, 2014.

[33] E Kaltenthaler, MD Withfield, SJ Walters, RL Akehurst, and S Paisley. Uk, usa and canada: how do their pressure ulcer prevalence and incidence data compare? *Journal of wound care*, 10(1):530–535, 2001.

[34] Michelle Briggs, Michelle Collinson, Lyn Wilson, Carly Rivers, Elizabeth McGinnis, Carol Dealey, Julia Brown, Susanne Coleman, Nikki Stubbs, Rebecca Stevenson, et al. The prevalence of pain at pressure areas and pressure ulcers in hospitalised patients. *BMC nursing*, 12(1):19, 2013.

[35] Kathryn R Vowden and Peter Vowden. The prevalence, management, equipment provision and outcome for patients with pressure ulceration identified in a wound care survey within one english health care district. *Journal of Tissue Viability*, 18(1):20–26, 2009.

[36] Rebecca Stevenson, Michelle Collinson, Val Henderson, Lyn Wilson, Carol Dealey, Elizabeth McGinnis, Michelle Briggs, E Andrea Nelson, Nikki Stubbs, Susanne Coleman, et al. The prevalence of pressure ulcers in community settings: an observational study. *International journal of nursing studies*, 50(11):1550–1557, 2013.

[37] C Dealey, J Posnett, and A Walker. The cost of pressure ulcers in the united kingdom. *Journal of wound care*, 21(6):261–266, 2012.

[38] Claudia Gorecki, Julia M Brown, E Andrea Nelson, Michelle Briggs, Lisette Schoonhoven, Carol Dealey, Tom Defloor, Jane Nixon, and European Quality of Life Pressure Ulcer Project group. Impact of pressure ulcers on quality of life in older patients: a systematic review. *Journal of the American Geriatrics Society*, 57(7):1175–1183, 2009.

[39] E McInnes, A Jammali-Blasi, SE Bell-Syer, JC Dumville, V Middleton, and N Cullum. Support surfaces for pressure ulcer prevention. *Cochrane Database of systematic Reviews*, 2015(9), 2015.

[40] J Nixon, E Nelson, C Rutherford, S Coleman, D Muir, J Keen, J McCabe, C Dealey, M Briggs, S Brown, et al. Pressure ulcer programme of research: using mixed methods (systematic reviews, prospective cohort, case-study, consensus and psychometrics) to identify patient and organisational risk and develop a risk assessment tool and patient reported outcome quality of life and health utility measures. *Programme Grants for Applied Research: NIHR Journals Library*, 2015.

[41] Amit Gefen. How much time does it take to get a pressure ulcer? integrated evidence from human, animal, and in vitro studies. *Ostomy Wound Manage*, 54(10):26–28, 2008.

[42] Susanne Coleman, Claudia Gorecki, E Andrea Nelson, S Jose Closs, Tom Defloor, Ruud Halfens, Amanda Farrin, Julia Brown, Lisette Schoonhoven, and Jane Nixon. Patient risk factors for pressure ulcer development: systematic review. *International journal of nursing studies*, 50(7):974–1003, 2013.

[43] Richard L Reed, Kenneth Hepburn, Richard Adelson, Bruce Center, and Patrick McKnight. Low serum albumin levels, confusion, and fecal incontinence: are these risk factors for pressure ulcers in mobility-impaired hospitalized adults? *Gerontology*, 49(4):255–259, 2003.

[44] Sheryl L Ramer. Site-ation pearl growing: methods and librarianship history and theory. *Journal of the Medical Library Association*, 93(3):397, 2005.

[45] Trisha Greenhalgh and Richard Peacock. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *Bmj*, 331(7524):1064–1065, 2005.

[46] ZEH Moore and S Cowman. Risk assessment tools for the prevention of pressure ulcers. *Cochrane Database of Systematic Reviews*, 2014(2), 2014.

[47] E McInnes, A Jammali-Blasi, SEM Bell-Syer, JC Dumville, V Middleton, and N Cullum. Support surfaces for pressure ulcer prevention. *Cochrane Database of Systematic Reviews*, 2015(9), 2015.

[48] ZEH Moore and J Webster. Dressings and topical agents for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2013(8), 2013.

[49] BM Gillespie, RM Walker, SL Latimer, L Thalib, JA Whitty, E McInnes, and WP Chaboyer. Repositioning for pressure injury prevention in adults. *Cochrane Database of Systematic Reviews*, 2020(6), 2020.

[50] Q Zhang, Z Sun, and J Yue. Massage therapy for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2015(6), 2015.

[51] Gero Langer and Astrid Fink. Nutritional interventions for preventing and treating pressure ulcers. *Cochrane Database of Systematic Reviews*, 2014(6), 2014.

[52] ZEH Moore, J Webster, and R Samuriwo. Wound-care teams for preventing and treating pressure ulcers. *Cochrane Database of Systematic Reviews*, 2015(9), 2015.

[53] ZEH Moore and J Webster. Dressings and topical agents for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2018(12), 2018.

[54] AP Porter-Armstrong, ZEH Moore, I Bradbury, and S McDonough. Education of healthcare professionals for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2018(5), 2018.

[55] T O'Connor, ZEH Moore, and D Patton. Patient and lay carer education for preventing pressure ulceration in at-risk populations. *Cochrane Database of Systematic Reviews*, 2021(2), 2021.

[56] LS Kao, D Meeks, VA Moyer, and KP Lally. Peri-operative glycaemic control regimens for preventing surgical site infections in adults. *Cochrane Database of Systematic Reviews*, 2009(3), 2009.

[57] ZEH Moore and D Patton. Risk assessment tools for the prevention of pressure ulcers. *Cochrane Database of Systematic Reviews*, 2019(1), 2019.

[58] C Shi, JC Dumville, N Cullum, S Rhodes, A Jammali-Blasi, and E McInnes. Alternating pressure (active) air surfaces for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2021(5), 2021.

[59] Chunhu Shi, Jo C Dumville, Nicky Cullum, Sarah Rhodes, Vannessa Leung, and Elizabeth McInnes. Reactive air surfaces for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2021(4), 2021.

[60] C Shi, JC Dumville, N Cullum, S Rhodes, and E McInnes. Foam surfaces for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2021(5), 2021.

[61] C Shi, JC Dumville, N Cullum, S Rhodes, and E McInnes. Alternative reactive support surfaces (non-foam and non-air-filled) for preventing pressure ulcers. *Cochrane Database of Systematic Reviews*, 2021(5), 2021.

[62] P Joyce, ZEH Moore, and J Christie. Organisation of health services for preventing and treating pressure ulcers. *Cochrane Database of Systematic Reviews*, 2018(12), 2018.

[63] Zena EH Moore, Joan Webster, and Ray Samuriwo. Wound-care teams for preventing and treating pressure ulcers. *Cochrane database of systematic reviews*, 2015(9), 2015.

[64] Rachel Walker, Leisa Huxley, Melanie Juttner, Elizabeth Burmeister, Justin Scott, and Leanne M Aitken. A pilot randomized controlled trial using prophylactic dressings to minimize sacral pressure injuries in high-risk hospitalized patients. *Clinical nursing research*, 26(4):484–503, 2017.

[65] Soundappan Kathirvel, Sukhpal Kaur, Mandeep Singh Dhillon, and Amarjeet Singh. Impact of structured educational interventions on the prevention of pressure ulcers in immobile orthopedic patients in india: A pragmatic randomized controlled trial. *Journal of Family Medicine and Primary Care*, 10(3):1267, 2021.

[66] Patriek Mistiaen, Wilco Achterberg, Andre Ament, Ruud Halfens, Janneke Huizinga, Ken Montgomery, Henri Post, Peter Spreeuwenberg, and Anneke L Francke. The effectiveness of the australian medical sheepskin for the prevention of pressure ulcers in somatic nursing home patients: a prospective multicenter randomized-controlled trial (isrctn17553857). *Wound Repair and Regeneration*, 18(6):572–579, 2010.

[67] Katrien Vanderwee, MHF Grypdonck, Dirk De Bacquer, and Tom Defloor. Effectiveness of turning with unequal time intervals on the incidence of pressure ulcer lesions. *Journal of advanced nursing*, 57(1):59–68, 2007.

[68] Tom Defloor and Maria FH Grypdonck. Pressure ulcers: validation of two risk assessment scales. *Journal of clinical nursing*, 14(3):373–382, 2005.

[69] Katrien Vanderwee, Maria Grypdonck, Dirk De Bacquer, and Tom Defloor. The identification of older nursing home residents vulnerable for deterioration of grade 1 pressure ulcers. *Journal of clinical nursing*, 18(21):3050–3058, 2009.

[70] Nancy Bergstrom et al. The braden scale for predicting pressure sore risk. *Nurs res*, 36(4):205–210, 1987.

[71] Doreen Norton, Arthur Norman Exton-Smith, and Rhoda McLaren. *An investigation of geriatric nursing problems in hospital.* National Corporation for the care of old people, 1962.

[72] JA Waterlow. A risk assessment card. *Nursing times*, 81:51–55, 1985.

[73] MR Bliss. Preventing pressure sores in elderly patients: a comparison of seven mattress overlays. *Age and ageing*, 24(4):297–302, 1995.

[74] David Brienza, Sheryl Kelsey, Patricia Karg, Ana Allegretti, Marian Olson, Mark Schmeler, Jeanne Zanca, Mary Jo Geyer, Marybeth Kusturiss, and Margo Holm. A randomized clinical trial on preventing pressure ulcers with wheelchair seat cushions. *Journal of the American Geriatrics Society*, 58(12):2308–2314, 2010.

[75] Richard G Bennett, Patricia J Baran, LaVeda DeVone, Hector Bacetti, Blaine Kristo, Matthew Tayback, and William B Greenough III. Low airloss hydrotherapy versus standard care for incontinent hospitalized patients. *Journal of the American Geriatrics Society*, 46(5):569–576, 1998.

[76] ME Collier. Pressure-reducing mattresses. *Journal of Wound Care*, 5(5):207–211, 1996.

[77] Victoria M Hoshowsky and Carol A Schramm. Intraoperative pressure sore prevention: an analysis of bedding materials. *Research in nursing & health*, 17(5):333–339, 1994.

[78] Marilyn J Rantz, Mary Zwygart-Stauffacher, Lanis Hicks, David Mehr, Marcia Flesner, Gregory F Petroski, Richard W Madsen, and Jill Scott-Cawiezell. Randomized multilevel intervention to improve outcomes of residents in nursing homes in need of improvement. *Journal of the American Medical Directors Association*, 13(1):60–68, 2012.

[79] Dimitri Beeckman, Lisette Schoonhoven, Jacqui Fletcher, Katia Furtado, Hilde Heyman, Louis Paquay, Dirk De Bacquer, and Tom Defloor. Pressure ulcers and incontinence-associated dermatitis: effectiveness of the pressure ulcer classification education tool on classification by nurses. *BMJ Quality & Safety*, 19(5):e3–e3, 2010.

[80] Colin Torrance. *Pressure sores: Aetiology, treatment, and prevention.* Taylor & Francis, 1983.

[81] AN Exton-Smith and RW Sherwin. The prevention of pressure sores significance of spontaneous bodily movements. *The Lancet*, 278(7212):1124–1126, 1961.

[82] J DARRELL Shea. Pressure sores: classification and management. *Clinical orthopaedics and related research*, 1975(112):89–100, 1975.

[83] United States. Pressure Ulcer Guideline Panel and Nancy Bergstrom. *Pressure ulcers in adults: prediction and prevention.* US Department of Health and Human Services, Public Health Service, Agency . . . , 1992.

[84] Marylou Guihan, Charles H Bombardier, Dawn M Ehde, Lauren M Rapacki, Thea J Rogers, Barbara Bates-Jensen, Florian P Thomas, Rama Parachuri, and Sally A Holmes. Comparing multicomponent interventions to improve skin care behaviors and prevent recurrence in veterans hospitalized for severe pressure ulcers. *Archives of physical medicine and rehabilitation*, 95(7):1246–1253, 2014.

[85] AN Exton-Smith, J Wedgwood, PW Overstall, and Gillian Wallace. Use of the'air wave system'to prevent pressure sores in hospital. *The Lancet*, 319(8284):1288–1290, 1982.

[86] Joan Webster, Kerrie Coleman, Alison Mudge, Louise Marquart, Glenn Gardner, Monica Stankiewicz, Julie Kirby, Catherine Vellacott, Margaret Horton-Breshears, and Alice McClymont. Pressure ulcers: effectiveness of risk-assessment tools. a randomised controlled trial (the ulcer trial). *BMJ quality & safety*, 20(4):297–306, 2011.

[87] Liesbet Demarré, Dimitri Beeckman, Katrien Vanderwee, Tom Defloor, Maria Grypdonck, and Sofie Verhaeghe. Multi-stage versus single-stage inflation and deflation cycle for alternating low pressure air mattresses to prevent pressure

ulcers in hospitalised patients: a randomised-controlled clinical trial. *International journal of nursing studies*, 49(4):416–426, 2012.

[88] Katrien Vanderwee, Maria HF Grypdonck, and Tom Defloor. Effectiveness of an alternating pressure air mattress for the prevention of pressure ulcers. *Age and ageing*, 34(3):261–267, 2005.

[89] RH Houwing, M Rozendaal, W Wouters-Wesseling, JWJ Beulens, E Buskens, and JR Haalboom. A randomised, double-blind assessment of the effect of nutritional supplementation on the prevention of pressure ulcers in hip-fracture patients. *Clinical Nutrition*, 22(4):401–405, 2003.

[90] Nancy Bergstrom, Susan D Horn, Mary Pat Rapp, Anita Stern, Ryan Barrett, and Michael Watkiss. Turning for ulcer reduction: a multisite randomized clinical trial in nursing homes. *Journal of the American Geriatrics Society*, 61(10):1705–1713, 2013.

[91] Martin Van Leen, Steven Hovius, Ruud Halfens, Jacques Neyens, and Jos Schols. Pressure relief with visco-elastic foam or with combined static air overlay? a prospective, crossover randomized clinical trial in a dutch nursing home. *Wounds*, 25(10):287–292, 2013.

[92] Martin van Leen, Ruud Halfens, and Jos Schols. Preventive effect of a microclimate-regulating system on pressure ulcer development: a prospective, randomized controlled trial in dutch nursing homes. *Advances in skin & wound care*, 31(1):1–5, 2018.

[93] Ronald Houwing, Wil Van der Zwet, Sweder van Asbeck, Ruud Halfens, and Willem Arends. an unexpected detrimental effect on the incidence of heel pressure ulcers after local 5% dmso cream application: A randomized, double-blind study in patients at risk for pressure ulcers. *Wounds: a compendium of clinical research and practice*, 20(4):84–88, 2008.

[94] S McGowan, K Montgomery, D Jolley, and R Wright. The role of sheepskins in preventing pressure ulcers in elderly orthopaedic patients. *Primary Intention*, 2000.

[95] Miles D Witham, Eleanor Anderson, Camille Carroll, Paul M Dark, Kim Down, Alistair S Hall, Joanna Knee, Rebecca H Maier, Gail A Mountain, Gary Nestor, et al. Developing a roadmap to improve trial delivery for under-served groups: results from a uk multi-stakeholder process. *Trials*, 21(1):1–9, 2020.

[96] Shaun Treweek, Katie Banister, Peter Bower, Seonaidh Cotton, Declan Devane, Heidi R Gardner, Talia Isaacs, Gary Nestor, Adepeju Oshisanya, Adwoa Parker, et al. Developing the include ethnicity framework—a tool to help trialists design trials that better reflect the communities they serve. *Trials*, 22(1):1–12, 2021.

[97] D Collett. Modelling binary data. 2nd (edn), 2002.

[98] TALI A Conine, CECIL Hershler, DAWN Daechsel, CAROL Peel, and ALISON Pearson. Pressure ulcer prophylaxis in elderly patients using polyurethane foam or jay wheelchair cushions. *International journal of rehabilitation research. Internationale Zeitschrift fur Rehabilitationsforschung. Revue internationale de recherches de readaptation*, 17(2):123–137, 1994.

[99] Nick Santamaria, Marie Gerdtz, Sarah Sage, Jane McCann, Amy Freeman, Theresa Vassiliou, Stephanie De Vincentis, Ai Wei Ng, Elizabeth Manias, Wei Liu, et al. A randomised controlled trial of the effectiveness of soft silicone multi-layered foam dressings in the prevention of sacral and heel pressure ulcers in trauma and critically ill patients: the border trial. *International Wound Journal*, 12(3):302–308, 2015.

[100] Gojiro Nakagami, Hiromi Sanada, Chizuko Konya, Atsuko Kitagawa, Etsuko Tadaka, and Yutaka Matsuyama. Evaluation of a new pressure ulcer preventive

dressing containing ceramide 2 with low frictional outer layer. *Journal of advanced nursing*, 59(5):520–529, 2007.

[101] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[102] C. H. Jackson. Multi-state models for panel data: The msm package for r. *Journal of Statistical Software*, 38(8):1–28, 2011.

[103] Daniel L Young, Nancy Estocado, Merrill R Landers, and Joyce Black. A pilot study providing evidence for the validity of a new tool to improve assignment of national pressure ulcer advisory panel stage to pressure ulcers. *Advances in skin & wound care*, 24(4):168–175, 2011.

[104] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC, 2006.

[105] Isabelle L Smith, Sarah Brown, Elizabeth McGinnis, Michelle Briggs, Susanne Coleman, Carol Dealey, Delia Muir, E Andrea Nelson, Rebecca Stevenson, Nikki Stubbs, et al. Exploring the role of pain as an early predictor of category 2 pressure ulcers: a prospective cohort study. *BMJ open*, 7(1):e013623, 2017.

[106] Jane Nixon, Sarah Brown, Isabelle L Smith, Elizabeth McGinnis, Armando Vargas-Palacios, E Andrea Nelson, Julia Brown, Susanne Coleman, Howard Collier, Catherine Fernandez, et al. Comparing alternating pressure mattresses and high-specification foam mattresses to prevent pressure ulcers in high-risk patients: the pressure 2 rct. *NIHR Journals Library*, 2019.

[107] David Collett. *Modelling survival data in medical research.* CRC press, 2015.

[108] Tom AB Snijders and Roel J Bosker. *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* sage, 2011.

[109] Theodor A Balan and Hein Putter. A tutorial on frailty models. *Statistical methods in medical research*, 29(11):3424–3454, 2020.

[110] Patrick Royston and Mahesh KB Parmar. A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. *Trials*, 21(1):1–17, 2020.

[111] Aidan G. O'Keeffe, Brian D. M. Tom, and Vernon T. Farewell. Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Statistics in medicine*, 32(4):600–619, 2013.

[112] Aidan G O'Keeffe, Brian DM Tom, and Vernon T Farewell. A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):675–699, 2011.

[113] U.S. Department of Health and Human Services. Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics, 2018.

[114] Howard HZ Thom, Christopher H Jackson, Daniel Commenges, and Linda D Sharples. State selection in markov models for panel data with application to psoriatic arthritis. *Statistics in medicine*, 34(16):2456–2475, 2015.

[115] Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

[116] K. E. Marqueen, N. Waingankar, J. Sfakianos, R. Mehrazin, S. A. Niglio, F. Audenet, R. Jia, M. Mazumdar, B. Ferket, and M. D. Galsky. Early mortality in patients with muscle-invasive bladder cancer undergoing cystectomy in the united states. *NCI Cancer Spectrum*, 2(4), 2018.

[117] D Heng, LD Sharples, K McNeil, S Stewart, T Wreghitt, and J Wallwork. Bronchiolitis obliterans syndrome: incidence, natural history, prognosis, and risk factors. *The Journal of heart and lung transplantation: the official publication of the International Society for Heart Transplantation*, 17(12):1255–1263, 1998.

[118] C. H. Jackson and L. D. Sharples. Hidden markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, 21(1):113–128, 2002.

[119] Luís Meira-Machado, Jacobo de Uña-Álvarez, Carmen Cadarso-Suárez, and Per K Andersen. Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222, 2009.

[120] Tiny Jaarsma, Martje HL van der Wal, Ivonne Lesman-Leegte, Marie-Louise Luttik, Jochem Hogenhuis, Nic J Veeger, Robbert Sanderman, Arno W Hoes, Wiek H van Gilst, Dirk JA Lok, et al. Effect of moderate or intensive disease management program on outcome in patients with heart failure: Coordinating study evaluating outcomes of advising and counseling in heart failure (coach). *Archives of internal medicine*, 168(3):316–324, 2008.

[121] Douwe Postmus, Dirk J van Veldhuisen, Tiny Jaarsma, Marie Louise Luttik, Johan Lassus, Alexandre Mebazaa, Markku S Nieminen, Veli-Pekka Harjola, James Lewsey, Erik Buskens, et al. The coach risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. *European journal of heart failure*, 14(2):168–175, 2012.

[122] F. Ieva, C. H. Jackson, and L. D. Sharples. Multi-state modelling of repeated hospitalisation and death in patients with heart failure: The use of large administrative databases in clinical epidemiology. *Statistical Methods in Medical Research*, 26(3):1350–1372, 2017.

[123] Marko Kornmann, Ludger Staib, Thomas Wiegel, Martina Kron, Doris Henne-Bruns, Karl-Heinrich Link, Andrea Formentini, Study Group Oncology of Gastrointestinal Tumors (FOGT, et al. Long-term results of 2 adjuvant trials reveal differences in chemosensitivity and the pattern of metastases between colon cancer and rectal cancer. *Clinical colorectal cancer*, 12(1):54–61, 2013.

[124] G. Manzini, T. J. Ettrich, M. Kremer, M. Kornmann, D. Henne-Bruns, D. A. Eikema, P. Schlattmann, and L. C. de Wreede. Advantages of a multi-state

approach in surgical research: how intermediate events and risk factor profile affect the prognosis of a patient with locally advanced rectal cancer. *Bmc Medical Research Methodology*, 18, 2018.

[125] J. G. Le-Rademacher, R. A. Peterson, T. M. Therneau, B. L. Sanford, R. M. Stone, and S. J. Mandrekar. Application of multi-state models in cancer clinical trials. *Clin Trials*, 15(5):489–498, 2018.

[126] JD Kalbfleisch and Jerald Franklin Lawless. The analysis of panel data under a markov assumption. *Journal of the american statistical association*, 80(392):863–871, 1985.

[127] Leilei Zeng, Richard J Cook, Lan Wen, and Audrey Boruvka. Bias in progression-free survival analysis due to intermittent assessment of progression. *Statistics in Medicine*, 34(24):3181–3193, 2015.

[128] L. L. Zeng, R. J. Cook, and K. A. Lee. Design of cancer trials based on progression-free survival with intermittent assessment. *Statistics in Medicine*, 37(12):1947–1959, 2018.

[129] J. Gruger, R. Kay, and M. Schumacher. The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47(2):595–605, 1991.

[130] Ardo Van Den Hout. *Multi-state survival models for interval-censored data.* CRC Press, 2017.

[131] Claire Williams, James D Lewsey, Daniel F Mackay, and Andrew H Briggs. Estimation of survival probabilities for use in cost-effectiveness analyses: a comparison of a multi-state modeling survival analysis approach with partitioned survival and markov decision-analytic modeling. *Medical Decision Making*, 37(4):427–439, 2017.

[132] Nicky J Welton and AE Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propa-

gation, evidence synthesis, and model calibration. *Medical Decision Making*, 25(6):633–645, 2005.

[133] Thomas Kneib and Andrea Hennerfeind. Bayesian semi parametric multi-state models. *Statistical Modelling*, 8(2):169–198, 2008.

[134] Orestis Efthimiou, Nicky Welton, Myrto Samara, Stefan Leucht, Georgia Salanti, and GetReal Work Package 4. A markov model for longitudinal studies with incomplete dichotomous outcomes. *Pharmaceutical statistics*, 16(2):122–132, 2017.

[135] M. Lauseker, J. Hasford, V. S. Hoffmann, M. C. Muller, R. Hehlmann, M. Pfirrmann, and German CML Study Grp. A multi-state model approach for prediction in chronic myeloid leukaemia. *Annals of Hematology*, 94(6):919–927, 2015.

[136] C. Cassarly, R. H. Martin, M. Chimowitz, E. A. Pena, V. Ramakrishnan, and Y. Y. Palesch. Assessing type i error and power of multistate markov models for panel data-a simulation study. *Commun Stat Simul Comput*, 46(9):7040–7061, 2017.

[137] Jon Michael Gran, Stein Atle Lie, Irene Øyeflaten, Ørnulf Borgan, and Odd O Aalen. Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation. *BMC public health*, 15(1):1–16, 2015.

[138] M. J. Crowther and P. C. Lambert. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Statistics in Medicine*, 36(29):4719–4742, 2017.

[139] L. C. de Wreede, M. Fiocco, and H. Putter. mstate: An r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7):1–30, 2011.

[140] Richard J Cook and Jerald F Lawless. *Multistate models for the analysis of life history data*. Chapman and Hall/CRC, 2018.

[141] Brian DM Tom and Vernon T Farewell. Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in medicine*, 30(30):3520–3531, 2011.

[142] National Institute for Health Research. Technology appraisal guidance, 2019.

[143] Patricia Guyot, Nicky J Welton, Mario JNM Ouwens, and AE Ades. Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. *Value in Health*, 14(5):640–646, 2011.

[144] ICH E9 working group et al. Ich e9 (r1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, 2020.

[145] Nicholas R Latimer, Keith R Abrams, Paul C Lambert, James P Morden, and Michael J Crowther. Assessing methods for dealing with treatment switching in clinical trials: a follow-up simulation study. *Statistical methods in medical research*, 27(3):765–784, 2018.

[146] Longyang Wu and Richard J Cook. The design of intervention trials involving recurrent and terminal events. *Statistics in Biosciences*, 5(2):261–285, 2013.

[147] A. H. Feiveson. Power by simulation. *The Stata Journal*, 2(2):107–124, 2002.

[148] Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

[149] Michael A Proschan and Myron A Waclawiw. Practical guidelines for multiplicity adjustment in clinical trials. *Controlled clinical trials*, 21(6):527–539, 2000.

[150] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[151] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[152] Mohammad F Huque. Validity of the hochberg procedure revisited for clinical trial applications. *Statistics in medicine*, 35(1):5–20, 2016.

[153] Timo B Brakenhoff, Marian Mitroiu, Ruth H Keogh, Karel GM Moons, Rolf HH Groenwold, and Maarten van Smeden. Measurement error is often neglected in medical literature: a systematic review. *Journal of clinical epidemiology*, 98:89–97, 2018.

[154] Linda Nab, Rolf HH Groenwold, Paco MJ Welsing, and Maarten van Smeden. Measurement error in continuous endpoints in randomised trials: Problems and solutions. *Statistics in medicine*, 38(27):5182–5196, 2019.

[155] Douglas G Altman and J Martin Bland. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943):1552, 1994.

[156] Peter J Godolphin, Philip M Bath, Christopher Partlett, Eivind Berge, Martin M Brown, Misha Eliasziw, Per Morten Sandset, Joaquín Serena, and Alan A Montgomery. Outcome assessment by central adjudicators in randomised stroke trials: Simulation of differential and non-differential misclassification. *European Stroke Journal*, page 2396987320910047, 2020.

[157] Brennan C Kahan, Brian Feagan, and Vipul Jairath. A comparison of approaches for adjudicating outcomes in clinical trials. *Trials*, 18(1):1–14, 2017.

[158] EA Eisenhauer and J Verweij. 11 new response evaluation criteria in solid tumors: Recist guideline version 1.1. *Ejc Supplements*, 2(7):5, 2009.

[159] Asbjørn Hróbjartsson, Ann Sofia Skou Thomsen, Frida Emanuelsson, Britta Tendal, Jørgen Hilden, Isabelle Boutron, Philippe Ravaud, and Stig Brorson. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *Bmj*, 344, 2012.

[160] S Hong, N Schmitt, A Stone, and J Denne. Attenuation of treatment effect due to measurement variability in assessment of progression-free survival. *Pharmaceutical statistics*, 11(5):394–402, 2012.

[161] Edward L Korn, Lori E Dodd, and Boris Freidlin. Measurement error in the timing of events: effect on survival analyses in randomized clinical trials. *Clinical Trials*, 7(6):626–633, 2010.

[162] Richard J Cook and Jerald F Lawless. Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences*, 6(1):127–161, 2014.

[163] Ardo van den Hout and Fiona E Matthews. Multi-state analysis of cognitive ability data: a piecewise-constant model and a weibull model. *Statistics in Medicine*, 27(26):5440–5455, 2008.

[164] Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

[165] Ronald E Gangnon, Kristine E Lee, Barbara EK Klein, Sudha K Iyengar, Theru A Sivakumaran, and Ronald Klein. Misclassification can explain most apparent regression of age-related macular degeneration: results from multistate models with misclassification. *Investigative ophthalmology & visual science*, 55(3):1780–1786, 2014.

[166] Peter Bacchetti, Ross Boylan, Jacquie Astemborski, Hui Shen, Shruti H Mehta, David L Thomas, Norah A Terrault, and Alexander Monto. Progression of biopsy-measured liver fibrosis in untreated patients with hepatitis c infection: non-markov multistate model analysis. *PLoS One*, 6(5):e20104, 2011.

[167] Peter Bacchetti, Ross D Boylan, Norah A Terrault, Alexander Monto, and Marina Berenguer. Non-markov multistate modeling using time-varying co-

variates, with application to progression of liver fibrosis due to hepatitis c following liver transplant. *The international journal of biostatistics*, 6(1), 2010.

[168] Peter Bacchetti and Ross Boylan. Estimating complex multi-state misclassification rates for biopsy-measured liver fibrosis in patients with hepatitis c. *The international journal of biostatistics*, 5(1), 2009.

[169] Rikesh Bhatt, Ardo van den Hout, and Nora Pashayan. A multistate survival model of the natural history of cancer using data from screened and unscreened population. *Statistics in Medicine*, 40(16):3791–3807, 2021.

[170] Josip Car, Gerald Choon-Huat Koh, Pin Sym Foong, and C Jason Wang. Video consultations in primary and specialist care during the covid-19 pandemic and beyond. *BMJ*, 371, 2020.

[171] Grace Y Yi, Wenqing He, and Feng He. Analysis of progressive multi-state models with misclassified states: likelihood and pairwise likelihood methods. *Biostatistics & Epidemiology*, 1(1):119–132, 2017.

[172] Bruce G Lindsay, Grace Y Yi, and Jianping Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, pages 71–105, 2011.

[173] Minhee Kang and Stephen W Lagakos. Statistical methods for panel data from a semi-markov process, with application to hpv. *Biostatistics*, 8(2):252–264, 2007.

[174] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[175] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[176] Vernon T Farewell and Brian DM Tom. The versatility of multi-state models for the analysis of longitudinal data with unobservable features. *Lifetime data analysis*, 20(1):51–75, 2014.

[177] Vernon T Farewell, Li Su, and Christopher Jackson. Partially hidden multi-state modelling of a prolonged disease state defined by a composite outcome. *Lifetime data analysis*, 25(4):696–711, 2019.

[178] James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

[179] James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

[180] Bernard F Cole, Marco Bonetti, Alan M Zaslavsky, and Richard D Gelber. A multistate markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Statistics in Medicine*, 24(15):2317–2334, 2005.

[181] Ardo Van Den Hout and Fiona E Matthews. Estimating stroke-free and total life expectancy in the presence of non-ignorable missing values. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):331–349, 2010.

[182] Baojiang Chen, Grace Y Yi, and Richard J Cook. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*, 29(11):1175–1189, 2010.

[183] Baojiang Chen, Y Yi Grace, and Richard J Cook. Progressive multi-state models for informatively incomplete longitudinal data. *Journal of Statistical Planning and Inference*, 141(1):80–93, 2011.

[184] Jane M Lange, Rebecca A Hubbard, Lurdes YT Inoue, and Vladimir N Minin. A joint model for multistate disease processes and random informative obser-

vation times, with applications to electronic medical records data. *Biometrics*, 71(1):90–101, 2015.

[185] Alessandro Gasparini, Keith R Abrams, Jessica K Barrett, Rupert W Major, Michael J Sweeting, Nigel J Brunskill, and Michael J Crowther. Mixed-effects models for health care longitudinal data with an informative visiting process: A monte carlo simulation study. *Statistica Neerlandica*, 74(1):5–23, 2020.

[186] Hongbin Zhang, Elizabeth A Kelvin, Arturo Carpio, and W Allen Hauser. A multistate joint model for interval-censored event-history data subject to within-unit clustering and informative missingness, with application to neurocysticercosis research. *Statistics in Medicine*, 39(23):3195–3206, 2020.

[187] Wenjie Lou, Lijie Wan, Erin L Abner, David W Fardo, Hiroko H Dodge, and Richard J Kryscio. Multi-state models and missing covariate data: expectation–maximization algorithm for likelihood estimation. *Biostatistics & epidemiology*, 1(1):20–35, 2017.

[188] Antonio Eleuteri, Azzam FG Taktak, Sarah E Coupland, Heinrich Heimann, Helen Kalirai, and Bertil Damato. Prognostication of metastatic death in uveal melanoma patients: A markov multi-state model. *Computers in biology and medicine*, 102:151–156, 2018.

[189] Isabelle L Smith, Jane E Nixon, and Linda Sharples. Power and sample size for multistate model analysis of longitudinal discrete outcomes in disease prevention trials. *Statistics in medicine*, 38(27):5182–5196, 2020.

[190] H Reulen. Simmsm: Simulation of event histories for multi-state models. *R package version*, 1:41, 2015.

[191] James K Lindsey et al. Models for repeated measurements. *OUP Catalogue*, 1999.

[192] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

Appendix A

Chapter 2: Literature review

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Andersen et al. 1983)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a; Shi, Dumville, et al. 2021b)	10 days	600	Hospital	Patients with acute conditions at risk of PU development, PU free	Shoulders, spine, sacral region, buttocks, hips and heels	Every other day	Study authors	Non-decubitus (Normal skin, redness & infiltration, Extravasation) ; Decubitus (Bullae, black necrosis, skin defect)	Decubitus	Chi square test	
(Beeckman et al. 2019)	(Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	14 days	308	Nursing homes	65y or older, bed bound or chair bound, at high risk of PUs (presence of non-blanchable erythema or according to Braden), excluded if PU category II to IV.	Not specified, but sacral area, trochanters, scapula, elbows and occiput discussed)	Daily	Ward nurses	NPUAP/EPUAP/PPPIA	PU Category II-IV, Category IV	Chi-square test, Fisher's exact test (for comparison of Cat IV and comparison at the sacrum), KM curves, log-rank test	Before the study started, all ward nurses attended training on PU classification, with the use of the Pressure ulcer classification (PUCLAS4), a validated e-learning tool. Researchers performed independent and unannounced skin assessments to assess IRR.
(Bennett et al. 1998)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021)	60 days	116	Hospital (acute and chronic)	Incontinence of urine and/or faeces; bedbound for 16h or more per day. Excluded if pre-existing stage III or IV PU with planned debridement/astringent dressing	Truncal region	3 times a week in week 1, at least 2 times a week and additionally as necessary thereafter	Study nurse and/or research technicians	NPUAP	New skin lesion; blanchable erythema; new stage II-IV	Chi square test or Fishers exact test; KM curves	
(Bergstrom et al. 2013)	(Gillespie et al. 2020)	21 days	967	Nursing homes	Braden scale either moderate (13 to 14) or high (11 to 12); limited mobility (≤ 3 on Braden subscale of mobility), PU free	Coccyx or sacrum, trochanter, heel	Weekly research assessments and daily ward records	NH staff/ licensed nurses	NPUAP	Stage 1 PU	Fishers exact test	Stage 1 PUs were recorded if they had been observed on 2 consecutive days to exclude false positives
(Bliss, McLaren, and Exton-Smith 1967)	(Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	16 days	83	Hospital	Patients liable to develop pressure sores according to Norton.	Trunk (sacrum, buttocks and hips) and heels	Every 2 days	Unclear	Trunk: 0, +, ++, +++, ++++; Heels: +, ++, +++	All PU grades reported	Descriptive only	Interesting plot showing progress of pressure areas over time for each patient..

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Bliss 1995)	(Shi, Dumville, et al. 2021a; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Shi, Dumville, et al. 2021b)	3 weeks	'457' trials – see notes	Hospital	Patients with pre-existing Grade 2-3 sores, excluding those >5cm or patients with discoloured areas >2cm	Not specified, but sites reported were pelvis, heels, other foot, other	Twice weekly for 6 assessments	Data collection forms were 'scored' blind at the end of the study by an independent observer who was a tissue viability nurse specialist.	Not specified	PU deterioration from grade 2-3	Chi-square test or Fisher's exact test	Patients could be re-randomised if their condition deteriorated prompting removal of the intervention, but then improved. At the 6th assessment patients could be re-randomised again if still eligible. If they had healed, they were followed up to assess effectiveness in preventing relapse, but if they deteriorated they were re-randomised again
(Bloemen - Vrencken et al. 2007)	(Joyce, Moore, and Christie 2018)	12 months	149	Rehab centres	<65y, spinal cord injury	Not specified, but heels, ankles, hips, buttocks, coccyx and other reported	12 months post discharge	Patient-reported	Not specified	PU Grade I+	Fisher's exact test	
(Bourdel-Marchass on et al. 2000)	(Langer and Fink 2014; Coleman et al. 2013)	15 days	672	Hospital	65y or older, acutely ill, unable to move or eat independently. PU free	Not specified	Daily	Trained ward nurses	AHCPR	Grade I+	Cox regression,	Weekly visits to assess the quality of reported measurements including recognition and scoring of PUs
(Brienza et al. 2010)	(McInnes et al. 2015)	6 months	232	Nursing home	Elderly residents who used wheelchairs as their primary means of seating and mobility, at risk of PUs,	Ischial tuberosities, sacral region	Weekly;If facility staff identified an ischial PU the RN was contacted and skin status verified within 24h	Research nurse trained in detecting and staging PUs	NPUAP	Siting acquired PU stage 1+ on the ischial tuberosities; PU stage 1+ on the sacrum; PUs stage 1+ across both skin regions	Chi square test or Fishers exact test; KM curves; log rank test	
(Bueno de Camargo et al. 2018)	(Shi, Dumville, et al. 2021a)	Not specified	62	ICU	Critically ill, moderate to high risk of PU (Braden 14 or less), PU free	Bony prominences	Daily	Not specified	NPUAP	PU Stage 2+	Chi-square test or Fishers exact test, KM curves, log rank test	
(Caplan et al. 1999)	(Joyce, Moore, and Christie 2018)	Unclear; 28 days post discharge	100	Tertiary referral hospital	65y or older encouraged	Not specified	Daily visits on average	Medical record review	Not specified	Not Specified	Fishers exact test	PU a complication
(Cassino, Ippolito, and Ricci 2013)	(Shi, Dumville, et al. 2021b)	150 days (primary EP: 60 days, secondary EP: 90 days later)	20	Long-term nursing facility	75y or older, Norton<9, intact skin without alterations due to pressure, skin disease or other causes	Sacrum, hips, buttocks, shoulders, spinal apophyses, heels, lateral malleoli and fifth metatarsal	Weekly (or more often if required)	Not specified	NPUAP, EPUAP	PU Stage 1, PU Stage 2	Method of analysis not reported, but statistical significance concluded	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Cavicchioli and Carella 2007)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	2 weeks	203	Hospital; acute, post-acute and long-term care	Braden score of <17 and mobility and activity sub scores of <3 respectively. Excluded if more than one PU, or pre-exsiting grade II+ PU	Not specified, but results reported PUs on sacrum, heels,	Fortnightly (once at baseline, once at 2 weeks)	Trained observer	EPUAP	PU grade I+	Chi square test or Fishers exact test	
(Chaboyer et al. 2016)	(O'Connor, Moore, and Patton 2021)	28 days	1598	Tertiary hospital	expected hospital length of stay of 48 hours; at risk of PU as measured by limited mobility	Not specified	Daily	Trained assessor	EPUAP	Hospital acquired PU any stage	Cox regression; cluster-adjusted Chi-squared test for severity	Inter-rater reliability was assessed for PU classification. Kappa was estimates as 0.923 for the presence of a new HAPU, and 0.635 for the specific PU grade
(Cobb 1995)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Coleman et al. 2013)	40 days	123	Hospital wards and ICU	High risk based on Braden, PU free	Not specified, but the skin assessment tool included a diagram of the human body to enable the investigator to draw the site of any PUs	Daily in ICU, every other day on wards	Investigator/project director	NPUAP; Shea Staging system	PU Stage I+	KM curves, Wilcoxon test;	*"There are two recognized limitations of the staging process. There is difficulty in the identification of Stage I ulcers in dark skinned individuals. When eschar (a thick, adherent tissue covering a wound) is present, accurate staging of the pressure ulcer is not possible until the eschar has sloughed or the wound is debrided."*
(Collier 1996)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	Not specified	99	General medical ward	Eligibility criteria not described	Not specified for skin assessments, but heels, sacrum, scapulae and trochanter were assessed for assessing mean interface pressures for mattresses.	At least weekly (Frequency was determined by each patient's condition)	Not specified, interface readings were obtained by the Principal Investigator	Not specified	skin deterioration, not defined	Not analysed - no skin deterioration observed	
(Conine, Daechsel, and Lau 1990)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Shi, Dumville, et al. 2021b)	Not specified	187	Extended care facility for neurological conditions	18-55 years, no evidence of skin breakdown for at least 2 weeks prior to study, at high risk according to Norton's scale	Not specified, but skin sites reported were: Coccyx, sacrum, trochanters, heels, malleoli and other	Daily by nurses and attendants; research assistant checked reports by nurses and attendants. Healing status updated weekly	Research assistant (nurse with tissue trauma experience), and nurses and attendants	Exton-Smith Scale	Grade 1+, possibility of more than one count per subject; healing of PUs also assessed; severity assessed	Not specified, but z statistic reported for patient level incidence; Chi-square test (incidence by skin sites), t-test	Exton-Smith scale chosen because it has excellent inter-rater reliability. One of the authors randomly checked 10% of patients at least once to assess IRR.

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Conine et al. 1994)	(McInnes et al. 2015)	3 months	163	Extended care hospital	High risk (Norton), PU free for at least 2 weeks prior to study, sitting in wheelchair for min 4 consecutive hours	Sacrococcyx, ischial tuberosities and trochanters (other sites noted in results mostly creases of buttocks and thighs)	Weekly	Research assistant (experienced registrered nurse)	Exton-Smith Scale	Grade 1+, possibility of more than one count per subject; healing of PUs also assessed; severity assessed	Z test (patient level incidence and by skin site), Chi squared test for severity	
(Cooper, Gray, and Mollison 1998)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021)	7 days	100	Emergency orthopaedic trauma	65y or older, PU free, Waterlow score of 15 or more	Not specified	Weekly (baseline and end of follow-up)	Not specified, but all nurses in the research wards were trained in the proper use and setting of equipment	Stirling Pressure Sore Severity Scale	Grade I+	Fishers exact test	
(Craig et al. 1998)	(Langer and Fink 2014)	12 weeks	34	Long-term care facility	50y or older, history of type 2 diabetes ,or documented hyperglycaemia	Not specified	Daily	Not specified	Not specified	PU grade not specified	Chi-square test of Fishers exact test	PUs were a safety outcome
(Daechsel and Conine 1985)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Shi, Dumville, et al. 2021b)	3 months	32	Long term care hospital	19-60y, no evidence of skin breakdown for at least 2 weeks prior to study, at high risk	Sacrum, Coccyx, iliac crest, greater trochanters, heels, malleoli, scapulae and elbows	Daily by attendants; weekly by research assistant	Research assistant and attendants	Exton-Smith Scale	Grade I+	Chi-squared test; median test for severity	Weekly checks were carried out 30 mins after turning so not to confuse blanching or reactive hyperthermia with persistent erythema
(Defloor, De Bacquer, and Grypdonck 2005)	(Gillespie et al. 2020)	28 days	838	Nursing homes	Geriatric residents with a Braden score of < 17 or a Norton score of < 12, PU free	sacrum, heels, shoulders, elbows, trochanters, malleoli, ischium	twice weekly	Trained nurses	AHCPR	Non-blanchable erythema (Grade 1); PU lesion (Grades II, III and IV)	Fisher's Exact test, logistic regression, log rank test, KM curves	
(Defloor and Grypdonck 2005)	(Coleman et al. 2013)	4 weeks	1772	Long-term care facility	All patients on selected ward	Sacrum, heels, shoulders, elbows, trochanters, malleoli, ischium	Daily	Trained nurses	EPUAP	PU Grade 1; PU Grade 2+	Chi-square test; logistic regression	Skin sites with pre-existing PUs were not considered in the endpoint derivation
(Delmi et al. 1990)	(Langer and Fink 2014)	6 months	59	Hospital	60y or older, femoral neck fracture,	Not specified	14 days, 21 days, 28 days, discharge, and at 6 months	Not specified	Not specified	PU grade not specified	Descriptive summaries only	PUs were a safety outcome
(Demarre et al. 2012)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	14 days	610	Hospitals	High risk according to Braden, No PU grade II-IV	Not specified, but results reported sacral area, hips, heels, ankles and other	Daily	Ward nurses (qualified nurses and nursing assistants under the supervision of a qualified nurse) trained in PU classification	EPUAP	Grade II-IV; anatomical site	Chi square test; logistic regression; KM curves; log rank test; Chi-square test or Fisher's exact test used to compare interventions in terms of each grade of PU at each skin site	Transparent disc method used to observe non-blanchable erythema IRR of skin observations monitored by the researcher and study nurse who did weekly assessments in a random sample of participants

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Dennis et al. 2005)	(Langer and Fink 2014)	6 months	4023	Hospital	Patients with recent stroke, able to swallow	Not specified	Not specified	Not specified	Not specified	PU grade not specified	KM curves, Log rank test	PUs were a safety outcome
(Diaz-Valenzuela et al. 2019)	(Moore and Webster 2018)	30 days	571	Nursing homes	Braden <14 points, no existing PUs or any category	Not specified, but results reported for sacrum, heels, gluteus, malleolus	weekly by researchers, twice daily by attending nurses	Researchers and attending nurses	EPUAP	Category 1+ PU	The non-inferiority hypothesis was tested by establishing the incidence difference between the groups, calculating the 95% confidence interval (CI) with the Newcombe method to estimate the intervals for the difference between independent proportions; KM curves; Cox regression	
(Donnelly et al. 2011)	(McInnes et al. 2015)	Not specified, but followed until discharge	239	Fracture trauma unit	65y or older, hip fracture, no existing heel pressure damage and no history of previous PUs	Not specified, but results reported sacrum, buttocks, heels, lateral malleolus, Achilles region, knees, toes	Daily	Lead author	NPUAP	Category 1+	Chi-squared test; KM curves; Cox regression	An independent tissue viability nurse assessed photographs of suspected pressure damage and intact skin to assess IRR
(Dutra et al. 2015)	(Moore and Webster 2018)	30 days	160	ICU, CCU, or medical clinic	moderate/high risk of PUs according to Braden, PU free	Not specified, but results reported for sacrum, heels, gluteus, malleolus	Daily	Nurse specialists in enterostomal therapy	Not specified explicitly	New PU (grade not specified)	Unclear, but Chi-square test or Fishers exact test	
(Economides et al. 1995)	(McInnes et al. 2015)	14 days	12	Not specified	Presence of stage IV PU needing myocutaneous flap closure	Site with pre-existing Stage IV	Daily; photographs were taken on days 3, 7, 11 and 14	Not specified	Not specified	Wound breakdown	Descriptive summaries only	
(Ek et al. 1991)	(Langer and Fink 2014; Coleman et al. 2013)	26 weeks	501	Long term medical ward	Newly admitted who remained hospitalised for 3 or more weeks	Not specified	Not specified	Not specified	PU defined as persistent discolouration or epithelial damage or damage to the full thickness of the skin with or without cavity and size, status and treatment.	Persistent discolouration or more severe (see left)	Chi-square test, regression analysis (method not specified)	
(Exton-Smith et al. 1982)	(McInnes et al. 2015)	14 days	66	Geriatric, long-stay and orthopaedic wards	Free from pressure sores or to have erythema of one or more pressure areas (because these are at high risk)	Not specified	3 times weekly	Nurse-investigator	Exton-Smith Scale	Progress (clear/improved/static/deteriorated)	Chi squared test; Wilcoxon test for matched samples to assess time to pressure sore	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Feuchtinger et al. 2006)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a; Coleman et al. 2013)	5 days	175	Cardiovascular surgery department	Scheduled for cardac surgery with extracorporal circulation	Not specified, but results reported on sacrum, buttocks, heels, elbows and shoulders	Pre-op, post op, day 1, day 3, day 5	First assessed by trained staff nurse, if skin alteration occurred a second assessment was made by the research nurse	EPUAP	Grade 1+	Chi square test; logistic regression;	IRR between clinical and research staff was assessed
(Finnegan et al. 2008)	(Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Unclear	40	Tertiary referral center	Admitted for reconstructive surgery to repair tissue deficit in the sacral-coccygeal, trochantic or ischial region	Other vulnerable anatomical sites to the original tissue deficit	Not specified	Surgical team	Not specified	Not specified	Pilot study, so descriptive, but no subjects developed new pressure damage	
(Forni et al. 2018)	(Moore and Webster 2018)	8 days	359	Orthopaedic hospital	Patients with fragility hip fracture aged >=65 years without PU in sacrum area	Sacrum was the primary, other skin sites as secondary but not specified	Daily	Nurse on health care team. Research nurses on each ward also had training in PU classification	NPUAP/EPUAP	PU any grade; Sacral PU Grade 2+	Chi-square test or Fishers exact test; logistic regression; KM curves; log rank test; Cox regression	highlighted limitation of grade I as outcome due to accuracy of diagnosis
(Gebhardt et al. 1996)	(McInnes et al. 2015)	3 months	23	ICU	At risk accoding to Norton (score <13), PU free	Not specified, but results reported for sacral area, spine and heels	Patients visited 4 times weekly, but later says progress of any pressure areas were recorded twice weekly	Research nurse	Exton-Smith Scale	Grade 1+	Chi square test	
(Gentilello et al. 1988)	(McInnes et al. 2015)	4 days after being allowed out of bed	65	Surgical ICU	Orthopaedic injury requiring traction, head injuries or spinal injuries	Not specified	Daily	Not specified	Not specified	Decubitus ulcer	Z statistic	PUs were collected as a complication
(Geyer et al. 2001)	(McInnes et al. 2015)	Not specified	32	Nursing homes	65y or older, Braden 18 or less, combined braden activity and movility or 5 or less, no sitting surface PUs, daily wheelchair use of 6 hours or more	Sitting surface	Weekly	Research staff	NPUAP	Lesions occurring on any aspect of the sitting surface, not just over bony prominences were classified as sitting acquired PU	Chi-square test, mean days until ulceration compared using t-test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Ghezeljeh et al. 2017)	(Gillespie et al. 2020)	3 days	120	ICU	hospitalised in ICU, undergoing mechanical ventilation support for 8 hours following hospitalisation, no spinal or unstable pelvic fractures	Not specified	2 hourly	Not specified	Not standardised - 5 questions to assess early PUs	Not specified	Not analysed - no PUs developed	
(Gilcreast et al. 2005)	(McInnes et al. 2015)	Not specified	338	Military tertiary-care academic medical centers	Moderate-high risk for heel PUs (Braden 14 or less). Excluded if pre-existing PU on the foot	“head to toe” including heels	Daily	Study registered nurses; training was delivered by the project director	NPUAP	PU stage I+; Heel PU	Analysis of variance; logistic regression (OR reported)	Stated that ulcers cannot be reverse staged, or down-staged, as they heal. An example given is that a stage IV does not become a stage III or II but will be a ‘healing stage IV’. The reason given is that scar tissue replaces the original tissue and is therefore different in structure and characteristics from the tissue that was lost. IRR for PU classification was assessed
(Goldstone et al. 1982)	(McInnes et al. 2015)	Not specified	75	A&E	60y and older, suspected fractured neck of femur	Not specified, although sacrum, buttocks, heels and elbows reported in results	Not specified	Not specified	Not specified	PU (grade not specified); width of lesions	Chi-squared test; Wilcoxon rank sum test (size of lesions)	
(Gray, Cooper, and Campbell 1998)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	10 days	54	Orthopaedic, trauma, vascular and medical oncology wards	Waterlow 15 and above, 65y or older, intact skin	Not specified	Every 5 days	Ward link nurse blind to whether patients were participating in the trial	Not specified	Superficial PU or more severe	Fisher’s exact test	
(Gray and Smith 2000)	(Shi, Dumville, et al. 2021a)	10days	100	Hospital	Emergency or list admission for bed rest or major surgery, intact skin, no existing skin conditions	Not specified, sacrum and heel reported in results	Every 5 days	Not specified	Torrance	Non-blanching redness or more severe	Fisher’s exact test	
(Gray et al. 2008)	(Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Not specified	100	Hospital	Emergency admissions at high risk of PU	Not specified, but PUs reported on sacrum and heels	Regular skin inspection carried out in line with routine care	PUs were graded by a member of the tissue viability department	EPUAP	PU Grade 2+	Descriptive summaries only	Findings from patient notes were compared with the ward PU incidence weekly reports to ensure no ulcers were missed
(Grindley and Acres 1996)	(Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	6 days (3 days each mattress)	20 (cross-over)	Hospice	Patients with existing grade 2 or above PU, or high to very high risk (Waterlow 15 or more)	Not specified, but buttocks, sacrum, heels, trochanter reported	Daily	Not specified	Torrance	Not specified	Not analysed	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Guihan et al. 2014)	(O'Connor, Moore, and Patton 2021)	6 months	144	Veterans Affairs (VA) SCI centres	admission for treatment of a severe (Stage III or IV) pelvic ulcer, 6 months post spinal cord injury	Ischium, trochanter, sacrum Coccyx, other	3 monthly	Photograph	PU stage not reported	Skin worsening (study specific definition)	Fishers exact test	
(Gunningberg et al. 2000)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	14 days	119 or 101 (unclear which were randomised)	A&E & Orthopaedic ward	65y or older, suspected hip fracture	Not specified, but results reported for sacrum, buttocks, back and heels	On admission to A&E, on admission to ward, 4 days post op, discharge/14 days post op	Registered nurse on duty for first two assessments, PU nurse for last two assessments. 25 PUs were photographed and graded by an expert nurse to assess IRR	EPUAP	PU Grade I+	Chi-Squared test; logistic regression	Most severe grade was used for patients with multiple or 'changing' PUs
(Hampton 1997)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Not specified	36	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	No PUs reported	
(Hartgrink et al. 1998)	(Langer and Fink 2014)	2 weeks	140	Hospital	Fracture of the hip, high risk of PU, excluded if pre-existing cat 2+	Sacrum, trochanters, heels and elsewhere	Weekly	Two physicians independently, any discrepancy in PU classification resolved by a third opinion	Conclusion from Dutch consensus study (Haalboom and Bakker 1992) 0 =Normal skin 1=Persistent Erythema 2=Blister formation 3=superficial (sub)cutaneous necrosis 4=Deep subcutaneous necrosis	PU grade 2+	Fisher's exact test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Hofman et al. 1994)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	2 weeks	44	Hospital	High risk of PU according to Pressure sore risk score (Bakker 1985), femoral-neck fracture, No PU cat 2+	Sacrum, trochanters, shoulders, heels, and 'other'	Weekly	2 independent physicians, disagreements resolved by third opinion	Conclusion from Dutch consensus study (Bakker 1985) 0=Normal skin,1=Persistent Erythema 2=Blister formation 3=superficial (sub)cutaneous necrosis 4=Deep subcutaneous necrosis	All grades including 0; PU Grade 2+	U test comparing median maximum Pressure sore grades; Fisher's Exact test	
(Hoshowsky and Schramm 1994)	(Shi, Dumville, et al. 2021a; Shi, Dumville, et al. 2021b)	1 day (preop vs post op)	505	Hospital	Placement in the supine or prone positions while undergoing surgery, 12y or older, possession of symmetrical lower limbs	Heels and knees	Pre-operatively, post operatively	researcher	NPUAP	Stage I+, skin changes less than Stage I	Logistic regression	
(Houwing et al. 2003)	(Langer and Fink 2014)	28 days	103	Not specified; 3 centres in the Netherlands	Fracture of the hip, high risk of PU	Tail-bone, heels, buttocks and other	Daily	Nursing staff	EPUAP	PU Stage I+, PU Stage II+, wound size and duration	Fishers exact test, t-test, ANOVA	Specific instructions given for Stage I using a plastic tongue depressor to assess blanching
(Houwing et al. 2008)	(Moore and Webster 2018)	28 days	79	Nursing homes	at high risk of developing PU according to the Braden scale using a cutt-off point of 20, PU free	heels and buttocks	Not specified	Not specified	EPUAP	PU Category 1 +	Logistic regression	Patients with dark skin were excluded because of difficult skin assessment; 2 additional observers were used to confirm the presence of a Category 1 PU because it was acknowledged that it can be difficult to differentiate between blanchable redness and grade 1 non-blanchable erythema
(Inman et al. 1993)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Coleman et al. 2013)	Not specified	100	Critical care trauma centre	APACHE II score >15; expected ICU stay of 3 or more days	13 bony prominances	Daily	Trained critical care research nurse	Shea	Single PU; multiple PUs; severe PU (severity score >1); resolution of PU	Logistic regression	
(Jiang et al. 2014)	(Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	5 days	1074	Hospital	Braden 16 or less, surgery >120 mins,	Head to toe; sites reported were sacrum, heels trochanter	2 hourly	Trained nurse	NPUAP	Stage I+ PU	Chi-square test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Jolley et al. 2004)	(McInnes et al. 2015; Shi, Dumville, et al. 2021b)	Not specified	539	Hospital	Patients at low to moderate risk of PU according to Braden. Excluded if no risk or high risk (requiring more complex interventions than the trial intervention), pre-existing PU, darkly pigmented skin making stage 1 PU difficult to detect	Not specified	Daily	Research nurses	Referenced Bergstrom 1992; Stage 1, Stage 2, Stage 3, Stage 4	PU Stage 1+ ; PU Stage 2+	Cumulative incidence risk; risk ratio; incidence rate per 100 bed days; incidence rate ratio; odds ratio reported; KM curves; Cox regression	Stage 1 diagnosed if non blanching erythema was still present after 30 minutes of pressure relief to the affected area; Intra-observer reliability was assessed for PU classification
(Kalowes, Messina, and Li 2016)	(Moore and Webster 2018)	6 months post discharge	366	ICU	Braden score30 of ≤ 13 and intact sacral skin	Sacrum was the primary, other skin sites as secondary but not specified	Daily in ICU, medical record review afterwards	trained study team member, independent ICU expert nurse independently verified outcome assessments	NPUAP	Hospital acquired PU Stage 1+	Fishers exact test; logistic regression; Cox regression; Poisson regression	IRR testing conducted before data collection to ensure data accuracy and consistency in PU assessment
(Karimi et al. 2018)	(O'Connor, Moore, and Patton 2021)	3 months	70	residents' home	stroke patients;absence of any signs suggesting bedsore or skin disorders in the patient at the beginning of the study; aged 45 to 75 years old; bedsore risk score of less than or equal to 14 (being moderately or severely at risk of bedsore, according to Braden scale).	Not specified	3 monthly	Patient reported	NPUAP/AWMA	PU grade not specified	Chi-square test	
(Kathirvel et al. 2021)	Additional reference identified in follow-up to a reference included in the Cochrane review (O'Connor, Moore, and Patton 2021) that was not eligible for this scoping review due to incompleteness	3 months post disharge	92	Orthopaedics	Braden score ≤ 12 or Stage I PU	Not specified	Daily until discharge, weekly for two weeks, fortnightly thereafter until 3m post discharge	Not specified	NPUAP	a PU event was defined it a patient moved from high/very high Braden score to stage I PU, or moving from stage I PU to stage II+ PU. Patients with more than one PU in a different anatomical site more than 7 days after the first PU were counted for a recurrent PU event	Cumulative incidence, incidence rate, rate ratio, preventive fraction; KM curves; Breslow test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Kaur et al. 2018)	(O'Connor, Moore, and Patton 2021)	12 months	92	Community	bedridden patient, defined as a patient above 12 years of age who had been confined to bed for 15 days or more, for 90% of the time during the day and who was unable to get out of bed or change position in bed without assistance	Not specified	Weekly for 1 month, fortnightly for 3 months, monthly for 6 months, 2 monthly for 1 year	Trained assessor	Not specified	PU grade I	Descriptive summaries only	
(Kemp et al. 1993)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a; Coleman et al. 2013)	1 month	84	General medicine unit, geriatric medicine unit, long-term care facility	65y or older; Braden 16 or less; PU free	Not specified but results include sacrum, buttocks, 'heels, toes or ankles', elbows and trochanters	3 times weekly	Trained research nurses	NPUAP	Stage I+	Log rank test; Cox regression	Training included assessment of 5 photographs to assess agreement on the PU stage. Discussion on the issue of panel data
(Keogh and Dealey 2001)	(McInnes et al. 2015)	10 days	100	Surgical and medical wards	Waterlow 15-25; no greater than Grade I PU	Not specified	Daily	Not specified	EPUAP	Grade I+; Resolution of pre-existing Grade I	Not specified; no PUs developed, and planned analysis not specified	
(Lim et al. 1988)	(McInnes et al. 2015)	5 months	62	Extended -care facility	60y or older; free of decubitus ulcer for at least 2weeks prior to study, high risk according to Norton's scale (14 o less), wheelchair use for 3h or more each day	Buttocks; sacrum; ischial tuberosities, trochanters	Weekly	Occupational therapist	Exton-Smith	PU score 1+; possibility of more than one count per subject if new PUs developed. Maximum score taken.	Chi-squared test, t-test; compared incidence at each skin site;	One of the authors monitored the assessor and randomly double-checked 10% of participants to ensure accuracy of measures taken.
(Lupianez -Perez et al. 2015)	(Moore and Webster 2018)	16 weeks	831	Home nursing service	patients at risk of suffering pressure ulcers, Braden <16, nutritional status<10 according to Mini Nutritional Assessment, No pre-existing PU	Sacrum, hips, heels	Weekly	Not specified	Not specified	PU Stage 2	Chi-Squared test, KM curves, log rank test; Absolute risk reduction and relative risk values reported	
(Malbrain et al. 2010)	(Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Not specified	16	ICU	High PU risk (Norton 8 or less) and requiring medical ventilation	Bony prominences – sacral area and heel reported	Daily in appropriate light	ICU nurse	EPUAP; PUSH (for healing)	PU Category 1, PU Category 2-4, deterioration of wounds (improved, unchanged, deteriorated), surface area, PUSH score (for healing)	Fishers exact test, t-test	PUs assessed independently once weekly by study nurse and study doctor to assess IRR

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Manzano et al. 2014)	(Gillespie et al. 2020)	60 days	330	ICU	Critically ill adults with no PI at ICU admission who received invasive mechanical ventilation for at least 24 hours	Not specified	Not specified	Trained study nurses	EPUAP	Grade II PU	Log rank test, KM curves, Cox PH regression	
(McGowan 2000)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021)	35 days according to KM plot	297	Hospital	60y or older, orthopaedic diagnosis, low or moderate risk of PU (Braden). Excluded if no risk or high risk; pre-existing PU, coloured skin patients were stage 1 ulcer detection is difficult	Not specified, but sacrum, heels and elbows reported	Daily	Trained research nurses	Described in paper: 4 stages	PU Stage 1+	Cumulative incidence risk ratio; KM curves; log-rank test	Regular inter-rater comparisons were conducted for PU classification
(Mistiaen et al. 2010)	(Shi, Dumville, et al. 2021b)	30 days	588	Nursing homes	Newly admitted for a primarily physical impairment; PU free on the sacrum; not having darkly pigmented skin (due to difficulty in diagnosing grade 1 PU)	Sacrum, plus 'other' skin sites	Daily	Attending nurse	EPUAP	Grade 1+	Chi-squared test, Logistic regression; difference in proportion of PU free days and mean number of PU days was tested by t-test	A photographic series of the various pressure ulcer grades was available on each ward as well as transparent disks that nurses pressed against erythema by hand to see whether the area blanched under pressure. Assessors contacted specialised nurse if they were uncertain about their observations. All PUs were reported to wound care specialist who verified the observation. Principal investigator conducted bimonthly observations to assess agreement in sacral PU assessment
(Moore, Cowman, and Conroy 2011)	(Gillespie et al. 2020)	28 days	213	long-term aged-care facilities	at risk of PI development using the activity and mobility components of Braden scale, no PI at time of recruitment to study	Not specified, but skin site data were collected with sacrum and buttocks reported as the most common site, on on the knee, and none on the heels	3 to 6 hourly	If any changes in skin integrity were noted, the researcher was informed which prompted skin assessments the assigned key staff member, clinical manager, and researcher. Agreement between assessors was	EPUAP	Grade 1+ PU	Chi-squared test, Poisson regression (adjusted for hospital clusters), Odds ratios also reported	The rationale for inclusion of grade 1 pressure ulcer damage was that it is considered to be an important indicator of risk for the development of more severe pressure ulcer development but was verified by consensus

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
								reached by comparing patients' skin condition to images of the EPUAP grading system				
(Nakagami et al. 2007)	(Moore and Webster 2018)	3 weeks	37	Hospital	Braden Scale score of <15, no pre-existing PU	L & R trochanters (paired comparison within patient)	Weekly	Trained RN + Separate Investigator to agree final skin condition by consensus	NPUAP	Persistent erythema or PU occurrence	MnNemar test; GEE for relative risk accounting intra-individual correlations	
(Nixon et al. 1998)	(McInnes et al. 2015; Shi, Dumville, et al. 2021b)	1 day	446	Hospital	Elective general, gynaecological or vascular surgery, 55y or older, excluded if PU grade 2a+, dark skin pigmentation which precludes reliable identification of Grade 1 and Grade 2a PUs; skin conditions over the sacrum, buttocks or heels which preclude reliable identification of grade 1 or grade 2a PUs	Sacrum, buttocks, heels	Pre-anaesthetic, up to 30 mins post op, 30mins-1h later, 1 day post op	Trained nursing staff	Torrance (adaptation)	Success/failure based on persistent deterioration of PU status	Odds ratio reported; logistic regression	IRR of skin assessment was assessed between clinical staff and the research nurses.
(Nixon et al. 2006)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Coleman et al. 2013)	60 days	1972	Hospital	Acute or elective surgery patients; limited activity and mobility according to Braden, or pre-existing grade 2 PU. Patients with Grade 3+ PU were excluded.	Sacrum, buttocks, hips and heels	Twice weekly for 30 days, once weekly for days 30-60	Clinical research nurses	Grade 0, Grade 1a, Grade 1b, Grade 2, Grade 3, Grade 4, Grade 5	Grade 2+ PU within 60 days (primary) and within 30 days	Chi-squared test, logistic regression; log rank test; Mann-Whitney U test to compare PU area. Cost effectiveness analysis used KM estimate of restricted mean time to PU development.	**Motivating dataset**
(Nixon et al. 2019)	(Shi, Dumville, et al. 2021a; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	90 days	2029	Hospital	acutely ill, bedfast/chairfast and/or Category 1 PU/pain at PU site. Excluded if current/previous Category 3 + PU	Sacrum, buttocks, heels, ankles, trochanters, ischial tuberosities, back, elbows	Twice weekly for 30 days, once weekly for day 30-60, final assessment at day 90	Research nurses	EPUAP	PU Category 2+, PU Category 1+, PU Category 3+	Fine and Gray model	Sub study to assess IRR (photographs and independent assessor), **Motivating trial dataset**

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recruited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Olofsson et al. 2007)	(Langer and Fink 2014)	4 months (PUs only assessed in hospital)	199	Hospital	Femoral neck fracture, 70y or older, patients bedridden prior to injury excluded	Not specified	Pre-and post-operatively, and once between 3 and 5 days post op, unclear thereafter	Study nurses with medical and social data collected from patients, relatives, staff and medical records	Not specified	PU grade not specified	Chi-square test	
(Otero et al. 2017)	(Moore and Webster 2018)	10 hours after removal of mask	171	HDU	Patients with acute respiratory failure requiring NIV, no lesion on the skin supporting respiratory interface	Facial areas	6 hourly	two assessors, highest category recorded in case of discrepancy	GNEAUPP	PUs on areas of skin in contact with mask; grade not specified	Analysis method not specified. Absolute risk reduction and number needed to treat reported	
(Pickham et al. 2018)	(Gillespie et al. 2020)	3 days	1312	ICU	critically ill medical, surgical, and trauma patients	Not specified, but "head to toe" described, sacrum and buttocks highlighted as most common regions	Daily	Shift nurse - any findings were assessed within 24h by independent Certified Wound, Ostomy, and Continence Nurse	NPUAP	Hospital acquired PU, grade not specified	Fisher's Exact test, logistic regression	
(Price et al. 1999)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	14 days	80	Hospital	Fractured neck of femur, 60y and older, very high risk according to Medley score>25	Sacrum, scapulas, elbows, buttocks, trochanters, calves, heels, medial and lateral malleoli	Not specified	Trained assessors	0=normal skin, 1=persistent erythema,2=blister formation, 3=superficial sub/cutaneous necrosis, 4=deep subcutaneous necrosis	All grades	ANCOVA	"No statistically significant difference between the groups at any time point or in terms of progression over assessment stages"
(Rafter 2011)	(Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	4 weeks	10	Hospital	No existing skin damage, or up to Category 2 PU. Excluded if re-admitted with PUs	Not specified, sacrum and heels reported	Daily by ward staff, 3 times a week by audit co-ordinator	Ward staff and audit co-ordinator. PUs categorised by tissue viability nurse specialist	EPUAP	PU Category 1+	Descriptive summaries only	
(Rantz et al. 2012)	(Porter-Armstrong et al. 2018)	2 years	58 centres	Nursing homes	Residents	Not specified	6 monthly		Not specified	PU stage I	Repeated measures Logistic regression	Nursing home level analysis
(Ricci et al. 2013)	(McInnes et al. 2015; Shi, Dumville, et al. 2021b)	4 weeks	50	Long-term care units	Braden 8-14, or Norton 6-12. PU stage 2+ excluded	Sacrum, heels, spinour process, trochanter, lateral and medial malleolus, ischium, elbows	Weekly	External physician	EPUAP/NPUAP	PU (Grade not specified); Change of ulcer size; PUs at sacrum and heels as well as other exposed anatomical areas	No PUs observed so no analysis conducted, and none pre-specified	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Rintala et al. 2008)	(O'Connor, Moore, and Patton 2021)	24 months	41	ICU/ Veterans Affairs Medical Center	veterans with spinal cord injury who have had surgical repair of a pressure ulcer.	trochanter, ischium, or sacrum/coccyx	monthly (3 monthly in one group)	Patient reported	Not specified	Ulcer recurrence (stage II +)	Cox regression; KM curves;	
(Dunlop and Care 1998)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Shi, Dumville, et al. 2021b)	7 days	195	Cardiotho racic surgery	Scheduled for cardiothoracic procedure, scheduled surgical time of 3+ hours, anaesthesia time of 4+ hours, PU free	Not specified	Pre-op assessment, 1 hour post op then daily	Not specified	NPUAP	PU Stage I+	Preliminary analysis, therefore descriptive only	
(Russell et al. 2003)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	Not specified – until death or discharge	1166	Hospital	65y and older, Waterlow 15-20, patients with small areas of blanching erythema were eligible		Daily by ward nurses, weekly by research nurses	Trained research nurses and ward nurses	Torrance	PU Grade II (primary), PU Grade I+ (secondary)	Chi-square test, KM curves, logistic regression	Training included specific definition of blanching and non-blanching erythema; Research nurses were notified immediately if the ward nurse reported skin deterioration to confirm the classification.If an ulcer healed and recurred, this was considered a new ulcer. Discussion noted that it is difficult to distinguish between blanching and non-blanching erythema. Adding that particularly difficult to assess blanching in pigmented skin. The study was mainly Caucasian raising concerns re generalisability. Also noted in discussion that a 5 day work week for the research nurses led to a sudden development of PUs at 7 days.
(Saleh, Anthony, and Parboteeah 2009)	(Moore and Patton 2019)	8 weeks	719	Military hospital	Braden score of less than or equal to 18	Not specified	Not specified	TVTN and trained staff nurses	NPUAP/EPUAP	PU grade 1+; PU grade 2+	Chi-square test; logistic regression	
(Sanada et al. 2003)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Not specified	123	Acute care unit	Braden 16 or less, bed bound, PU free, and required heel elevation	Not specified, but coccyx, sacrum heels and trochanter reported in results	Daily	Trained nurses	NPUAP	Pu Stage I+	Chi-square test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Santamaria et al. 2015)	(Moore and Webster 2018)	Unclear, but survival curves go up to 25 days	440	ICU	ED and ICU admission for critical illness and/or trauma. No pre-existing sacral or heel PU	sacrum and heels	Daily	study team member	AWMA	PUs any grade, also reported and compared by anatomical site	Fishers exact test; Cox regression; Absolute risk reduction and number needed to treat reported	All members of the research team underwent inter-rater reliability testing in September 2010 prior to data collection to ensure consistency in pressure ulcer identification and staging.
(Sauvage et al. 2017)	(Shi, Dumville, et al. 2021a; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	30days	76	Medium and long term care facilities	70y and over, bedridden for 15+ hours per day, no PUs, medium to high risk of PUs (Braden 14 or less)	Not specified, sacrum and heels reported in results	Daily	Not specified	EPUAP, NPUAP, PPPIA	Category I+	Km curves, log rank test, Cox regression	
(Schultz et al. 1999)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a; Coleman et al. 2013)	6 days post op	413	Tertiary care centre	Surgical patients with scheduled surgery for 2+ hours in the lithotomy or supine position, excluded if pre-existing PU or severe chronic skin problems	Sacral/coccyx, heels, elbows	Pre-operative, and then daily post operatively	Trained research assistants	Described in paper: Stage I, Stage II, Stage III, Stage IV, Unable to stage	Pu Stage I+	Chi-square test; logistic regression	Minimum agreement for identification of PUs was set for new assessors, and re-established part way through the study
(Siderank o et al. 1992)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021; Shi, Dumville, et al. 2021b)	Not specified	57	Surgical ICU	Excluded if pre-existing skin breakdown	Not specified, but 'pressures' were assessed at sacrum and heels	Not specified	Not specified	Not specified	PU (grade not specified)	Chi-square test	Main outcome was 'pressures' and used a repeated measures analysis
(Summer et al. 1989)	(McInnes et al. 2015)	Not specified	83	ICU	One of 6 conditions	Not specified	Not specified	Study nurse	Not specified	Classic decubitus ulcer (grade not specified)	No classic ulcer developed	Ulcers were secondary outcome
(Taylor 1999)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	Not specified (discharge/ death)	44	Hospital	PU free, required a pressure relief support due to their medical condition	Not specified, but 'pressures' measured at coccygeal/sacral area and ischial tuberosities	Every 4 days	Not specified	Not specified	Blanching erythema (or more severe)	Descriptive summaries only	Suggestion of pressures as a surrogate for prevention of PUs
(Theaker, Kuper, and Soni 2005)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Jammali-Blasi, et al. 2021)	2 weeks post ICU discharge	62	ICU	High risk of PU,	Not specified, but PUs reported on heels and sacrum	Every 8h	Attending nurse	Lowthian scale	Not specified, but observed PUs were Grade 2 and grade 3	Chi-square test, two way ANOVA, Mann-Whitney U-test (development of PU and duration of time on the study bed)	If PU was identified by attending nurse, it was photographed and assessed by two independent tissue viability nurses
(Theilla et al. 2007)	(Langer and Fink 2014)	7 days	100	ICU	Acute lung injury	Not specified	Daily	Researcher	NPUAP	PU Grade 1+	Chi-square test	

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Torra i Bou et al. 2005)	(Moore and Webster 2018)	30 days	380	Unclear	medium, high or very high risk of PU development, no more than 3 PUs	not specified, but could assume at least sacrum, trochanter and heels as this is where the intervention was applied	not specified	not specified	not specified	PU grade not specified	Chi-square test; KM curves; log rank test; Cox regression; Relative risk, predictable fraction and number needed to treat reported	
(Tymec, Pieper, and Vollman 1997)	(McInnes et al. 2015)	14 days	52	Hospital	Braden 16 or less, intact skin on lower extremities	Lower extremities – knees to toes	Not specified	Not specified	AHCPR	PU stage I+	Fisher's exact test, KM curves, log rank test	Main outcome was 'pressures'. In the discussion acknowledged that research error may have occurred in this study during the assessment of erythema
(van Leen et al. 2011)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, et al. 2021a)	6 months	83	Nursing home	Norton score 5-12, excluded if PU in previous 6 months,	Sacral/hip region and heels	Weekly	Independent nurse	EPUAP	PU grade 2-4	Fisher's exact test	
(van Leen et al. 2013)	(Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, et al. 2021a)	1 year (6m each mattress)	40 (cros s over)	Nursing home	>65y, Braden 6-19, PU free	Not specified, but heels and pelvic region reported	Weekly	Not specified	EPUAP	Category 2,3,4	Fishers exact test	Category 1PU excluded from outcome due to inconclusiveness of the diagnosis
(van Leen, Halfens, and Schols 2018)	(Shi, Dumville, et al. 2021a; Shi, Dumville, et al. 2021b)	12 weeks	206	Nursing homes	60y or older, Medium to high risk of PUs (Braden <16), PU free for at least 3 months	Not specified, but sacral, heels, other reported	weekly	Research nurses	EPUAP, NPUAP, PPPIA	PU Category 2,3 or 4	Chi-squared test	Category 1 PU excluded from outcome because of potential equivocal diagnosis
(Vanderwee, Grypdonck, and Defloor 2005)	(McInnes et al. 2015)	Not specified	447	Hospital	No Grade 2+ PU, in need of PU preventative measures according to Braden or the presence of a grade 1 PU	Not specified, but sacral and heel PUs highlighted in results	Daily	Ward nurses	EPUAP	PU grade 2-4	Fisher's exact test, logistic regression, KM curves, log rank test	Random sample of patients observed at unexpected moments by both the researcher and data nurse to assess IRR. Assessment of blanchable erythema was assessed using transparent disk
(Vanderwee et al. 2007)	Additional reference identified in follow-up to Vanderwee 2009 included in the systematic review (Coleman et al. 2013)	5 weeks	235	Nursing homes	Non-blanchable erythema but no pre-existing Grade 2,3,4 PU	Sacral area, heels, ankles, trochanter and hips	Daily	Trained nurses (PUCLAS)	EPUAP	PU grade 2-4	Fisher's exact test, logistic regression, KM curves, log-rank test	Plastic disc used to differentiate between blanching and non-blanching erythema. IRR of PU classification assessed through weekly assessments by researcher and study nurse.
(Vanderwee et al. 2009)	(Coleman et al. 2013)										Cox regression	Secondary analysis of trial above.

Author and year	Source (e.g. Cochrane review, systematic review, or additional identified paper)	Maximum Follow-up period	N recr uited	Setting	Patient inclusion criteria	Skin sites	Assessment schedule	Assessor information	Classification scale	Skin status endpoints	Analysis methods	Other information
(Vermette , Reeves, and Lemaire 2012)	(McInnes et al. 2015; Shi, Dumville, Cullum, Rhodes, Leung, et al. 2021; Shi, Dumville, et al. 2021b)	2 weeks	110	Hospital/ICU	Moderate to very high risk of PU (Braden<14), no skin lesions	‘head to toe’	3 times weekly	Research nurse	NPUAP	PU stage I+	Fisher’s exact test, Logistic regression	
(Vyhlidal et al. 1997)	(McInnes et al. 2015; Shi, Dumville, et al. 2021a)	21 days	40	Hospital	At risk of PU (Braden <18), PU free	28 sites pre-specified, plus option for others. Most common sites reported as sacrum/coccyx, ankles and heels	3 times weekly	Research team member	Bergstrom skin assessment tool	PU stage I+	Chi-square test, t-test for mean number of days to PU	IRR assessed by a pair of investigators monthly
(Ward et al. 2004)	4	12 months	114	Patient's homes	adults aged over 15 years with 1 of the following possible recorded diagnoses: multiple sclerosis, Parkinson's disease and other causes of progressive parkinsonism, motor neurone disease, Huntington's disease, and other degenerative disorders affecting the central nervous system, muscles, or peripheral nerves	Not specified	two monthly	patient reported	Not specified	Patient report of skin sores	logistic regression	
(Webster et al. 2011)	9	Until discharge	1231	Internal medicine or oncology wards	Length stay >=3 days, and recruited within 24h of admission	Not specified	Daily (except weekends)	Trained research assistants	NPUAP	PU event defined as new PU or any increase in existing PU stage	Logistic regression	Highlighted that there are issues in the assessment of Grade 1 pressure ulcers. Photographic review and/or nurse practitioner was asked for judgement. IRR was assessed for research assistants
(Young 2004)	1	1 day	46	acute hospital	at risk of developing a PI using Waterlow score, no existing PI	Not specified, but "full skin inspection" was conducted, and the sites with PUs were reported (sacrum, trochanters and heels)	once at 24h	Researcher	EPUAP	Non-blanching erythema	Fishers exact test	

Source

Coleman, S., C. Gorecki, E. A. Nelson, S. J. Closs, T. Defloor, R. Halfens, A. Farrin, J. Brown, L. Schoonhoven, and J. Nixon. 2013. 'Patient risk factors for pressure ulcer development: systematic review', *Int J Nurs Stud*, 50: 974-1003.

Gillespie, B. M., R. M. Walker, S. L. Latimer, L. Thalib, J. A. Whitty, E. McInnes, and W. P. Chaboyer. 2020. 'Repositioning for pressure injury prevention in adults', *Cochrane Database Syst Rev*, 6: CD009958.

Joyce, P., Z. E. Moore, and J. Christie. 2018. 'Organisation of health services for preventing and treating pressure ulcers', *Cochrane Database Syst Rev*, 12: CD012132.

Langer, G., and A. Fink. 2014. 'Nutritional interventions for preventing and treating pressure ulcers', *Cochrane Database Syst Rev*: CD003216.

McInnes, E., A. Jammali-Blasi, S. E. Bell-Syer, J. C. Dumville, V. Middleton, and N. Cullum. 2015. 'Support surfaces for pressure ulcer prevention', *Cochrane Database Syst Rev*: CD001735.

Moore, Z. E. H., and D. Patton. 2019. 'Risk assessment tools for the prevention of pressure ulcers', *Cochrane Database of Systematic Reviews*.

Moore, Z. E., and J. Webster. 2018. 'Dressings and topical agents for preventing pressure ulcers', *Cochrane Database Syst Rev*, 12: CD009362.

O'Connor, T., Z. E. Moore, and D. Patton. 2021. 'Patient and lay carer education for preventing pressure ulceration in at-risk populations', *Cochrane Database Syst Rev*, 2: CD012006.

Porter-Armstrong, A. P., Z. E. Moore, I. Bradbury, and S. McDonough. 2018. 'Education of healthcare professionals for preventing pressure ulcers', *Cochrane Database Syst Rev*, 5: CD011620.

Shi, C., J. C. Dumville, N. Cullum, S. Rhodes, V. Leung, and E. McInnes. 2021. 'Reactive air surfaces for preventing pressure ulcers', *Cochrane Database Syst Rev*, 5: CD013622.

Shi, C., J. C. Dumville, N. Cullum, S. Rhodes, and E. McInnes. 2021a. 'Foam surfaces for preventing pressure ulcers', *Cochrane Database Syst Rev*, 5: CD013621.

Shi, C. H., J. C. Dumville, N. Cullum, S. Rhodes, A. Jammali-Blasi, and E. McInnes. 2021. 'Alternating pressure (active) air surfaces for preventing pressure ulcers', *Cochrane Database of Systematic Reviews*.

Shi, C. H., J. C. Dumville, N. Cullum, S. Rhodes, and E. McInnes. 2021b. 'Alternative reactive support surfaces (non-foam and non-air-filled) for preventing pressure ulcers', *Cochrane Database of Systematic Reviews*.

Reviewed articles

Andersen, K. E., O. Jensen, S. A. Kvorning, and E. Bach. 1983. 'Decubitus prophylaxis: a prospective trial on the efficiency of alternating-pressure air-mattresses and water-mattresses', *Acta Derm Venereol*, 63: 227-30.

Beeckman, D., B. Serraes, C. Anrys, H. Van Tiggelen, A. Van Hecke, and S. Verhaeghe. 2019. 'A multicentre prospective randomised controlled clinical trial comparing the effectiveness and cost of a static air mattress and alternating air pressure mattress to prevent pressure ulcers in nursing home residents', *International Journal of Nursing Studies*, 97: 105-13.

Bennett, R. G., P. J. Baran, L. V. DeVone, H. Bacetti, B. Kristo, M. Tayback, and W. B. Greenough, 3rd. 1998. 'Low airloss hydrotherapy versus standard care for incontinent hospitalized patients', *J Am Geriatr Soc*, 46: 569-76.

Bergstrom, N., S. D. Horn, M. P. Rapp, A. Stern, R. Barrett, and M. Watkiss. 2013. 'Turning for Ulcer ReductioN: a multisite randomized clinical trial in nursing homes', *J Am Geriatr Soc*, 61: 1705-13.

Bliss, M. R. 1995. 'Preventing pressure sores in elderly patients: a comparison of seven mattress overlays', *Age Ageing*, 24: 297-302.

Bliss, M. R., R. McLaren, and A. N. Exton-Smith. 1967. 'Preventing pressure sores in hospital: controlled trial of a large-celled ripple mattress', *Br Med J*, 1: 394-7.

Bloemen-Vrencken, J. H., L. P. de Witte, M. W. Post, C. Pons, F. W. van Asbeck, L. H. van der Woude, and W. J. van den Heuvel. 2007. 'Comparison of two Dutch follow-up care models for spinal cord-injured patients and their impact on health problems, re-admissions and quality of care', *Clin Rehabil*, 21: 997-1006.

Bourdel-Marchasson, I., M. Barateau, V. Rondeau, L. Dequae-Merchadou, N. Salles-Montaudon, J. P. Emeriau, G. Manciet, and J. F. Dartigues. 2000. 'A multi-center trial of the effects of oral nutritional supplementation in critically ill older inpatients. GAGE Group. Groupe Aquitain Geriatrique d'Evaluation', *Nutrition*, 16: 1-5.

Brienza, D., S. Kelsey, P. Karg, A. Allegretti, M. Olson, M. Schmeler, J. Zanca, M. J. Geyer, M. Kusturiss, and M. Holm. 2010. 'A randomized clinical trial on preventing pressure ulcers with wheelchair seat cushions', *J Am Geriatr Soc*, 58: 2308-14.

Bueno de Camargo, W. H. , R. D. Pereira, M. T. Tanita, L. Heko, I. C. Grion, J. Festti, A. L. Mezzaroba, and C. M. C. Grion. 2018. 'The Effect of Support Surfaces on the Incidence of Pressure Injuries in Critically Ill Patients: A Randomized Clinical Trial', *Critical Care Research and Practice*, 2018.

Caplan, G. A., J. A. Ward, N. J. Brennan, J. Coconis, N. Board, and A. Brown. 1999. 'Hospital in the home: a randomised controlled trial', *Med J Aust*, 170: 156-60.

Cassino, R, AM Ippolito, and E %J Acta Vulnologica Ricci. 2013. 'Comparison of two mattress overlays in the prevention of pressure ulcers', 11: 15-21.

Cavicchioli, A., and G. Carella. 2007. 'Clinical effectiveness of a low-tech versus high-tech pressure-redistributing mattress', *J Wound Care*, 16: 285-9.

Chaboyer, W., T. Bucknall, J. Webster, E. McInnes, B. M. Gillespie, M. Banks, J. A. Whitty, L. Thalib, S. Roberts, M. Tallott, N. Cullum, and M. Wallis. 2016. 'The effect of a patient centred care bundle intervention on pressure ulcer incidence (INTACT): A cluster randomised trial', *Int J Nurs Stud*, 64: 63-71.

Cobb, Gladys A. 1995. *Pressure ulcers: patient outcomes on a KinAir bed or EHOB waffle mattress* (Nursing Research Service, Department of Nusring, Brooke Army Medical Center).

Collier, M. E. 1996. 'Pressure-reducing mattresses', *J Wound Care*, 5: 207-11.

Conine, T. A., D. Daechsel, and M. S. Lau. 1990. 'The role of alternating air and Silicore overlays in preventing decubitus ulcers', *Int J Rehabil Res*, 13: 57-65.

Conine, T. A., C. Hershler, D. Daechsel, C. Peel, and A. Pearson. 1994. 'Pressure ulcer prophylaxis in elderly patients using polyurethane foam or Jay wheelchair cushions', *Int J Rehabil Res*, 17: 123-37.

Cooper, P. J., D. G. Gray, and J. Mollison. 1998. 'A randomised controlled trial of two pressure-reducing surfaces', *J Wound Care*, 7: 374-6.

Craig, L. D., S. Nicholson, F. A. SilVerstone, and R. D. Kennedy. 1998. 'Use of a reduced-carbohydrate, modified-fat enteral formula for improving metabolic control and clinical outcomes in long-term care residents with type 2 diabetes: results of a pilot trial', *Nutrition*, 14: 529-34.

Daechsel, D., and T. A. Conine. 1985. 'Special mattresses: effectiveness in preventing decubitus ulcers in chronic neurologic patients', *Arch Phys Med Rehabil*, 66: 246-8.

Defloor, T., D. De Bacquer, and M. H. F. Grypdonck. 2005. 'The effect of various combinations of turning and pressure reducing devices on the incidence of pressure ulcers', *International Journal of Nursing Studies*, 42: 37-46.

Defloor, T., and M. F. H. Grypdonck. 2005. 'Pressure ulcers: validation of two risk assessment scales', *Journal of Clinical Nursing*, 14: 373-82.

Delmi, M., C. H. Rapin, J. M. Bengoa, P. D. Delmas, H. Vasey, and J. P. Bonjour. 1990. 'Dietary Supplementation in Elderly Patients with Fractured Neck of the Femur', *Lancet*, 335: 1013-16.

Demarre, L., D. Beeckman, K. Vanderwee, T. Defloor, M. Grypdonck, and S. Verhaeghe. 2012. 'Multi-stage versus single-stage inflation and deflation cycle for alternating low pressure air mattresses to prevent pressure ulcers in hospitalised patients: a randomised-controlled clinical trial', *Int J Nurs Stud*, 49: 416-26.

Dennis, M., S. C. Lewis, C. Warlow, and FOOD Trial Collaboration. 2005. 'Routine oral nutritional supplementation for stroke patients in hospital (FOOD): a multicentre randomised controlled trial', *Lancet*, 365: 755-63.

Diaz-Valenzuela, A., F. P. Garcia-Fernandez, P. J. C. Fernande, M. J. V. Canete, and P. L. Pancorbo-Hidalgo. 2019. 'Effectiveness and safety of olive oil preparation for topical use in pressure ulcer prevention: Multicentre, controlled, randomised, and double-blinded clinical trial', *International Wound Journal*, 16: 1314-22.

Donnelly, J., J. Winder, W. G. Kernohan, and M. Stevenson. 2011. 'An RCT to determine the effect of a heel elevation device in pressure ulcer prevention post-hip fracture', *Journal of Wound Care*, 20: 309-+.

Dunlop, Wince %J Advances in Skin, and Wound Care. 1998. 'Preliminary results of a randomized, controlled study of a pressure ulcer prevention system', 11: 14.

Dutra, R. A. A., G. M. Salome, J. R. Alves, V. O. S. Pereira, F. D. Miranda, V. B. Vallim, M. J. A. de Brito, and L. M. Ferreira. 2015. 'Using transparent polyurethane film and hydrocolloid dressings to prevent pressure ulcers', *Journal of Wound Care*, 24: 268-75.

Economides, N. G., V. A. Skoutakis, C. A. Carter, and V. H. Smith. 1995. 'Evaluation of the effectiveness of two support surfaces following myocutaneous flap surgery', *Adv Wound Care*, 8: 49-53.

Ek, A. C., M. Unosson, J. Larsson, H. Von Schenck, and P. Bjurulf. 1991. 'The development and healing of pressure sores related to the nutritional state', *Clin Nutr*, 10: 245-50.

Exton-Smith, A. N., P. W. Overstall, J. Wedgwood, and G. Wallace. 1982. 'Use of the 'air wave system' to prevent pressure sores in hospital', *Lancet*, 1: 1288-90.

Feuchtinger, J., R. de Bie, T. Dassen, and R. Halfens. 2006. 'A 4-cm thermoactive viscoelastic foam pad on the operating room table to prevent pressure ulcer during cardiac surgery', *Journal of Clinical Nursing*, 15: 162-7.

Finnegan, M. J., L. Gazzerro, J. O. Finnegan, and P. Lo. 2008. 'Comparing the effectiveness of a specialized alternating air pressure mattress replacement system and an air-fluidized integrated bed in the management of post-operative flap patients: a randomized controlled pilot study', *J Tissue Viability*, 17: 2-9.

Forni, C., F. D'Alessandro, P. Gallerani, R. Genco, A. Bolzon, C. Bombino, S. Mini, L. Rocchegiani, T. Notarnicola, A. Vitulli, A. Amodeo, G. Celli, and P. Taddia. 2018. 'Effectiveness of using a new polyurethane foam multi-layer dressing in the sacral area to prevent the onset of pressure ulcer in the elderly with hip fractures: A pragmatic randomised controlled trial', *International Wound Journal*, 15: 383-90.

Gebhardt, K. S., M. R. Bliss, P. L. Winwright, and J. Thomas. 1996. 'Pressure-relieving supports in an ICU', *J Wound Care*, 5: 116-21.

Gentilello, L., D. A. Thompson, A. S. Tonnesen, D. Hernandez, A. S. Kapadia, S. J. Allen, B. A. Houtchens, and M. E. Miner. 1988. 'Effect of a rotating bed on the incidence of pulmonary complications in critically ill patients', *Crit Care Med*, 16: 783-6.

Geyer, M. J., D. M. Brienza, P. Karg, E. Trefler, and S. Kelsey. 2001. 'A randomized control trial to evaluate pressure-reducing seat cushions for elderly wheelchair users', *Adv Skin Wound Care*, 14: 120-9; quiz 31-2.

Ghezeljeh, Tahereh Najafi, Leila Kalhor, MOGHADAM OMID MORADI, LAHIJI MOHAMMAD NIYAKAN, and Hamid Haghani. 2017. 'The comparison of the effect of the head of bed elevation to 30 and 45 degreess on the incidence of ventilator associated pneumonia and the risk for pressure ulcers: A controlled randomized clinical trial'.

Gilcreast, D. M., J. B. Warren, L. H. Yoder, J. J. Clark, J. A. Wilson, and M. Z. Mays. 2005. 'Research comparing three heel ulcer-prevention devices', *J Wound Ostomy Continence Nurs*, 32: 112-20.

Goldstone, L. A., M. Norris, M. O'Reilly, and J. White. 1982. 'A clinical trial of a bead bed system for the prevention of pressure sores in elderly orthopaedic patients', *J Adv Nurs*, 7: 545-8.

Gray, D. G., and M. Smith. 2000. 'Comparison of a new foam mattress with the standard hospital mattress', *J Wound Care*, 9: 29-31.

Gray, David, Pam Cooper, Melvyn Bertram, Kirsten Duguid, and Gail %J Wounds UK Pirie. 2008. 'A clinical audit of the Softform Premier Active™ mattress in two acute care of the elderly wards', 4: 124-8.

Gray, David G, Pamela J Cooper, and Marion %J Journal of Tissue Viability Campbell. 1998. 'A study of the performance of a pressure reducing foam mattress after three years of use', 8: 9-13.

Grindley, A., and J. Acres. 1996. 'Alternating pressure mattresses: comfort and quality of sleep', *Br J Nurs*, 5: 1303-10.

Guihan, M., C. H. Bombardier, D. M. Ehde, L. M. Rapacki, T. J. Rogers, B. Bates-Jensen, F. P. Thomas, R. Parachuri, and S. A. Holmes. 2014. 'Comparing multicomponent interventions to improve skin care behaviors and prevent recurrence in veterans hospitalized for severe pressure ulcers', *Arch Phys Med Rehabil*, 95: 1246-53 e3.

Gunningberg, L., C. Lindholm, M. Carlsson, and P. O. Sjoden. 2000. 'Effect of visco-elastic foam mattresses on the development of pressure ulcers in patients with hip fractures', *J Wound Care*, 9: 455-60.

Hampton, S. 1997. 'Evaluation of the new Cairwave Therapy System in one hospital trust', *Br J Nurs*, 6: 167-70.

Hartgrink, H. H., J. Wille, P. Konig, J. Hermans, and P. J. Breslau. 1998. 'Pressure sores and tube feeding in patients with a fracture of the hip: a randomized clinical trial', *Clin Nutr*, 17: 287-92.

Hofman, A., R. H. Geelkerken, J. Wille, J. J. Hamming, J. Hermans, and P. J. Breslau. 1994. 'Pressure sores and pressure-decreasing mattresses: controlled clinical trial', *Lancet*, 343: 568-71.

Hoshowsky, V. M., and C. A. Schramm. 1994. 'Intraoperative pressure sore prevention: an analysis of bedding materials', *Res Nurs Health*, 17: 333-9.

Houwing, R. H., M. Rozendaal, W. Wouters-Wesseling, J. W. Beulens, E. Buskens, and J. R. Haalboom. 2003. 'A randomised, double-blind assessment of the effect of nutritional supplementation on the prevention of pressure ulcers in hip-fracture patients', *Clin Nutr*, 22: 401-5.

Houwing, Ronald, Wil Van der Zwet, Sweder van Asbeck, Ruud Halfens, Willem %J Wounds: a compendium of clinical research Arends, and practice. 2008. ' An Unexpected Detrimental Effect on the Incidence of Heel Pressure Ulcers After Local 5% DMSO Cream Application: A Randomized, Double-blind Study in Patients at Risk for Pressure Ulcers', 20: 84-88.

Inman, K. J., W. J. Sibbald, F. S. Rutledge, and B. J. Clark. 1993. 'Clinical utility and cost-effectiveness of an air suspension bed in the prevention of pressure ulcers', *JAMA*, 269: 1139-43.

Jiang, Q. X., X. H. Li, A. Q. Zhang, Y. X. Guo, Y. H. Liu, H. Y. Liu, X. L. Qu, Y. J. Zhu, X. J. Guo, L. Liu, L. Y. Zhang, S. P. Bo, J. Jia, Y. J. Chen, R. Zhang, and J. D. Wang. 2014. 'Multicenter comparison of the efficacy on prevention of pressure ulcer in postoperative patients between two types of pressure-relieving mattresses in China', *International Journal of Clinical and Experimental Medicine*, 7: 2820-27.

Jolley, D. J., R. Wright, S. McGowan, M. B. Hickey, D. A. Campbell, R. D. Sinclair, and K. C. Montgomery. 2004. 'Preventing pressure ulcers with the Australian Medical Sheepskin: an open-label randomised controlled trial', *Med J Aust*, 180: 324-7.

Kalowes, P., V. Messina, and M. Li. 2016. 'Five-Layered Soft Silicone Foam Dressing to Prevent Pressure Ulcers in the Intensive Care Unit', *American Journal of Critical Care*, 25: E108-E19.

Karimi, Fateme, Fariba Yaghoubinia, Aliakbar Keykhah, and Hassan %J Medical-Surgical Nursing Journal Askari. 2018. 'Investigating the effect of home-based training for family caregivers on the incidence of bedsore in patients with stroke in Ali Ebne Abitaleb Hospital, Zahedan, Iran: a clinical trial study', 7.

Kathirvel, Soundappan, Sukhpal Kaur, Mandeep Singh Dhillon, Amarjeet %J Journal of Family Medicine Singh, and Primary Care. 2021. 'Impact of structured educational interventions on the prevention of pressure ulcers in immobile orthopedic patients in India: A pragmatic randomized controlled trial', 10: 1267.

Kaur, S., A. Singh, M. K. Tewari, and T. Kaur. 2018. 'Comparison of Two Intervention Strategies on Prevention of Bedsores among the Bedridden Patients: A Quasi Experimental Community-based Trial', *Indian Journal of Palliative Care*, 24: 28-34.

Kemp, M. G., D. Kopanke, L. Tordecilla, L. Fogg, S. Shott, V. Matthiesen, and B. Johnson. 1993. 'The role of support surfaces and patient attributes in preventing pressure ulcers in elderly patients', *Res Nurs Health*, 16: 89-96.

Keogh, A., and C. Dealey. 2001. 'Profiling beds versus standard hospital beds: effects on pressure ulcer incidence outcomes', *J Wound Care*, 10: 15-9.

Lim, R, R Sirett, TA Conine, D %J Journal of rehabilitation research Daechsel, and development. 1988. 'Clinical trial of foam cushions in the prevention of decubitis ulcers in elderly patients', 25: 19-26.

Lupianez-Perez, I., S. K. Uttumchandani, J. Morilla-Herrera, F. J. Martin-Santos, M. C. Fernandez-Gallego, F. J. Navarro-Moya, Y. Lupianez-Perez, E. Contreras-Fernandez, and J. M. Morales-Asencio. 2015. 'Topical Olive Oil Is Not Inferior to Hyperoxygenated Fatty Aids to Prevent Pressure Ulcers in High-Risk Immobilised Patients in Home Care. Results of a Multicentre Randomised Triple-Blind Controlled Non-Inferiority Trial', *Plos One*, 10.

Malbrain, M., B. Hendriks, P. Wijnands, D. Denie, A. Jans, J. Vanpellicom, and B. De Keulenaer. 2010. 'A pilot randomised controlled trial comparing reactive air and active alternating pressure mattresses in the prevention and treatment of pressure ulcers among medical ICU patients', *J Tissue Viability*, 19: 7-15.

Manzano, F., M. Colmenero, A. M. Perez-Perez, D. Roldan, M. Jimenez-Quintana Mdel, M. R. Manas, M. A. Sanchez-Moya, C. Guerrero, M. A. Moral-Marfil, E. Sanchez-Cantalejo, and E. Fernandez-Mondejar. 2014. 'Comparison of two repositioning schedules for the prevention of pressure ulcers in patients on mechanical ventilation with alternating pressure air mattresses', *Intensive Care Med*, 40: 1679-87.

McGowan, S; Montgomery, K; Jolley, D; Robyn,W. 2000. 'The role of sheepskins in preventing pressure ulcers in elderly orthopaedic patients'.

Mistiaen, P., W. Achterberg, A. Ament, R. Halfens, J. Huizinga, K. Montgomery, H. Post, P. Spreeuwenberg, and A. L. Francke. 2010. 'The effectiveness of the Australian Medical Sheepskin for the prevention of pressure ulcers in somatic nursing home patients: a prospective multicenter randomized-controlled trial (ISRCTN17553857)', *Wound Repair Regen*, 18: 572-9.

Moore, Zena, Seamus Cowman, and Ronán M %J Journal of clinical nursing Conroy. 2011. 'A randomised controlled clinical trial of repositioning, using the 30 tilt, for the prevention of pressure ulcers', 20: 2633-44.

Nakagami, G., H. Sanada, C. Konya, A. Kitagawa, E. Tadaka, and Y. Matsuyama. 2007. 'Evaluation of a new pressure ulcer preventive dressing containing ceramide 2 with low frictional outer layer', *J Adv Nurs*, 59: 520-9.

Nixon, J., S. Brown, I. L. Smith, E. McGinnis, A. Vargas-Palacios, E. A. Nelson, J. Brown, S. Coleman, H. Collier, C. Fernandez, R. Gilberts, V. Henderson, C. McCabe, D. Muir, C. Rutherford, N. Stubbs, B. Thorpe, K. Wallner, K. Walker, L. Wilson, and C. Hulme. 2019. 'Comparing alternating pressure mattresses and high-specification foam mattresses to prevent pressure ulcers in high-risk patients: the PRESSURE 2 RCT', *Health Technol Assess*, 23: 1-176.

Nixon, J., D. McElvenny, S. Mason, J. Brown, and S. Bond. 1998. 'A sequential randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores', *Int J Nurs Stud*, 35: 193-203.

Nixon, J., E. A. Nelson, G. Cranny, C. P. Iglesias, K. Hawkins, N. A. Cullum, A. Phillips, K. Spilsbury, D. J. Torgerson, S. Mason, and Pressure Trial Group. 2006. 'Pressure relieving support surfaces: a randomised evaluation', *Health Technol Assess*, 10: iii-iv, ix-x, 1-163.

Olofsson, B., M. Stenvall, M. Lundstrom, O. Svensson, and Y. Gustafson. 2007. 'Malnutrition in hip fracture patients: an intervention study', *Journal of Clinical Nursing*, 16: 2027-38.

Otero, D. P., D. V. Dominguez, L. H. Fernandez, A. S. Magarino, V. J. Gonzalez, J. V. Klepzing, and J. V. Montesinos. 2017. 'Preventing facial pressure ulcers in patients under non-invasive mechanical ventilation: a randomised control trial', *J Wound Care*, 26: 128-36.

Pickham, D., N. Berte, M. Pihulic, A. Valdez, B. Mayer, and M. Desai. 2018. 'Effect of a wearable patient sensor on care delivery for preventing pressure injuries in acutely ill adults: A pragmatic randomized clinical trial (LS-HAPI study)', *Int J Nurs Stud*, 80: 12-19.

Price, P., S. Bale, R. Newcombe, and K. Harding. 1999. 'Challenging the pressure sore paradigm', *J Wound Care*, 8: 187-90.

Rafter, L. 2011. 'Evaluation of patient outcomes: pressure ulcer prevention mattresses', *Br J Nurs*, 20: S32, S34-8.

Rantz, M. J., M. Zwygart-Stauffacher, L. Hicks, D. Mehr, M. Flesner, G. F. Petroski, R. W. Madsen, and J. Scott-Cawiezell. 2012. 'Randomized multilevel intervention to improve outcomes of residents in nursing homes in need of improvement', *J Am Med Dir Assoc*, 13: 60-8.

Ricci, Ellia, Cassino Roberto, Annamaria Ippolito, Andrea Bianco, and Maria Teresa %J EWMA journal Scalise. 2013. 'A NEW PRESSURE-RELIEVING MATTRESS OVERLAY', 13.

Rintala, D. H., S. L. Garber, J. D. Friedman, and S. A. Holmes. 2008. 'Preventing recurrent pressure ulcers in veterans with spinal cord injury: impact of a structured education and follow-up intervention', *Arch Phys Med Rehabil*, 89: 1429-41.

Russell, Linda J, Tim M Reynolds, Carol Park, Shyam Rithalia, M Gonsalkorale, Jan Birch, David Torgerson, Cynthia Iglesias, PPUS-1 Study Group %J Advances in skin, and wound care. 2003. 'Randomized clinical trial comparing 2 support surfaces: results of the Prevention of Pressure Ulcers Study', 16: 317-27.

Saleh, M., D. Anthony, and S. Parboteeah. 2009. 'The impact of pressure ulcer risk assessment on patient outcomes among hospitalised patients', *Journal of Clinical Nursing*, 18: 1923-9.

Sanada, H., J. Sugama, Y. Matsui, C. Konya, A. Kitagawa, M. Okuwa, and S. Omote. 2003. 'Randomised controlled trial to evaluate a new double-layer air-cell overlay for elderly patients requiring head elevation', *J Tissue Viability*, 13: 112-4, 16, 18 passim.

Santamaria, N., M. Gerdtz, S. Sage, J. McCann, A. Freeman, T. Vassiliou, S. De Vincentis, A. W. Ng, E. Manias, W. Liu, and J. Knott. 2015. 'A randomised controlled trial of the effectiveness of soft silicone multi-layered foam dressings in the prevention of sacral and heel pressure ulcers in trauma and critically ill patients: the border trial', *International Wound Journal*, 12: 302-8.

Sauvage, P., M. Touflet, C. Pradere, F. Portalier, J. M. Michel, P. Charru, Y. Passadori, R. Fevrier, A. M. Hallet-Lezy, F. Beauchene, and B. Scherrer. 2017. 'Pressure ulcers prevention efficacy of an alternating pressure air mattress in elderly patients: E(2)MAO a randomised study', *J Wound Care*, 26: 304-12.

Schultz, A., M. Bien, K. Dumond, K. Brown, and A. Myers. 1999. 'Etiology and incidence of pressure ulcers in surgical patients', *AORN J*, 70: 434, 37-40, 43-9.

Sideranko, S., A. Quinn, K. Burns, and R. D. Froman. 1992. 'Effects of position and mattress overlay on sacral and heel pressures in a clinical population', *Res Nurs Health*, 15: 245-51.

Summer, Warren R, Phyllis Curry, Edward F Haponik, Steve Nelson, and Robert %J Journal of Critical Care Elston. 1989. 'Continuous mechanical turning of intensive care unit patients shortens length of stay in some diagnostic-related groups', 4: 45-53.

Taylor, L. 1999. 'Evaluating the Pegasus Trinova: a data hierarchy approach', *Br J Nurs*, 8: 771-4, 76-8.

Theaker, C., M. Kuper, and N. Soni. 2005. 'Pressure ulcer prevention in intensive care - a randomised control trial of two pressure-relieving devices', *Anaesthesia*, 60: 395-9.

Theilla, M., P. Singer, J. Cohen, and F. Dekeyser. 2007. 'A diet enriched in eicosapentanoic acid, gamma-linolenic acid and antioxidants in the prevention of new pressure ulcer formation in critically ill patients with acute lung injury: A randomized, prospective, controlled study', *Clin Nutr*, 26: 752-7.

Torra i Bou, J. E., T. Segovia Gomez, J. Verdu Soriano, A. Nolasco Bonmati, J. Rueda Lopez, and M. Arboix i Perejamo. 2005. 'The effectiveness of a hyperoxygenated fatty acid compound in preventing pressure ulcers', *J Wound Care*, 14: 117-21.

Tymec, A. C., B. Pieper, and K. Vollman. 1997. 'A comparison of two pressure-relieving devices on the prevention of heel pressure ulcers', *Adv Wound Care*, 10: 39-44.

van Leen, M., R. Halfens, and J. Schols. 2018. 'Preventive Effect of a Microclimate-Regulating System on Pressure Ulcer Development: A Prospective, Randomized Controlled Trial in Dutch Nursing Homes', *Adv Skin Wound Care*, 31: 1-5.

van Leen, M., S. Hovius, R. Halfens, J. Neyens, and J. Schols. 2013. 'Pressure relief with visco-elastic foam or with combined static air overlay? A prospective, crossover randomized clinical trial in a dutch nursing home', *Wounds*, 25: 287-92.

van Leen, M., S. Hovius, J. Neyens, R. Halfens, and J. Schols. 2011. 'Pressure relief, cold foam or static air? A single center, prospective, controlled randomized clinical trial in a Dutch nursing home', *J Tissue Viability*, 20: 30-4.

Vanderwee, K., M. H. Grypdonck, D. De Bacquer, and T. Defloor. 2007. 'Effectiveness of turning with unequal time intervals on the incidence of pressure ulcer lesions', *J Adv Nurs*, 57: 59-68.

Vanderwee, K., M. H. Grypdonck, and T. Defloor. 2005. 'Effectiveness of an alternating pressure air mattress for the prevention of pressure ulcers', *Age Ageing*, 34: 261-7.

Vanderwee, Katrien, Maria Grypdonck, Dirk De Bacquer, and Tom %J Journal of clinical nursing Defloor. 2009. 'The identification of older nursing home residents vulnerable for deterioration of grade 1 pressure ulcers', 18: 3050-58.

Vermette, S., I. Reeves, and J. Lemaire. 2012. 'Cost effectiveness of an air-inflated static overlay for pressure ulcer prevention: a randomized, controlled trial', *Wounds*, 24: 207-14.

Vyhlidal, S. K., D. Moxness, K. S. Bosak, F. G. Van Meter, and N. Bergstrom. 1997. 'Mattress replacement or foam overlay? A prospective study on the incidence of pressure ulcers', *Appl Nurs Res*, 10: 111-20.

Ward, C. D., G. Turpin, M. E. Dewey, S. Fleming, B. Hurwitz, S. Ratib, M. von Fragstein, and M. Lymbery. 2004. 'Education for people with progressive neurological conditions can have negative effects: evidence from a randomized controlled trial', *Clin Rehabil*, 18: 717-25.

Webster, J., K. Coleman, A. Mudge, L. Marquart, G. Gardner, M. Stankiewicz, J. Kirby, C. Vellacott, M. Horton-Breshears, and A. McClymont. 2011. 'Pressure ulcers: effectiveness of risk-assessment tools. A randomised controlled trial (the ULCER trial)', *BMJ Qual Saf*, 20: 297-306.

Young, T. 2004. 'The 30 degree tilt position vs the 90 degree lateral and supine positions in reducing the incidence of non-blanching erythema in a hospital inpatient population: a randomised controlled trial', *J Tissue Viability*, 14: 88, 90, 92-6.

Appendix B

Chapter 4: Multi-state models

Table B.1: Observed state occupancies for all skin sites in the PRESSURE dataset, by skinsite

Sacrum				
From	**To**			
	1	2	3	4
1	1,763	233	19	9
2	0	1,627	95	17
3	0	0	639	25

Left Buttock				
From	**To**			
	1	2	3	4
1	1,862	254	29	4
2	0	1,614	92	20
3	0	0	585	27

Right buttock				
From	**To**			
	1	2	3	4
1	1,883	267	20	6
2	0	1,630	90	18
3	0	0	555	35

Left Hip				
From	**To**			
	1	2	3	4
1	3,873	28	1	1
2	0	146	2	1
3	0	0	9	0

Right hip				
From	**To**			
	1	2	3	4
1	3,921	34	3	1
2	0	161	3	1
3	0	0	20	0

Left heel				
From	**To**			
	1	2	3	4
1	1,481	323	26	1
2	0	2,149	101	6
3	0	0	986	15

Right heel				
From	**To**			
	1	2	3	4
1	1,478	326	19	0
2	0	2,190	93	8
3	0	0	955	10

Table B.2: Observed occupancies for all skin sites in the PRESSURE2 dataset, by skinsite

Sacrum				
From	**To**			
	1	2	3	4
1	3,044	352	16	4
2	0	3,504	63	20
3	0	0	552	12

Back				
From	**To**			
	1	2	3	4
1	6,520	151	3	2
2	0	1,402	3	3
3	0	0	39	1

Left buttock				
From	**To**			
	1	2	3	4
1	2,980	309	22	7
2	0	3,701	49	17
3	0	0	459	14

Right buttock				
From	**To**			
	1	2	3	4
1	3,093	328	17	5
2	0	3,709	51	16
3	0	0	419	13

Left ischial				
From	**To**			
	1	2	3	4
1	6,088	272	2	2
2	0	1,755	5	1
3	0	0	38	1

Right ischial				
From	**To**			
	1	2	3	4
1	6,092	251	2	3
2	0	1,746	3	1
3	0	0	53	0

Left hip				
From	**To**			
	1	2	3	4
1	7,144	79	1	2
2	0	889	0	0
3	0	0	7	0

Right hip				
From	**To**			
	1	2	3	4
1	6,987	94	2	1
2	0	1,010	1	0
3	0	0	15	0

Left heel				
From	**To**			
	1	2	3	4
1	2,097	345	11	5
2	0	5,138	34	11
3	0	0	250	4

Right heel				
From	**To**			
	1	2	3	4
1	2,062	345	7	6
2	0	5,134	24	9
3	0	0	248	3

Left ankle				
From	**To**			
	1	2	3	4
1	3,921	315	3	0
2	0	3,620	2	0
3	0	0	21	0

Right ankle				
From	**To**			
	1	2	3	4
1	3,853	307	4	0
2	0	3,583	9	4
3	0	0	63	0

Left elbow				
From	**To**			
	1	2	3	4
1	4,381	356	4	0
2	0	3,513	5	2
3	0	0	65	0

Right elbow				
From	**To**			
	1	2	3	4
1	4,514	370	2	3
2	0	3,374	11	4
3	0	0	76	2

Appendix C

Chapter 5: Power and sample size requirements

C.1 Code

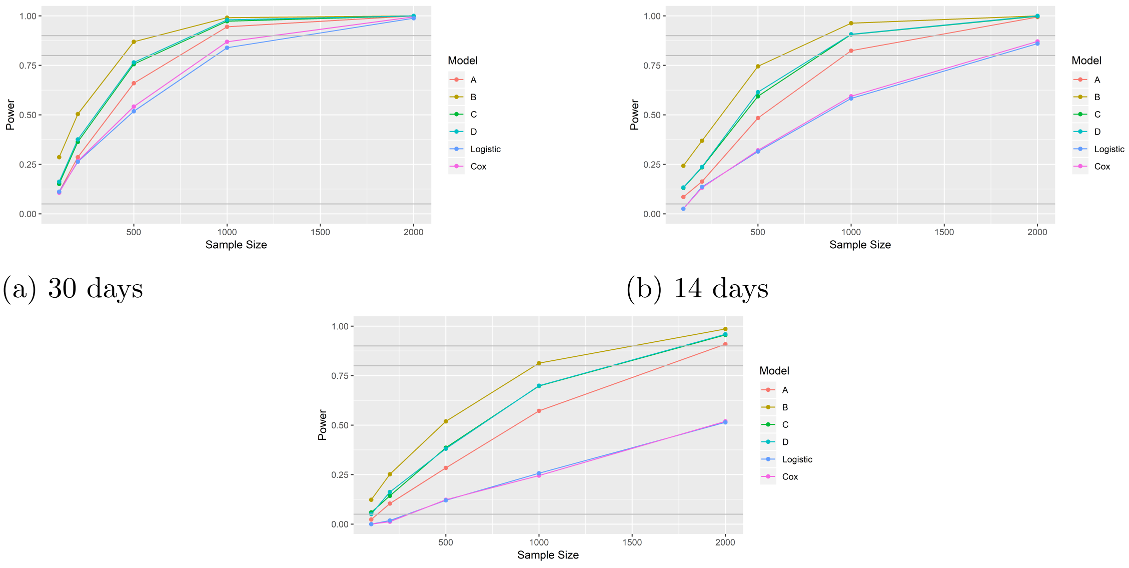
```
## Code for first simulation study from NIHR Doctoral Research Fellowship

## In this study I want to simulate data from a 4-state model with the
## following states:

## State 1 : Healthy
## State 2 : Altered
## State 3 : Category 1
## State 4 : Category 2+

## There is a maximum follow-up time specified in days, and patients may leave
## the system (be censored) at anytime.

## There will be 20 pre-specified inputs for each simulation scenario and
## these will be set up in an input matrix as descibed here:

## Column 1  = Seed
## Column 2  = Maximum length of follow-up (maxfup)
## Column 3  = Assessment frequencies e.g. 1=daily, 2=every other day (VisitFreq)
## Column 4  = The sample size (n=nc+ne)
## Column 5  = The number of simulations (N)
## Column 6  = Baseline transition rate from state 1 to state 2 (lambda0)
## Column 7  = Baseline transition rate from state 2 to state 3 (lambda0)
## Column 8  = Baseline transition rate from state 3 to state 4 (lambda0)
## Column 9  = hazard ratio corresponding to treatment effect on transition from
##             state 1 to state 2 (hr0)
## Column 10 = hazard ratio corresponding to treatment effect on transition from
##             state 2 to state 3 (hr0)
## Column 11 = hazard ratio corresponding to treatment effect on transition from
##             state 3 to state 4 (hr0)
## Column 12 = censoring rate from state 1 for control
## Column 13 = censoring rate from state 2 for control
## Column 14 = censoring rate from state 3 for control
## Column 15 = censoring rate from state 1 for intervention
## Column 16 = censoring rate from state 2 for intervention
## Column 17 = censoring rate from state 3 for intervention
## Column 18 = proportion of patients starting in state 1 (assumed the same for both
groups)
## Column 19 = proportion of patients starting in state 2 (assumed the same for both
groups)
## Column 20 = proportion of patients starting in state 3 (assumed the same for both
groups)

##Load required packages

library(survival)
library(plyr)
library(dplyr)
library(broom)
library(msm)
library(tictoc)

##################################################################################
##Firstly, this is a function to simulate transition times based on input matrix##
###
##################################################################################

simexp<-function(lambda0, hr0, censor0, censor1, nc, ne, maxfup,start1,start2,start3){

  #Determine starting states
  startc_0<-runif(nc,0,1)
  starte_0<-runif(ne,0,1)

  startc<-rep(1,nc)
  startc[startc_0<=start1]<-1
  startc[startc_0>start1 & startc_0<=1-start3]<-2
  startc[startc_0>1-start3]<-3
  starte<-rep(1,ne)
  starte[starte_0<=start1]<-1
  starte[starte_0>start1 & starte_0<=1-start3]<-2
  starte[starte_0>1-start3]<-3

  ##lambda0 is a vector of transition rates for control arm (c)

  ##lambda1 is a vector of transition rates for experimental arm (e)
  lambda1<-lambda0*hr0

  ##random generation of exponential times to transition out of no PU state

  t12c<-rexp(nc, lambda0[1])
  t12e<-rexp(ne, lambda1[1])

  ##random generation of exponential times to censoring from no PU state

  t199c<-rexp(nc, censor0[1])
  t199e<-rexp(ne, censor1[1])

  ##calculate observed transition times and other outcomes

  exittime1c<-pmin(t12c,t199c,maxfup)
  exittime1e<-pmin(t12e,t199e,maxfup)

  ##calculate how many days it took until first transition ie at which daily assessment
was the 1st transition

  exitday1c=ceiling(exittime1c)
  exitday1e=ceiling(exittime1e)
  exitday1c[startc>1]<-0
  exitday1e[starte>1]<-0

  exitstate1c<-rep(1, nc);
  exitstate1c[exittime1c==t12c]<-2
  exitstate1c[exittime1c==t199c]<-99
  exitstate1c[startc>1]<-9876
  #exitstate1c[startc==2|startc==3]<-2
  exitstate1e<-rep(1, ne);
  exitstate1e[exittime1e==t12e]<-2
  exitstate1e[exittime1e==t199e]<-99
  exitstate1e[starte>1]<-9876
  #exitstate1e[starte==2|starte==3]<-2

  ##random generation of exponential times to transition out of altered state

  t23c<-rexp(nc, lambda0[2])
  t23e<-rexp(ne, lambda1[2])

  ##random generation of exponential times to censoring from no PU state

  t299c<-rexp(nc, censor0[2])
  t299e<-rexp(ne, censor1[2])

  ##calculate observed transition times and other outcomes

  exittime2c<-pmin(t23c+exitday1c,t299c+exitday1c,maxfup)
  exittime2e<-pmin(t23e+exitday1e,t299e+exitday1e,maxfup)

  #exittime2c[t23c<exitday1c]<-maxfup #set to max follow-up as transition out of state
2 occurred before entry to state 2
  #exittime2e[t23e<exitday1e]<-maxfup #set to max follow-up as transition out of state
2 occurred before entry to state 2

  exittime2c[exitstate1c==1|exitstate1c==99]<-0 #set to zero as they have already
stopped follow-up
```

```
  exittime2e[exitstate1e==1|exitstate1e==99]<-0 #set to zero as they have already
stopped follow-up

  ##calculate how many days it took until second transition ie at which daily
assessment was the 2nd transition

  exitday2c<-round(ceiling(exittime2c))
  exitday2c[startc>2]<-0

  exitstate2c<-rep(2,length(exittime2c))
  exitstate2c[exittime2c==t23c+exitday1c]<-3
  exitstate2c[exittime2c==t299c+exitday1c]<-99
  exitstate2c[startc>2]<-9876
  # exitstate1c[startc==3]<3
  exitstate2c[exittime2c==0]<-NA

  exitday2e<-round(ceiling(exittime2e))
  exitday2e[starte>2]<-0

  exitstate2e<-rep(2,length(exittime2e))
  exitstate2e[exittime2e==t23e+exitday1e]<-3
  exitstate2e[exittime2e==t299e+exitday1e]<-99
  exitstate2e[starte>2]<-9876
  # exitstate1e[starte==3]<-3
  exitstate2e[exittime2e==0]<-NA

  ##random generation of exponential times to transition out of Category 1 state

  t34c<-rexp(nc, lambda0[3])
  t34e<-rexp(ne, lambda1[3])

  ##random generation of exponential times to censoring from Category 1 state

  t399c<-rexp(nc, censor0[3])
  t399e<-rexp(ne, censor1[3])

  ##calculate observed transition times and other outcomes

  exittime3c<-pmin(t34c+exitday2c,t399c+exitday2c,maxfup)
  exittime3e<-pmin(t34e+exitday2e,t399e+exitday2e,maxfup)

  #exittime3c[t34c<exitday2c|t34c<exitday1c]<-maxfup #set to max follow-up as
transition out of state 2 occurred before entry to state 2
  #exittime3e[t34e<exitday2e|t34e<exitday1e]<-maxfup #set to max follow-up as
transition out of state 2 occurred before entry to state 2

  exittime3c[exitstate1c==1|exitstate1c==99|exitstate2c==2|exitstate2c==99]<-0 #set to
zero as they have already stopped follow-up
  exittime3e[exitstate1e==1|exitstate1e==99|exitstate2e==2|exitstate2e==99]<-0 #set to
zero as they have already stopped follow-up

  ##calculate how many days it took until third transition ie at which daily assessment
was the 3rd transition

  exitday3c<-round(ceiling(exittime3c))

  exitstate3c<-rep(3,length(exittime3c))
  exitstate3c[exittime3c==t34c+exitday2c]<-4
  exitstate3c[exittime3c==t399c+exitday2c]<-99
  exitstate3c[exittime3c==0]<-NA

  exitday3e<-round(ceiling(exittime3e))

  exitstate3e<-rep(3,length(exittime3e))
  exitstate3e[exittime3e==t34e+exitday2e]<-4
  exitstate3e[exittime3e==t399e+exitday2e]<-99
  exitstate3e[exittime3e==0]<-NA
  ##create data for binary outcomes

  PU2c<-rep(0,nc)
  PU2c[exitstate3c==4]<-1

  PU2e<-rep(0,ne)
  PU2e[exitstate3e==4]<-1

  ##Now create data for the TTE outcomes

  PUtime2c<-pmax(exitday1c,exitday2c,exitday3c)
  PUtime2e<-pmax(exitday1e,exitday2e,exitday3e)

  groupc<-rep(0,nc)
  groupe<-rep(1,ne)

  tempc<-
cbind(groupc,startc,t12c,t23c,t34c,t199c,t299c,t399c,exittime1c,exitday1c,exitstate1c,e
xittime2c,exitday2c,exitstate2c,exittime3c,exitday3c,exitstate3c,maxfup,PU2c,PUtime2c)
  tempe<-
cbind(groupe,starte,t12e,t23e,t34e,t199e,t299e,t399e,exittime1e,exitday1e,exitstate1e,e
xittime2e,exitday2e,exitstate2e,exittime3e,exitday3e,exitstate3e,maxfup,PU2e,PUtime2e)

  array.colnames<-
sapply(strsplit(colnames(tempc),split='c',fixed=TRUE),function(x)(x[1]))

  array.test<-array(cbind(tempc,tempe),
                    dim=c(ne,ncol(tempc),2),
                    dimnames=list(NULL,array.colnames,c("control","Experimental")))

  dataset.test<-rbind(as.data.frame(array.test[,,1]),as.data.frame(array.test[,,2]))

  dataset.test<-cbind(Patnum=seq(1:(nc+ne)),dataset.test)

  return(dataset.test)

}

##################################################################################
## Next, we build a function that uses the simexp function and translates the   ##
## simulated transition times into a multi-state data frame. The function then  ##
## analyses the multi-state data and the time to event/binary data and stores   ##
## the estimands in an output matrix.                                           ##
##################################################################################

simanalyse<-function(inputmat){
  modelout2=list()
  for (k in 1:nrow(inputmat)){
    set.seed(inputmat[k,1])

    ##Set maximum follow-up
    maxfup<-inputmat[k,2]

    ##Set follow-up schedule
    fup<-inputmat[k,3]

    ##set total sample sizes n to be considered
    n<-inputmat[k,4]

    ##set number of simulations N

    N<-inputmat[k,5]

    #lambda0
    lambda<-c(inputmat[k,6],inputmat[k,7],inputmat[k,8])
    #h0
    h<-c(inputmat[k,9],inputmat[k,10],inputmat[k,11])

    #censor0
    censor0<-c(inputmat[k,12],inputmat[k,13],inputmat[k,14])

    #censor1
    censor1<-c(inputmat[k,15],inputmat[k,16],inputmat[k,17])

    #start1
    start1<-inputmat[k,18]
    #start2
    start2<-inputmat[k,19]
    #start3
    start3<-inputmat[k,20]

    #Simulate dataset of transition times
    data.simul=list()
    for (i in 1:N){data.simul[[i]]<-simexp(lambda,h, censor0,censor1, n/2, n/2,
maxfup,start1,start2,start3)}
    #data.simul is a list of N datasets with each one containing n observations as
expected

    #Re-format the dataset to be structured like a multi-state model
    data.format.fct<-function(patid,data){

      pat.1<-list()

      pat.1$id<-data[patid,"Patnum"]
      pat.1$grp<-data[patid,"group"]

      ndays<-data[pat.1$id,"maxfup"]

      mat<-matrix(NA,nrow=ndays+1,ncol=5)

      colnames(mat)<-c("PatID","AssessDay","State","Group","delete")
      mat[,1]<-pat.1$id
      mat[,2]<-0:ndays

      if (data[pat.1$id,"start"]==1){
        mat[,3]<-c(1,rep(1,data[pat.1$id,"exitday1"]-
1),rep(data[pat.1$id,"exitstate1"],ndays-data[pat.1$id,"exitday1"]+1))

        if
(data[pat.1$id,"exittime2"]!=0){mat[c(data[pat.1$id,"exitday2"]:ndays+1),3]<-
data[pat.1$id,"exitstate2"]}

        if
(data[pat.1$id,"exittime3"]!=0){mat[c(data[pat.1$id,"exitday3"]:ndays+1),3]<-
data[pat.1$id,"exitstate3"]}
      } else if (data[pat.1$id,"start"]==2){
        mat[,3]<-c(2,rep(2,data[pat.1$id,"exitday2"]-
1),rep(data[pat.1$id,"exitstate2"],ndays-data[pat.1$id,"exitday2"]+1))

        if
(data[pat.1$id,"exittime3"]!=0){mat[c(data[pat.1$id,"exitday3"]:ndays+1),3]<-
data[pat.1$id,"exitstate3"]}
      }else if (data[pat.1$id,"start"]==3){
        mat[,3]<-c(3,rep(3,data[pat.1$id,"exitday3"]-
1),rep(data[pat.1$id,"exitstate3"],ndays-data[pat.1$id,"exitday3"]+1))}

      mat[,4]<-pat.1$grp

      if(fup!=1){pat.1$df<-as.data.frame(mat) %>%
        filter(row_number() %% fup==1) %>%
        mutate(delete=lag(State), order_by=AssessDay) %>%
        filter(State<=99 & (delete<4|is.na(delete)))}
      else {pat.1$df<-as.data.frame(mat) %>%
        mutate(delete=lag(State), order_by=AssessDay) %>%
        filter(State<=99 & (delete<4|is.na(delete)))}
      return(pat.1$df)
    }

    datasimul.MSM=list()

    for (i in 1:N){datasimul.MSM[[i]]<-
do.call("rbind",lapply(1:(max(data.simul[[i]][1])),function(x)
      data.format.fct(x,data=data.simul[[i]])))}


    #Re-format the dataset to obtain time to event and binary outcomes
    datasimul.BINCOX<-list()

    for(i in 1:N){datasimul.BINCOX[[i]]<-datasimul.MSM[[i]] %>%
      mutate(delete=lead(PatID), order_by=AssessDay) %>%
      filter(PatID!=delete|is.na(delete)) %>%
      mutate(PU2=ifelse(State==4,1,0)) %>%
      mutate(PUtime2=AssessDay)}

    ##NOW ANALYSE DATA

    modelout=list()
    for (i in 1:N){
    ## Fit a logistic model for the binary outcome for each simulated dataset
     logisticmod<-
glm(PU2~1+Group,data=datasimul.BINCOX[[i]],family=binomial(link="logit"))
     logisticout<-as.data.frame(coef(summary(logisticmod)))
     logisticout<-subset(logisticout, rownames(logisticout) %in% "Group",
colnames(logisticout) %in% "Pr(>|z|)")
     logisticout2<-as.data.frame(exp(cbind(OR = coef(logisticmod),
confint.default(logisticmod))))
     logisticout2<-subset(logisticout2, rownames(logisticout2) %in% "Group")
     logisticout$model<-"Logistic"
     logisticout2$model<-"Logistic"
     logistic_or<-merge(logisticout,logisticout2,by=c("model"))
     colnames(logistic_or)[colnames(logistic_or)=="2.5 %"]<-"Lower95"
     colnames(logistic_or)[colnames(logistic_or)=="97.5 %"]<-"Upper95"
     colnames(logistic_or)[colnames(logistic_or)=="OR"]<-"Estimate"
     colnames(logistic_or)[colnames(logistic_or)=="Pr(>|z|)"]<-"Pvalue"

    ## Fit a Cox model for the TTE outcome for each simulated dataset

    coxmod<-coxph(Surv(PUtime2,PU2)~Group, data=datasimul.BINCOX[[i]])
    if (is.na(coef(coxmod))){coxout$Estimate<-NA
                             coxout$Lower95<-NA
                             coxout$Upper95<-NA
                             coxout$Pvalue<-NA
                             }
    {coxout<-coxmod %>%
        tidy %>%
        mutate(
          Estimate=exp(estimate),
          Lower95=exp(conf.low),
          Upper95=exp(conf.high),
          Pvalue=p.value
        ) %>%
        filter(term=="Group") %>%
        select(Estimate, Lower95, Upper95, Pvalue)}

    coxout$model<-"Cox"

    ## Now let's look at multi-state models

    #Set up Q matrix based on input matrix
    Q<-rbind(c(-lambda[1],lambda[1],0,0),
             c(0,-lambda[2],lambda[2],0),
             c(0,0,-lambda[3],lambda[3]),
             c(0,0,0,0))


    msmdata<-datasimul.MSM[[i]]

    #unconstrained model
    msmmod_uncon_hr<-data.frame(model=as.character(c("msm_con")),
                              transition=as.factor(c("State 1 - State 2","State 2 -
State 3","State 3 - State 4")),
                              Estimate=as.double(NA,NA,NA))
    tryCatch({msmmod_uncon<-msm(State~AssessDay, subject=PatID, data=msmdata,
qmatrix=Q,
                                covariates=~Group, censor=99, censor.states=c(1,2,3))

    msmmod_uncon_95<-as.data.frame(hazard.msm(msmmod_uncon, cl=0.95))
    msmmod_uncon_95$model<-"msmmod_uncon"
    msmmod_uncon_95$transition<-as.factor(rownames(msmmod_uncon_95))

    colnames(msmmod_uncon_95)[colnames(msmmod_uncon_95)=="Group.L"]<-"Lower95"
    colnames(msmmod_uncon_95)[colnames(msmmod_uncon_95)=="Group.U"]<-"Upper95"
    colnames(msmmod_uncon_95)[colnames(msmmod_uncon_95)=="Group.HR"]<-"Estimate"

    msmmod_uncon_975<-as.data.frame(hazard.msm(msmmod_uncon, cl=0.975))
    msmmod_uncon_975$model<-"msmmod_uncon"
    msmmod_uncon_975$transition<-as.factor(rownames(msmmod_uncon_975))

    colnames(msmmod_uncon_975)[colnames(msmmod_uncon_975)=="Group.L"]<-"Lower975"
    colnames(msmmod_uncon_975)[colnames(msmmod_uncon_975)=="Group.U"]<-"Upper975"
    colnames(msmmod_uncon_975)[colnames(msmmod_uncon_975)=="Group.HR"]<-"Estimate"

    msmmod_uncon_9833<-as.data.frame(hazard.msm(msmmod_uncon, cl=0.9833))
    msmmod_uncon_9833$model<-"msmmod_uncon"
    msmmod_uncon_9833$transition<-as.factor(rownames(msmmod_uncon_9833))

    colnames(msmmod_uncon_9833)[colnames(msmmod_uncon_9833)=="Group.L"]<-"Lower9833"
    colnames(msmmod_uncon_9833)[colnames(msmmod_uncon_9833)=="Group.U"]<-"Upper9833"
    colnames(msmmod_uncon_9833)[colnames(msmmod_uncon_9833)=="Group.HR"]<-"Estimate"

    msmmod_uncon_hr_a<-
merge(msmmod_uncon_95,msmmod_uncon_975,by=c("model","transition","Estimate"))
    msmmod_uncon_hr<-
merge(msmmod_uncon_hr_a,msmmod_uncon_9833,by=c("model","transition","Estimate"))
    msmmod_uncon_hr$SE<-c(msmmod_uncon$QmatricesSE$Group[1,2],
                          msmmod_uncon$QmatricesSE$Group[2,3],
                          msmmod_uncon$QmatricesSE$Group[3,4])
    msmmod_uncon_hr$WaldTS<-log(msmmod_uncon_hr$Estimate)/msmmod_uncon_hr$SE

    }, error=function(e){})

    #Completely constrained model
    msmmod_con_hr<-data.frame(model=as.character(c("msm_con")),
                            transition=as.factor(c("State 1 - State 2","State 2 - State
3","State 3 - State 4")),
                            Estimate=as.double(NA,NA,NA))
    tryCatch({msmmod_con<-msm(State~AssessDay, subject=PatID, data=msmdata, qmatrix=Q,
                              covariates=~Group, censor=99, censor.states=c(1,2,3),
                              constraint = list(Group=c(1,1,1)))

    msmmod_con_95<-as.data.frame(hazard.msm(msmmod_con, cl=0.95))
    msmmod_con_95$model<-"msmmod_con"
    msmmod_con_95$transition<-as.factor(rownames(msmmod_con_95))

    colnames(msmmod_con_95)[colnames(msmmod_con_95)=="Group.L"]<-"Lower95"
    colnames(msmmod_con_95)[colnames(msmmod_con_95)=="Group.U"]<-"Upper95"
    colnames(msmmod_con_95)[colnames(msmmod_con_95)=="Group.HR"]<-"Estimate"

    msmmod_con_975<-as.data.frame(hazard.msm(msmmod_con, cl=0.975))
    msmmod_con_975$model<-"msmmod_con"
    msmmod_con_975$transition<-as.factor(rownames(msmmod_con_975))


    colnames(msmmod_con_975)[colnames(msmmod_con_975)=="Group.L"]<-"Lower975"
    colnames(msmmod_con_975)[colnames(msmmod_con_975)=="Group.U"]<-"Upper975"
    colnames(msmmod_con_975)[colnames(msmmod_con_975)=="Group.HR"]<-"Estimate"

    msmmod_con_9833<-as.data.frame(hazard.msm(msmmod_con, cl=0.9833))
    msmmod_con_9833$model<-"msmmod_con"
    msmmod_con_9833$transition<-as.factor(rownames(msmmod_con_9833))

    colnames(msmmod_con_9833)[colnames(msmmod_con_9833)=="Group.L"]<-"Lower9833"
    colnames(msmmod_con_9833)[colnames(msmmod_con_9833)=="Group.U"]<-"Upper9833"
    colnames(msmmod_con_9833)[colnames(msmmod_con_9833)=="Group.HR"]<-"Estimate"

    msmmod_con_hr_a<-
merge(msmmod_con_95,msmmod_con_975,by=c("model","transition","Estimate"))
    msmmod_con_hr<-
merge(msmmod_con_hr_a,msmmod_con_9833,by=c("model","transition","Estimate"))
    msmmod_con_hr$SE<-c(msmmod_con$QmatricesSE$Group[1,2],
                       msmmod_con$QmatricesSE$Group[2,3],
                       msmmod_con$QmatricesSE$Group[3,4])
    msmmod_con_hr$WaldTS<-log(msmmod_con_hr$Estimate)/msmmod_con_hr$SE

    }, error=function(e){})

    #Partially constrained model - beta12=beta23
    msmmod_123con_hr<-data.frame(model=as.character(c("msm_con")),
                              transition=as.factor(c("State 1 - State 2","State 2 -
State 3","State 3 - State 4")),
                              Estimate=as.double(NA,NA,NA))
    tryCatch({msmmod_123con<-msm(State~AssessDay, subject=PatID, data=msmdata,
qmatrix=Q,
                                covariates=~Group, censor=99, censor.states=c(1,2,3),
                                constraint = list(Group=c(1,1,2)))

    msmmod_123con_95<-as.data.frame(hazard.msm(msmmod_123con, cl=0.95))
    msmmod_123con_95$model<-"msmmod_123con"
    msmmod_123con_95$transition<-as.factor(rownames(msmmod_123con_95))

    colnames(msmmod_123con_95)[colnames(msmmod_123con_95)=="Group.L"]<-"Lower95"
    colnames(msmmod_123con_95)[colnames(msmmod_123con_95)=="Group.U"]<-"Upper95"
    colnames(msmmod_123con_95)[colnames(msmmod_123con_95)=="Group.HR"]<-"Estimate"

    msmmod_123con_975<-as.data.frame(hazard.msm(msmmod_123con, cl=0.975))
    msmmod_123con_975$model<-"msmmod_123con"
    msmmod_123con_975$transition<-as.factor(rownames(msmmod_123con_975))


    colnames(msmmod_123con_975)[colnames(msmmod_123con_975)=="Group.L"]<-"Lower975"
    colnames(msmmod_123con_975)[colnames(msmmod_123con_975)=="Group.U"]<-"Upper975"
    colnames(msmmod_123con_975)[colnames(msmmod_123con_975)=="Group.HR"]<-"Estimate"

    msmmod_123con_9833<-as.data.frame(hazard.msm(msmmod_123con, cl=0.9833))
    msmmod_123con_9833$model<-"msmmod_123con"
    msmmod_123con_9833$transition<-as.factor(rownames(msmmod_123con_9833))

    colnames(msmmod_123con_9833)[colnames(msmmod_123con_9833)=="Group.L"]<-"Lower9833"
    colnames(msmmod_123con_9833)[colnames(msmmod_123con_9833)=="Group.U"]<-"Upper9833"
    colnames(msmmod_123con_9833)[colnames(msmmod_123con_9833)=="Group.HR"]<-"Estimate"

    msmmod_123con_hr_a<-
merge(msmmod_123con_95,msmmod_123con_975,by=c("model","transition","Estimate"))
```

```
    msmmod_123con_hr<-
merge(msmmod_123con_hr_a,msmmod_123con_9833,by=c("model","transition","Estimate"))
    msmmod_123con_hr$SE<-c(msmmod_123con$QmatricesSE$Group[1,2],
                     msmmod_123con$QmatricesSE$Group[2,3],
                     msmmod_123con$QmatricesSE$Group[3,4])
    msmmod_123con_hr$WaldTS<-log(msmmod_123con_hr$Estimate)/msmmod_123con_hr$SE

    }, error=function(e){})

    #Partially constrained model - beta23=beta34
    msmmod_234con_hr<-data.frame(model=as.character(c("msm_con")),
                              transition=as.factor(c("State 1 - State 2","State 2 -
State 3","State 3 - State 4")),
                              Estimate=as.double(NA,NA,NA))
    tryCatch({msmmod_234con<-msm(State~AssessDay, subject=PatID, data=msmdata,
qmatrix=Q,
                                 covariates=~Group, censor=99, censor.states=c(1,2,3),
                                 constraint = list(Group=c(1,2,2)))

    msmmod_234con_95<-as.data.frame(hazard.msm(msmmod_234con, cl=0.95))
    msmmod_234con_95$model<-"msmmod_234con"
    msmmod_234con_95$transition<-as.factor(rownames(msmmod_234con_95))

    colnames(msmmod_234con_95)[colnames(msmmod_234con_95)=="Group.L"]<-"Lower95"
    colnames(msmmod_234con_95)[colnames(msmmod_234con_95)=="Group.U"]<-"Upper95"
    colnames(msmmod_234con_95)[colnames(msmmod_234con_95)=="Group.HR"]<-"Estimate"

    msmmod_234con_975<-as.data.frame(hazard.msm(msmmod_234con, cl=0.975))
    msmmod_234con_975$model<-"msmmod_234con"
    msmmod_234con_975$transition<-as.factor(rownames(msmmod_234con_975))


    colnames(msmmod_234con_975)[colnames(msmmod_234con_975)=="Group.L"]<-"Lower975"
    colnames(msmmod_234con_975)[colnames(msmmod_234con_975)=="Group.U"]<-"Upper975"
    colnames(msmmod_234con_975)[colnames(msmmod_234con_975)=="Group.HR"]<-"Estimate"

    msmmod_234con_9833<-as.data.frame(hazard.msm(msmmod_234con, cl=0.9833))
    msmmod_234con_9833$model<-"msmmod_234con"
    msmmod_234con_9833$transition<-as.factor(rownames(msmmod_234con_9833))

    colnames(msmmod_234con_9833)[colnames(msmmod_234con_9833)=="Group.L"]<-"Lower9833"
    colnames(msmmod_234con_9833)[colnames(msmmod_234con_9833)=="Group.U"]<-"Upper9833"
    colnames(msmmod_234con_9833)[colnames(msmmod_234con_9833)=="Group.HR"]<-"Estimate"

    msmmod_234con_hr_a<-
merge(msmmod_234con_95,msmmod_234con_975,by=c("model","transition","Estimate"))
    msmmod_234con_hr<-
merge(msmmod_234con_hr_a,msmmod_234con_9833,by=c("model","transition","Estimate"))
    msmmod_234con_hr$SE<-c(msmmod_234con$QmatricesSE$Group[1,2],
                     msmmod_234con$QmatricesSE$Group[2,3],
                     msmmod_234con$QmatricesSE$Group[3,4])
    msmmod_234con_hr$WaldTS<-log(msmmod_234con_hr$Estimate)/msmmod_234con_hr$SE

    }, error=function(e){})

    #Combine all analysis output datasets
    overall<-join_all(list(logistic_or, coxout, msmmod_uncon_hr,msmmod_con_hr,
msmmod_123con_hr, msmmod_234con_hr), by="model", type="full")
    overall$msmpvalue<-2*pnorm(-abs(overall$WaldTS))
    modelout[[i]]<-overall

    modelout[[i]]$SimNo<-i
    modelout[[i]]$SampleSize<-n}

    modelout2[[k]]<-do.call(rbind.data.frame,modelout)
    modelout2[[k]]$Scenario<-k
    modelout2[[k]]$seed<-inputmat[k,1]
    modelout2[[k]]$maxfup<-inputmat[k,2]
    modelout2[[k]]$fupsched<-inputmat[k,3]
    modelout2[[k]]$lambda<-toString(c(inputmat[k,6],inputmat[k,7],inputmat[k,8]))
    modelout2[[k]]$h0<-toString(c(inputmat[k,9],inputmat[k,10],inputmat[k,11]))
    modelout2[[k]]$censor0<-toString(c(inputmat[k,12],inputmat[k,13],inputmat[k,14]))
    modelout2[[k]]$censor1<-toString(c(inputmat[k,15],inputmat[k,16],inputmat[k,17]))
    modelout2[[k]]$startstateprop<-
toString(c(inputmat[k,18],inputmat[k,19],inputmat[k,20]))
    }

    return(do.call(rbind.data.frame,modelout2))}

##Example inputmat for base case

BC_inputmat<-matrix(NA,1,20)
BC_inputmat[,1]<-c(25680)
BC_inputmat[,2]<-60 #Maximum follow-up is 60 days
BC_inputmat[,3]<-1 #Daily assessments
BC_inputmat[,4]<-c(100) #Total sample size
BC_inputmat[,5]<-1000 #1000 simulations
BC_inputmat[,6:7]<-0.05 #High risk of moving from state 1 to 2, and from state 2 to 3
BC_inputmat[,8]<-0.03 #Moderate risk of moving from state 3 to state 4
BC_inputmat[,9:11]<-0.67 #Moderate treatment effect on all transitions
BC_inputmat[,12:17]<-0.05 #Censoring transition rate same for all state and all groups
BC_inputmat[,18]<-0.15 # Proportion starting in state 1
BC_inputmat[,19]<-0.70 # Proportion starting in state 1
BC_inputmat[,20]<-0.15 # Proportion starting in state 1
BC_inputmat

BCModel<-simanalyse(BC_inputmat)
```

C.2 Power and Type I error

C.2.1 Length of follow-up



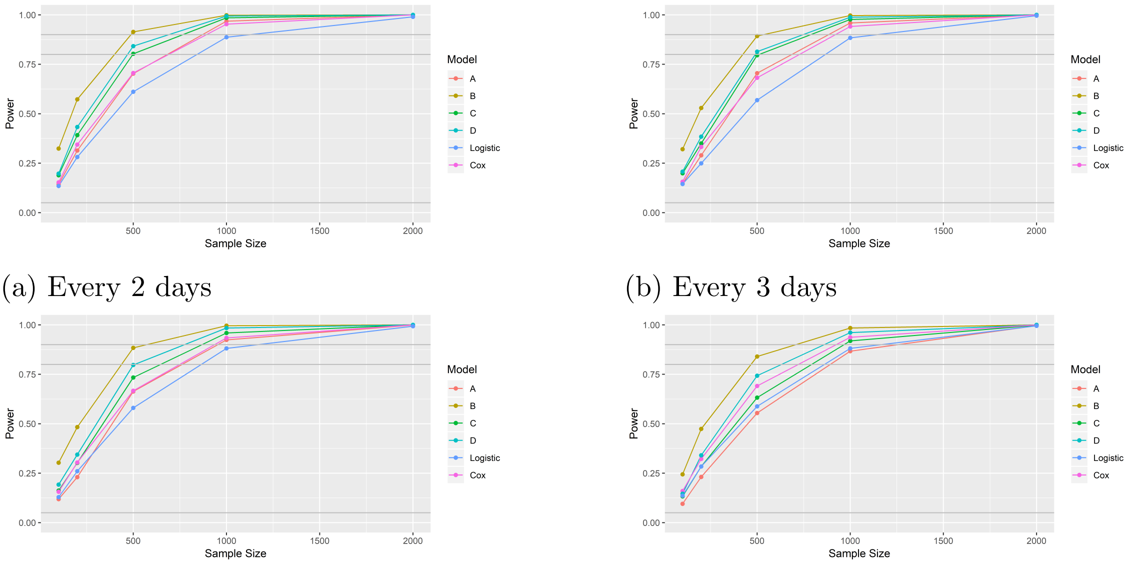
(a) 30 days

(b) 14 days

(c) 7 days

Figure C.1: Power of detecting a significant treatment effect overall according to sample size for different lengths of follow-up (Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.2.2 Assessment intervals



(a) Every 2 days

(b) Every 3 days

(c) Every 7 days

(d) Every 14 days

Figure C.2: Power of detecting a significant treatment effect overall according to sample size for different Assessment intervals (Maximum length of follow-up= 60 days, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.3 Monte Carlo Standard Errors for estimates of power

Table C.1: MCSE for estimates of power in the Null case (Maximum length of follow-up=60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (1, 1, 1)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

N	**Model A**	**Model B**	**Model C**	**Model D**	**Cox PH**	**Logistic**
100	0.006	0.007	0.006	0.006	0.006	0.006
200	0.007	0.008	0.007	0.007	0.008	0.008
500	0.007	0.006	0.007	0.006	0.007	0.007
$1,000$	0.007	0.007	0.006	0.006	0.007	0.007
$2,000$	0.007	0.007	0.006	0.006	0.007	0.007

Table C.2: MCSE for estimates of power in the Base case (Maximum length of follow-up=60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

N	**Model A**	**Model B**	**Model C**	**Model D**	**Cox PH**	**Logistic**
100	0.011	0.015	0.013	0.012	0.012	0.011
200	0.015	0.016	0.015	0.016	0.015	0.014
500	0.014	0.009	0.013	0.012	0.015	0.016
$1,000$	0.006	0.002	0.004	0.003	0.007	0.010
$2,000$	0.000	0.000	0.000	0.000	0.000	0.002

Table C.3: MCSE for estimates of power for different lengths of follow-up (Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Length of Follow-up	N	Model A	Model B	Model C	Model D	Cox PH	Logistic
30 Days	100	0.010	0.014	0.011	0.012	0.010	0.010
	200	0.014	0.016	0.015	0.015	0.014	0.014
	500	0.015	0.010	0.014	0.013	0.016	0.016
	1,000	0.007	0.003	0.005	0.004	0.011	0.012
	2,000	0.000	0.000	0.000	0.000	0.002	0.003
14 Days	100	0.009	0.014	0.011	0.011	0.005	0.005
	200	0.012	0.015	0.013	0.013	0.011	0.011
	500	0.016	0.014	0.016	0.015	0.015	0.015
	1,000	0.012	0.006	0.009	0.009	0.016	0.016
	2,000	0.002	0.000	0.001	0.000	0.011	0.011
7 Days	100	0.005	0.010	0.007	0.007	0.000	0.000
	200	0.010	0.014	0.011	0.012	0.003	0.004
	500	0.014	0.016	0.015	0.015	0.010	0.010
	1,000	0.016	0.012	0.015	0.015	0.014	0.014
	2,000	0.009	0.004	0.006	0.006	0.016	0.016

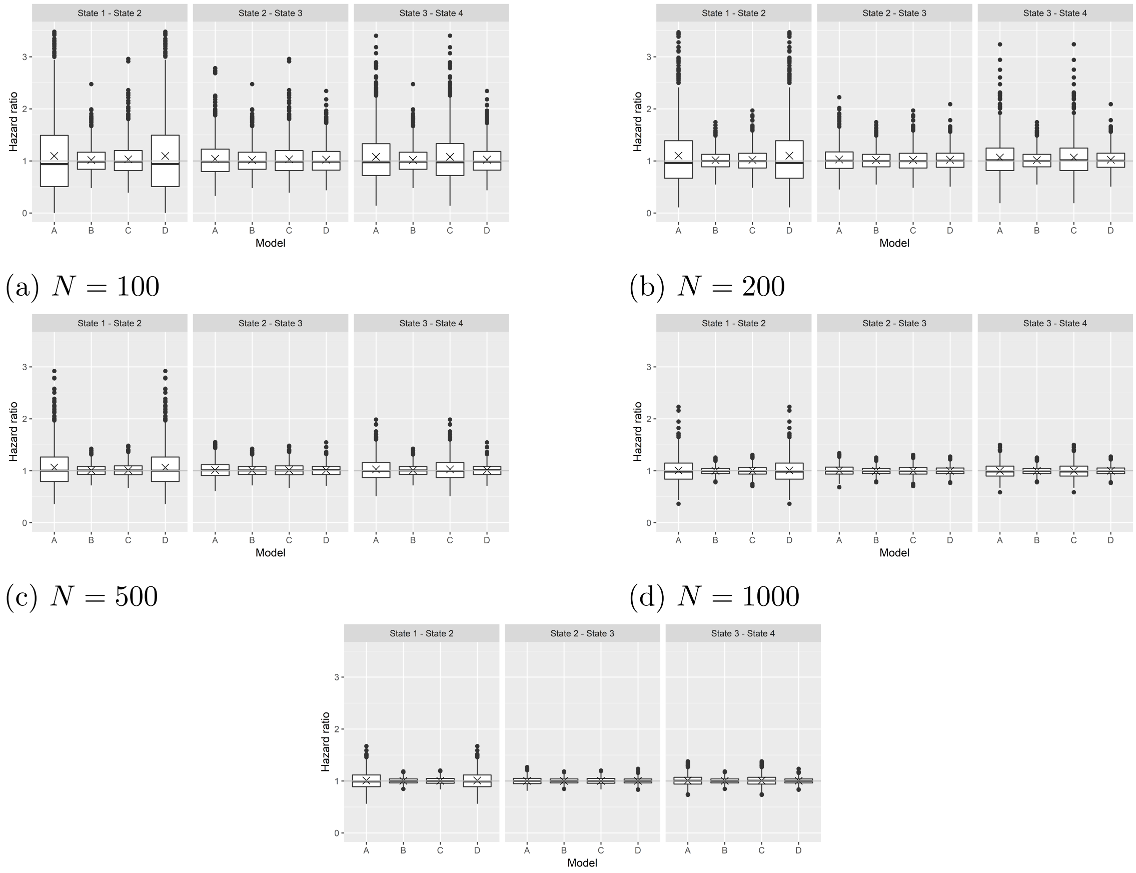
C.4 Box plots of estimated hazard ratios

C.4.1 Null case

Table C.4: MCSE for estimates of power for different assessment intervals (Maximum length of follow-up= 60 Days, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Assessment frequency	N	Model A	Model B	Model C	Model D	Cox PH	Logistic
Every 2 Days	100	0.011	0.015	0.012	0.013	0.011	0.011
	200	0.015	0.016	0.015	0.016	0.015	0.014
	500	0.014	0.009	0.013	0.012	0.014	0.015
	$1,000$	0.006	0.001	0.004	0.002	0.007	0.010
	$2,000$	0.000	0.000	0.000	0.000	0.000	0.003
Every 3 Days	100	0.011	0.015	0.013	0.013	0.011	0.011
	200	0.014	0.016	0.015	0.015	0.015	0.014
	500	0.014	0.010	0.013	0.012	0.015	0.016
	$1,000$	0.006	0.002	0.005	0.003	0.007	0.010
	$2,000$	0.000	0.000	0.001	0.000	0.000	0.002
Every 7 Days	100	0.010	0.015	0.012	0.012	0.011	0.011
	200	0.013	0.016	0.015	0.015	0.015	0.014
	500	0.015	0.010	0.014	0.013	0.015	0.016
	$1,000$	0.008	0.002	0.006	0.004	0.008	0.010
	$2,000$	0.001	0.000	0.000	0.000	0.000	0.003
Every 14 Days	100	0.009	0.014	0.011	0.011	0.012	0.011
	200	0.013	0.016	0.014	0.015	0.015	0.014
	500	0.016	0.012	0.015	0.014	0.015	0.016
	$1,000$	0.011	0.004	0.009	0.006	0.008	0.010
	$2,000$	0.002	0.000	0.001	0.000	0.000	0.002

C.4.2 Base case



(a) $N = 100$

(b) $N = 200$

(c) $N = 500$

(d) $N = 1000$

(e) $N = 2000$

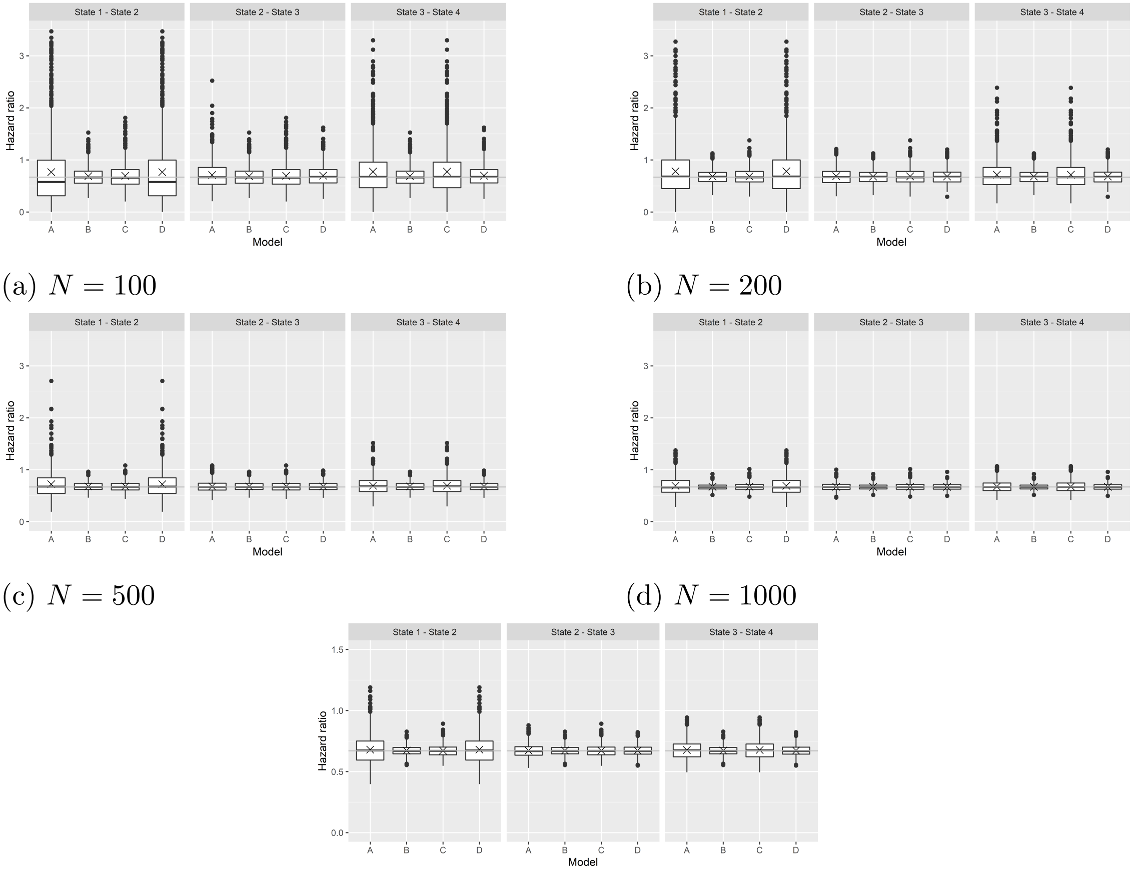
Figure C.3: Point estimates of treatment effects for the null case (Maximum length of follow-up=60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (1, 1, 1)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Table C.5: MCSE for estimates of power for different baseline transition intensities (Maximum length of follow-up= 60 Days, Assessment frequency= daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

$\boldsymbol{q_0}$	N	**Model A**	**Model B**	**Model C**	**Model D**	**Cox PH**	**Logistic**
$(0.01, 0.01, 0.01)$	100	0.004	0.009	0.007	0.008	0.006	0.000
	200	0.007	0.012	0.010	0.010	0.004	0.005
	500	0.013	0.015	0.014	0.014	0.011	0.011
	$1,000$	0.015	0.008	0.016	0.016	0.014	0.014
	$2,000$	0.014	0.007	0.012	0.011	0.016	0.016
$(0.01, 0.01, 0.05)$	100	0.005	0.011	0.006	0.008	0.007	0.008
	200	0.009	0.014	0.011	0.012	0.011	0.011
	500	0.015	0.016	0.015	0.015	0.015	0.015
	$1,000$	0.016	0.011	0.015	0.014	0.016	0.016
	$2,000$	0.009	0.003	0.007	0.005	0.011	0.012
$(0.01, 0.05, 0.01)$	100	0.010	0.014	0.011	0.013	0.005	0.004
	200	0.013	0.016	0.014	0.015	0.010	0.010
	500	0.016	0.013	0.015	0.015	0.015	0.015
	$1,000$	0.010	0.004	0.008	0.006	0.016	0.016
	$2,000$	0.001	0.000	0.001	0.000	0.011	0.012
$(0.05, 0.01, 0.01)$	100	0.005	0.011	0.009	0.007	0.005	0.001
	200	0.009	0.014	0.011	0.011	0.004	0.004
	500	0.014	0.016	0.015	0.015	0.011	0.011
	$1,000$	0.016	0.012	0.014	0.015	0.015	0.015
	$2,000$	0.010	0.005	0.007	0.008	0.016	0.016
$(0.01, 0.05, 0.05)$	100	0.010	0.014	0.012	0.012	0.012	0.011
	200	0.014	0.016	0.015	0.016	0.015	0.015
	500	0.014	0.009	0.013	0.011	0.014	0.015
	$1,000$	0.006	0.002	0.004	0.002	0.006	0.010
	$2,000$	0.000	0.000	0.000	0.000	0.000	0.002
$(0.05, 0.01, 0.05)$	100	0.007	0.011	0.009	0.009	0.007	0.007
	200	0.011	0.015	0.013	0.013	0.011	0.011
	500	0.015	0.015	0.016	0.016	0.015	0.015
	$1,000$	0.015	0.009	0.013	0.013	0.016	0.016
	$2,000$	0.006	0.001	0.004	0.003	0.010	0.011
$(0.05, 0.05, 0.01)$	100	0.010	0.014	0.012	0.012	0.005	0.005
	200	0.014	0.016	0.015	0.015	0.011	0.010
	500	0.015	0.011	0.014	0.013	0.015	0.015
	$1,000$	0.007	0.003	0.005	0.005	0.015	0.016
	$2,000$	0.000	0.000	0.000	0.000	0.010	0.011
$(0.05, 0.05, 0.05)$	100	0.012	0.015	0.013	0.013	0.013	0.012
	200	0.015	0.016	0.015	0.016	0.016	0.015
	500	0.014	0.008	0.012	0.011	0.013	0.015

Table C.6: MCSE for estimates of power for different treatment effects (Maximum length of follow-up= 60 Days, Assessment frequency= daily, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

$\exp(\boldsymbol{q_0})$	N	**Model A**	**Model B**	**Model C**	**Model D**	**Cox PH**	**Logistic**
$(0.50.0.50, 0.67)$	100	0.015	0.016	0.016	0.016	0.013	0.013
	200	0.014	0.010	0.013	0.013	0.016	0.015
	500	0.004	0.002	0.003	0.004	0.011	0.013
	1,000	0.000	0.000	0.000	0.000	0.003	0.005
	2,000	0.000	0.000	0.000	0.000	0.000	0.000
$(0.67.0.67, 0.50)$	100	0.012	0.016	0.013	0.014	0.014	0.014
	200	0.016	0.015	0.016	0.016	0.016	0.016
	500	0.010	0.005	0.009	0.008	0.009	0.011
	1,000	0.002	0.000	0.001	0.000	0.001	0.002
	2,000	0.000	0.000	0.000	0.000	0.000	0.000
$(0.90, 0.90, 0.67)$	100	0.008	0.010	0.008	0.008	0.010	0.009
	200	0.011	0.012	0.010	0.010	0.013	0.012
	500	0.015	0.015	0.015	0.014	0.016	0.015
	1,000	0.015	0.015	0.015	0.016	0.013	0.015
	2,000	0.009	0.010	0.009	0.012	0.007	0.010
$(0.67, 0.67, 0.90)$	100	0.010	0.013	0.012	0.011	0.008	0.007
	200	0.014	0.016	0.015	0.014	0.011	0.010
	500	0.015	0.013	0.014	0.015	0.014	0.014
	1,000	0.008	0.005	0.005	0.009	0.016	0.016
	2,000	0.001	0.000	0.001	0.001	0.012	0.014



(a) $N = 100$

(b) $N = 200$

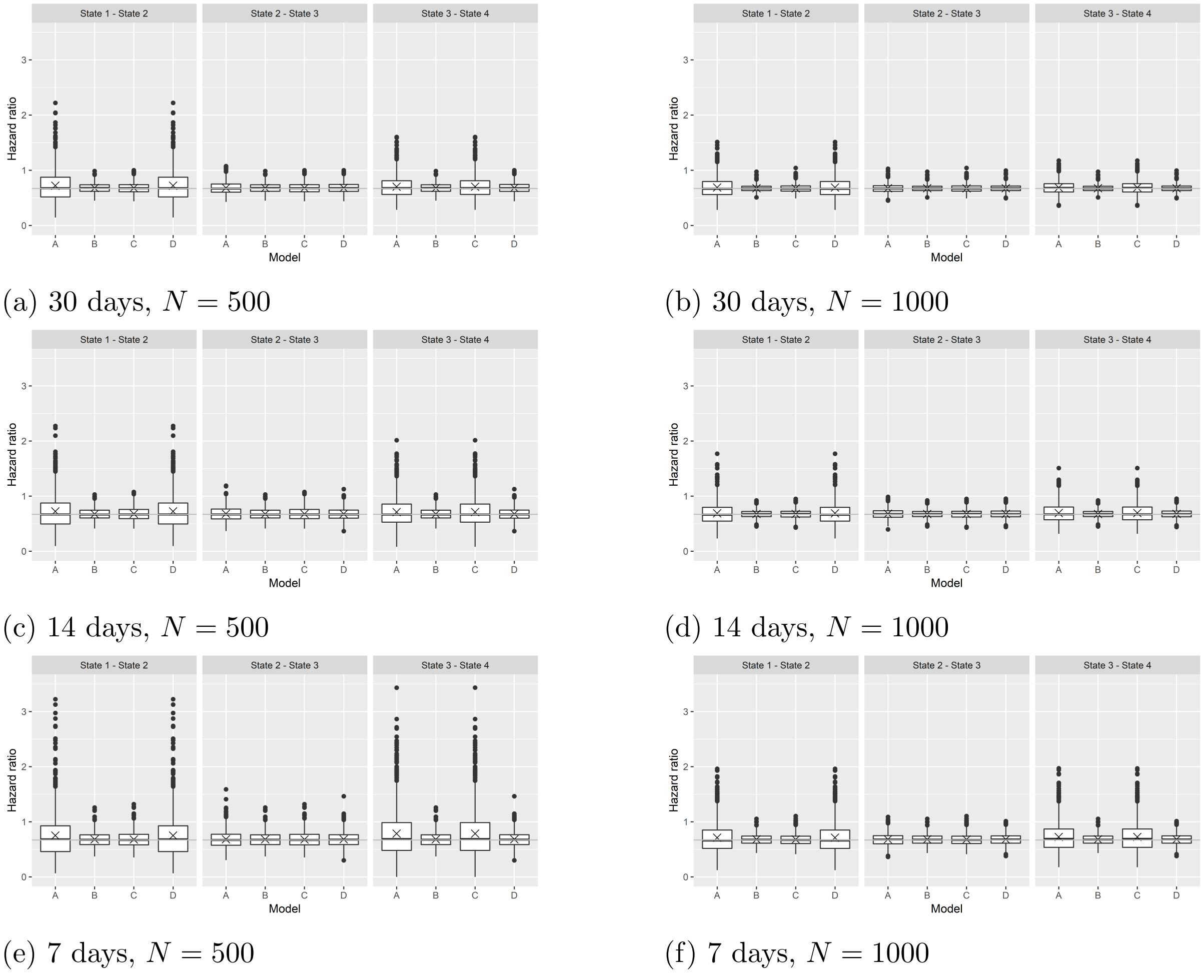
(c) $N = 500$

(d) $N = 1000$

(e) $N = 2000$

Figure C.4: Point estimates of treatment effects for the base case (Maximum length of follow-up=60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.4.3 Length of follow-up



(a) 30 days, $N = 500$

(b) 30 days, $N = 1000$

(c) 14 days, $N = 500$

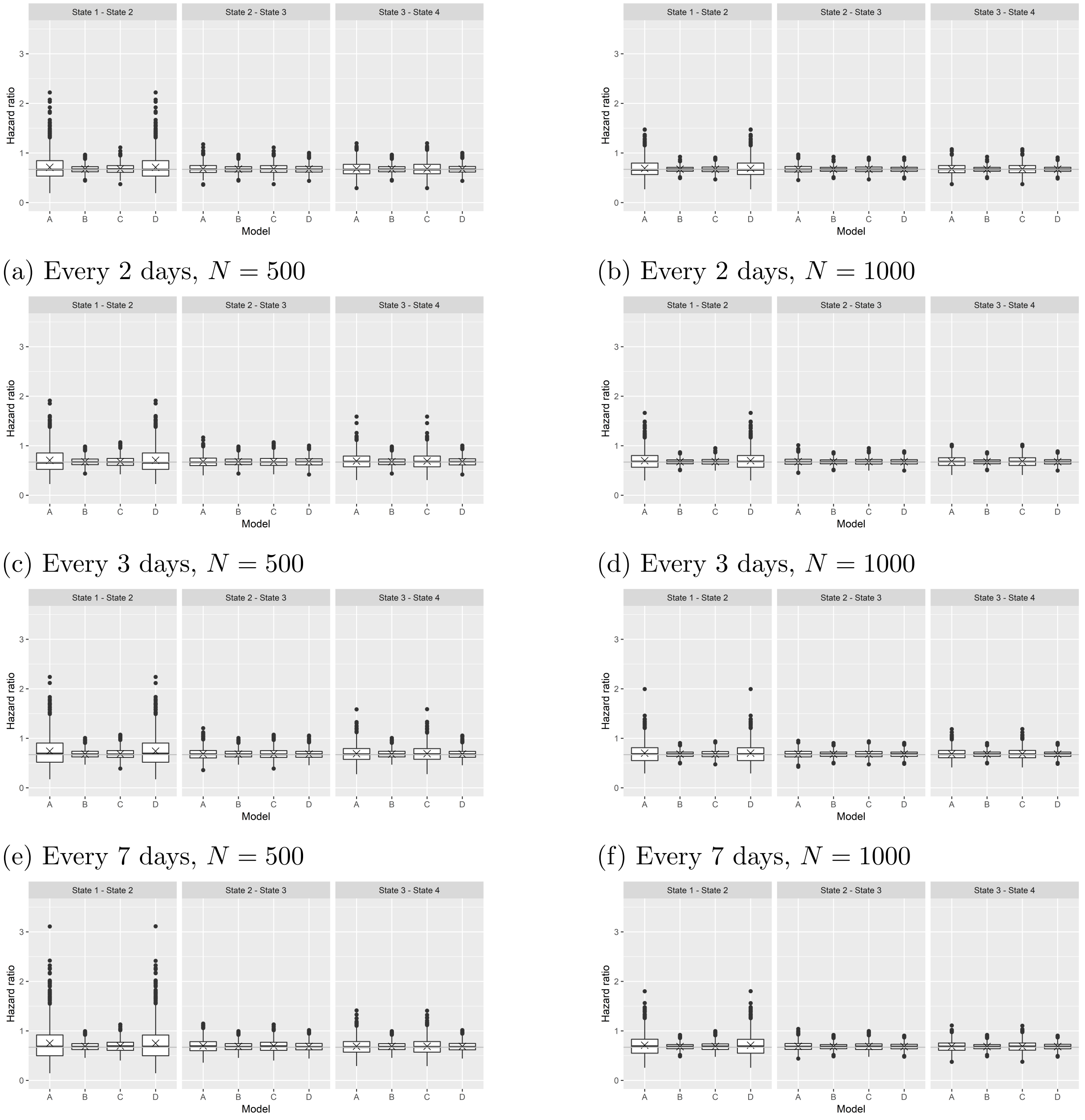
(d) 14 days, $N = 1000$

(e) 7 days, $N = 500$

(f) 7 days, $N = 1000$

Figure C.5: Point estimates of treatment effects for different lengths of follow-up (Assessment interval= Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.4.4 Assessment intervals



(a) Every 2 days, $N = 500$

(b) Every 2 days, $N = 1000$

(c) Every 3 days, $N = 500$

(d) Every 3 days, $N = 1000$

(e) Every 7 days, $N = 500$

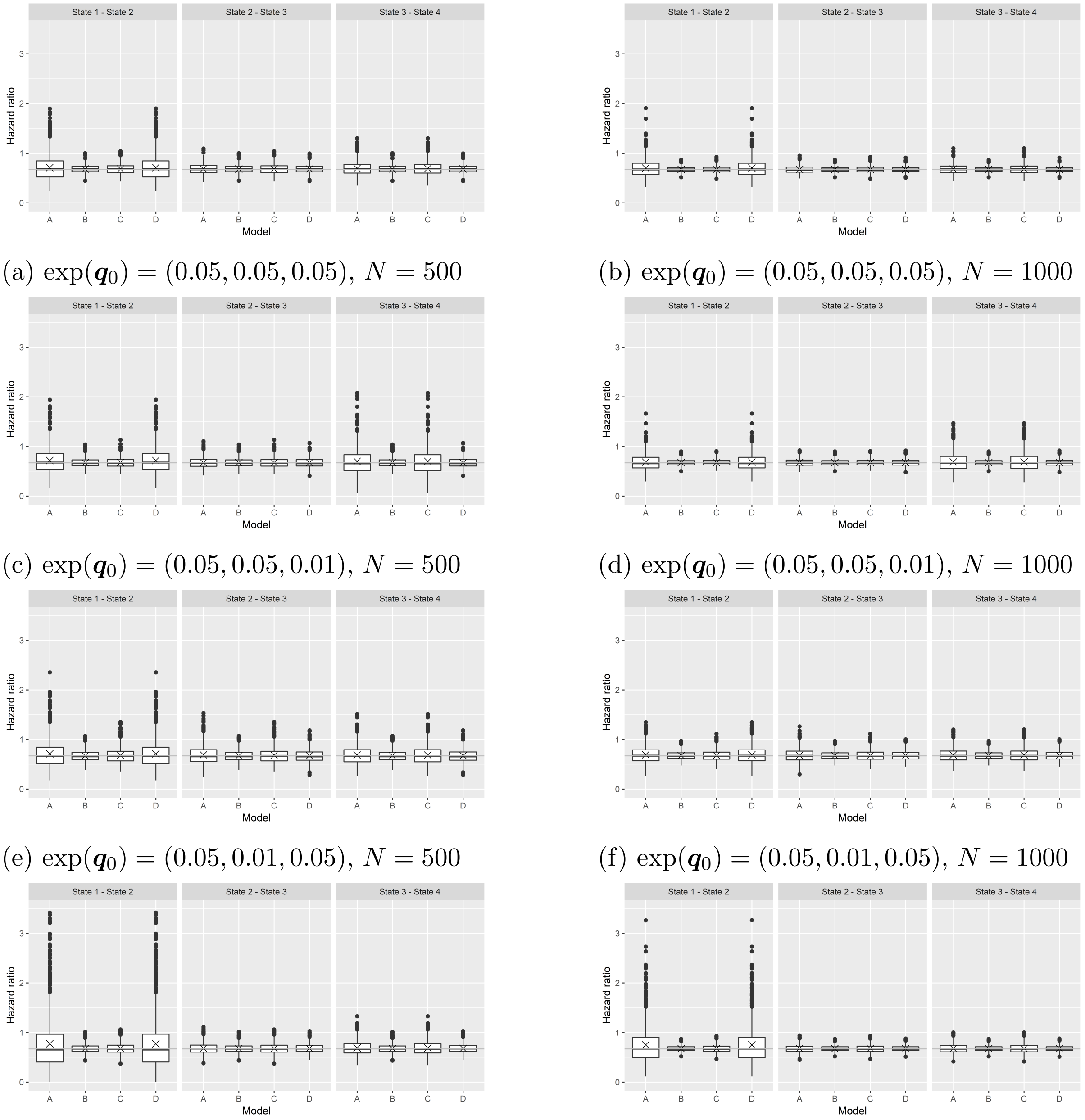
(f) Every 7 days, $N = 1000$

(g) Every 14 days, $N = 500$

(h) Every 14 days, $N = 1000$

Figure C.6: Point estimates of treatment effects for different assessment intervals (Maximum length of follow-up=60 days, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.4.5 Baseline transition intensities



(a) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.05)$, $N = 500$

(b) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.05)$, $N = 1000$

(c) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.01)$, $N = 500$

(d) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.01)$, $N = 1000$

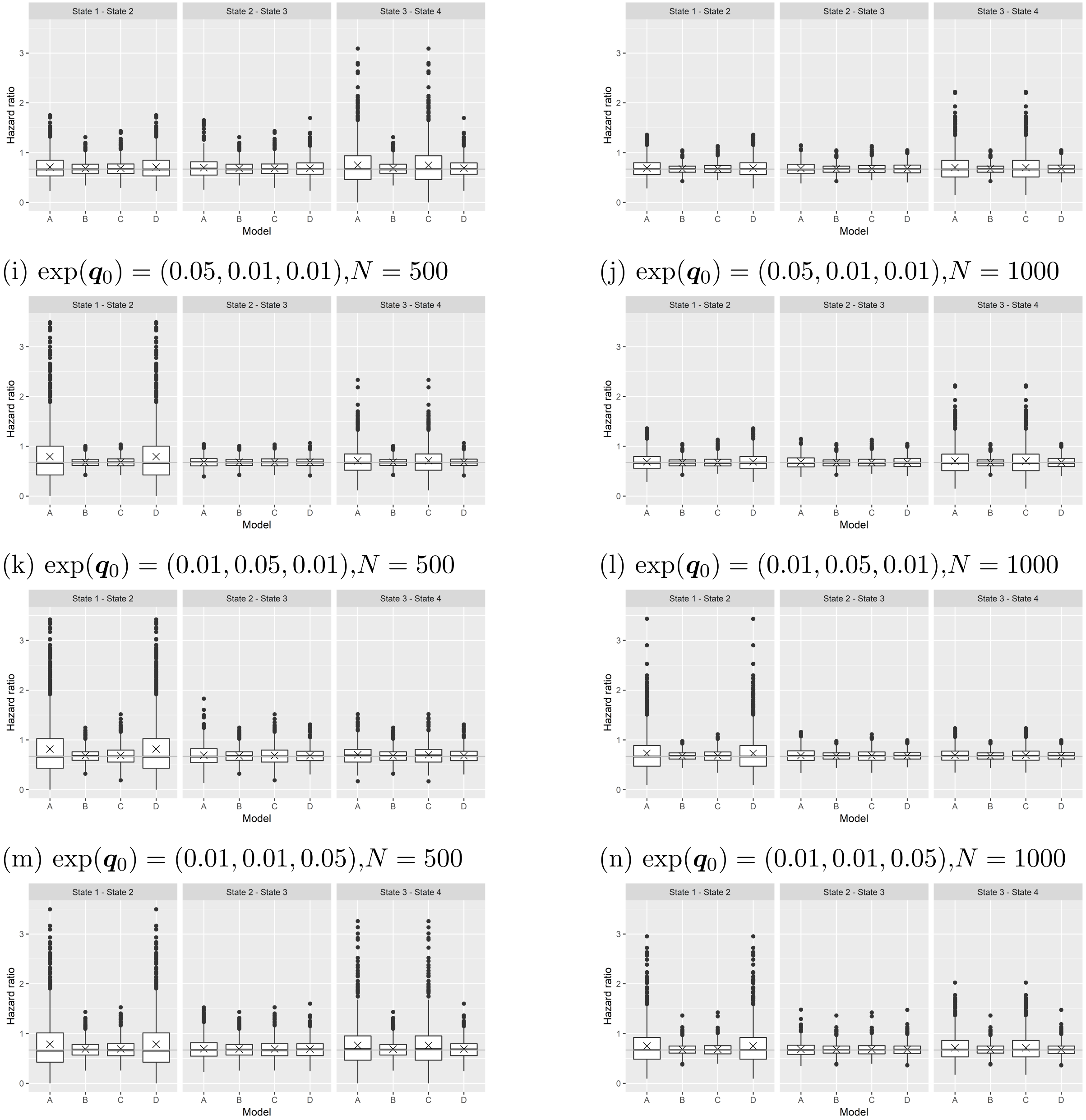
(e) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.05)$, $N = 500$

(f) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.05)$, $N = 1000$

(g) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.05)$,$N = 500$

(h) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.05)$,$N = 1000$

Figure C.7: Part 1: Point estimates of treatment effects for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)



(i) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.01)$,$N = 500$

(j) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.01)$,$N = 1000$

(k) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.01)$,$N = 500$

(l) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.01)$,$N = 1000$

(m) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.05)$,$N = 500$

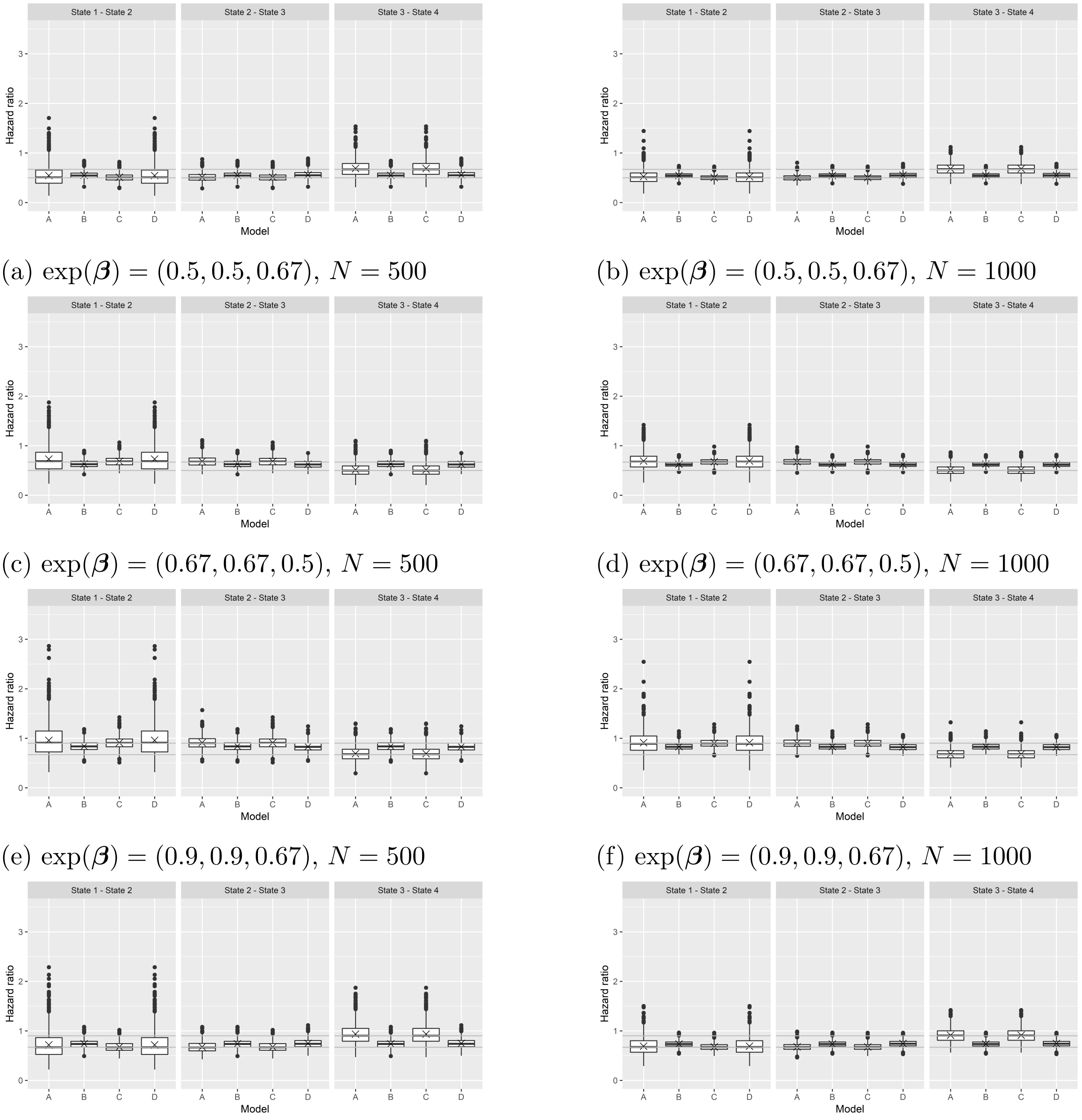
(n) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.05)$,$N = 1000$

(o) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.01)$,$N = 500$

(p) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.01)$,$N = 1000$

Figure C.7: Point estimates of treatment effects for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$) (cont.)

C.4.6 Treatment effects



(a) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$, $N = 500$

(b) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$, $N = 1000$

(c) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.5)$, $N = 500$

(d) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.5)$, $N = 1000$

(e) $\exp(\boldsymbol{\beta}) = (0.9, 0.9, 0.67)$, $N = 500$

(f) $\exp(\boldsymbol{\beta}) = (0.9, 0.9, 0.67)$, $N = 1000$

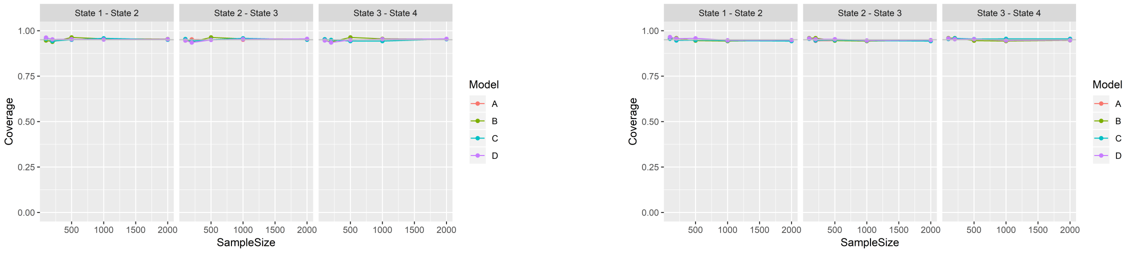
(g) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.9)$, $N = 500$

(h) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.9)$, $N = 1000$

Figure C.8: Point estimates of treatment effects for different treatment effects (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.5 Coverage

C.5.1 Null and Base case

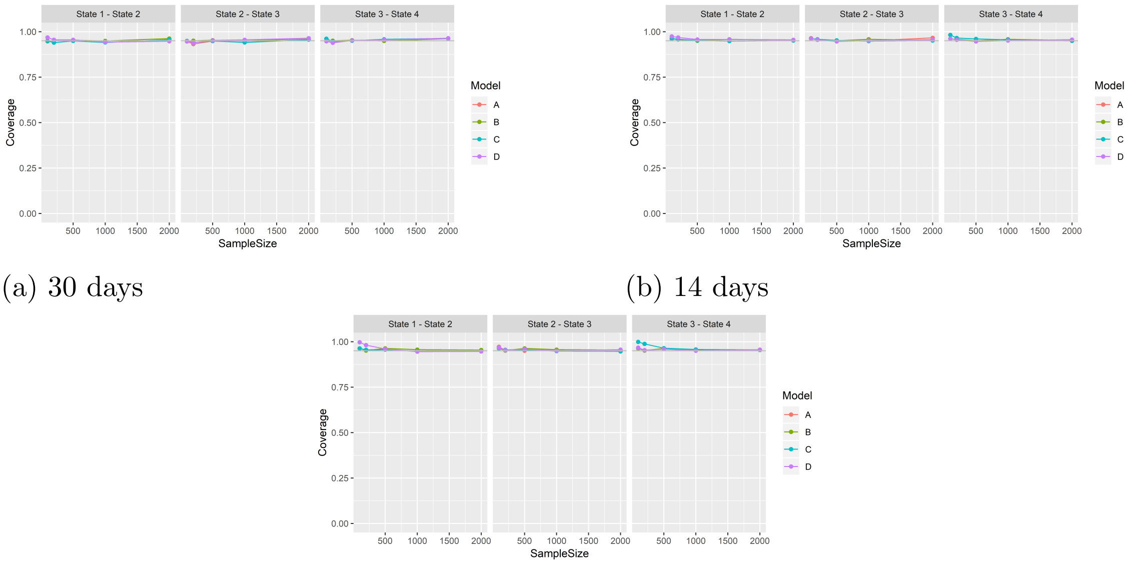


(a) Null $\exp(\boldsymbol{\beta}) = (1, 1, 1)$

(b) Base case $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$

Figure C.9: Coverage of treatment effect estimates for the null and base case (Assessment frequency=Daily, Maximum length of follow-up= 60 days, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.5.2 Length of follow-up



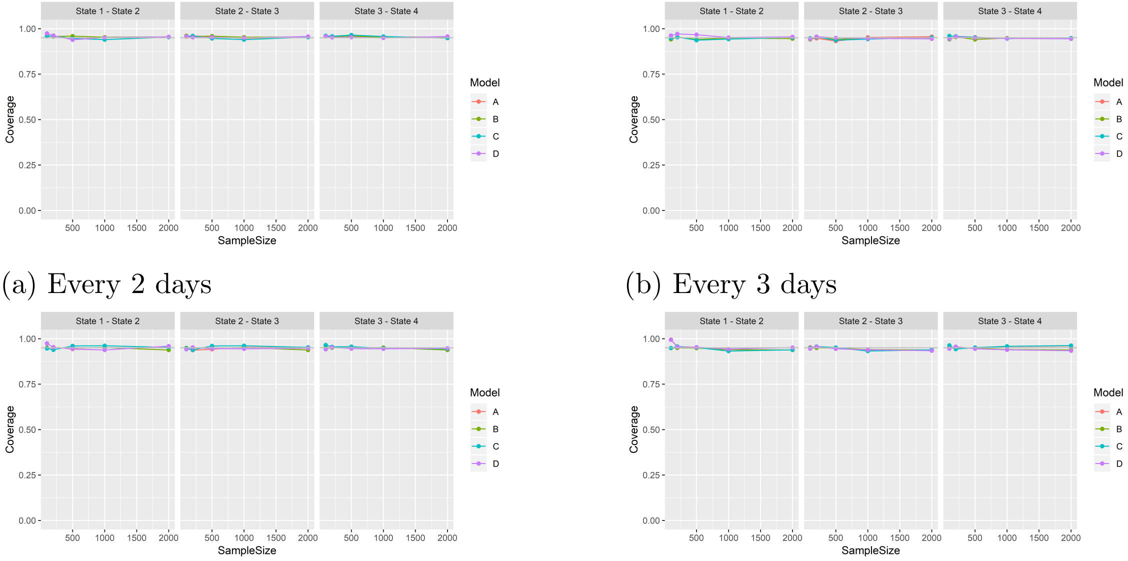
(a) 30 days

(b) 14 days

(c) 7 days

Figure C.10: Coverage of treatment effect estimates for different lengths of follow-up (Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.5.3 Assessment intervals



(a) Every 2 days

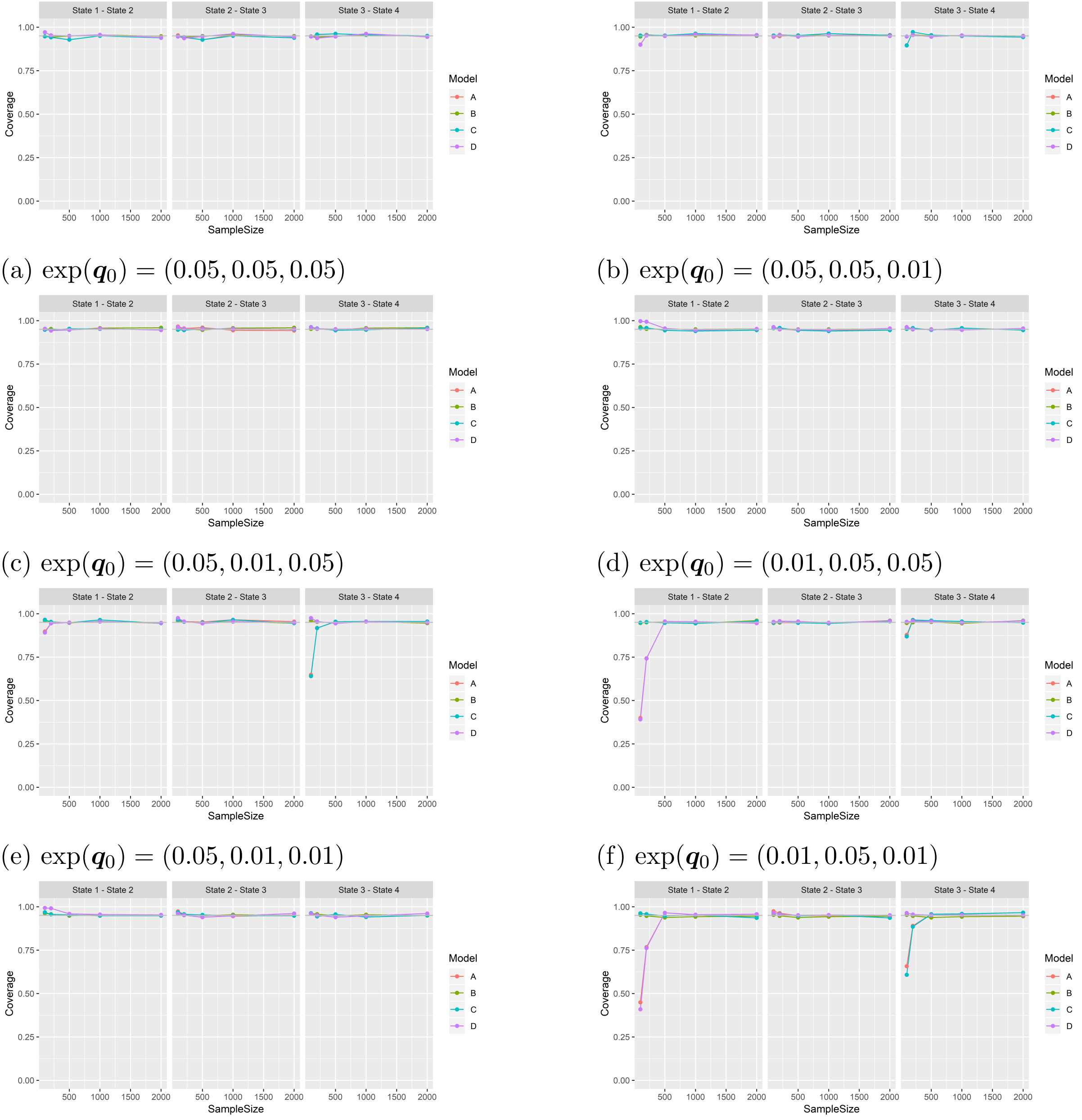
(b) Every 3 days

(c) Every 7 days

(d) Every 14 days

Figure C.11: Coverage of treatment effect estimates for different assessment intervals (Maximum length of follow-up= 60 days, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.5.4 Baseline transition intensities



(a) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.05)$

(b) $\exp(\boldsymbol{q}_0) = (0.05, 0.05, 0.01)$

(c) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.05)$

(d) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.05)$

(e) $\exp(\boldsymbol{q}_0) = (0.05, 0.01, 0.01)$

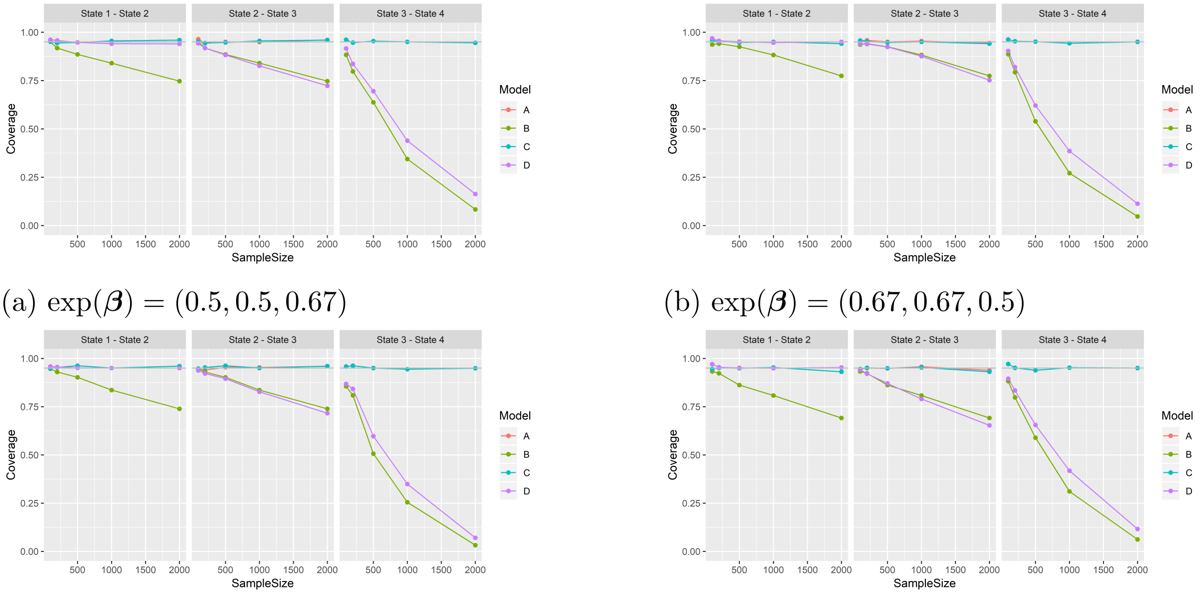
(f) $\exp(\boldsymbol{q}_0) = (0.01, 0.05, 0.01)$

(g) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.05)$

(h) $\exp(\boldsymbol{q}_0) = (0.01, 0.01, 0.01)$

Figure C.12: Coverage of treatment effect estimates for different baseline transition intensities (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

C.5.5 Treatment effects



(a) $\exp(\boldsymbol{\beta}) = (0.5, 0.5, 0.67)$

(b) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.5)$

(c) $\exp(\boldsymbol{\beta}) = (0.9, 0.9, 0.67)$

(d) $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.9)$

Figure C.13: Coverage of treatment effect estimates for different treatment effects (Maximum length of follow-up= 60 days, Assessment frequency=Daily, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

Appendix D

Chapter 7: Impact of Misclassification on power, bias and coverage

D.1 Monte Carlo Standard Errors for estimates of power

Table D.1: MCSE for estimates of power in Part I (Misclassification of all transient states, misclassification of the absorbing state with the adjacent state at most)

Scenario	**Model A**	**Model B**	**Model C**
No misclassification of absorbing state			
1	0.005	0.008	0.011
2	0.006	0.010	0.012
3	0.006	0.011	0.012
4	0.007	0.011	0.014
5	0.008	0.012	0.014
6	0.008	0.013	0.014
Under-reporting of absorbing state			
7	0.006	0.009	0.014
8	0.006	0.011	0.014
9	0.006	0.011	0.014
10	0.007	0.010	0.015
11	0.007	0.012	0.015
12	0.007	0.012	0.015
Over-reporting of absorbing state			
13	0.006	0.009	0.013
14	0.006	0.013	0.015
15	0.006	0.008	0.015
16	0.007	0.013	0.015
17	0.007	0.013	0.015
18	0.008	0.010	0.015
Both under- and over-reporting of absorbing state			
19	0.006	0.010	0.014
20	0.006	0.009	0.015
21	0.006	0.016	0.015
22	0.007	0.012	0.015
23	0.008	0.012	0.016
24	0.008	0.015	0.016

Table D.2: MCSE for estimates of power in Part II (Misclassification of all states with the adjacent state at most)

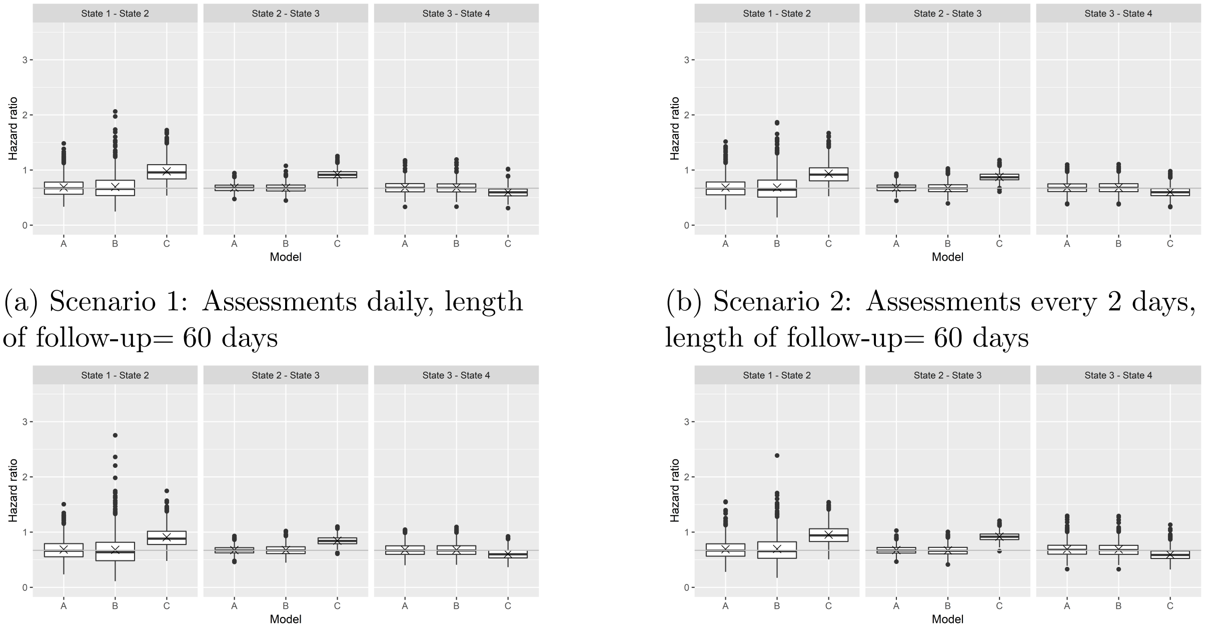
Scenario	**Model A**	**Model B**	**Model C**
25	0.006	0.008	0.011
26	0.006	0.008	0.013
27	0.005	0.009	0.013
28	0.006	0.011	0.014

Table D.3: MCSE for estimates of power in Part III (similar to PU case studies)

Scenario	**Model A**	**Model B**	**Model C**
29	0.006	0.009	0.013
30	0.005	0.009	0.015
31	0.006	0.010	0.013

D.2 Box plots of estimated hazard ratios

D.2.1 Part I: Misclassification of transient states only (Scenarios 1 to 6)



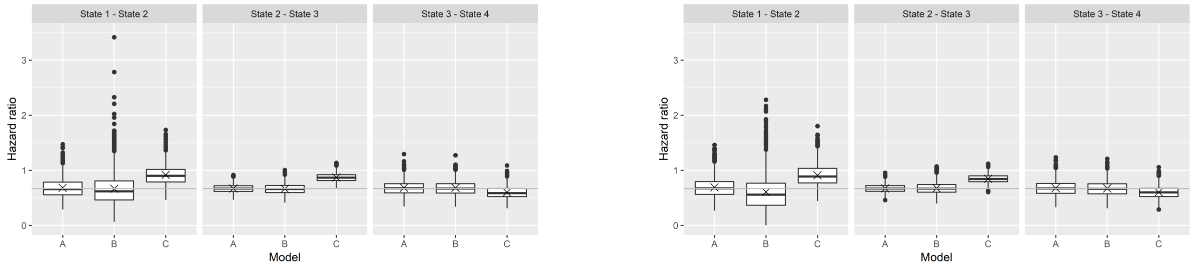
(a) Scenario 1: Assessments daily, length of follow-up= 60 days

(b) Scenario 2: Assessments every 2 days, length of follow-up= 60 days

(c) Scenario 3: Assessments every 3 days, length of follow-up= 60 days

(d) Scenario 4: Assessments daily, length of follow-up= 30 days

Figure D.1: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)

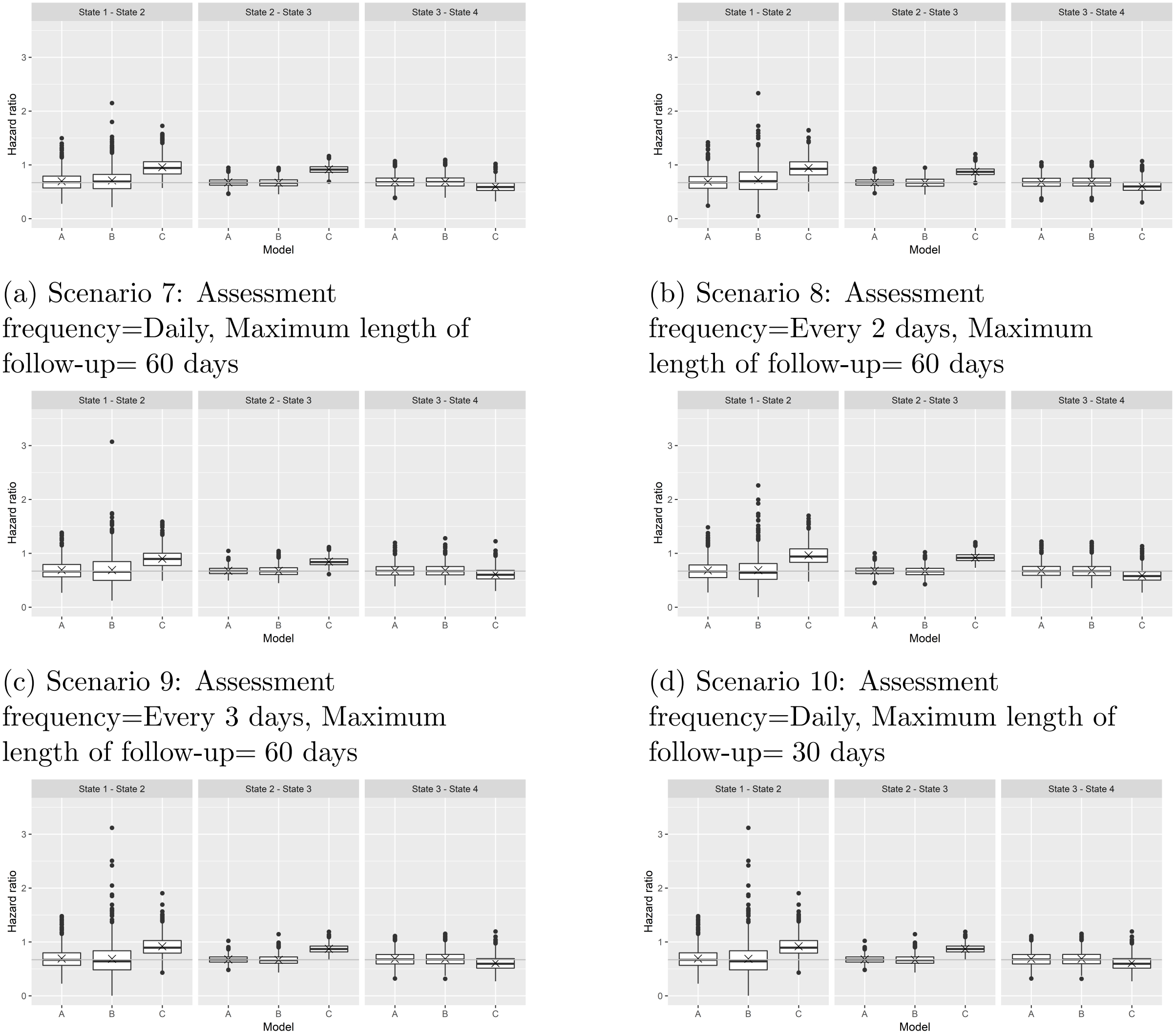


(e) Scenario 5: Assessments every 2 days, length of follow-up= 30 days

(f) Scenario 6: Assessments every 3 days, length of follow-up= 30 days

Figure D.1: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$), Model A: $\beta_{12} \neq \beta_{23} \neq \beta_{34}$, Model B: $\beta_{12} = \beta_{23} = \beta_{34}$, Model C: $\beta_{12} = \beta_{23} \neq \beta_{34}$, Model D: $\beta_{12} \neq \beta_{23} = \beta_{34}$)(cont.)

D.2.2 Part I: Misclassification of transient states and under-reporting of the absorbing state (Scenarios 7 to 12)



(a) Scenario 7: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 8: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

(c) Scenario 9: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

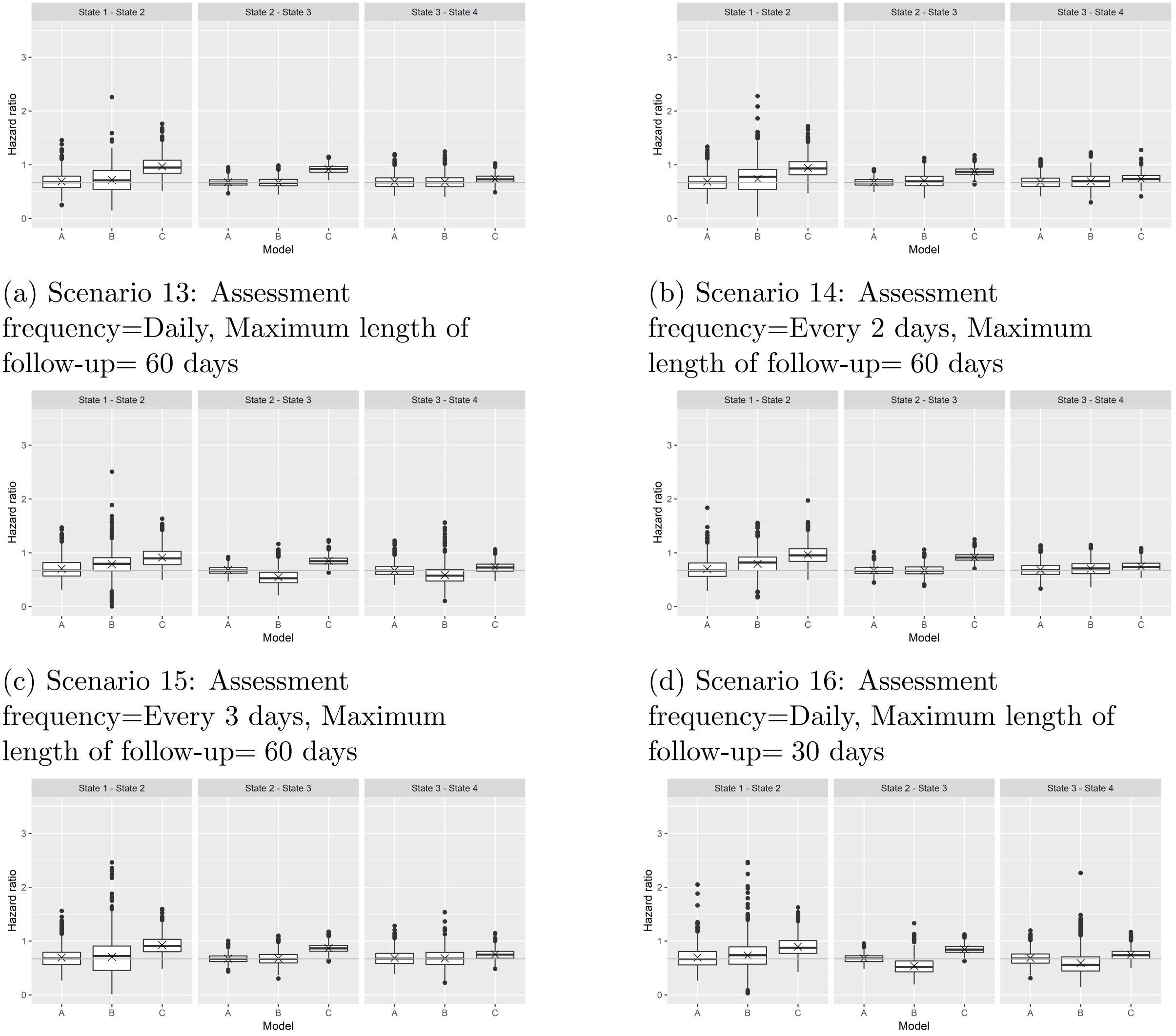
(d) Scenario 10: Assessment frequency=Daily, Maximum length of follow-up= 30 days

(e) Scenario 11: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days

(f) Scenario 12: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.2: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.2.3 Part I: Misclassification of transient states and over-reporting of the absorbing state (Scenarios 13 to 18)



(a) Scenario 13: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 14: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

(c) Scenario 15: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

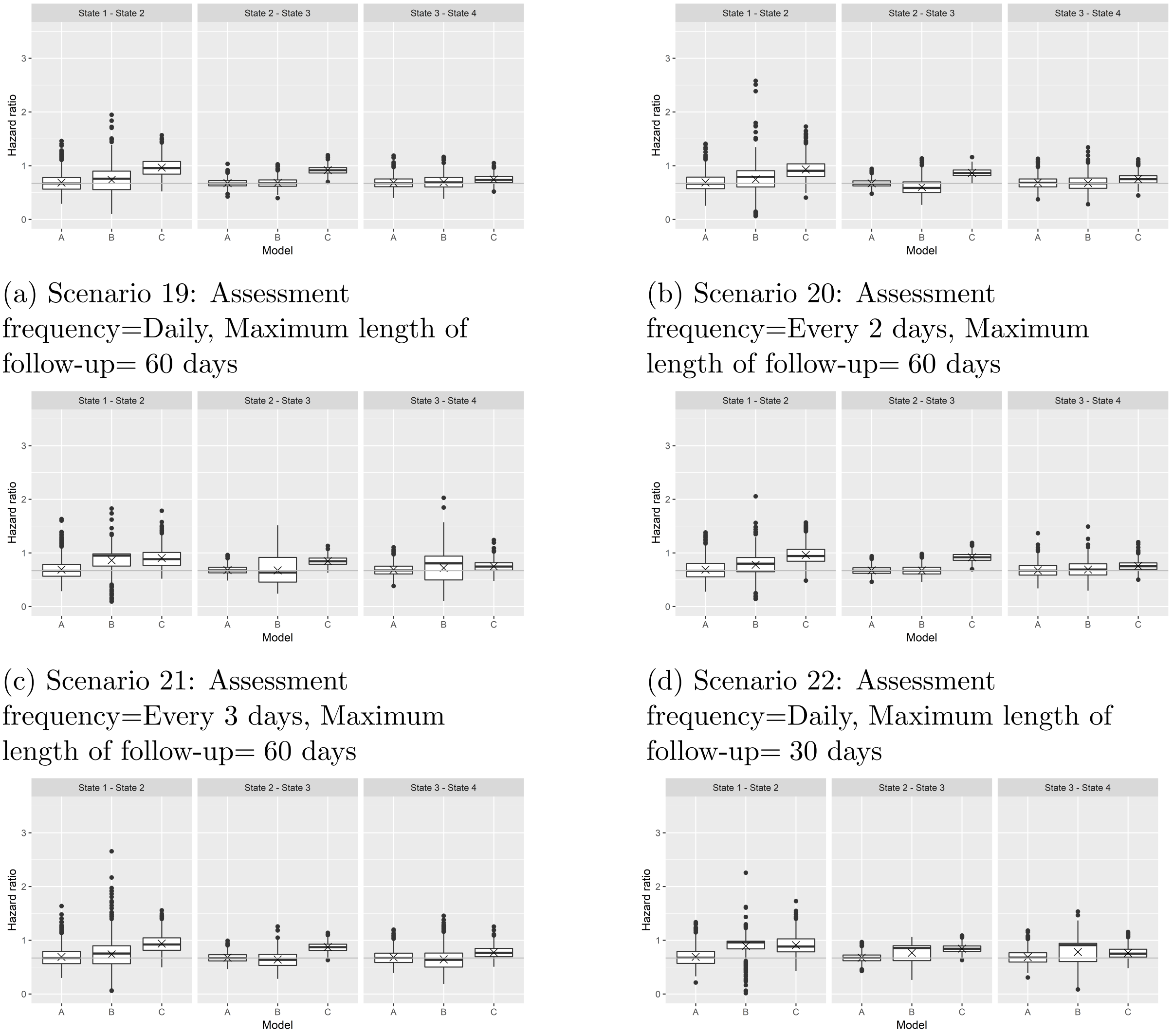
(d) Scenario 16: Assessment frequency=Daily, Maximum length of follow-up= 30 days

(e) Scenario 17: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days

(f) Scenario 18: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.3: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.2.4 Part I: Misclassification of transient states and both under- and over-reporting of the absorbing state (Scenarios 19 to 24)



(a) Scenario 19: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 20: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

(c) Scenario 21: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

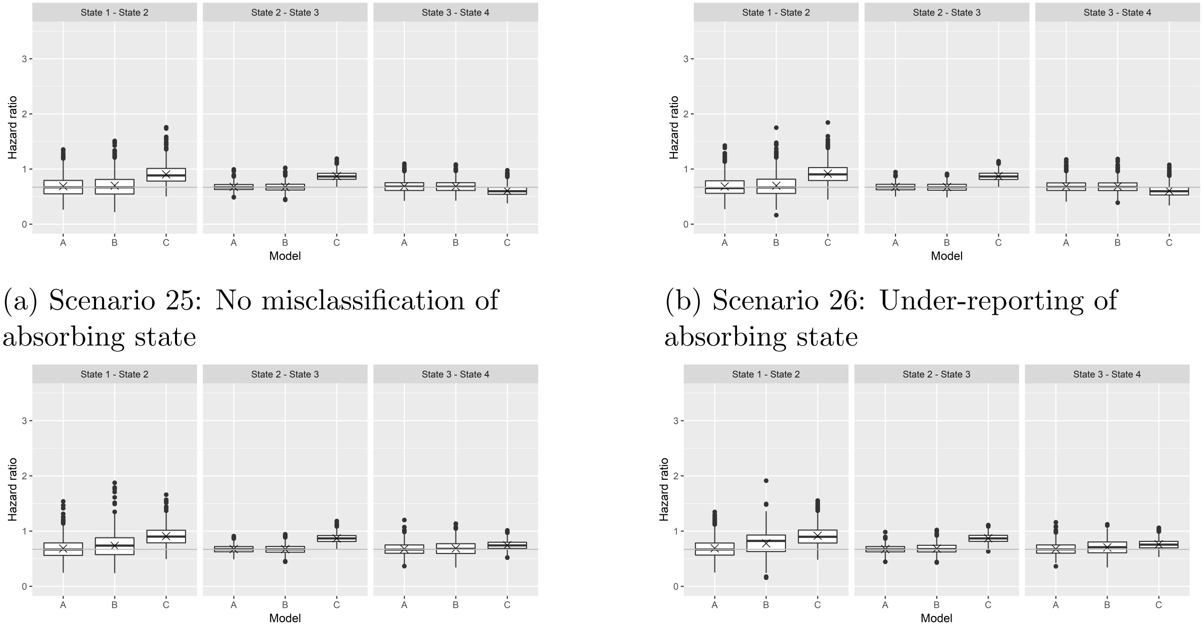
(d) Scenario 22: Assessment frequency=Daily, Maximum length of follow-up= 30 days

(e) Scenario 23: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days

(f) Scenario 24: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.4: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.2.5 Part II: Misclassification of adjacent states only (Scenarios 25 to 28)



(a) Scenario 25: No misclassification of absorbing state

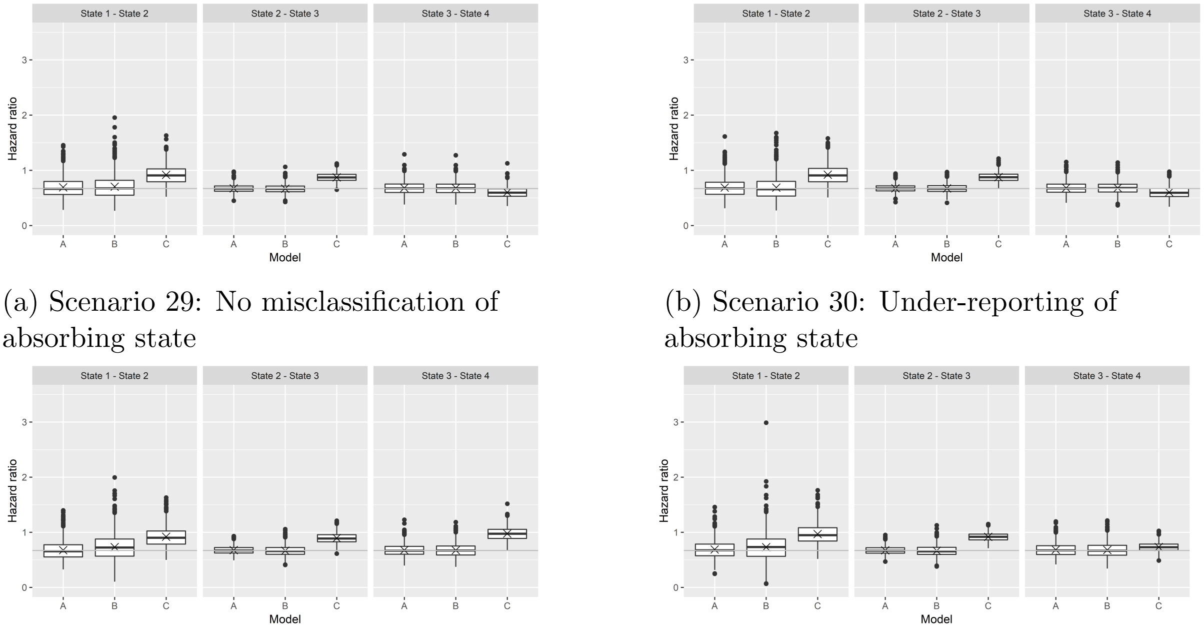
(b) Scenario 26: Under-reporting of absorbing state

(c) Scenario 27: Over-reporting of absorbing state

(d) Scenario 28: Both under- and over-reporting of absorbing state

Figure D.5: Point estimates for hazard ratios ($N = 1000$, Length of follow-up= 60 days, Assessment frequency = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.2.6 Part III: Pressure ulcer setting (Scenarios 29 to 32)



(a) Scenario 29: No misclassification of absorbing state

(b) Scenario 30: Under-reporting of absorbing state

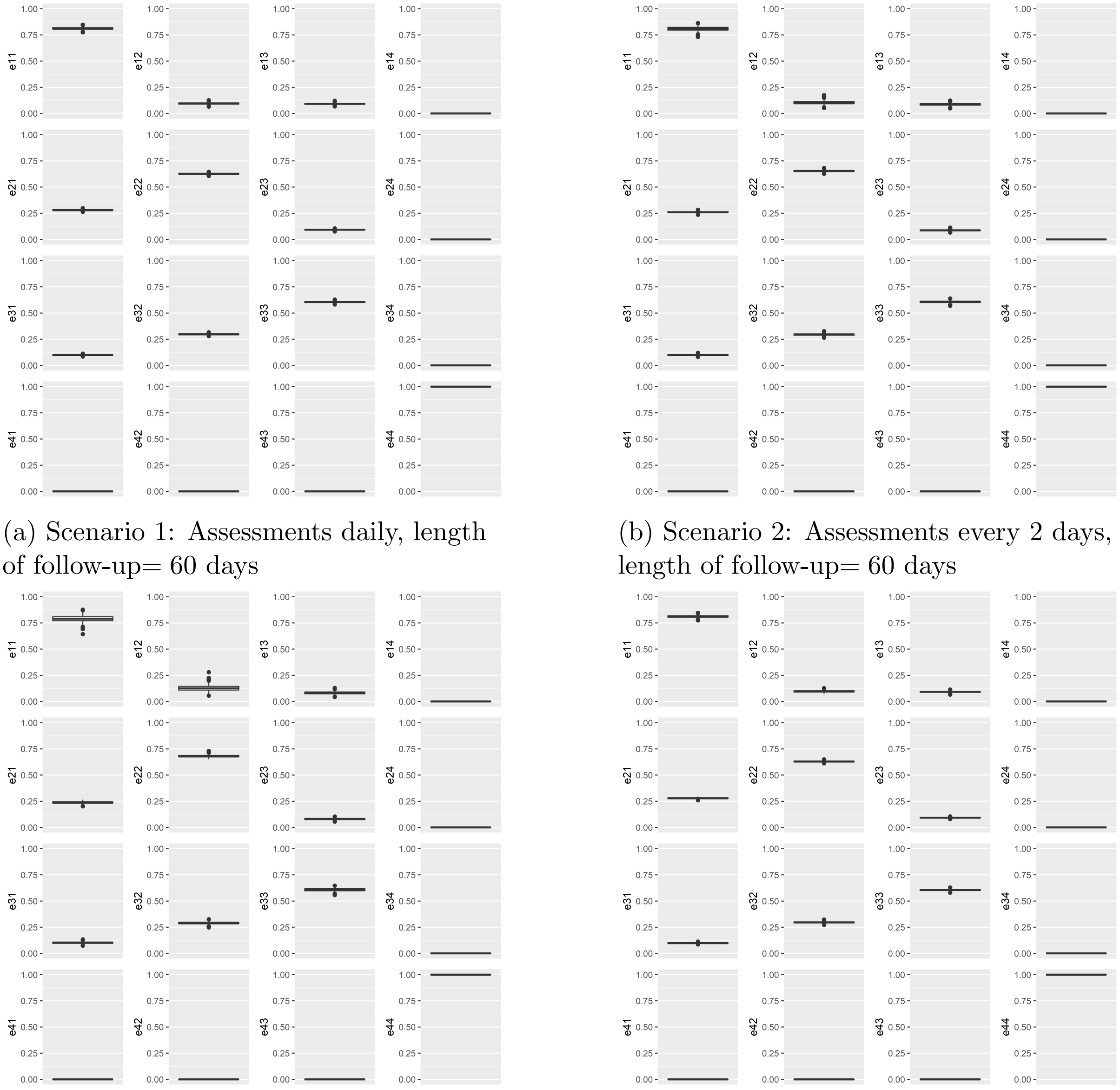
(c) Scenario 31: Over-reporting of absorbing state

(d) Scenario 32: Both under- and over-reporting of absorbing state

Figure D.6: Point estimates for hazard ratios ($N = 1000$, Length of follow-up= 60 days, Assessment frequency = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.3 Estimated misclassification probabilities

D.3.1 Part I: Misclassification of transient states only (Scenarios 1 to 6)



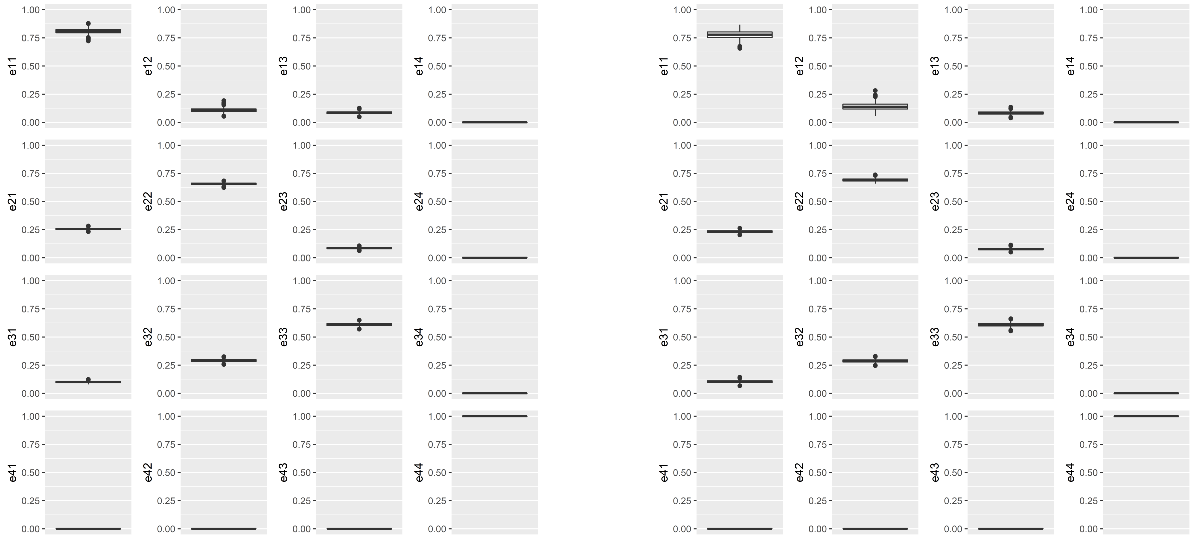
(a) Scenario 1: Assessments daily, length of follow-up= 60 days

(b) Scenario 2: Assessments every 2 days, length of follow-up= 60 days

(c) Scenario 3: Assessments every 3 days, length of follow-up= 60 days

(d) Scenario 4: Assessments daily, length of follow-up= 30 days

Figure D.7: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

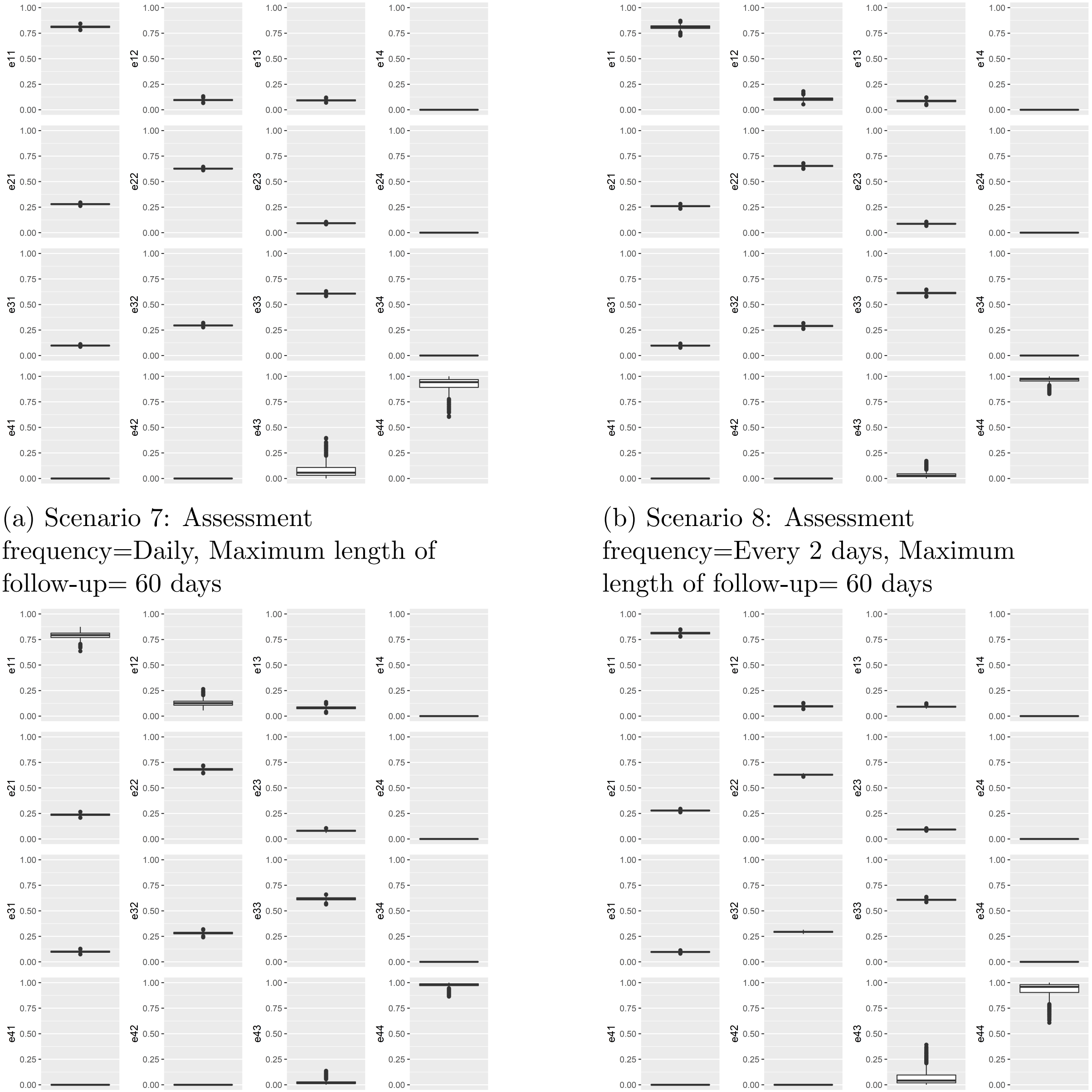


(e) Scenario 5: Assessments every 2 days, length of follow-up= 30 days

(f) Scenario 6: Assessments every 3 days, length of follow-up= 30 days

Figure D.7: Point estimates for hazard ratios ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$) (cont.)

D.3.2 Part I: Misclassification of transient states and under-reporting of the absorbing state (Scenarios 7 to 12)



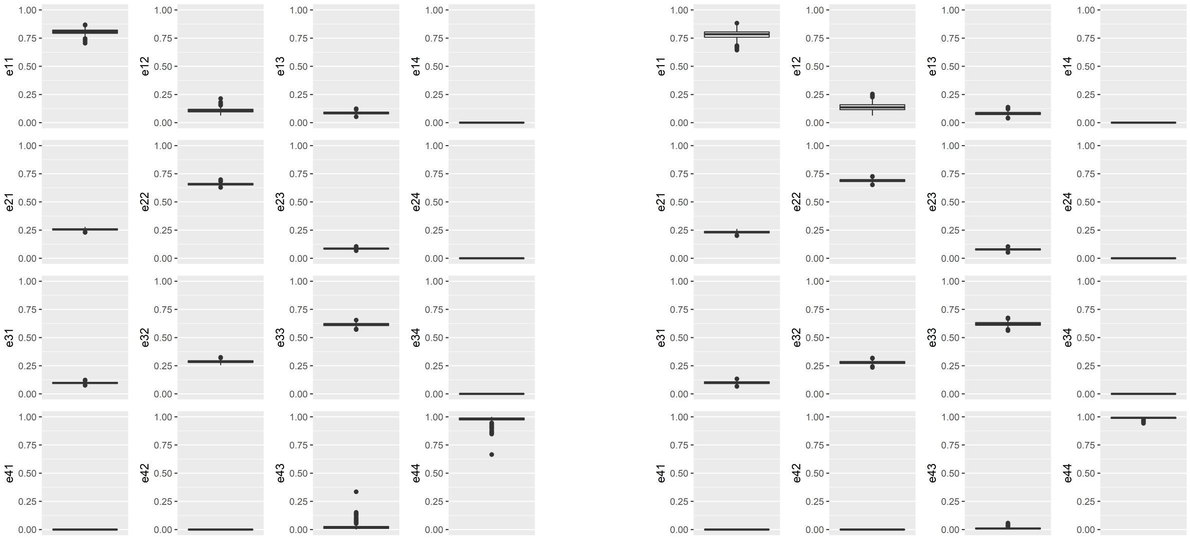
(a) Scenario 7: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 8: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

(c) Scenario 9: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

(d) Scenario 10: Assessment frequency=Daily, Maximum length of follow-up= 30 days

Figure D.8: Point estimates for misclassification probabilities ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

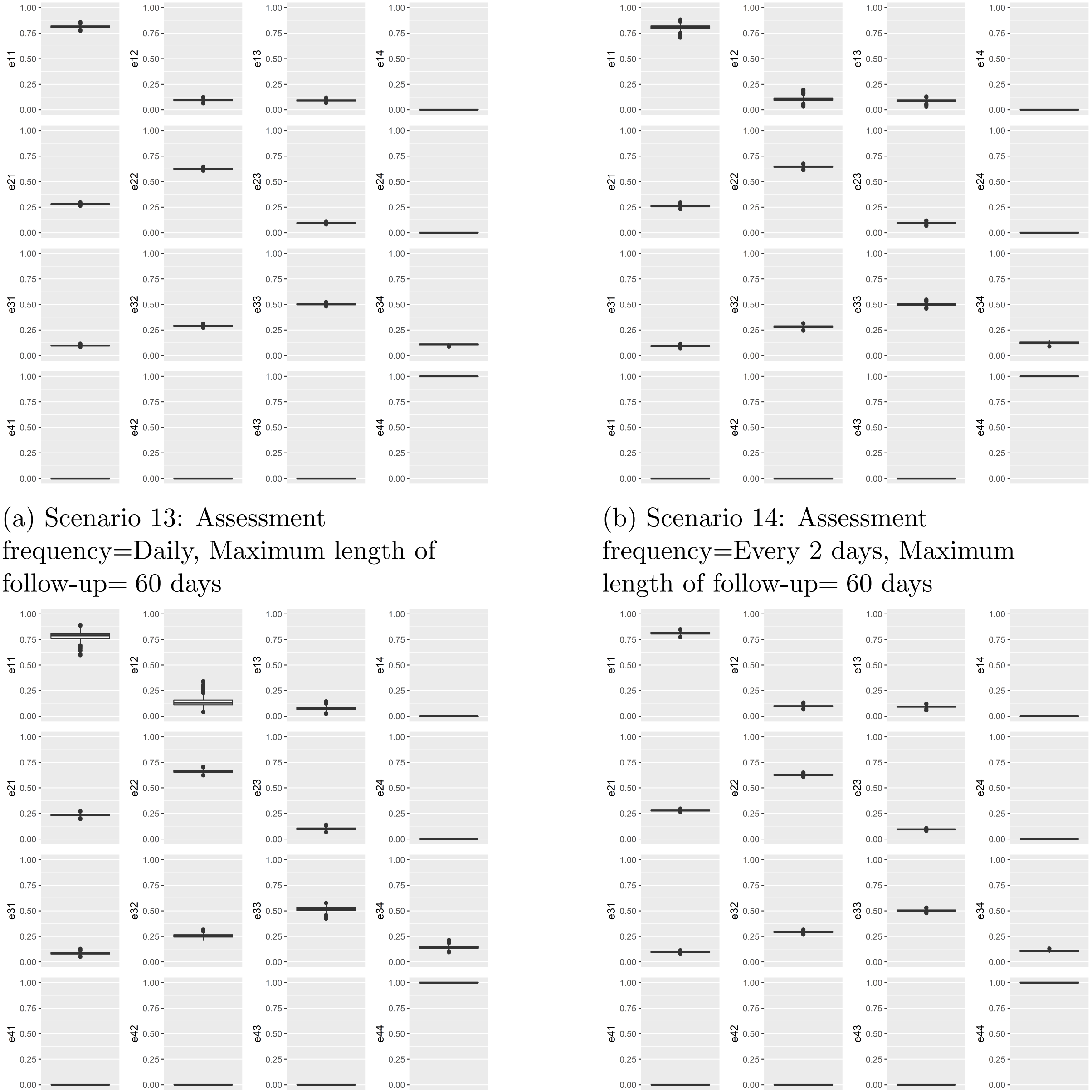


(e) Scenario 11: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days

(f) Scenario 12: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.8: Point estimates for misclassification probabilities ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.3.3 Part I: Misclassification of transient states and over-reporting of the absorbing state (Scenarios 13 to 18)



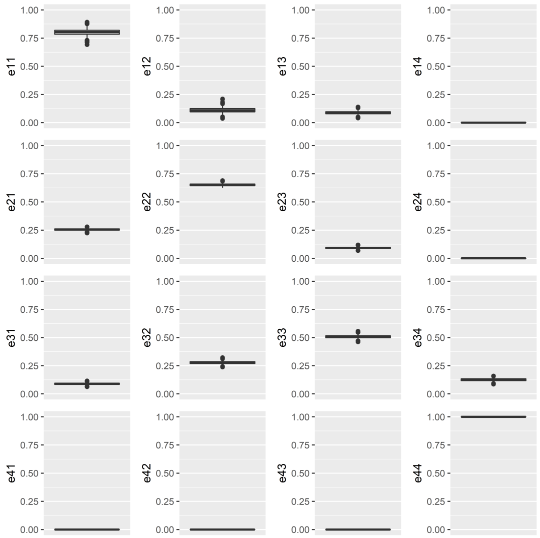
(a) Scenario 13: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 14: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

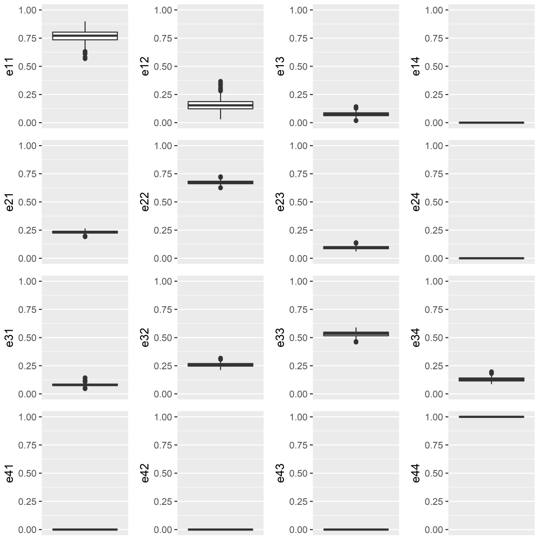
(c) Scenario 15: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

(d) Scenario 16: Assessment frequency=Daily, Maximum length of follow-up= 30 days

Figure D.9: Point estimates for misclassification probabilities ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)



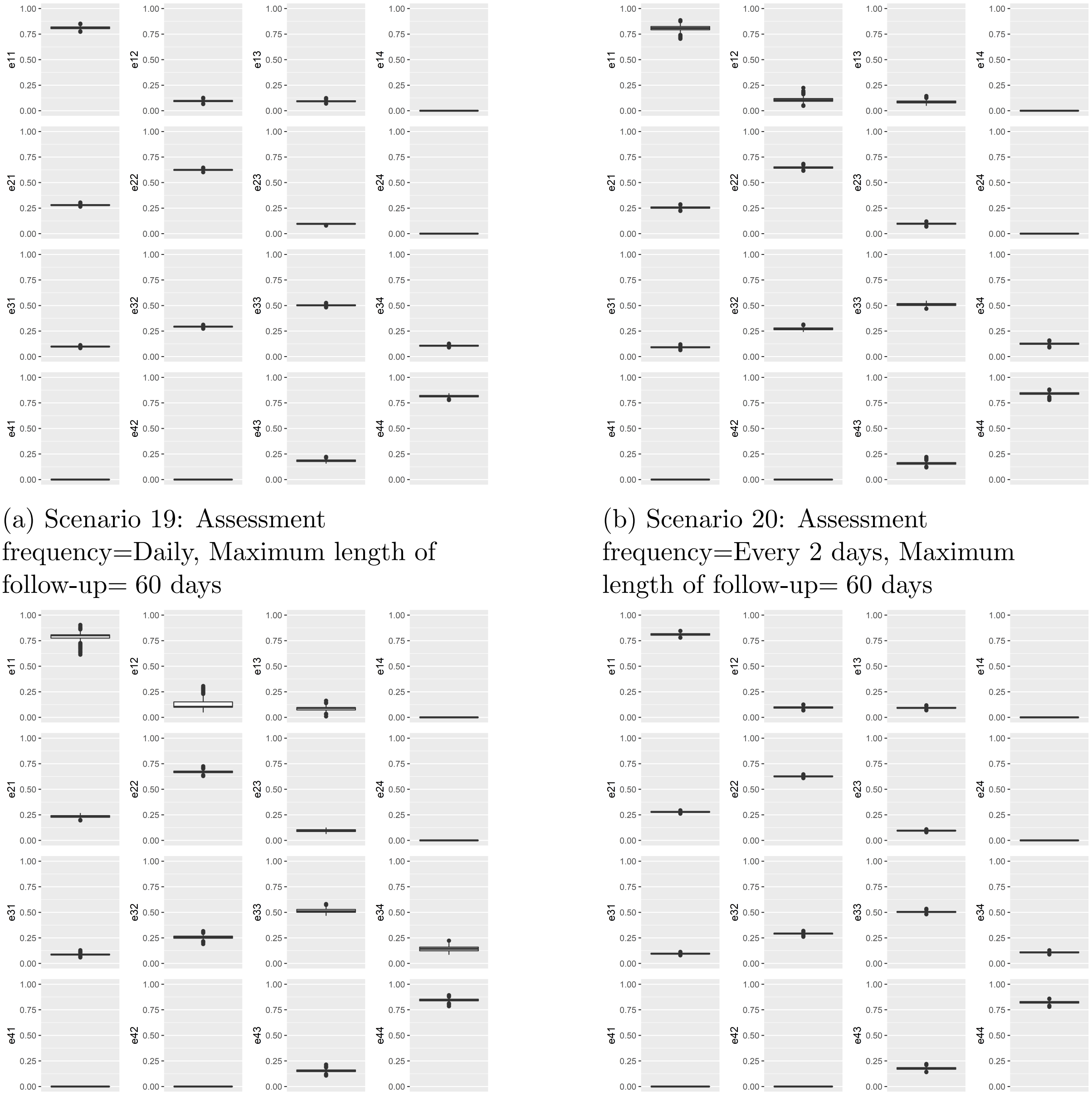
(e) Scenario 17: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days



(f) Scenario 18: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.9: Point estimates for misclassification probabilities ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.3.4 Part I: Misclassification of transient states and both under- and over-reporting of the absorbing state (Scenarios 19 to 24)



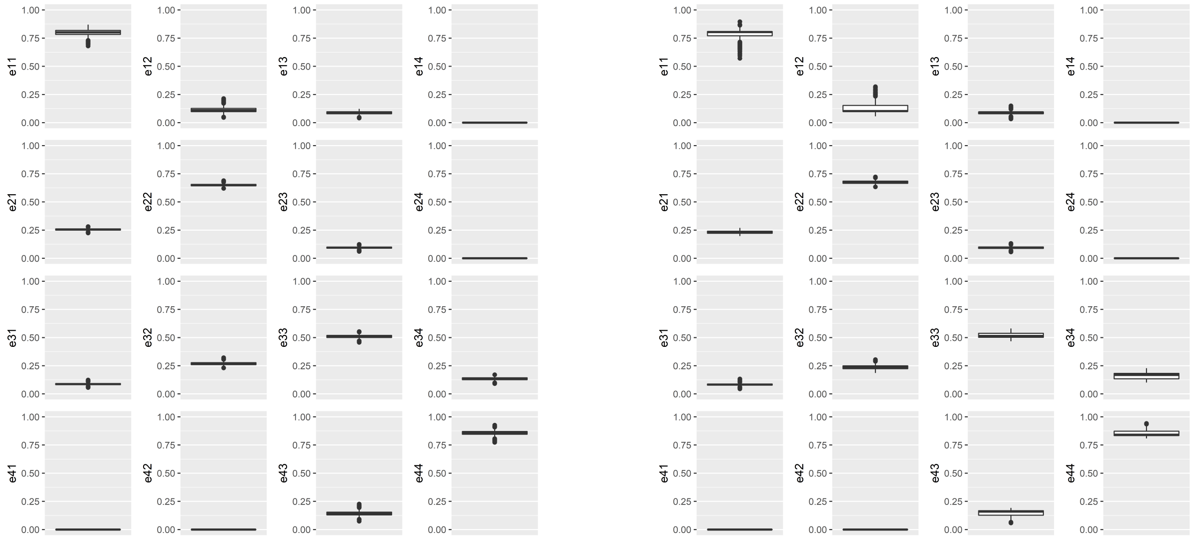
(a) Scenario 19: Assessment frequency=Daily, Maximum length of follow-up= 60 days

(b) Scenario 20: Assessment frequency=Every 2 days, Maximum length of follow-up= 60 days

(c) Scenario 21: Assessment frequency=Every 3 days, Maximum length of follow-up= 60 days

(d) Scenario 22: Assessment frequency=Daily, Maximum length of follow-up= 30 days

Figure D.10: Point estimates for misclassification probabilities (N = 1000, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

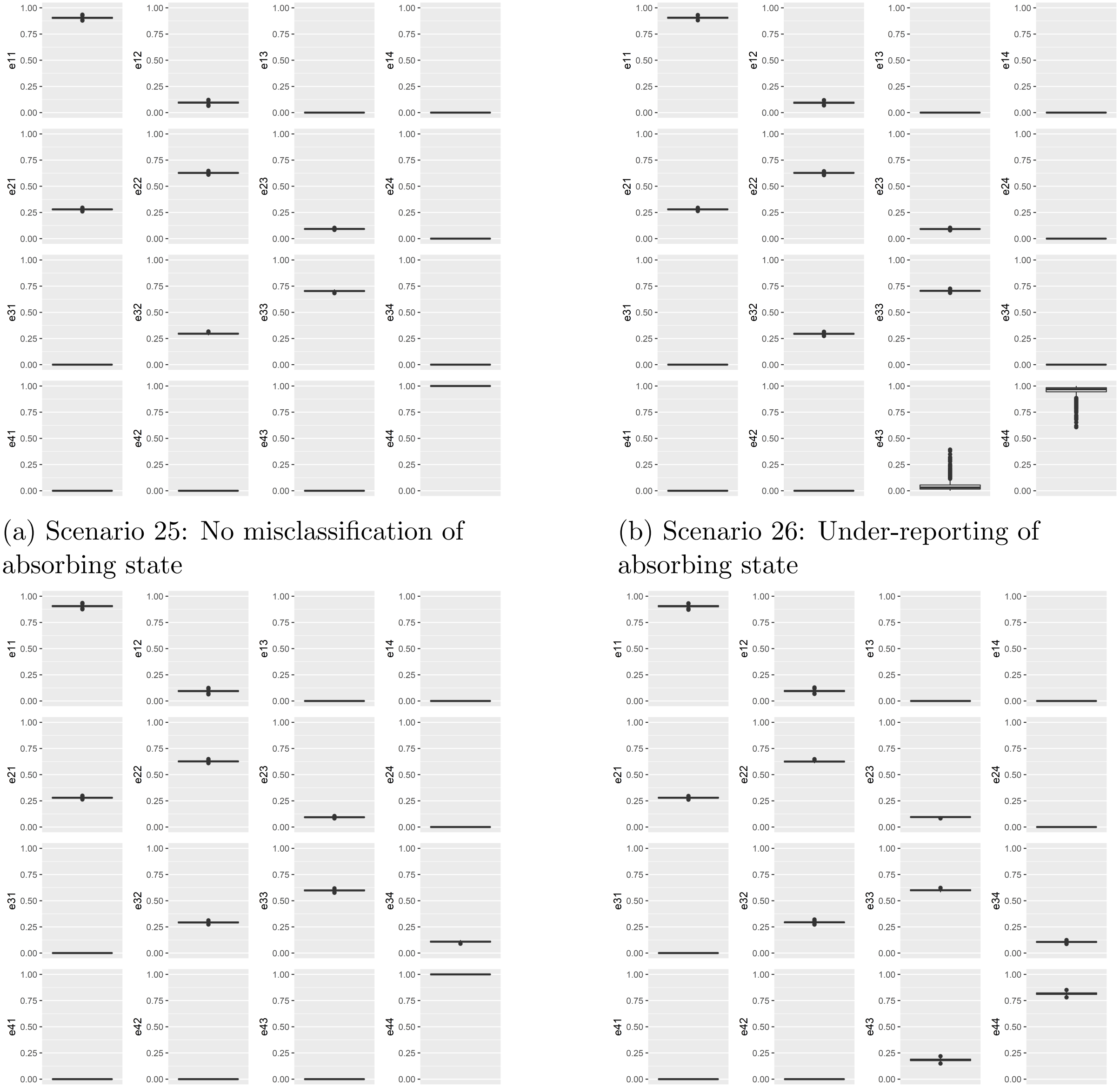


(e) Scenario 23: Assessment frequency=Every 2 days, Maximum length of follow-up= 30 days

(f) Scenario 24: Assessment frequency=Every 3 days, Maximum length of follow-up= 30 days

Figure D.10: Point estimates for misclassification probabilities ($N = 1000$, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.3.5 Part II: Misclassification of adjacent states only (Scenarios 25 to 28)



(a) Scenario 25: No misclassification of absorbing state

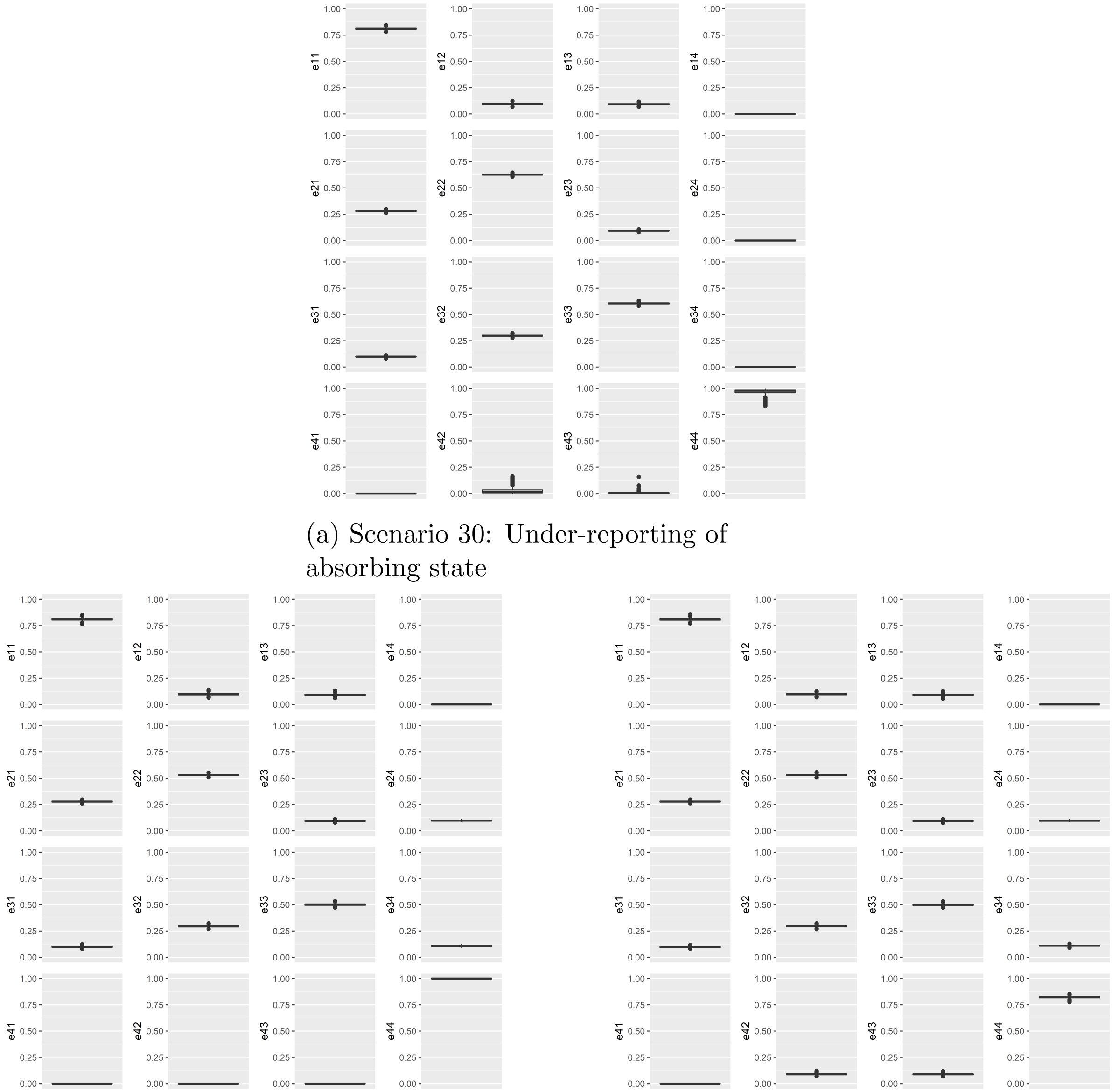
(b) Scenario 26: Under-reporting of absorbing state

(c) Scenario 27: Over-reporting of absorbing state

(d) Scenario 28: Both under- and over-reporting of absorbing state

Figure D.11: Point estimates for misclassification probabilities (N = 1000, Length of follow-up= 60 days, Assessment frequency = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

D.3.6 Part III: Pressure ulcer setting (Scenarios 29 to 31)



(a) Scenario 30: Under-reporting of absorbing state

(b) Scenario 31: Over-reporting of absorbing state

(c) Scenario 32: Both under- and over-reporting of absorbing state

Figure D.12: Point estimates for misclassification probabilities ($N = 1000$, Length of follow-up= 60 days, Assessment frequency = Daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\boldsymbol{q_0} = (0.05, 0.05, 0.03)$)

Appendix E

Chapter 8 Missing data

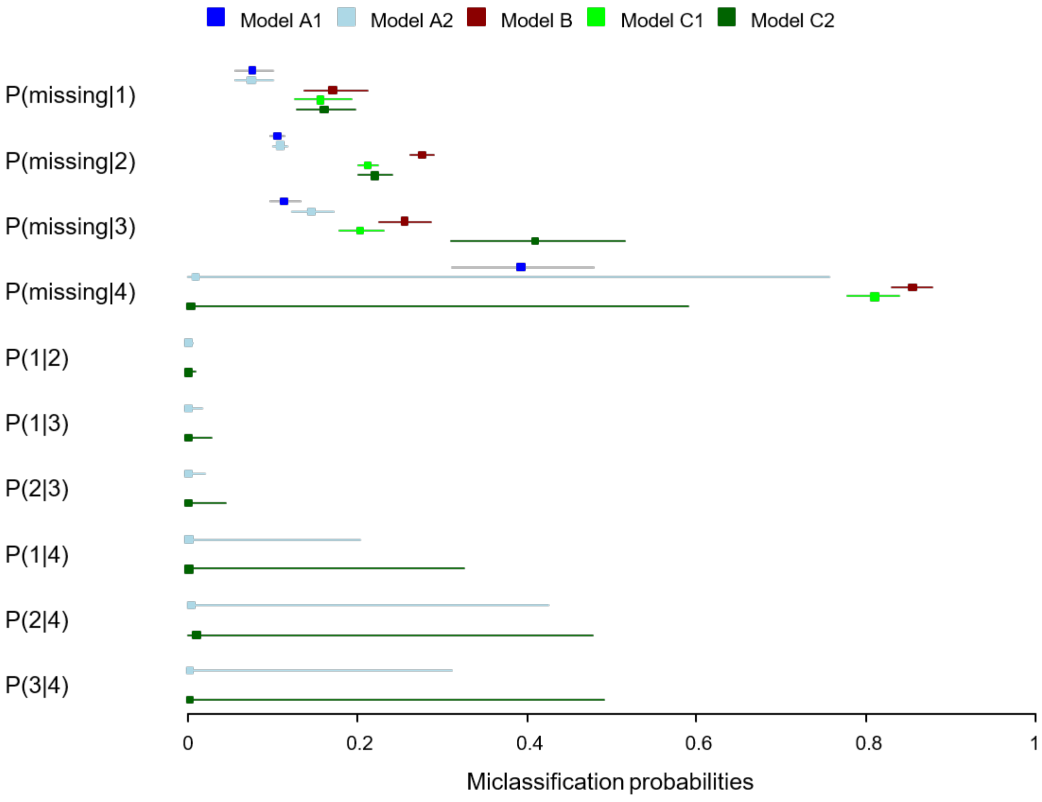


Figure E.1: Forest plot of estimated misclassification probabilities including Models $A2$ and $C2$