UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD thesis

Treatment effect measures for recurrent event endpoints with and without presence of terminal events

Julie Kjærulff Furberg

University supervisor: Per Kragh Andersen
Company supervisor: Henrik Ravn
Company co-supervisor: Trine Saugstrup

Submitted: March 31, 2023

This thesis has been submitted to the Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen

Treatment effect measures for recurrent event endpoints with and without presence of terminal events

PhD thesis

Julie Kjærulff Furberg

Section of Biostatistics
Department of Public Health
Faculty of Health and Medical Sciences
University of Copenhagen
Øster Farimagsgade 5
DK-1014 Copenhagen K

Biostatistics OSCD & Outcomes 1
Biostatistics
Novo Nordisk
Vandtårnsvej 114
DK-2860 Søborg

Advisors
Academic supervisor: Per Kragh Andersen, University of Copenhagen
Industrial supervisors: Henrik Ravn and Trine Saugstrup, Novo Nordisk

Preface

This work was conducted as an Industrial PhD with collaboration between the Section of Biostatistics, University of Copenhagen and Biostatistics OSCD & Outcomes 1, Novo Nordisk. Innovation Fund Denmark and the STAR programme at Novo Nordisk has funded the Industrial PhD project.

I would like to thank my always-inspiring main supervisors, Henrik and Per. The learnings I have obtained from them has helped me grow as a statistician and educated me in "their school" which I will strive to pass on in the future. Thank you to Trine Saugstrup for ensuring smoothness with practical aspects of the project.

Thank you to Section of Biostatistics at the University of Copenhagen for providing me an atmosphere with appreciation of biostatistical rigor and application. It has been a great place to be, especially when working within life history analysis. Thank you to Thomas Scheike for useful discussions. I am grateful for Helene Charlotte Wiese Rytgaard's thorough comments and kind inputs to my dissertation. A large thank you to Novo Nordisk for funding my PhD project and providing several interesting biostatistical topics. I look forward to continuing my career at Novo Nordisk in the company of a great team of colleagues. Thank you to Innovation Fund Denmark for funding my PhD project.

Thank you to my family members and friends for supporting me during the last months in my country side escape with a growing belly. A special thanks to the bundles of joy in my life: Lasse, Theodor, and soon August.

Summary

Recurrent events are events that can happen more than once for a person during their lifetime. Such events occur naturally within disease progression, e.g., recurrent heart failures. For randomised trials, the effect of treatment on recurrent events may be explored. This treatment effect, the estimand, may be formulated in several ways depending on the scientific question of interest.

This thesis discusses various estimands for recurrent events and explores shortcomings and benefits of each. It is recommended to focus on marginal models for recurrent events, such as the proportional means model. This model has a simple interpretation and allows for modelling flexibility. In the presence of terminal events, it is important to also evaluate the effect of treatment on mortality. Thus, the effect of treatment on both recurrent events and death should be explored jointly. A bivariate procedure based on pseudo-observations is proposed in this thesis. This method can model the effect of treatment on recurrent events and death using marginal models simultaneously. The construction of a bivariate procedure may enable more powerful sequential tests compared to exploiting univariate features.

Recurrent event methods may be more powerful than methods based on first events. Moreover, analysing recurrent events can provide new insights into treatment effects. Thus, Novo Nordisk could benefit from having the possibility of including recurrent event analyses as either primary or confirmatory analyses when designing randomised controlled trials. A simulation-based sample size estimation procedure based on the recommended marginal models for recurrent events and deaths is proposed in this thesis. This allows the user to make power estimation of such marginal models when treatment may influence mortality and recurrent events in various ways. Such a tool is key in order to dimension the clinical trials to ensure that claims may also be made on recurrent events.

Finally, the recommended marginal models for recurrent events assume independent censoring. This implies that an unbiased causal treatment effect may not be extracted from these models if this assumption is violated. Discussion on how to modify the marginal models to a situation with dependent censoring is presented. This would be useful for pre-specifying sensitivity analyses for recurrent events which explores the assumption of independent censoring.

In summary, this thesis provides recommendations for how to analyse recurrent events for randomised trials, especially with the additional complication of terminal events. Moreover, it suggests how to design such clinical trials with confirmatory recurrent event endpoints with high mortality rates. Software has been made publicly available for both analysing and planning clinical trials. Finally, relevant sensitivity analyses for recurrent events is proposed. Following these recommendations, ensures that clinically relevant treatment effects may be extracted from data on recurrent events, with or without terminal events. The contributions of this thesis are both methodological and practical.

Resumé

Gentagne hændelser er begivenheder, som kan ske flere gange for et individ i løbet af dets levetid. Sådanne begivenheder opstår naturligt indenfor sygdomsprogression, fx gentagne hjertesvigt. Behandlingseffekten på gentagne hændelser kan undersøges i randomiserede studier. Afhængig af det videnskabelige spørgsmål kan denne behandlingseffekt, estimanden, formuleres på flere måder.

Denne afhandling diskuterer fordele og ulemper ved diverse estimander for gentagne hændelser. Det anbefales at fokusere på marginale modeller for gentagne hændelser såsom modellen for proportionelle middelværdier. Modellen har en simpel fortolkning og tillader fleksibilitet i modelleringen. I tilstedeværelse af terminale hændelser, fx død, skal behandlingseffekten på dødelighed tillige evalueres. Dermed skal behandlingseffekten på gentagne hændelser og terminale hændelser undersøges samtidigt. En to-dimensionel procedure baseret på pseudo-observationer er foreslået i denne afhandling. Metoden kan modellere behandlingseffekten på gentagne hændelser og død samtidigt. Den to-dimensionelle procedure kan tillade sekventielle hypotesetest med større styrke end tests baseret på en-dimensionelle egenskaber.

Metoder for gentagne hændelser kan have større styrke end metoder baseret på de første hændelser. Desuden kan analysen af gentagne hændelser give nyt indblik i behandlingseffekter. Når Novo Nordisk planlægger randomiserede kliniske studier, vil det være nyttigt at kunne inkludere analyser af gentagne hændelser som enten primære eller bekræftende sekundære analyser (dvs. analyser med overordnet type 1 fejls kontrol). Afhandlingen foreslår en procedure til udregningen af stikprøvestørrelse via simulation baseret på de anbefalede marginale modeller for gentagne hændelser og død. Proceduren tillader brugeren at lave en styrkeberegning under disse marginale modeller, når behandling både kan påvirke dødelighed og gentagne hændelser i flere retninger. Et sådant værktøj er vigtigt for at kunne dimensionere de kliniske studier således, at man kan undersøge påstande omkring gentagne hændelser.

De anbefalede marginale modeller for gentagne hændelser er baseret på en antagelse omkring uafhængig censurering. Hvis denne antagelse er overtrådt, kan man ikke regne med at estimere en retvisende behandlingseffekt fra modellerne. Afhandlingen diskuterer måder, hvorpå man kan justere de marginale modeller til situationen med afhængig censurering. Denne udvidelse er brugbar for at kunne planlægge sensitivitetsanalyser for gentagne hændelser, som undersøger antagelsen omkring uafhængig censurering.

Afhandling anbefaler, hvordan man skal analysere gentagne hændelser for randomiserede studier, specielt med den ekstra kompleksitet som introduceres ved tilstedeværelsen af terminale hændelser. Desuden foreslår afhandlingen, hvordan man kan planlægge studier med gentagne hændelser, når dødelighedsraten er høj. Software er gjort offentligt tilgengængeligt for såvel analyse og planlægning af kliniske studier. Endeligt er relevante sensitivitetsanalyser blevet foreslået. Hvis man følger disse anbefalinger, sikrer man at behandlingseffekter med en klinisk relevant fortolkning kan uddrages fra data omkring gentagne hændelser - med eller uden terminale hændelser. Afhandlingens bidrag er både metodologiske og praktiske.

Contents

Abbreviations

ACE Average Causal Effect

AG Andersen and Gill

CHMP Committee for Medicinal Products for Human Use

EM Expectation-Maximization

EMA European Medicines Agency

FDA Food and Drug Administration

GEE Generalised Estimating Equations

GL Ghosh and Lin

ICH International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

IPCW Inverse Probability of Censoring Weights

IPSW Inverse Probability of Survival Weights

IPTW Inverse Probability of Treatment Weights

LWYY Lin, Wei, Yang, and Ying

MACE Major Adverse Cardiovascular Events

PWP Prentice, Williams, and Peterson

RCT Randomised Controlled Trial

RMST Restricted Mean Survival Time

WLW Wei, Lin, and Weissfeld

Notation

dt An instantaneously small time interval

t^- A time that is infinitesimally smaller than time t

$E_P(X)$ The expectation of the random variable X under the distribution P: $E_P(X) = \int X \, dP$

$\mathcal{H}(t)$ The state occupation history at time t

I The identity matrix

$I(X \in A)$ The indicator function, $I(\cdot)$, where $I(X \in A) = 1$ if $X \in A$ and 0 otherwise

$\overset{as}{\sim}$ Asymptotic distribution as $n \to \infty$

x^T The transposed of the vector or matrix, x

Chapter 1

Introduction and overview

The analysis of recurrent events is a field of increasing interest for clinicians and pharmaceutical companies. Recurrent events are events that may happen more than once for an individual during the course of their life, such as hospitalisations. In randomised controlled trials, information on such repeated events is often collected. However, due to convention and convenience, it happens that only first events are considered for statistical analysis. Cox's proportional hazards model has become a standard tool in the statistical toolbox for analysing time-to-first events within survival analysis (Cox (1972)). The effect of a randomised treatment on a pre-specified survival outcome may be addressed using this model. Unless the considered outcome is terminal, e.g., death, subsequent events could occur for an individual. From a clinical point-of-view, it is of interest to understand the effect of treatment on, not only, the first event but also subsequent events. This would capture another treatment effect and perhaps address the total burden of a disease better. When events are in fact recurrent, discarding events beyond first events is an insufficient use of data. Thus, from an efficiency and clinical perspective, analysing recurrent events is valuable. As it turns out, the statistical analysis of recurrent events is more complicated than the analysis of first events. Moreover, if mortality rates are high, the complexity is increased.

Novo Nordisk collects life history data in several large randomised controlled trials. Both the clinical and statistical society within heart failure are discussing whether it is *"time to move on from 'time-to-first'"*, that is, consideration of recurrent events as opposed to first events (Anker and McMurray (2012)). Such a transition requires both clinicians, sponsors and regulators to develop their understanding of which recurrent event endpoints may be of interest and how these should be analysed. Hence, from Novo Nordisk's viewpoint it will be key to develop an understanding and methodology within this field. The aim of this thesis is to discuss and suggest clinically relevant treatment effect measures for recurrent event endpoints with and without the presence of terminal events.

1.1 Motivation

In 2016, the Committee for Medicinal Products for Human Use (CHMP) within European Medicines Agency (EMA) issued a reflection paper on the assessment of cardiovascular risk of medicinal products for the treatment of cardiovascular and metabolic diseases (EMA (2016)). This states that the approval of certain medicinal products within cardiovascular and metabolic diseases (e.g., diabetes and obesity) requires an in-depth evaluation of cardiovascular safety. A possibility is to design a ded-

icated cardiovascular outcomes trial which can demonstrate cardiovascular safety. They state that the preferred cardiovascular outcome is the *"composite of all major cardiovascular events (MACE); i.e., cardiovascular death, non-fatal myocardial infarction, and non-fatal stroke"*. Specifically, they require that the upper limit of the confidence interval for the hazard ratio is below 1.8 for a comparison between an active treatment and placebo (or standard of care). This wording clearly suggests that the sponsor should consider the timing until first events which could be analysed using a Cox model. EMA states that additional outcomes, like hospitalisation for cardiovascular causes, could also be included in a "MACE plus" endpoint. However, such endpoints would always be presented separately as supportive analyses. Consequently, the main focus for analysing cardiovascular data has been to use approaches within traditional survival analysis that targets first events. It is known that statistical analysis of individual MACE components, such as non-fatal stroke, introduces the problem of competing risks while only considering first events (Andersen, Geskus, et al. (2012)). In 2008, the Food and Drug Administration (FDA) released a similar recommendation for the assessment of cardiovascular risk associated with treatment of type 2 diabetes with respect to bounds placed on the hazard ratios. This recommendation was withdrawn in 2020 and replaced with a newer and less specific draft guideline on the assessment of cardiovascular risk (FDA (2020)). FDA (2019) has issued a draft guidance on relevant endpoints for the treatment of heart failure. Both mortality and morbidity endpoints are of interest. In this document, both number of hospitalisations for heart failure and time to recurrent hospitalisation qualify as relevant morbidity endpoints. This highlights the acceptance of recurrent event endpoints by the agency.

During recent years, recurrent events has gained focus for the analysis of life history data collected in clinical trials (Anker and McMurray (2012)). This would ensure usage of all collected data and may estimate other treatment effects than those for first events. Moreover, the analysis of recurrent events may characterise the total burden of a disease more accurately from a patient perspective. Several large outcome trials with primary time-to-event endpoints have been re-analysed with a focus on recurrent events (e.g., PARADIGM-HF (2014) (2018), CHARM-Preserved (2014)). However, the treatment effect of interest for analysis of recurrent events lacks clarity in these papers. This is illustrated by the use of forest plots to illustrate parameter estimates from different models that target different (and non-compatible) parameters. In addition, the recent clinical trial PARAGON-HF (2019) was designed with primary recurrent event endpoint; composite of total (first and recurrent) hospitalizations for heart failure and death from cardiovascular causes. This illustrates the increased interest for recurrent events.

In a Qualification Opinion (EMA (2020b)), EMA expressed the regulatory interest in recurrent events. It is mentioned that experience with recurrent events with high mortality rates is more limited. Moreover, they highlight the methodological and clinical challenges in adequately describing the effect of treatment on recurrent events if treatment affects mortality. However, they agree that statistical efficiency may be gained if data on recurrent events can be used. In conclusion, they argue that methods which capture a clinically relevant treatment effect is worth while for recurrent events with terminal events but emphasise the present methodological challenges.

Novo Nordisk is conducting several large outcome trials where the endpoints of interest are time-to-event outcomes. Examples include the randomised controlled trials: SELECT, SOUL, and FLOW. SELECT investigates the effect of the drug semaglutide versus placebo on cardiovascular outcomes in people with obesity and established cardiovascular disease. SOUL also investigates the drug semaglutide versus placebo on cardiovascular outcomes in people with type 2 diabetes and established cardiovascular disease. Whereas FLOW investigates semaglutide versus placebo on renal outcomes in people with type 2 diabetes and chronic kidney disease. The primary endpoints in these trials are time-to-first events, but many of the components can occur several times per in-

dividual. For example, time-to-first non-fatal myocardial infarction is pre-specified for analysis in SELECT. Clearly, this endpoint is recurrent in nature and each individual may have more than one event. Analyses of recurrent myocardial infarction may highlight new aspects of disease progression and treatment benefits. Such analyses may be used for regulatory discussion, labelling claims, and additional exploratory analyses for publication.

Further, potential power gains with recurrent events compared to first events has been explored and discussed within the scientific community (Claggett et al. (2018), Fritsch et al. (2021)). Using recurrent events as opposed to first events can result in lower sample sizes. For large and expensive clinical trials, a sponsor has an interest in reducing costs as much as possible. This could be offered by analysing recurrent events instead. Moreover, from an ethical point-of-view, the sample size should be as small as possible (EMA (1998)). However, focusing on recurrent events with or without terminal events adds complexity to the clinical question and statistical modelling. Both regulators and practitioners are exploring various options without any go-to solution. The concept of estimands proposes a framework to describe treatment effects clearly (EMA (2020a)). This thesis will discuss and suggest estimands as well as analysis methods suitable for such scenarios. These analyses should provide treatment effect measures, estimands, on recurrent event endpoints which has a clinical interpretation both with and without terminal events. Ideally, it should be simple for both statisticians, regulators, general practitioners, and patients to understand what treatment effect such analyses describe. Emphasis will be placed on inference on treatment effects drawn from randomised controlled trials. Extensions to observational data will be discussed briefly.

1.2 Objectives

The overall objective of this PhD project is to investigate, develop, and implement statistical methods with clinically interpretable treatment effect measures for the analysis of recurrent event endpoints with and without presence of terminal events.

In detail, the project has the following objectives:

1. Development and characterisation of statistical methods for recurrent event endpoints, particularly in the presence of terminal events
2. Recommendation of statistical methods for recurrent event endpoints in clinical trials with a focus of clinically interpretable treatment effect measures, i.e. estimands
3. Development of statistical methods for recurrent events that are used in statistical analysis plans for Novo Nordisk's randomised controlled trials
4. Implementation of new statistical methods for recurrent events in statistical software that is made publicly available

These objectives will be met through the research conducted in the relation to the manuscripts included in this thesis.

1.3 Overview of thesis

The thesis consists of four manuscripts and an `R`-package with the overall aim of investigating and developing methods for analysing recurrent events with or without presence of terminal events. The following manuscripts are included in the thesis,

Manuscript I *Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER*

Manuscript II *Bivariate pseudo-observations for recurrent event analysis with terminal events*

Manuscript IIx `R`-package 'recurrentpseudo': *Creates Pseudo-Observations and Analysis for Recurrent Event Data*

Manuscript III *Simulation-based sample size calculations for recurrent events with competing risks*

Manuscript IV *Marginal models for recurrent events under covariate dependent censoring*

All manuscripts describe various aspects of obtaining clinically relevant treatment effects for analysing recurrent events with or without competing risks in the setting of randomised controlled trials. Manuscript IV discusses dependent censoring in the realm of randomised trials, but some aspects could be extended to observational data. The publication status of each manuscript is displayed in Table 1.1.

Manuscript	Publication status
I	Published in *Pharmaceutical Statistics* (January 2022)
II	Published in *Lifetime Data Analysis* (April 2023)
IIx	Published on `CRAN` (September 2022)
III	Under revision for *Pharmaceutical Statistics* (March 2023)
IV	In preparation

Table 1.1: Publication status of each manuscript.

The thesis consists of seven chapters followed by the manuscripts. Chapters 2-4 introduce key concepts for the analysis of recurrent events both with and without terminal events. Chapter 5 introduces pseudo-observations within survival analysis and recurrent events. Chapter 6 discusses various important aspects for planning randomised controlled trials with confirmatory recurrent event endpoints. Chapter 7 introduces causality and addresses analysis of recurrent events under dependent censoring. Table 1.2 outlines the overall structure of the thesis.

Moreover, the following describes the background for each manuscript,

Manuscript I The background is found in Chapters 2-4

Manuscript II(x) The background is found in Chapters 2-5

Manuscript III The background is found in Chapters 2-4, and 6

Manuscript IV The background is found in Chapters 2-5, and 7

Chapters 2-4 discuss and characterise current statistical methods for recurrent events with or without

terminal events. In relation to these chapters, recommendations of statistical methods for analysing recurrent events in randomised trials that ensures a clinically relevant treatment effect will be made (objectives 1-2). The development of new statistical methods for recurrent events that can be used in Novo Nordisk's clinical trials is covered by the Manuscripts II-IV, which is introduced in Chapters 5-7 (objective 3). In relation to the new statistical methods for recurrent events, software has been made available on `CRAN` and `GitHub` (objective 4). Vignettes which describe the content of the `R` software are available on `GitHub` and has been included after the relevant manuscripts. The focus in this thesis will be inference drawn from randomised trials. Extensions to observational data will be introduced in Chapter 7.

Chapter	Content
2	introduces multi-state models for survival data and recurrent events with or without competing risks.
3	discusses relevant regression models for modelling recurrent events with or without competing risks.
4	explores conditional and marginal parameters of interest for quantifying a treatment effect from randomised controlled trials.
5	introduces pseudo-observation within survival analysis. Relevant parameters to compute pseudo-observations on based recurrent events are discussed.
6	discusses general concepts for planning clinical trials with a confirmatory recurrent event endpoint.
7	introduces causality and issues related to dependent censoring. Marginal models for recurrent events are discussed under dependent censoring.

Table 1.2: Overall structure of thesis.

Chapter 2

Multi-state models

Multi-state models provide a general framework for describing and analysing life history data. Life history data governs information on individuals collected over time. In the analysis of life history data, e.g., survival analysis, the waiting time until occurrence of a specific event is often of interest. This data collection, which observes and records events for individuals over time, is characterised by *right-censoring*. Right-censoring occurs when the observed individuals do not have the event of interest within the period of observation. Thus, all that is known, is that the individual did not have the event of interest. Due to this special feature, the field of survival analysis emerged, which handles censoring by focusing on transition intensities.

The type of life history data that we are interested in, is collected on individuals in (almost) continuous time. Usually, event information is available in days. Thus, attention is restricted to multi-state models in continuous time as opposed to discrete time as this resembles the continuous collection.

The multi-state models discussed in this chapter provides the overall conceptual and mathematical background for Manuscripts I, II(x), III and IV. Knowledge of general inference from survival, competing risks, and recurrent events data is required to read these manuscripts. This chapter focuses on visualizing the concepts of interest, formulating relevant quantities as well as non-parametric inference using multi-state models. Regression models will be discussed in Chapter 3. Treatment effects for randomised trials will be elaborated in Chapter 4.

Cook and Lawless (2018) provide a comprehensive introduction to multi-state models as well as detailed modelling considerations. A multi-state process in continuous time is a stochastic process $X(t)$, $t \in [0, \tau]$, $\tau < \infty$ with a finite state space $S = \{0, \ldots, K\}$. The occupied state at time t is represented by $X(t)$. The *transition intensity function* between states $\{k, l\} \in S$, $k \neq l$, is defined as

$$\lambda_{kl}(t \mid \mathcal{H}(t^-)) = \lim_{dt \to 0} \frac{P(X(t + dt) = l \mid X(t) = k, \mathcal{H}(t^-))}{dt}. \tag{2.1}$$

Here t^- denotes a time that is infinitesimally smaller than time t. The state occupation history over $[0, t)$ is represented by $\mathcal{H}(t^-)$, where $\mathcal{H}(t) = \{X(s), 0 \leq s \leq t\}$. The state occupation history for individual i is given by $\mathcal{H}_i(t) = \{X_i(s), 0 \leq s \leq t\}$, which describes the states that individual $i = 1, \ldots, n$ has occupied. Compressed notation will be used and the i's will often be suppressed in the notation.

A *Markov* multi-state process satisfies that $\lambda_{kl}(t \mid \mathcal{H}(t^-)) = \lambda_{kl}(t)$. In other words, this implies that the dependence on the history is given only through the current state. This corresponds to so-called *total time models*: evaluating the time since, e.g., randomisation.

For a *semi-Markov* multi-state process, it holds that $\lambda_{kl}(t \mid \mathcal{H}(t^-)) = h_{kl}(B(t))$, where $h_{kl}(t)$ denotes the transition intensity function and $B(t)$ denotes the time since entry into the current state k. That is, the dependence on the history depends alone on the duration in the current state. This corresponds to so-called *gap time models*: modelling the duration of the stay in a given state.

For Markov models, the transition probabilities may be elegantly presented using product integration (Cook and Lawless (2018)). The $K \times K$ *transition probability matrix*, $P(s,t)$, may be obtained by product integration as

$$P(s,t) = \prod_{[s,t)} \{I + Q(u)\, du\}, \tag{2.2}$$

where $Q(t)$ denotes the $K \times K$ matrix with $\lambda_{kl}(t)$ in the off-diagonal entries ($k \neq l$) and $-\sum_{k \neq l} \lambda_{kl}(t)$ in the diagonal entries ($k = l$). In $P(s,t)$, the (k,l)'th element is $P(X(t) = l \mid X(s) = k)$.

Data for the multi-state process is characterised by the presence of right-censoring. We distinguish between *independent* and *dependent* censoring. Independent censoring occurs when the censoring process is unrelated to the event (or failure) process of interest. Administrative censoring occurring due to study closure is typically perceived as being independent. On the other hand, censoring is said to be dependent if censoring is related to the event process in some way. An example would be if individuals tend to drop out of the study for reasons that are related to the failure process.

2.1 Multi-state models for survival data

First, multi-state models are illustrated using traditional survival data, e.g., time-to-death data. Let the survival time be given by D^*. Due to right-censoring, D^* is incompletely observed, since either a censoring or survival time is observed. To that end, the censoring indicator is defined as $\delta = I(D^* \leq C)$. Moreover, let $D = D^* \wedge C$. Figure 2.1 displays the relevant multi-state model with the states 'alive' and 'dead'. The instantaneous probability of death, $\lambda_{01}(t)$, may be considered. Here,

$$\lambda_{01}(t) = \lim_{dt \to 0} \frac{P(X(t+dt) = 1 \mid X(t) = 0)}{dt}.$$

The *cumulative hazard* of dying is given by

$$\Lambda_{01}(t) = \int_0^t \lambda_{01}(u)\, du.$$

The probability of surviving past time t, the *survival function*, is defined by,

$$S(t) = P(D^* > t) = \exp(-\Lambda_{01}(t)).$$

The cumulative hazard of death may be estimated using the non-parametric Nelson-Aalen estimator (Aalen (1978), Nelson (1969), Nelson (1972)). This estimator is given by,

$$\widehat{\Lambda}_{01}(t) = \int_0^t \frac{1}{Y(u)}\, dN(u),$$

where $Y(t) = I(t \leq D^*)$ denotes the at-risk process and $N(t) = I(D^* \leq t)$ denotes the event process. See Andersen, Borgan, et al. (1993) for further details on counting processes and stochastic integration. The Nelson-Aalen estimator is a step function with jumps at each observed failure time. Under independent censoring, the Nelson-Aalen estimator is an asymptotically unbiased estimator of $\Lambda_{01}(t)$ (Andersen, Borgan, et al. (1993), Martinussen and Scheike (2006)). Similarly, a non-parametric estimator, the Kaplan-Meier estimator, exists for estimation of the survival probabilities (Kaplan and Meier (1958)). The Kaplan-Meier estimator is given by,

$$\widehat{S}(t) = \prod_{u \leq t} \left(1 - d\widehat{\Lambda}_{01}(u)\right) = \prod_{u \leq t} \left(1 - \frac{dN(u)}{Y(u)}\right).$$

The Kaplan-Meier estimator is also a step function with jumps at the observed failure times. This estimator, $\widehat{S}(t)$, is an asymptotically unbiased estimator of the survival probability, $S(t)$, under independent censoring (Andersen, Borgan, et al. (1993)).

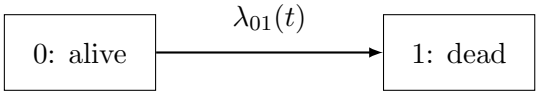


Figure 2.1: Multi-state model for a two-state survival model.

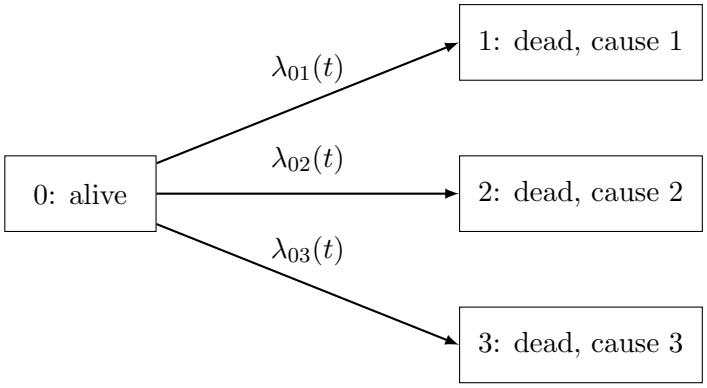


Figure 2.2: Multi-state model for a competing risks model.

In some situations, it may be possible to die from more than one cause. The causes of death are 'competing'. Such data is within the realm of *competing risks*. Now, let $\Delta = \{1, 2, \ldots\}$ be a cause-of-death indicator, which indicates the cause of death (cause 1, cause 2, and so forth). Figure 2.2 represents a competing risks setting with three competing causes of death. The *cause-specific hazards* (intensities) are given by $\lambda_{01}(t)$, $\lambda_{02}(t)$ and $\lambda_{03}(t)$, with

$$\lambda_{0h}(t) = \lim_{dt \to 0} \frac{P(X(t + dt) = h \mid X(t) = 0)}{dt}, \quad \text{for} \quad h \in \Delta = \{1, 2, 3\}.$$

The *cause-specific cumulative hazard* for cause h is given by,

$$\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(u)\, du.$$

The overall probability of being alive at time t, the survival probability, is given by

$$S(t) = \exp\left(-\int_0^t \sum_h \lambda_{0h}(u)\, du\right).$$

The overall survival probability may still be estimated using the Kaplan-Meier estimator while considering all causes of death as a failure. The probability of dying from cause h before time t, the *cumulative incidence*, is given by

$$F_h(t) = \int_0^t S(u)\lambda_{0h}(u)\, du.$$

As seen, the cumulative incidence depends on the overall survival probability and the cause-specific hazard. Hence, the probability of dying of cause h prior to time t depends both on; how likely it is that the individual has survived until time t and the hazard of dying of cause h. Thus, the cumulative incidence depends on all cause-specific hazard functions. Consequently, the one-to-one relationship between rates and risks is lost when moving away from studying all-cause mortality, as discussed in Andersen, Geskus, et al. (2012). The Aalen-Johansen estimator is a non-parametric estimator of the cumulative incidence function (Aalen and Johansen (1978)). This estimator is given by,

$$\widehat{F}_h(t) = \int_0^t \widehat{S}(u)\, d\widehat{\Lambda}_{0h}(u).$$

Large sample behaviour of this estimator is discussed in Andersen, Borgan, et al. (1993). Specifically, it is asymptotically unbiased and the variance can be consistently estimated under independent censoring. Technically, the cause-specific cumulative hazard for cause h, $\Lambda_{0h}(t)$, may be estimated by considering events of cause h and censoring for the competing causes using the Nelson-Aalen estimator.

2.2 Multi-state models for recurrent events

Recurrent events are events that can happen several times during an individual's life. Examples could include hospitalisation or bruises. Such events are collected for all individuals during the period of observation. These events are still characterised by the presence of right-censoring. Figure 2.3 illustrates recurrent events and right-censoring for six individuals.

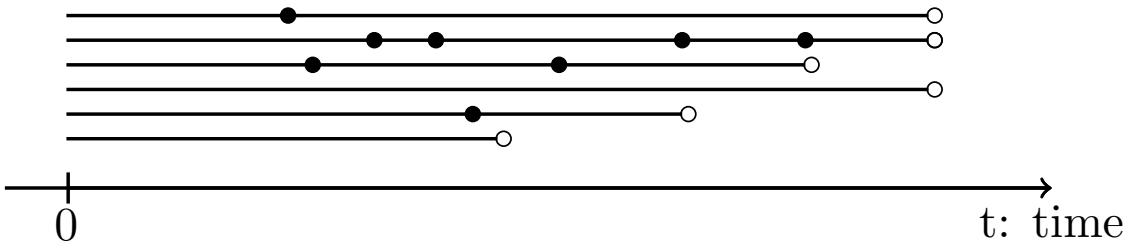


Figure 2.3: Illustration of recurrent events for six individuals. The filled and unfilled circles represent events and censoring, respectively.

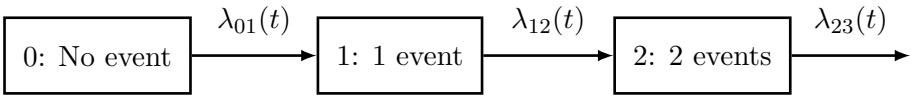


Figure 2.4: A recurrent event process.

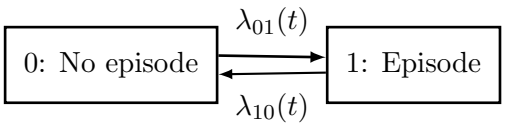


Figure 2.5: A recurrent event process with episodes (risk-free periods).

After experiencing such events, individuals may or may not be at risk of experiencing an event immediately after. For example, if subjects are hospitalised, they may not be at risk of being hospitalised again until admission from the hospital. Here, there is 'gaps' between at-risk periods. We will denote such events by 'episodes' to indicate the risk-free periods. If instead bruises were considered, subject may be at risk of getting a new bruise shortly after experiencing the first bruise. In this scenario, there would be no 'gaps' between the at-risk periods. Figure 2.4 illustrates a multi-state model for recurrent events with no 'gaps' between the at-risk periods. Alternatively, Figure 2.5 illustrates a multi-state model for recurrent events with 'gaps' between the at-risk periods.

At times, recurrent events such as hospitalisations occur in a population with a high mortality rate. This could happen if a very ill population is followed over time. Figure 2.6 illustrates such a situation for six individuals. When mortality is non-negligible, both a recurrent event process and death process may be of interest in a multi-state model. Figure 2.7 illustrates a multi-state model for recurrent events, terminal events, and no 'gaps' between at-risk periods. Figure 2.8 illustrates a

multi-state model for recurrent events, terminal events, and 'gaps' between at-risk periods.

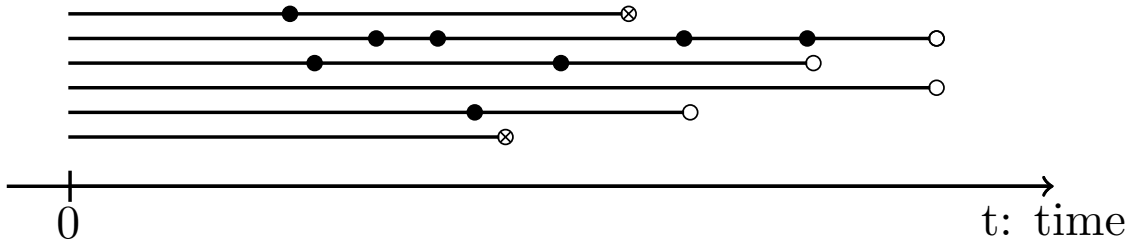


Figure 2.6: Illustration of recurrent events and deaths for six individuals. The filled, unfilled, and crossed-out circles represent events, censoring, and deaths, respectively.

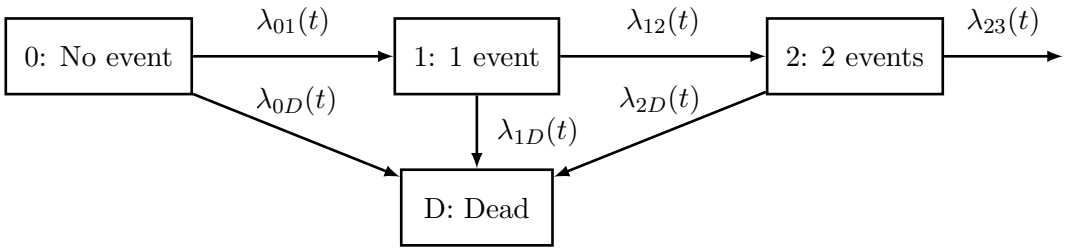


Figure 2.7: A recurrent event process with a terminal event.

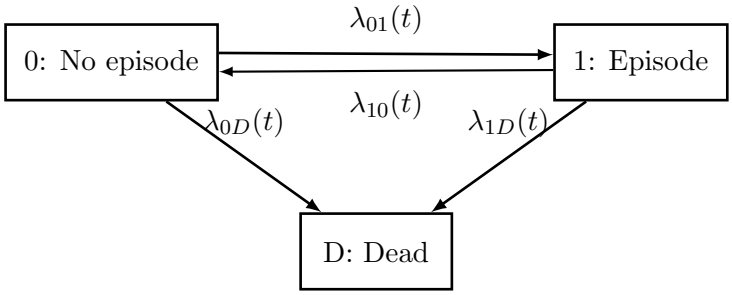


Figure 2.8: A recurrent event process with a terminal event and with episodes (risk-free periods).

The number of recurrent events by time t, $N^*(t)$, could be an interesting quantity when focusing on recurrent events. Figure 2.9 displays an example of the behaviour of $N^*(t)$ and $X(t)$ when a terminal event exists, and (a) subjects are immediately at risk of experiencing a new event or (b) when subjects have periods where they are not at risk of experiencing the event of interest. This corresponds to the multi-state models in Figures 2.7 and 2.8 (either risk-free periods or not). Here, $X(t) \in S = \{0, 1, 2, 3, 4\}$ and the states $0, 1, 2, 3, 4$ represents 'No event', '1 event', '2 events', '3 events', and 'Death', respectively, for situation (a). Conversely, $X(t) \in S = \{0, 1, 2\}$ and the states $0, 1, 2$ represents 'No episode', 'Episode', and 'Death', respectively, for situation (b). The expected

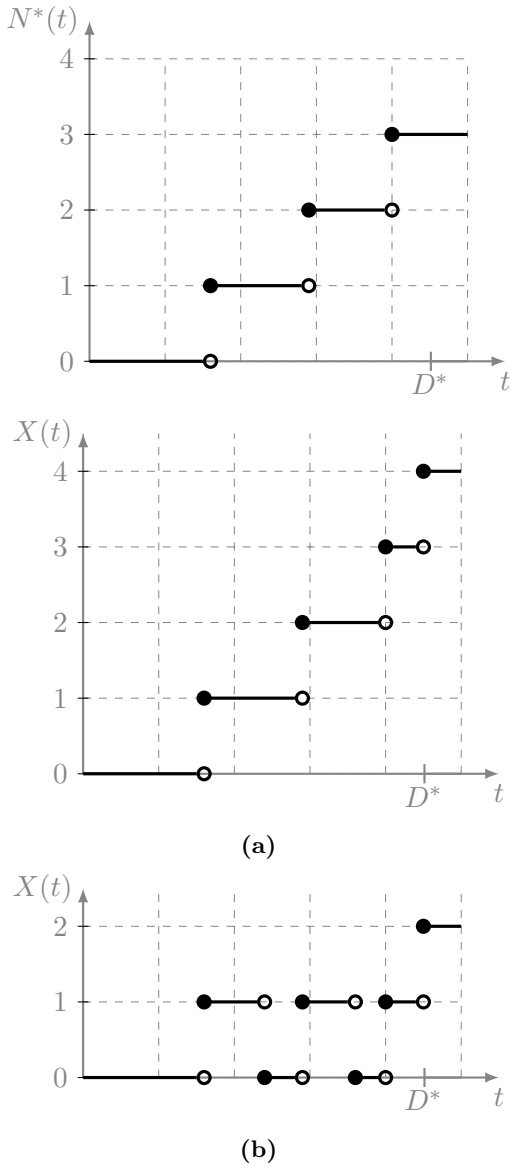


Figure 2.9: Examples of $N^*(t)$ and $X(t)$ processes for a recurrent event process with a terminal event, D^*, and (a) no gaps between at-risk periods or (b) gaps between at-risk periods.

number of recurrent events is given by

$$\mu(t) = E(N^*(t)) = \int_0^t dE(N^*(u)).$$

When there are no terminal events, the Nelson-Aalen estimator may be used to estimate the marginal mean function as suggested by Lawless and Nadeau (1995). That is, the marginal mean may be

non-parametrically estimated by

$$\widehat{\mu}(t) = \int_0^t \frac{1}{Y(u)} dN^*(u).$$

The marginal mean estimator is a step function with jumps at the observed recurrent event times. This estimator will be an asymptotically unbiased estimator of $\mu(t)$ under independent censoring (Lawless and Nadeau (1995)).

Cook and Lawless (1997) argued that the marginal mean should be calculated differently in the presence of terminal events. Specifically, it holds that,

$$\mu(t) = E(N^*(t)) = \int_0^t P(D^* > u) E(dN^*(u) \mid D^* \geq u). \tag{2.3}$$

The Nelson-Aalen estimator will be an upwards biased estimator of $\mu(t)$ if there is terminal events (Cook and Lawless (1997), Ghosh and Lin (2000)). Instead, Cook and Lawless, subsequented by Ghosh and Lin, suggest considering the plug-in estimator,

$$\widehat{\mu}(t) = E(N^*(t)) = \int_0^t \widehat{S}(u)\, d\widehat{R}(u \mid D^* \geq u), \tag{2.4}$$

where $\widehat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$ and $\widehat{R}(t \mid D^* \geq t)$ is the Nelson-Aalen estimator of $R(t \mid D^* \geq t)$. Technically, $\widehat{R}(t \mid D^* \geq t)$ is estimated by considering the occurrence of recurrent events and censoring for death using the Nelson-Aalen estimator. From equation (2.3), it is apparent that the expected number of recurrent events with terminal events will be mixture of; the probability of being alive and the expected number of recurrent events while alive. Two-sample tests and large sample properties were discussed by Ghosh and Lin (2000). The loss of the one-to-one correspondence between rates and risks for general competing risks data was discussed in Section 2.1. Analogously, this correspondence is also lost for relationship between the marginal mean and the rates, as seen through equation (2.3). In summation, it will be important to consider both rates of recurrent events and mortality when analysing recurrent events with terminal events.

Other marginal features of $X(t)$ than $N^*(t)$ may be of interest when considering the recurrent event process. Examples could include the *state occupation probability* in a given state, e.g., $P(X(t) = 1)$, which corresponds to the probability of being in the 'Episode' state at time t for the multi-state model in Figure 2.5. The state occupation probabilities are given by,

$$\psi_k(t) = P(X(t) = k) = \sum_{l=0}^{K} P(X(t) = k \mid X(0) = l) P(X(0) = l).$$

If everyone starts in state 0, this simplifies to $P(X(t) = k) = P(X(t) = k \mid X(0) = 0) = P_{0k}(0, t)$. For Markov models, the cumulative intensities between the states k and l, $\Lambda_{kl}(t) = \int_0^t d\Lambda_{kl}(u) = \int_0^t \lambda_{kl}(u)\, du$ may be estimated non-parametrically using the Nelson-Aalen estimator,

$$\widehat{\Lambda}_{kl}(t) = \int_0^t \frac{1}{Y_k(u)}\, dN_{kl}(u),$$

where $Y_k(t)$ denotes the at-risk process for state k and $N_{kl}(t)$ the event process counting transitions from k to l. Subsequently, the transition probabilities may be estimated using the structure in the

transition probability matrix in equation (2.2),

$$\widehat{P}(s,t) = \prod_{[s,t)} \{I + \widehat{Q}(u)\, du\}, \tag{2.5}$$

where the entries in $\widehat{Q}(u)\, du$ are $d\widehat{\Lambda}_{kl}(u)$ for the off-diagonal terms ($k \neq l$) and $-\sum_{k \neq l} d\widehat{\Lambda}_{kl}(u)$ in the diagonal ($k = l$). The estimator in equation (2.5) is an Aalen-Johansen type product limit estimator (Aalen and Johansen (1978)). Large sample results of this estimator is discussed in Andersen, Borgan, et al. (1993). In summary, the state occupation probability, $P(X(t) = k)$, may be estimated from $\widehat{P}(0,t)$. For example, for the multi-state model in Figure 2.5, assume that we wish to calculate $P(X(t) = 1)$. Then, if $P(X(0) = 0) = 1$,

$$P_{01}(0,t) = P(X(t) = 1 \mid X(0) = 0) = P(X(t) = 1).$$

Hence, $P(X(t) = 1)$ may be estimated by $\widehat{P}_{01}(0,t)$ through equation (2.5) where

$$\widehat{Q}(t)\, dt = \begin{pmatrix} -d\widehat{\Lambda}_{01}(t) & d\widehat{\Lambda}_{01}(t) \\ d\widehat{\Lambda}_{10}(t) & -d\widehat{\Lambda}_{10}(t) \end{pmatrix}.$$

An alternative measure may be obtained by considering the average time spent in state k by time t, $\phi_k(t)$. This may be computed based on the state occupation probabilities, since

$$\phi_k(t) = E\left(\int_0^t I(X(u) = k)\, du\right) = \int_0^t P(X(u) = k)\, du, \quad \text{for} \quad k \in S.$$

Hence, the area under the curve is the time spent in state k until time t. That is,

$$\widehat{\phi}_k(t) = \int_0^t \widehat{P}(X(u) = k)\, du, \quad \text{for} \quad k \in S.$$

Thus, for example, for the multi-state model in Figure 2.5, the average time spent in the 'Episode' state (average time of an episode) may be estimated by $\widehat{\phi}_1(t) = \int_0^t \widehat{P}_{01}(0,u)\, du$ if $P(X(0) = 0) = 1$.

One way of handling the situation with both recurrent events and terminal events is to focus on the composite endpoint which consists of both recurrent events and death. Time-to-first composite event, either first recurrence or death, could be considered. This composite event could be examined using the multi-state model in Figure 2.1. The composite event could be analysed using the proportional hazards model of Cox (1972). In this model, the instantaneous probability of getting an event (either first recurrent or death) given covariates, $\lambda(t \mid Z)$, is given by,

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta^T Z),$$

where $\lambda_0(t)$ denotes the unspecified baseline intensity and β the effect of the Z covariates. In order to utilise more data, all recurrences may be considered for such a composite endpoint. That is, the recurrent event process consists of both actual recurrent events and death. An example could be the composite recurrence of hospitalisation for heart failure and death. Interpretation of a recurrent event process with a terminal component is less clear, since occurrence of a death prevents further recurrences. This will be elaborated and discussed in more detail later.

In the following chapters, there may be situations in which it can be beneficial to split the terminal event into two or more components. The multi-state models in Figures 2.11 and 2.10 illustrate such scenarios with two death types. These multi-state models naturally occur when contemplating to formulate a recurrent event process with a type of terminal event component but with other competing terminal events. An example could be recurrent heart failure hospitalisation and cardiovascular death with competing non-cardiovascular death. For modelling such data, it would be beneficial to split mortality into two components.

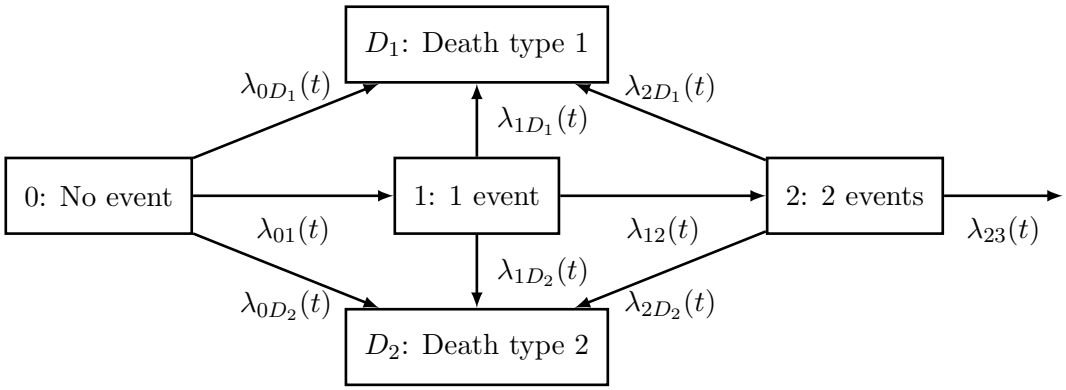


Figure 2.10: A recurrent event process with two types of terminal events, death type 1 and 2.

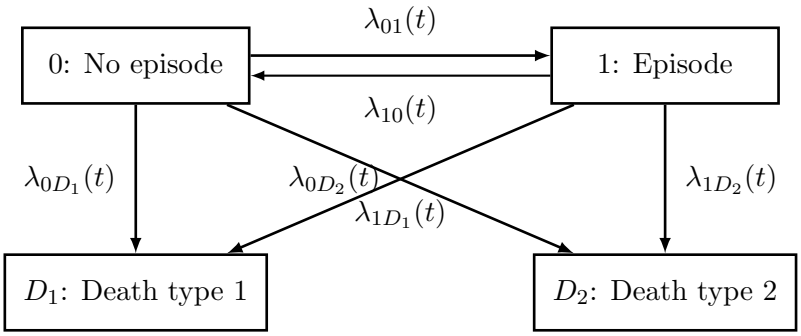


Figure 2.11: A recurrent event process with two types of terminal events, death type 1 and 2, and with episodes (risk-free periods).

Chapter 3

Regression models for recurrent events

Often it will be desirable to formulate a regression model that captures the effect of covariates on recurrent events. The following chapter will introduce the most common models for recurrent events with or without competing risks. The regression models either target *conditional* or *marginal* features of the recurrent event process as discussed by Cook and Lawless (2007). The conditional models focus on specifying the transition intensities in multi-state models for recurrent events with or without terminal events. On contrary, the marginal models focus on marginal features of the stochastic process, such as the expected number of events by time t, state occupation probabilities and more, with or without the influence of terminal events.

The following chapter will focus strictly on semi-parametric multiplicative regression models. Thus, the chapter will not cover additive regression models for recurrent events or fully parametric regression models for recurrent events. For additional literature on additive regression models for recurrent events the reader is referred to Schaubel et al. (2006), Zeng and Cai (2010), and Sun et al. (2020). Fully parametric models may of interest, but the flexibility of semi-parametric models is a general strength. However, a correctly specified parametric model will be more powerful than the semi-parametric counterpart. Model diagnostics and goodness-of-fit techniques are not covered in this chapter but should be explored for each model. Various residual processes may be explored for departures from the models (see, e.g., Lin et al. (2000), and Ghosh and Lin (2002)).

The regression models introduced in this chapter provides the background for models used or discussed in Manuscripts I, II(x), III, and IV. Manuscript I presents several methods for the analysis of recurrent events and focus on the challenges in extracting treatment effects from recurrent event data with terminal events through a case study. Treatment effect measures for these models will be elaborated further in Chapter 4. Manuscripts II, III, and IV discuss marginal regression models modelling expected number of recurrent events which are introduced in this chapter.

3.1 Conditional models for recurrent events

Assume now that there is a single event intensity such that

$$\lambda(t) = \lambda_{01}(t) = \lambda_{12}(t) = \dots$$

Further, assume that there are no competing risks. Several regression models may be formulated for $\lambda(t)$ and we will now discuss the most common models. Andersen and Gill (1982) proposed a counting process style extension of the Cox model which allow the event intensity to depend on the past through time-dependent covariates,

$$\lambda(t \mid Z(t)) = \lambda_0(t) \exp(\beta^T Z(t)),$$

where $\lambda_0(t)$ is a unspecified baseline intensity function and $Z(t)$ denotes potentially time-varying covariates. For recurrent events, $Z(t)$ could contain information on the previous number of events as well as other covariates. The covariates, $Z(t)$, represent the relevant history of the recurrent event process, $\mathcal{H}(t^-)$, as per equation (2.1). The model assumes that observations are independent among subjects, but robust standard errors may be used to account for dependency within subjects.

Alternatively, each transition may be governed by a baseline intensity function and potentially own covariate effect. This can be accommodated by considering the Cox type model with time-dependent strata suggested by Prentice, Williams, and Peterson (1981),

$$\lambda(t \mid Z) = \lambda_{k,k+1}(t) \exp(\beta_k^T Z), \quad k = 0, 1, \dots$$

where $\lambda_{k,k+1}(t)$ is the unspecified baseline intensity function for the transition between states k and $k+1$ and Z denotes the covariates. The effect of Z can be allowed to differ with each transition, k, or to be the same across all transitions, corresponding to $\beta^T = \beta_k^T$.

One may believe that some subjects are more prone ("frail") to experience recurrent events. This naturally introduces the notion of a frailty model, where the recurrent event intensity is assumed to depend on the *frailty* of each individual,

$$\lambda(t \mid Z(t), \nu) = \nu \lambda_0(t) \exp(\beta^T Z(t)),$$

where ν is a random and unobserved frailty which is assumed to follow some distribution, $\lambda_0(t)$ is the unspecified baseline intensity function and $Z(t)$ denotes the possibly time-varying covariates (Vaupel et al. (1979)). Often, ν is assumed to be Gamma distributed with mean 1 and variance σ^2. The larger σ, the more heterogeneous the population is.

In the presence of terminal events, the Andersen and Gill (AG) model, and the Prentice, Williams and Peterson (PWP) model may be used to model the cause-specific hazard of recurrent events, as depicted in Figures 2.7 and 2.8. Technically, this can be achieved by treating the competing cause as a censoring in estimation. This is analogous to the way in which cause-specific hazards can be estimated in a time-to-first event situation by technically censoring for the competing causes (Andersen, Geskus, et al. (2012)). The frailty model will also model the cause-specific hazard of recurrent events in the presence of terminal events. However, if individuals that are frail to experience recurrent events also are frail to die, the assumption of independent censoring given frailty is violated as discussed by Nielsen et al. (1992). This would imply that the frailty model is mis-specified in such a scenario and inference on the cause-specific hazard function derived from such an analysis should be considered twice. The joint frailty model suggested by Liu et al. (2004) tackles this issue

by introducing a common frailty which is allowed to affect both recurrent events and death. The recurrent event intensity is given by $\lambda(t \mid Z(t), \nu)$ and the death intensity is given by $\lambda^D(t \mid Z(t), \nu)$, where it is assumed that

$$\lambda(t \mid Z(t), \nu) = \nu \lambda_0(t) \exp(\beta^T Z(t)),$$
$$\lambda^D(t \mid Z(t), \nu) = \nu^\gamma \lambda_0^D(t) \exp(\alpha^T Z(t)).$$

Here, ν is a random and unobserved frailty with some distribution, $\lambda_0(t)$ is the unspecified baseline intensity function for recurrent events, $\lambda_0^D(t)$ is the unspecified baseline intensity function for terminal events, and $Z(t)$ denotes the possibly time-varying covariates. Further, the parameter γ expresses the relationship between the frailty on recurrent events and death. If $\gamma > 0$, an increased frailty will increase the hazard of both recurrent events and deaths. Covariate effects are expressed through β (effect on recurrent events) and α (effect on terminal events).

Inference

Inference for the intensity models may be based on a maximum likelihood approach. See Chapter 3 and 6 of Cook and Lawless (2007) for further technical details on likelihood estimation for each model type. Consider the ordered m event times, $t_1 < \ldots < t_m$, occurring over the time interval $[\tau_0, \tau]$. Conditional on the history at time τ_0, $\mathcal{H}(\tau_0)$, the probability density of having m events at the ordered times with an intensity of the form in equation (2.1), is

$$\prod_{j=1}^{m} \lambda(t_j \mid \mathcal{H}(t_j^-)) \cdot \exp\left(- \int_{\tau_0}^{\tau} \lambda(u \mid \mathcal{H}(u-)) \right),$$

which can be obtained through product integration, see Cook and Lawless (2007). The likelihood of an event is directly obtainable when the censoring process is completely independent of the recurrent event process. With model parameters, θ, the likelihood may be written as,

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta),$$

where i denotes the individual ($i = 1, \ldots, n$). The likelihood is based on the conditional probability of observing m_i events at $t_{i1} < \ldots < t_{im_i}$ times for individual i. Under the Andersen and Gill model, the likelihood takes the form

$$L_i(\theta) = \prod_{j=1}^{m_i} \{\lambda_0(t_{ij}) \exp(\beta^T Z(t_{ij}))\} \cdot \exp\left(- \int_{\tau_0}^{\tau} Y_i(u) \lambda_0(u) \exp(\beta^T Z_i(u))\, du \right),$$

where $\tau = \max\{\tau_1, \ldots, \tau_n\}$. If the censoring process is conditionally independent of the recurrent event process, $L(\theta)$ is a partial likelihood. As long as the observation process, $Y(t)$, only depends on information in the history, $\mathcal{H}(t^-)$, the partial likelihood may be treated as a standard likelihood. The partial maximum likelihood estimator, $\widehat{\theta}$, may be obtained by maximizing the partial likelihood, $L(\theta)$. Under some regularity conditions, $\widehat{\theta}$ has well-behaved large sample properties. Breslow-type estimators are available for estimating the cumulative baseline hazard.

The inference for frailty models is more complex and is discussed in Chapter 3 and 6 in Cook and Lawless (2007) and in detail in Hougaard (2000). Of note, the likelihood contribution for individual

i for the (shared) frailty model conditional on the frailty, ν_i, takes the form,

$$L_i(\theta \mid \nu_i) = \prod_{j=1}^{m_i} \{\nu_i \lambda_0(t_{ij}) \exp(\beta^T Z(t_{ij}))\} \cdot \exp\left(- \int_{\tau_0}^{\tau} \nu_i Y_i(u) \lambda_0(u) \exp(\beta^T Z_i(u))\, du \right).$$

For semi-parametric frailty models, inference is usually based on the Expectation-Maximization (EM) algorithm or penalised likelihood as the likelihood cannot be maximised directly (Balan and Putter (2020)).

3.2 Marginal models for recurrent events

Assume first that there are no terminal events. A marginal Cox model for the waiting time to the k'th recurrent event was suggested by Wei, Lin, and Weissfeld (1989). This models the waiting time from randomisation until each event. It is assumed that the marginal intensities, $\lambda_k(t \mid Z)$, are given by

$$\lambda_k(t \mid Z(t)) = \lambda_{0k}(t) \exp(\beta_k^T Z(t)), \quad k = 1, 2, \ldots,$$

where $\lambda_{0k}(t)$ denotes the unspecified baseline intensity function for the event k and β_k is event-specific effect of covariates $Z(t)$. Cook and Lawless (2007) discuss shortcomings of this model; the specification of the hazard requires a choice of the maximum number of recurrent events, individuals are at risk for all events from the beginning and it is implausible that a proportional hazards model will hold for each recurrent event. The model is known as the WLW model.

A regression model for the rate of recurrent events, $d\mu(t \mid Z(t))$, given possibly time-varying external covariates, $Z(t)$, was formulated by Lin, Wei, Yang, and Ying (2000), where

$$d\mu(t \mid Z(t)) = E(dN^*(t) \mid Z(t)) = \exp(\beta^T Z(t)) d\mu_0(t). \tag{3.1}$$

Here, $d\mu_0(t)$ denotes the unspecified baseline rate function and β^T is the effect of covariates $Z(t)$. This model is a proportional rates model. If there are only time-fixed covariates, such that $Z(t) = Z$, the model simplifies to

$$\mu(t \mid Z) = E(N^*(t) \mid Z) = \exp(\beta^T Z) \mu_0(t), \tag{3.2}$$

where $\mu_0(t)$ is the baseline mean function. This is referred to as the proportional means model.

The Lin, Wei, Yang, and Ying (LWYY) model will be mis-specified in the presence of terminal events. This is similar to the situation for the non-parametric estimators of the expected number of recurrent events with terminal events discussed in Chapter 2. In the presence of competing risks, Ghosh and Lin (2002) suggested an adjustment of the proportional means and rates model. The fact that no recurrent events can occur after death is technically handled in the estimation using inverse probability of censoring weights (IPCW) or inverse probability of survival weights (IPSW). IPCW gives more weight to the alive subjects and thus compensates for the fact that the censoring times are unknown for dead subjects. Equations (3.1) and (3.2) are assumed to be fulfilled corresponding to either the proportional rates or means model. Here, $N^*(t)$ has no jumps after D^* has occurred. Hence, with no time-varying covariates, the Ghosh and Lin (GL) model targets the expected number of recurrent events by time t in a world where individuals may die. This implies that number

of events cannot be interpreted alone without considering the distribution in survival among the selected covariates. A large value of $\mu(\cdot \mid Z)$ could be the result of an increased survival for Z, an increased event rate while alive or a combination of the two. The same pattern is clear when considering the non-parametric estimator of $\mu(t)$ in the presence of competing risks as given in equation (2.4) in Section 2. A similar analogy holds when consideration marginal inference, e.g., cumulative incidences, in a competing risks model. Therefore, Andersen, Geskus, et al. (2012) highlight that both rates and risks for all competing events are useful when studying models for competing risks.

An alternative approach is to define a composite endpoint consisting of both recurrent events and death. Mao and Lin (2016) suggested to model the weighted composite endpoint which consists of one or more recurrent event types and has death component. Each k'th event type, including death, is given a weight, w_k, which indicates the severity for $k = 1, \ldots, K$. For each event type, $N_k^*(t)$ counts the number of events by time t. The overall number of severity-weighted events by time t, $N^*(t)$, is given by

$$N^*(t) = \sum_{k=1}^{K} w_k N_k^*(t). \tag{3.3}$$

It is assumed that $Z(t)$ has a multiplicative effect on the marginal rate funtion of $N^*(t)$, such that

$$d\mu(t \mid Z(t)) = E(dN^*(t) \mid Z(t)) = \exp(\beta^T Z(t))\, d\mu_0(t).$$

This reduces to the proportional means model if there is no time-varying covariates, such that $E(N^*(t) \mid Z) = \exp(\beta^T Z)\, \mu_0(t)$. Thus, the inference is now on the weighted composite counting process. The resulting estimates will be a mixture of the covariate's effects on the composite endpoint. Difference in mortality for covariates will be an explicit part of the model in contrast to the Ghosh and Lin model. The introduction of weights, allows the user to quantify *how* much worse experiencing a death is compared to the recurrent events.

The model suggested by Mao and Lin models a composite recurrent event where death is a component. In the context of clinical relevance of the composite recurrent event endpoint, it may be of interest to model a composite event which only has a specific type of death included as a component. An example could be composite recurrent heart failure hospitalisation and cardiovascular death (see Figure 2.10). Here, the effect of the covariates, $Z(t)$, on this composite cardiovascular endpoint is of interest. Naturally, the model of Mao and Lin is no longer fulfilled, since a competing risk of non-cardiovascular death occurs. Furberg, Rasmussen, et al. (2022) suggest a modification of the procedure which accommodates this situation (Manuscript I). Any dependent censoring introduced by death is handled using IPCW, that is, for both cardiovascular and non-cardiovascular death in this example. Inference remains to only focus on the marginal rate of recurrent heart failure hospitalisation and cardiovascular death.

Inference

Inference from the WLW model relies on maximizing the partial likelihood for the k'th failure, as discussed in Section 3.1. For the remaining marginal models, inference is based on estimating equations. The inference is based on specifying the following type of estimating equations,

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t \{Z_i(u) - \bar{Z}(\beta, u)\}\, dN_i(u).$$

For the LWYY model, $\bar{Z}(\beta, u)$, takes the form,

$$\bar{Z}(\beta, u) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(t) Z_i(t) \exp(\beta^T Z_i(t))}{\frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp(\beta^T Z_i(t))}$$

The solution to the estimating equations, $U(\beta, \tau) = 0$, is denoted $\widehat{\beta}$. Under independent censoring and some additional regularity conditions, the estimator $\widehat{\beta}$ is consistent and asymptotically normal. Moreover, an asymptotically well-behaved Breslow-type estimator may be constructed for $\mu_0(t)$. The specification of $\bar{Z}(\beta, u)$ depends on the specific model.

3.3 Compatibility of conditional and marginal models

In general, a conditional intensity-based model does not ensure that a known marginal model for recurrent events holds. This is equivalent to the loss in one-to-one correspondence between risks and rates for competing risk models (Andersen, Geskus, et al. (2012)). For example, the LWYY model with a binary time-fixed covariate implies that

$$\mu(t \mid Z) = \mu_0(t) \exp(\beta Z).$$

Whereas, the AG model with a binary time-fixed covariate assumes that

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z).$$

The point estimates from these models will be the same but standard error estimates will remain different if there is dependence among recurrent events for an individual as discussed by Lin et al. (2000) and Amorim and Cai (2015).

Hougaard (2000) discusses frailty-type models for recurrent events extensively. In Chapter 7 (pg. 232), it is emphasised that proportional conditional hazards (e.g., a frailty model) does not generally correspond to proportional marginal hazards. An exception happens when using a positive stable frailty distribution. Moreover, he highlights that the benefit of marginal models is that the dependence between recurrent events does not need to be specified. However, as a result, a marginal model will not always enable inference on such dependency parameters (Chapter 13.3, pg. 431).

The concept of collapsibility was discussed by Greenland et al. (1999). Due to non-linearity, certain adjusted and non-adjusted effects may not be the same. If a treatment effect was collapsible, the same effects would be obtainable while adjusting for additional predictive covariates or not. Daniel et al. (2021) discuss collapsible effects for binary and time-to-event outcomes.

Chapter 4

Treatment effect measures for recurrent events

The goal of this chapter is to recommend and suggest clinically relevant treatment effect measures for analysing recurrent events with and without terminal events for randomised controlled trials. Therefore, the concept of *estimands* will be introduced. Estimands quantify the treatment effect of interest. The focus of this chapter will restrict attention to setting of randomised trials with a single covariate of interest, namely the randomised treatment, Z. Without of loss of generality, we assume that the treatment is binary, such that $Z \in \{0, 1\}$. In this setting, this chapter will discuss what inference may be drawn using the regression models described in the previous chapter. Moreover, relevant non-parametric estimates will be briefly discussed. We will distinguish between conditional and marginal parameters in line with the distinction between conditional (intensity-based) and marginal models discussed in the previous chapter. Furthermore, the relevance of other treatment effect measures will be discussed.

A key lesson from analysing randomised trials is not to condition on anything occurring post-baseline as this may bias the treatment effect. Clearly, the same holds when analysing recurrent events. Thus, intensity models which condition on the previous number of recurrent events may bias the treatment effect, as argued by Cook and Lawless (2007). If the treatment has an effect on the number of recurrent events, conditioning both on recurrent events and the previous number of events will condition some of the treatment effect away. Analogous considerations apply to modelling gap times as individuals will be differently selected into gap times after the first gap time (Cook and Lawless (2007), Hougaard (2022)). Thus, the interest will mainly be on Markov multi-state models modelling total time. Due to these facts, Cook and Lawless (2007) and Bühler et al. (2023) argue that focus should be placed on marginal models as it can be hard to correctly specify the intensity models when analysing randomised controlled trials. The following chapter will discuss interpretation of treatment effects for recurrent event models presented in the Chapter 3. Due to randomisation, all models will only consider a single binary treatment covariate. We distinguish between the situation with competing terminal events or not and their impact on interpretation.

The interpretation of treatment effects for randomised controlled trials discussed in this chapter provides background for Manuscripts I, II(x), III, and IV. Manuscript I focus on challenges in obtaining clinically relevant treatment effects from recurrent event data with terminal events through a case study using a randomised controlled trial. Manuscripts II(x) suggest a bivariate procedure to

overcome some of the difficulties discussed in this chapter and Manuscript I. The discussions in this chapter form the basis of the choice of marginal models considered in Manuscript III and IV.

4.1 Estimand

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) issued an addendum to ICH E9 (1998) on the importance of estimands and related statistical concepts in randomised trials which was adopted by both EMA (2020a) and FDA (2021). The description of a treatment effect should be facilitated by construction of the estimand. This estimand corresponds to a clinical question of interest and is specified to aid communication. The estimand consists of five key components,

1. **Treatment** describes the treatment of interest and the comparator
2. **Population** describes patients targeted by the clinical question
3. **Variable** describes the endpoint to be obtained for each patient to address the clinical question
4. **Intercurrent events** describes any events that may influence or prevent the observation of the variable of interest. Precise definition of treatment, population, and variable will likely address any intercurrent events
5. **Population-level summary** describes a population-level summary between the treatments based on the variable

The specification of the above should make it possible to clearly distinguish *"between the target of estimation (trial objective, estimand), the method of estimation (estimator), the numerical result (estimate), ..."* (EMA (2020a), FDA (2021)). Moreover, they explain that the goal of the treatment effect, the estimand, should be to summarise *"at a population level what the outcomes would be in the same patients under different treatment conditions being compared"*. This statement exactly coincides with the definition of a causal treatment effect. Causal treatment effects will be discussed in more detail in Chapter 7.

The variable describes the endpoint of interest for each patient. Such variables may be formulated in several ways. Relevant examples for recurrent events include total number of hospitalisations, times until hospitalisations, and length of stays in the hospital.

The addendum mentions that several *strategies* may be considered for addressing intercurrent events upon defining the clinical question of interest. Five main strategies are highlighted in the addendum: *treatment policy*, *hypothetical*, *composite variable*, *while on treatment*, and *principal stratum*. For the treatment policy strategy, the occurrence of the intercurrent event is considered irrelevant for defining the treatment effect of interest. For the hypothetical strategy, the treatment effect is estimated in the hypothetical scenario where the intercurrent event could not occur. For the composite variable strategy, the occurrence of the intercurrent event is considered important for the patient's outcome and is thus incorporated into the definition of the endpoint. Hence, the endpoint will be a composite of the event of interest and the intercurrent event. The while on treatment strategy tries to quantify the effect on the variable prior to the occurrence of the intercurrent effect. If the intercurrent event was death, the strategy would be called a while alive strategy. For the principal stratum strategy, emphasis is placed on a certain subgroup of patients in which the intercurrent event would or would not occur.

A terminal event is an intercurrent event of importance since it precludes the observation of any subsequent recurrent events. A hypothetical strategy to analysing recurrent events in the presence

of terminal events could be to estimate the expected number of events while assuming that deaths cannot occur. Specifically, this would happen if terminal events were treated as censorings while estimating the expected number of recurrent events using a LWYY model. This scenario will be discussed later. In general, the addendum states that even though the variable of interest cannot be measured after death, this data should not be considered missing. For recurrent events with terminal events, the composite strategy is popular. Examples include composite heart failure hospitalisation and cardiovascular death or composite non-fatal stroke, non-fatal myocardial infarction, or cardiovascular death. Each of these components represent cardiovascular burdens for an individual. The treatment effect on a composite recurrent event and death endpoint will generally differ from the treatment effect on recurrent events alone. That is, if treatment affects mortality, the effect on the composite event will be a mixture of such effects.

In the setting of recurrent events, several estimands may be formulated. An applicant group of statisticians from the pharmaceutical industry (Akacha et al. (2018)) sought a qualification opinion on EMA's thoughts on the statistical analysis of recurrent events. In this qualification opinion, the applicant group claimed that clinically relevant treatment effect measures may be extracted from analysing recurrent events. Additionally, they state that the statistical analyses are more efficient compared to those based on first events. In their simulations, they mainly focus on the negative binomial, the LWYY, the PWP, and the WLW models comparing them to a Cox model for first events. The qualification opinion from EMA stated that the presented recurrent event methods provide clinically interpretable treatment effects if there are no terminal events. If there are terminal events, the suggested methods still lack justification (EMA (2020b)). EMA state that an analysis method which provides separate estimates for the treatment effect on recurrent event while alive and the terminal event would be of interest for confirmatory decision making. They state that such an analysis method should have transparent modelling assumptions and be unbiased (regardless of the impact of terminal events). Moreover, the assessment of the treatment effect on the terminal event should precede the assessment of the treatment effect on recurrent events.

Schmidli et al. (2021) discuss causal estimands for recurrent events with terminal events and how to estimate these using statistical tools. Their discussion on which statistical models to use, however, lacks clarity.

4.2 Recurrent events without terminal events

The regression models presented in the previous chapter may be used in order to draw inference on the effect of treatment on recurrent events either through conditional or marginal models. In the present section, the interpretation of the models for recurrent events is discussed in the situation without terminal events. This corresponds to the multi-state models in Figures 2.4 and 2.5.

Conditional parameters

Intensity-based models may be used to characterise process dynamics within a multi-state model. Conditional parameters can be extracted from intensity-based models.

The Andersen and Gill model with a single binary treatment covariate models the recurrent event intensity given treatment. Hence, the regression parameter β from this model expresses the log-hazard ratio when comparing $Z = 1$ to $Z = 0$. A Poisson model with the logarithm of the risk time

as an offset corresponds to the AG model with a constant baseline intensity. If this assumption is reasonable, the two models will provide similar results.

Similarly, the PWP model will also provide a log-hazard ratio for comparing $Z = 1$ to $Z = 0$ under another model for the recurrent event intensity function. The model, however, will quickly be mis-specified due to the stratification according to the previous number of events. A treatment effect per transition, β_k, or an overall treatment effect, β, could be impacted by a treatment effect on the first events through the stratification, analogous to the impact on the AG model conditional on both treatment and previous number of events.

Inference based on the frailty model will have a subject specific interpretation, that is, β expresses the subject specific log-hazard ratio between $Z = 1$ and $Z = 0$ for a given frailty level. Since individuals in randomised controlled trials are (usually) only given one of two treatments, it is ambitious to try to extract this contrast. The addendum to ICH E9 highlights that the treatment effect should have a population-level interpretation (EMA (2020a)). Due to non-collapsibility of the frailty Cox-type model, the subject specific treatment estimates will not correspond to a population-level treatment effect (Martinussen and Vansteelandt (2013)). However, the frailty models allow for a dependence between recurrent events through the frailty variance. That is, some individuals may be more prone to experience recurrent events than others. This dependence is imposed without too many modelling assumptions. Hence, the frailty model may be good at describing process dynamics intended for simulation or further multi-state modelling. A negative binomial model with the logarithm of the risk time as an offset corresponds to the gamma frailty model with a constant baseline intensity. Hence, if the assumption of a constant baseline intensity is reasonable, the two models will provide similar results.

Inference based on generalised linear models may be adequate if there is no specific desire to model the entire time. Piece-wise constant versions of both the Poisson and negative binomial model may be used to relax the assumption of constant baseline intensities. In the presence of terminal events, the interpretation of inference drawn from a Poisson, negative binomial or Quasi-Poisson models changes. This will be discussed in the next section.

The Nelson-Aalen estimator may be used to obtain non-parametric estimates of cumulative hazard of recurrent events as discussed in Chapter 2. An estimate may be compared for each treatment group and compared using a log-rank test.

Marginal parameters

The WLW model with an overall treatment effect models the marginal waiting time from randomisation to the k'th event, when it is assumed that the treatment affects each waiting time in the same manner. Thus, the regression parameter, β, from this model expresses the marginal log-hazard ratio when comparing $Z = 1$ to $Z = 0$. Instead with a treatment effect per transition, β_k expresses the marginal log-hazard ratio on event k when comparing $Z = 1$ to $Z = 0$. However, as noted by Metcalfe and Thompson (2007), the construction of the risk sets implies that the treatment effects will be 'carried over' from event k to event $k + 1$, as a delayed waiting time for event k will imply a delayed waiting time for event $k + 1$. Hence, treatment effects may be impacted by the constructed risk sets using the WLW model with an overall effect or an effect per event.

The LWYY model provides inference on the expected number of events per treatment group. Thus, the regression parameter, β expresses the log-mean ratio for comparing $Z = 1$ to $Z = 0$. A benefit of the model is that the dependence structure between recurrent events for an individual does not

need to be specified.

The Nelson-Aalen estimator may be used to obtain non-parametric estimates of marginal mean of recurrent events as discussed in Chapter 2. With only a binary treatment variable and no terminal events, the estimate of the cumulative hazard of recurrent events and the expected number of recurrent events by time t coincides. The expected number of events per treatment group can be compared using a log-rank type test where robust standard errors are used (Lawless and Nadeau (1995)).

4.3 Recurrent events with terminal events

In the following section, the interpretation of the models for recurrent events with terminal events is discussed. This corresponds to the multi-state models in Figures 2.8, 2.7, 2.10, and 2.11. The interpretation of the conditional and marginal models based on multi-state models with terminal events is different from the interpretation without terminal events. This subtleties will be discussed in the following section.

Conditional parameters

In the presence of terminal events, the Andersen and Gill model with a binary treatment covariate models the cause-specific hazard function for recurrent events. Thus, the regression parameter, β, from this model expresses the logarithm of the cause-specific hazard ratio when comparing $Z = 1$ to $Z = 0$. A Poisson model for the rate of recurrent events will be equivalent to the Andersen and Gill model with a constant baseline intensity. The PWP model will also provide inference on the cause-specific log-hazard ratio comparing $Z = 1$ to $Z = 0$.

Analogously, the frailty model targets the cause-specific hazard function conditional on the subject specific frailty. Again, the subject specific interpretation will still be present with terminal events. Moreover, the frailty model assumes that censoring is independent of frailty. If a high frailty of recurrent events implies that the frailty of death is high or vice versa, this censoring assumption may be violated. That is, the inference drawn from the cause-specific hazard function if there is dependence between frailty-levels for recurrent events and deaths, can be questionable as discussed in Chapter 3. A negative binomial model for the rate of recurrent events will be equivalent to a gamma frailty model with a constant baseline hazard. Hence, inference from such a model can also be questionable if the censoring assumption is violated by treating terminal events as censorings.

The joint frailty model specifies two cause-specific hazard functions; one for recurrent events and one for terminal events. Hence, two treatment effects are obtained. With only a binary treatment variable, the cause-specific hazards of recurrent events and deaths are modelled. For recurrent events, β expresses the logarithm of the conditional cause-specific hazard ratio comparing $Z = 1$ to $Z = 0$ for an individual. For terminal events, α also expresses the conditional cause-specific hazard ratio comparing $Z = 1$ to $Z = 0$. An assumption is still that the censoring is independent of the frailty. Since the hazards of both recurrent and terminal events are modelled explicitly, it is less likely that this assumption is violated. Again, the model parameters have a subject-specific interpretation which is not compatible with population-level estimates.

Marginal parameters

The WLW model with a single binary treatment variable targets the marginal cause-specific hazard. That is, with an overall treatment effect, β, expresses the marginal cause-specific hazard ratio comparing $Z = 1$ to $Z = 0$. Li and Lagakos (1997) emphasised that the WLW model estimates the marginal hazard of recurrent events in a world where it is not possible to die with terminal events due to an infringement of the independent censoring assumption. Neither interpretation seems relevant for describing a treatment effect from a randomised trial in a simple and understandable manner. Moreover, the risk sets are still constructed in an undesirable way.

The LWYY model with a single treatment covariate may be used to estimate the expected number of recurrent events. However, with terminal events, the model will estimate the expected number of events in a hypothetical population where it is not possible to die. Hence, the log-mean ratio extracted from this model compares $Z = 1$ to $Z = 0$ for the hypothetical population.

The Ghosh and Lin model with a single treatment covariate models the expected number of recurrent events by time t. Here, the regression parameter, β, expresses the log-mean ratio between the two treatment groups. The expected number of events in this model is the number of recurrent events that would be observed in a world where it is possible to die from other causes. Hence, it is also important to analyse the impact of treatment on mortality, e.g., using a Cox model.

The Mao and Lin model formulated using a single treatment covariate models the expected number of weighted composite events including death by time t. Again, the parameter of interest is a log-mean ratio comparing $Z = 1$ to $Z = 0$. However, now it expresses the effect of each treatment on the number of composite events. Since death is included in the composite endpoint, the impact of treatment on both recurrent events and death is explored simultaneously. Conversely, this fact also makes it hard to disentangle the effect of treatment on each component since the result will be a mixture of both. The Mao and Lin model may be formulated using a larger weight for the death component and hierarchically smaller weights for recurrent events according to severity. This can reflect that experiencing death is a much worse clinical outcome for a patient. However, it might be hard to choose the actual weights in practice.

The updated Mao and Lin model suggested by Furberg, Rasmussen, et al. (2022) (Manuscript I) analogously models the expected number of weighted composite events, only including some mortality in the composite event, by time t. With a single treatment covariate, the regression parameter, β, still expresses the log-mean ratio comparing $Z = 1$ to $Z = 0$. This parameter models the effect of treatment on the composite event with some type of mortality included. Since other causes of mortality act as competing risks, it should be considered whether there could be an imbalance between treatments. Else, the skewed distribution between the two treatment groups in terms of the competing causes should be considered when drawing inference from this model.

4.4 Other estimands

Other estimands of interest may be considered for the analysis of recurrent events with or without terminal events. This chapter has discussed inference from conditional or marginal multiplicative regression models.

State occupation probabilities and average time spent in a state may also be marginal parameters of interest for analysing recurrent events with or without terminal events. A treatment effect could be based on such a comparison either through non-parametric estimates or regression models.

Claggett et al. (2018) suggest to focus on the area under the marginal mean function until time τ, $\int_0^\tau \mu(t)\,dt$. The interpretation of this measure may be less clear, however. A while-alive estimator which focuses on the ratio between the marginal mean function and the restricted mean survival time has been suggested as a relevant measure for analysing recurrent events in the presence of terminal events. See Wei, Mütze, et al. (2022) and Mao (2022) for discussion of such an estimand. The ratio is given by,

$$\frac{\mu(t)}{\varepsilon(t)} = \frac{\mu(t)}{\int_0^t S(u)\,du}.$$

Again, the aim of this is to extract a single measure from a two-dimensional problem. The expected number of recurrent events will be down-scaled according to mortality.

Cook and Lawless (2007, Chapter 8) advocate for the use of a mixed Poisson model for the rate or mean of recurrent events to capture a treatment effect from a randomised trial in a simple way. If ν_i denotes a subject-specific random effect with $E(\nu_i) = 1$, the rate model states that

$$E(dN_i^*(t) \mid \nu_i, Z_i) = \nu_i d\mu_0(t) \exp(\beta Z_i) dt,$$

which may be reformulated as a mean model since

$$E(N_i^*(t) \mid \nu_i, Z_i) = \nu_i \mu_0(t) \exp(\beta Z_i).$$

This model has the benefit of facilitating a simple estimand, a rate or mean ratio, comparing treatment $Z = 1$ to $Z = 0$. For this model, the subject-specific and population-average treatment effects, $\exp(\beta)$, are the same, since

$$E(N_i^*(t) \mid Z_i) = \mu_0(t) \exp(\beta Z_i).$$

Additionally, model parameters may be estimated robustly and the multiplicative model can be extended to include time-varying effects. We recommend marginal models that do not specify the dependence between individuals to limit the influence of the frailty distribution and make the model as general as possible (e.g., Lin et al. (2000) and Ghosh and Lin (2002)).

Pocock et al. (2012) suggested the *win ratio* method for the analysis of composite endpoints. The method quantifies ratio of 'winners' versus 'losers' according to a pre-specified order of endpoints in a matched or unmatched manner. An example could be first a comparison of cardiovascular death, followed by a comparison of hospitalisation for heart failure. This ranking reflects the fact that cardiovascular death is more severe than hospitalisation for heart failure. The parameter of interest is the win-ratio which measures the number of 'winners' with treatment equal to $Z = 1$ versus the number of 'losers' with $Z = 1$. The interpretation of this estimand is less clear even though clinical relevance may be ensured through ranking. The win-ratio method has become popular for the analysis of recurrent events with terminal events as it condenses the treatment effect into a single measure. Oakes (2016) highlighted that the win-ratio parameter will depend on the distribution of follow-up times. If treatment affects mortality and therefore follow-up times, this is an undesirable feature. In the simulation studies of Fritsch et al. (2021), the win-ratio approach performed similar in terms of power to first composite events analysed using a Cox model.

Another approach is to focus on utilities as discussed in Glasziou et al. (1990) and Cook, Lawless, and Lee (2003). In the utility-based approach, an utility is assigned to each state k in the multi-state model. Through this framework, estimates of the cumulative cost or quality of life may be derived.

The win ratio approach may be derived as a special case within the cumulative utility framework as argued by Bühler et al. (2023). Usefulness of the utility framework relies on the ability to specify the utilities associated with each state. This might be easy, but if not, they suffer from the same shortcomings as methods based on assigning weights. Within the field of clinical trials for heart failure, Anker, Schroeder, et al. (2016) state that there is a lack of consensus on how to define the relative weights making statistical models based on a weighting scheme less useful. As mentioned, the Mao and Lin (2016) model is a type of utility model that can be formulated using different utilities or weights for the components of the composite endpoint.

Chapter 5

Pseudo-observations

Andersen, Klein, and Rosthøj (2003) suggested to use *pseudo-observations* within the field of life history data. This is an alternative to the common methods within survival analysis that address right-censoring. In summary, the original incomplete data is used to compute pseudo-observations. The pseudo-observations ensures that each individual has a valid observation. The new data is subsequently used as response in a generalised linear model conditioning on a selection of covariates. Thus, this mimics a traditional regression approach, however, with a pre-step involving the computation of pseudo-observations.

The method of pseudo-observations introduced in this chapter forms the basis for Manuscripts II(x) and IV. Manuscripts II(x) suggest a bivariate procedure based on pseudo-observations which models recurrent events and terminal events simultaneously. Manuscript IV suggest a modified pseudo-observation method based on estimators that accounts for presence of covariate dependent censoring. Inference may be based on pseudo-observations for several parameters of interest as introduced as examples in this chapter. These parameters may be combined to make higher dimensional inference simultaneously. This idea was explored in Manuscript II(x).

The general idea behind pseudo-observations is as follows. Let $(X_i, i = 1, \ldots, n)$, be independent and identically distributed random variables with distribution P on the sample space Ω. Here, each X_i may be multivariate, either a vector $X_i = (X_{il}, l = 1, \ldots, k_i)$ or a process $X_i = \{X_i(t), t \geq 0\}$. Imagine that we are interested in the expectation, θ, of some function, f, of $X = (X_1, \ldots, X_n)$, i.e.

$$\theta = E_P(f(X)) = \int_\Omega f(x)\, P(dx).$$

Moreover, assume that an unbiased estimator, $\widehat{\theta}$, of θ exists, such that

$$E_P(\widehat{\theta}) = \int_\Omega \widehat{\theta}(x) P(dx) = \theta.$$

Assume that the covariates $Z = (Z_1, \ldots, Z_n)$ are independent and identically distributed random variables with distribution Q on Ω. Let R be the joint distribution of (X, Z). Then,

$$\theta = \int_\Omega f(x)\, P(dx) = \int_\Omega \int_\Omega f(x)\, R(dx, dz) = \int_\Omega E(f(X) \mid Z)\, Q(dz).$$

Let

$$\theta_i = \theta_i(Z_i) = E(f(X_i) \mid Z_i), \quad i = 1, \ldots, n.$$

Moreover, define their average as

$$\tilde{\theta} = \frac{1}{n} \sum_i \theta_i(Z_i).$$

Define the pseudo-observation for individual i as

$$\widehat{\theta}_i = \widehat{\theta}_i(X) = n\widehat{\theta}(X) - (n-1)\widehat{\theta}_{-i}(X),$$

where $\widehat{\theta}(X)$ is the estimate based on the entire sample and $\widehat{\theta}_{-i}(X)$ is the leave-one-out estimator based on the entire sample where the observations from individual i is omitted. Here,

$$E_R(\widehat{\theta}_i(X)) = E_R(\tilde{\theta}_i(Z_i)),$$

since $E_R(\widehat{\theta}(X)) = E_R(\tilde{\theta}(Z)) = \theta$. Consequently, the aim is to perform a regression analysis of $\widehat{\theta}_i$ on Z_i. Thus, it is assumed that,

$$g(\theta_i) = g(E(f(X_i) \mid Z_i)) = \beta^T Z_i, \tag{5.1}$$

for some link function g. The mean function, or inverse link function, is defined as

$$\xi(\beta^T Z_i) = g^{-1}(\beta^T Z_i) = \theta_i.$$

A regression of $\widehat{\theta}_i$ on Z_i with mean function $\xi(\cdot)$ is performed in order to estimate the regression parameters β in equation (5.1). This can be achieved using generalised estimating equations (GEE), see Liang and Zeger (1986). Estimates may be obtained by solving the following estimating equations,

$$U(\beta) = \sum_i U_i(\beta) = \sum_i \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right)^T V_i^{-1} (\widehat{\theta}_i - g^{-1}(\beta^T Z_i)) = 0,$$

here V_i is a working covariance matrix. A sandwich estimator may be used to estimate the variance of $\widehat{\beta}$. To that end define,

$$I(\beta) = \sum_i \left(\frac{\partial g^{-1}(\beta^T Z_i)}{\partial \beta} \right)^T V_i^{-1} \left(\frac{\partial g^{-1}(\beta^T Z_i)}{\partial \beta} \right),$$
$$\widehat{\mathrm{var}}(U(\beta)) = \sum_i U_i(\widehat{\beta})^T U_i(\widehat{\beta}),$$

so that

$$\widehat{\mathrm{var}}(\widehat{\beta}) = I(\widehat{\beta})^{-1} \widehat{\mathrm{var}}(U(\beta)) I(\widehat{\beta})^{-1}.$$

Andersen, Klein, and Rosthøj (2003) argued that $\widehat{\beta}$ followed a large sample normal distribution under some regularity conditions within the GEE framework. Hence,

$$\widehat{\beta} \overset{as}{\sim} \mathcal{N}\left(\beta, \sigma^2\right),$$

such that $\widehat{\beta}$ is asymptotically normal with mean β and a variance σ^2 which can be consistently estimated using $\widehat{\text{var}}(\widehat{\beta})$. Theoretical work by Jacobsen and Martinussen (2016), Overgaard, Parner, et al. (2017), and Overgaard (2019a) indicate that the pseudo-observations are approximately conditionally unbiased given covariates for a set of situations. Overgaard, Parner, et al. (2017) discuss general requirements for various estimators which ensures large sample normality for the estimates derived from GEE based on pseudo-observations. In particular, this relies on the assumption of independence between censoring and covariates to ensure large sample unbiasedness as discussed in Overgaard, Parner, et al. (2019). Models for recurrent events based on pseudo-observations with covariate dependent censoring will be discussed in Chapter 5 and in Manuscript IV. Moreover, work by the authors indicate that standard errors based on the suggested sandwich variance estimator tend to be slightly conservative. However, this bias is usually small unless there are very strong covariate effects. No general framework for the large sample behaviour of the pseudo-observations exist at the moment and theoretical properties should be explored on a case-to-case situation.

Inference on the pseudo-observations can be based on generalised estimating equations for these situations. The pseudo-observations may be computed at more than one time-point, such that $\widehat{\theta}_i$ can be multi-dimensional. This allows one to specify interactions between time and covariates. The correlation between observations from the same individuals can be specified using the GEE framework.

Introductory examples of pseudo-observations will be given in this chapter and information on model diagnostic tools for pseudo-observations will not be explained further. Goodness-of-fit for pseudo-observation models is discussed in Andersen and Perme (2010) as well as in Pavlič et al. (2019). Examples of appearances of individual pseudo-observations over time may also be found in Andersen and Perme (2010) and Furberg, Andersen, et al. (2023).

5.1 Pseudo-observations for survival data

The following section provides some examples of interesting parameters within the field of survival analysis. The parameters apply to situations with or without competing risks (see multi-state models in Figures 2.1 and 2.2).

Survival probabilities

Let D^* denote a survival time and let $f(D^*) = I(D^* > t)$, such that

$$\theta = E(f(D^*)) = E(I(D^* > t)) = P(D^* > t) = S(t),$$

where $S(t)$ denotes the survival probability at time t. The Kaplan-Meier estimator, $\widehat{S}(t)$ of $S(t)$ is an approximately unbiased estimator under independent censoring (see Chapter 2). For a fixed $t \in [0, \tau]$, the pseudo-observation for individual i is given by

$$\widehat{S}_i(t) = n\widehat{S}(t) - (n-1)\widehat{S}_{-i}(t).$$

Here, a specific t or a selection of time points may be considered. Several link functions may be considered for the regression model of the survival probabilities, e.g., $g(x) = x$, $g(x) = \log(x)$, or

$g(x) = \log(-\log(x))$. This suggests the following three models,

$$\begin{aligned} &\text{Model 1:} && S(t \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(S(t \mid Z)) = \beta^T Z, \\ &\text{Model 3:} && \log(-\log(S(t \mid Z))) = \beta^T Z. \end{aligned}$$

Model 1 targets risk differences, model 2 targets risk ratios, and model 3 targets (cumulative) hazard ratios.

Restricted mean life time

An alternative quantity of interest within classical survival analysis could be restricted mean life time or the restricted mean survival time (RMST). Let $\varepsilon(t)$ denote the RMST until time $t \geq 0$. For a fixed $\tau \in (0, \infty)$, this is given by

$$\varepsilon(\tau) = \int_0^\tau S(t)\, dt.$$

Due to censoring, it will typically not be feasible to consider this integral on $(0, \infty)$, as the tails of the survival distribution is not observed. Thus, attention is typically restricted to time τ. The RMST until time τ, $\varepsilon(\tau)$, provides an estimate of the expected event-free time until τ. Here,

$$\theta = E(D^* \wedge \tau) = \int_0^\tau S(t)\, dt,$$

with $f(D^*) = D^* \wedge \tau$. Consider the estimator, $\widehat{\varepsilon}(\tau)$, given by

$$\widehat{\varepsilon}(\tau) = \int_0^\tau \widehat{S}(t)\, dt,$$

where $\widehat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$. This estimator, $\widehat{\varepsilon}(\tau)$, is approximately unbiased under independent censoring as the Kaplan-Meier estimator is approximately unbiased (see Chapter 2). Thus, the pseudo-observation for individual i may be computed as,

$$\begin{aligned} \widehat{\varepsilon}_i(\tau) &= n\widehat{\varepsilon}(\tau) - (n-1)\widehat{\varepsilon}_{-i}(\tau) \\ &= n \int_0^\tau \widehat{S}(t)\, dt - (n-1) \int_0^\tau \widehat{S}_{-i}(t)\, dt \\ &= \int_0^\tau \widehat{S}_i(t)\, dt. \end{aligned}$$

In order to formulate regression models for the restricted mean survival time, one could consider the link functions $g(x) = x$ and $g(x) = \log(x)$. For a given τ, this suggests the following models,

$$\begin{aligned} &\text{Model 1:} && \varepsilon(\tau \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(\varepsilon(\tau \mid Z)) = \beta^T Z. \end{aligned}$$

Model 1 targets differences in restricted mean survival times and model 2 targets ratios of restricted mean survival times at time τ.

Cumulative incidences

Let (D^*, Δ) denote survival time and cause-of-death indicator corresponding to competing risks model with Δ causes of death (see Figure 2.2). Then, the parameter of interest for cause h, is

$$\theta_h = E(I(D^* \leq t, \Delta = h)) = P(D^* \leq t, \Delta = h) = F_h(t) \quad \text{for} \quad h \in \Delta,$$

where $F_h(t)$ denotes the cumulative incidence for cause h. The Aalen-Johansen estimator, $\widehat{F}_h(t)$, is an approximately unbiased estimator of the cumulative incidence, $F_h(t)$, under independent censoring (see Chapter 2). For a fixed $t \in [0, \tau]$, the pseudo-observation for individual i is given by

$$\widehat{F}_{ih}(t) = n\widehat{F}_h(t) - (n-1)\widehat{F}_{-ih}(t).$$

We consider the three link functions, $g(x) = x$, $g(x) = \log(x)$, or $g(x) = \log(-\log(1-x))$, for the regression models for the cumulative incidences. This suggests the following three models,

$$\begin{aligned} &\text{Model 1:} && F_h(t \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(F_h(t \mid Z)) = \beta^T Z, \\ &\text{Model 3:} && \log(-\log(1 - F_h(t \mid Z))) = \beta^T Z. \end{aligned}$$

Model 1 targets risk differences for cause h, model 2 targets risk ratios for cause h, and model 3 targets (cumulative) sub-distribution hazard ratios for cause h. Model 3 corresponds to a Fine-Gray model (Fine and Gray (1999)).

5.2 Pseudo-observations for recurrent events

Regression models for recurrent events could also be formulated using pseudo-observations. This could be models for recurrent events with or without the influence of competing deaths.

Expected number of events

Let $N^*(t)$ denote the number of recurrent events by time t. Then the parameter of interest is

$$\theta = \mu(t) = E(N^*(t)).$$

Under independent censoring and when there are no competing deaths, the Nelson-Aalen estimator is an approximately unbiased estimator of $\mu(t)$ (see Chapter 2). In the presence of competing risks, the estimator suggested by Cook and Lawless (1997) may be considered instead, see equation (2.4). Under independent censoring, this estimator is approximately unbiased. Let $\widehat{\mu}(t)$ denote the relevant estimator depending on whether there is competing risks or not. Thus, for a fixed $t \in [0, \tau]$, the pseudo-observation for individual i is given by

$$\widehat{\mu}_i(t) = n\widehat{\mu}(t) - (n-1)\widehat{\mu}_{-i}(t).$$

We consider the link functions, $g(x) = x$ or $g(x) = \log(x)$, to formulate regression models for the expected number of recurrent events. These link functions suggest the following two models,

$$\begin{aligned} &\text{Model 1:} && \mu(t \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(\mu(t \mid Z)) = \beta^T Z. \end{aligned}$$

Model 1 targets mean differences and model 2 targets mean ratios. Model 2 corresponds to the proportional means model as formulated by Lin, Wei, Yang, and Ying (no competing risks) or Ghosh and Lin (with competing risks) (Lin et al. (2000), Ghosh and Lin (2002)).

State occupation probabilities

Andersen and Klein (2007) and Andersen, Angst, et al. (2019) studied pseudo-observations of state occupation probabilities. Let $X(t)$ denote the state occupied by time t and let $S = \{0, \dots, K\}$ denote the state space in a continuous multi-state model. Then the parameter of interest is given by

$$\theta_k = E(I(X(t) = k)) = P(X(t) = k) = \psi_k(t) \quad \text{for} \quad k \in S.$$

Under independent censoring, the Aalen-Johansen estimator is an approximately unbiased estimator of $P(X(t) = k)$ (Datta and Satten (2001), Overgaard (2019b)). Thus, for a fixed $t \in [0, \tau]$, the pseudo-observation for individual i is given by

$$\widehat{\psi}_{ik}(t) = n\widehat{\psi}_k(t) - (n-1)\widehat{\psi}_{-ik}(t).$$

We consider the following link functions, $g(x) = x$ or $g(x) = \log(x)$, which suggests the models,

$$\begin{aligned} &\text{Model 1:} && \psi_k(t \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(\psi_k(t \mid Z)) = \beta^T Z. \end{aligned}$$

Model 1 targets risk differences and model 2 targets risk ratios.

Average time spent in a state

Alternatively, the average time spent in a given state until time τ may be of interest for a given multi-state model. Following the notation of the previous example, the parameter of interest is

$$\theta_{k\tau} = E\left(\int_0^\tau I(X(u) = k)\, du\right) = \int_0^\tau P(X(t) = k)\, dt = \phi_k(\tau) \quad \text{for} \quad k \in S,$$

Thus, it can be utilised that the Aalen-Johansen estimator is an approximately unbiased estimator of $P(X(t) = k)$ under independent censoring. Then the continuous mapping theorem and linearity of the integral implies that $\widehat{\phi}_k(\tau) = \int_0^\tau \widehat{\psi}_k(u)\, du$ also will be an approximately unbiased estimator of $\phi_k(\tau)$. Thus, the pseudo-observation for individual i and state k may be computed as

$$\widehat{\phi}_{ik}(\tau) = n\widehat{\phi}_k(\tau) - (n-1)\widehat{\phi}_{-ik}(\tau).$$

We consider the following link functions, $g(x) = x$ or $g(x) = \log(x)$, which suggests the models,

$$\begin{aligned} &\text{Model 1:} && \phi_k(\tau \mid Z) = \beta^T Z, \\ &\text{Model 2:} && \log(\phi_k(\tau \mid Z)) = \beta^T Z. \end{aligned}$$

Model 1 targets differences in average time spent in state k before time τ and model 2 targets ratios of average time spent in state k before time τ. Model 2 will ensure that the model predictions remain positive.

Chapter 6

Trial planning

Randomised controlled trials (RCTs) are planned to provide knowledge on a specific scientific question. The study population, collected data, and pre-defined hypotheses are carefully selected in order to ensure that the desired knowledge can be extracted. Well-defined interventions and their effects on endpoints are studied. Without loss of generality, we assume that we focus on a comparison between an active and a placebo treatment. Typically, hypotheses involving superiority or non-inferiority of the active treatment ($Z = 1$) versus the placebo treatment ($Z = 0$) is investigated. Let μ_1 and μ_0 denote the mean response of the active and placebo treatment, respectively. Let ω denote the treatment effect, which could be difference, $\omega = \mu_0 - \mu_1$, or a log-ratio, $\omega = \log(\mu_0/\mu_1) = \log(\mu_0) - \log(\mu_1)$ or something similar. Consider the superiority hypotheses,

$$H_0 : \omega \leq 0, \quad H_A : \omega > 0.$$

For a superiority trial, the aim is to claim that the active treatment is better than placebo. The randomised trial is planned to ensure that the hypothesis of interest can be met using a sample size calculation. A decision rule determines whether the null hypothesis will be rejected. A decision rule could be associated with critical values of a test statistic. If the test statistic is higher than some threshold, the decision is to reject the null hypothesis. Upon this decision, either a correct or one of two wrong choices has been made. The error of the first kind, a type I error, is the decision to reject the null hypothesis when it is true. The error of the second kind, a type II error, is the decision to accept the null hypothesis when the alternative hypothesis is true. This is visualised in Table 6.1.

Testing scenarios		Reality	
		H_0 true	H_a true
Decision	Accept H_0	Correct	Type II error
	Reject H_0	Type I error	Correct

Table 6.1: Display of type I and type II errors.

In general, tests are constructed such as to minimize the risk of committing both error types. Conventionally, a bound of the probability of committing a type I error is set, and the probability of committing a type II error is minimised. Hence, the significance level is chosen as a number, $\alpha \in (0, 1)$, such that

$$P(\text{Type I error}) = P_{H_0 \text{ true}}(\text{Reject } H_0) \leq \alpha.$$

Then, under the above condition, the target is to minimise

$$P(\text{Type II error}) = P_{H_a \text{ true}}(\text{Accept } H_0).$$

Equivalently, we may choose to maximise the power, which is given as $P(\text{Reject } H_0)$ when H_a is true. In the following chapter, we will discuss planning a randomised trial where the primary interest is a recurrent event endpoint.

The concepts surrounding trial design introduced in this chapter forms the basis for Manuscript III. Manuscript III suggest a simulation-based sample size estimation procedure for recurrent events with terminal events based on marginal models.

6.1 Design

The sponsors design their randomised trials to unveil the scientific hypotheses of interest. From a statistical point-of-view, this implies that the hypotheses can be explored using statistics. Clearly, this depends on the choice of estimand as well as other design characteristics. One of the first choices is to decide which recurrent event endpoint and estimand to consider (see Chapter 4 for an elaborate discussion of this).

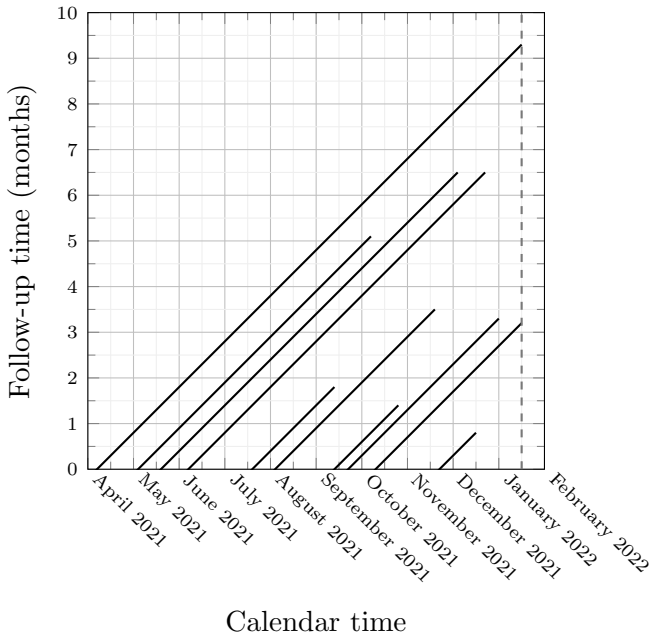


Figure 6.1: Lexis diagram: Illustration of calendar time versus follow-up time for 10 individuals.

Enrolment and study length

Randomised trials which collect life history data are dependent on the duration of follow-up. That is, the longer individuals are observed, the more events are collected. From a practical and economical

perspective, only trials of a certain length may be feasible. The closure of a trial according to some fixed day introduces administrative censoring for the event-free individuals. Additionally, it will be unrealistic to enrol all patients in the trials from the same day. Thus, subjects will enrol with delayed entry according to when they were randomised. Figure 6.1 illustrates a Lexis diagram visualising calendar time versus follow-up for 10 individuals according to their enrolment. A long period of enrolment, or accrual, will result in less time to observe life history data. Thus, both study length and accrual are important ingredients when planning a randomised trial with a survival type outcome. Figure 6.2 illustrates uniform accrual and study closure for six individuals.

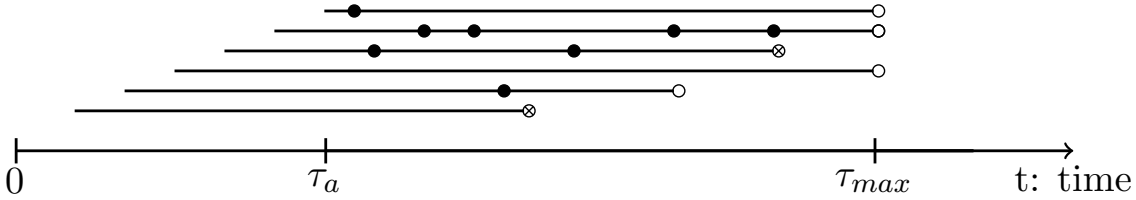


Figure 6.2: Illustration of enrolment and study closure for six individuals. The filled, unfilled, and crossed-out circles represent recurrent events, censoring, and death, respectively. The subjects are uniformly accrued until τ_a. The study had a duration of τ_{max}. The fifth individual was lost-to-follow-up prior to the study closure, τ_{max}. Individual three and six died prior to study closure.

The number of enrolled individuals, the sample size, is chosen to be high enough to investigate the hypothesis of interest for a fixed significance level and power. This will depend on enrolment, length of follow-up, effect size(s), and event rates. Once a sample size is determined, expected number recurrent events will typically be obtainable through the assumed event rate.

Testing procedures

In order to secure safety of future patients, error control is a topic of importance. In the superiority setting, a type I error corresponds to deeming a worse treatment better. Approval of such a drug could jeopardise the health of future patients. Thus, regulatory agencies require type I error control in the strong sense (EMA (1998)). This implies that all confirmatory endpoints, that is, endpoints which the sponsor wish to make a claim on, should all have type I error control on the same overall α-level. If this error rate is not controlled, testing more than one hypothesis will lead to error inflation due to multiplicity.

Hierarchical testing procedures ensure error control by pre-specifying a fixed sequence of ordered hierarchically hypotheses. Each ordered hypothesis is tested at level α in hierarchical order until the first non-rejection. Alternatively, graphical testing procedures may be used to govern type I error as suggested by Bretz et al. (2009). Here, the hypotheses are specified using a graph and α may be split and borrowed depending on the chosen graph.

Stopping rules

The clinical trial may be closed down upon pre-defined stopping rules. Trials that collect life history data for meeting the primary endpoints are usually closed upon reaching a total duration of a certain length, a minimum length of follow-up per individual or an overall minimum number of primary events. The trial may be stopped upon a mixture of these.

Clinical trials with a group sequential design allows for interim analyses and possibly early study

termination with a control over type I error. Group sequential design of trials with recurrent event endpoints is discussed by Cook and Lawless (2007), Mütze et al. (2019a), and Mütze et al. (2019b). A group sequential design will typically require a higher sample size compared to a traditional design due to α-cost associated with the possibility of stopping at interim.

Chapter 7

Causality and dependent censoring

Generally, it is believed that data from a randomised trial allows for a *causal* interpretation of the randomised treatment effect. Observational data is not protected by randomisation, but the desire is to extract causal relationships from this data. In order to do so, the exact meaning of causality should be explained. This chapter will introduce the concept of causality and causal treatment effects.

Classical inference based on survival methods often relies on an assumption of independent censoring. That is, the fact that an individual is censored does not leave a unrepresentative sample of the population. If systematic reasons which introduce censoring exist, inference may be affected. The statistical methods may be modified to accommodate dependent censoring (see, e.g., Cook and Lawless (2007)). These ideas are closely related to those from causality. Marginal models for recurrent events are a powerful analysis tool as argued in the previous chapters. However, these models (including models based on pseudo-observations) rely heavily on the assumption of independent censoring with or without impact of terminal events. Marginal models that accommodate dependent censoring will be introduced and discussed in this chapter. Marginal models for recurrent events under dependent censoring will be useful in order to make sensitivity analyses (EMA (2020a)).

The concepts of causality and dependent censoring introduced in this chapter forms the basis for Manuscript IV. Manuscript IV suggest extensions of marginal models for recurrent events, with or without terminal events, which can accommodate covariate dependent censoring. The emphasis is placed on adjustment using baseline covariates. This would answer the following question "what is the effect of being randomised to treatment $A = 1$ versus $A = 0$?". This corresponds to a treatment policy estimand as discussed in Chapter 4. However, this may not be the estimand of interest. Instead, one might wish to answer the following question, "what is the effect of taking treatment $A = 1$ versus $A = 0$ as prescribed?". With perfect treatment adherence this question is answered by focusing on the randomised treatment. If not, concepts from causality and causal treatment effects are vital. Dependent censoring could be introduced by censoring caused by treatment discontinuation where a rigorous definition of target estimand would be important.

7.1 Introduction to causality

Causality will be introduced using the concept of counterfactuals as introduced by Splawa-Neyman et al. (1923 [1990]) and Rubin (1974). For more details, see Hernán and Robins (2020). For simplicity, we assume that we are focusing on a binary treatment variable, $A \in \{0, 1\}$. Due to convention within causal inference, A is now used to denote treatment as opposed to Z in the previous chapters. The outcome of interest is denoted by Y. Hence, we wish to extract a causal treatment effect contrasting treatment $a = 1$ versus $a = 0$. Following the concept of counterfactuals, each individual i has two potential outcomes; one if they were given treatment $a = 0$, $Y_i^{a=0}$, and one if they were given treatment $a = 1$, $Y_i^{a=1}$. Since each individual is only given one of the treatments, either $Y_i^{a=0}$ or $Y_i^{a=1}$ is observed, and the other remains counterfactual. The individual causal effect is given by

$$Y_i^{a=1} - Y_i^{a=0},$$

but since only one of the treatments are given, this can never be observed or estimated. Instead, focus is placed on the *average causal effect* (ACE), denoted η, which is given by

$$\eta = E(Y^{a=1} - Y^{a=0}) = E(Y^{a=1}) - E(Y^{a=0}), \tag{7.1}$$

here $E(Y^{a=1})$ and $E(Y^{a=0})$ denotes the expected outcome if everyone in the population were given treatment $a = 1$ or $a = 0$, respectively. As seen from equation (7.1), η targets the treatment effect which compares the difference in expected outcome if everyone were given treatment $a = 1$ versus if everyone were given treatment $a = 0$. Without a loss of generality, η may target ratios instead of differences. Moreover, with a binary outcome η may compare probabilities.

Assume that we have L measured baseline covariates. In order to draw a causal effect from observational data, the following assumptions are usually required,

1. Consistency: $E(Y_a \mid L, A = a) = E(Y \mid L, A = a)$,
2. Exchangeability: $Y_a \perp\!\!\!\perp A = a \mid L$,
3. Positivity: $P(A) > 0$.

Consistency relates to the fact that the considered treatments should correspond to well-defined interventions. Exchangeability means that the conditional probability of receiving a certain treatment should only depend on measured covariates. Exchangeability is related to the concept of "no unmeasured confounders". Positivity implies that every individual in the population should have a positive probability of receiving both treatments.

7.2 Causal marginal models

Hernán and Robins (2020) introduce causal survival analysis in Chapter 17 of their book "Causal inference: What If?". Here, the lack of a causal interpretation of hazard ratios is discussed. Thus, the Cox model has been criticised for providing a parameter which does not have a causal interpretation (Hernán (2010) and Aalen, Cook, et al. (2015)). Thus, inference based on cause-specific hazard functions for recurrent events will also lack a causal interpretation. This is a disadvantage of the parameters derived from intensity-based methods. Luckily, marginal parameters estimated from a randomised trial usually allows for a causal interpretation. If these parameters can be updated to potential dependent censoring, a causal interpretation will also be possible in such a setting.

G-formula

The parametric g-formula was introduced by Robins (1986). The estimated ACE from the g-formula is given by

$$\widehat{\eta} = \bar{Y}_1^* - \bar{Y}_0^*,$$

where

$$\bar{Y}_a^* = \frac{1}{n}\sum_{i=1}^{n} \widehat{E}\left(Y_i \mid L_i, A_i = a\right),$$

here $\widehat{E}\left(Y_i \mid L_i, A_i = a\right)$ corresponds to the predicted outcome for individual i for his or her covariate values, L_i, with treatment fixed at $A_i = a$. Thus, regardless of the actual treatment assignment, the response is predicted for each individual given their covariate values and with both $a = 1$ and $a = 0$, leading to two predicted values (regardless of the observed outcome). This prediction requires that an outcome model, $E(Y \mid L, A)$, is specified. Thus, in order to make correct inference it is required that the outcome model is correct.

The g-formula adjusts for any imbalances that may be present due to a different distribution of covariate values (L) per treatment. If randomisation works well, the distribution of covariate values should be similar across the treatment, and this adjustment will not have an effect.

Su et al. (2022) discuss causal inference for recurrent event data focusing on the comparison of mean functions using the g-formula and other approaches. In line with this, we may place emphasis on a proportional means model for recurrent events as the outcome model, which models the expected number of recurrent events by time t. Here, for $t \in [0, \tau]$,

$$Y_a^*(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i(t \mid L_i, A_i = a) = \frac{1}{n}\sum_{i=1}^{n} \widehat{E}(N_i^*(t) \mid L_i, A_i = a).$$

Here, $\widehat{\mu}_i(t \mid L_i, A_i = a)$ can be derived from a regression model for the proportional mean. If there are no competing risks, the LWYY model may be used to estimate $\widehat{\mu}_i(\cdot)$ (Lin et al. (2000)). In the presence of competing risks, the GL model may be used to estimate $\widehat{\mu}_i(\cdot)$ (Ghosh and Lin (2002)). Subsequently, the average causal effect may be targeted as a mean-difference, mean-ratio or log-mean ratio for a fixed t, i.e.

$$\begin{aligned}
1 &: \widehat{\eta}_1 = Y_1^*(t) - Y_0^*(t), \\
2 &: \widehat{\eta}_2 = \frac{Y_1^*(t)}{Y_0^*(t)}, \\
3 &: \widehat{\eta}_3 = \log\left(\frac{Y_1^*(t)}{Y_0^*(t)}\right) = \log(Y_1^*(t)) - \log(Y_0^*(t)).
\end{aligned}$$

The choice of scale for the average causal effect depends on the scientific question of interest.

Inverse probability of treatment weights

An alternative way to adjust for any imbalances in covariate values per treatment is to consider inverse probability of treatment weights (IPTW). Here, an individual is weighted according to their

probability of receiving the treatment that they actually did receive. The probability of receiving treatment $a = 1$ given covariates is denoted by,

$$p(l) = P(A = 1 \mid L = l).$$

This probability is also known as *the propensity score*. Then, the inverse probability of treatment weights are given by $w_i = \frac{1}{p(L_i)}$ for an individual with $A_i = 1$ and $w_i = \frac{1}{1-p(L_i)}$ for an individual with $A_i = 0$. Then, $E(Y \mid A)$ may be fitted using the estimated individual weights $\widehat{w}_i$, where

$$\widehat{w}_i = \frac{1}{\widehat{p}(L_i)} \quad \text{if} \quad A_i = 1,$$
$$\widehat{w}_i = \frac{1}{1 - \widehat{p}(L_i)} \quad \text{if} \quad A_i = 0.$$

The estimated $\widehat{ACE}$ can be extracted by estimating β_1 from the following type of regression model,

$$E(Y \mid A = a) = \beta_0 + \beta_1 a,$$

fitted using the inverse probability of treatment weights. In order for this method to provide correct inference, the propensity model, $P(A \mid L)$, must be correctly specified. The scale of the average causal effect is chosen by specification of the outcome model.

For recurrent events, an appropriate propensity model may be formulated, $P(A \mid L)$. Subsequently, the outcome model may be a proportional means model fitted using the inverse probability of treatment weights, i.e. a model for $E(N^*(t) \mid A)$ with the weights $\widehat{w}_i$.

Note that both the g-formula and IPTW may only be of relevance if a randomised comparison cannot be based on the randomised treatment but should be additionally adjusted according to skewness in distribution of other baseline covariates. These approaches do not per se allow for adjustment according to a differential censoring pattern. Su et al. (2022) discuss using IPTW and g-formula for inference on the cumulative rate of recurrent events without terminal events. They also discuss an estimator based on pseudo-observations as well as a doubly robust estimator (robust towards mis-specifying either the outcome model or the propensity model). Alternatively, these causal methods would be relevant if focusing on a treatment effect different from the effect of being randomised, e.g., "efficacy" versus "effectiveness" estimand (EMA (2020a)).

7.3 Marginal models accommodating dependent censoring

Dependent censoring which compromises randomisation can be an issue. Marginal models that explicitly accommodate dependent censoring are thus of interest. These models can be used as sensitivity analyses of the primary analyses, where the censoring process is assumed to be independent of the recurrent event process. Primary and confirmatory endpoints are expected to have pre-specified sensitivity analyses which explore various assumptions from the primary or confirmatory analyses. Marginal models are recommended for the analysis of recurrent events in randomised trials. Thus, the focus of this section is to extend the marginal models discussed in Chapter 3 and Chapter 5 to include a degree of dependent censoring. The Ghosh and Lin model (discussed in Chapter 3) could be a candidate for a primary analysis of recurrent events with terminal events.

Missing data should be limited to the extent possible as it represents a source of potential bias as discussed in the ICH E9 guideline for statistical principles for the analysis of clinical trials (see

EMA (1998)). Similar considerations apply to the presence of censored observations during follow-up. Administrative censoring caused by study closure is usually not perceived as a source of bias or relevant to consider for potential dependent censoring. Any relevant information should be collected from censored individuals. At best, the statistical models can try to adjust for potential censoring patterns.

Janvin et al. (2022) discuss causal estimands for recurrent event with terminal events with a special focus on separable treatment effects as suggested by Stensrud et al. (2022). They also suggest focusing on the expected number of recurrent events.

Inverse probability of censoring weights

The method of inverse probability of censoring weights (IPCW) was first suggested by Robins and Rotnitzky (1992) and Robins (1993). This was intended as a tool to reduce the bias that may be introduced if individuals switch treatment arms as may be done in cancer trials due to ethical reasons. Treatment switching compromises the randomisation and at such, a randomised comparison is no longer ensured. This may be remedied by using IPCW.

The usability of IPCW methods will be illustrated using an example of non-parametric inference on the marginal rate or mean of recurrent events without terminal events. Assume that we wish to make inference on the recurrent event counting process, $N(t)$, where $N_i(t)$ denotes the process for individual i. The mean function is denoted $\mu(t)$ and the rate function $d\mu(t) = \rho(t)dt$. Following the notation of Cook, Lawless, Lakhal-Chaieb, et al. (2009), we consider the estimating equations,

$$\sum_{i=1}^{n} U_i(t) = \sum_{i=1}^{n} C_i(t)\{dN_i(t) - d\mu(t)\} = 0, \tag{7.2}$$

where $C_i(t) = I(t \leq C_i)$ indicates whether individual i is under observation at time t. Lawless and Nadeau (1995) suggested to focus on the non-parametric estimator,

$$d\widehat{\mu}(t) = \frac{\sum_{i=1}^{n} C_i(t)\, dN_i(t)}{\sum_{i=1}^{n} C_i(t)},$$

where $\widehat{\mu}(t) = \int_0^t d\widehat{\mu}(u)$ is the Nelson-Aalen estimate (see Chapter 2). Large sample unbiasedness relies on the following independence assumption

$$E(dN_i(t) \mid C_i \geq t) = E(dN_i(t)) = d\mu(t),$$

that is, the recurrent event counting process is completely independent of the censoring process. Intensity models will typically only require conditional independence, $E(dN_i(t) \mid C_i \geq t, \mathcal{H}_i(t^-)) = E(dN_i(t) \mid \mathcal{H}_i(t^-))$, to make valid inference (Cook, Lawless, Lakhal-Chaieb, et al. (2009)). Equation (7.2) may be modified using the inverse probability of being censored to the following,

$$\sum_{i=1}^{n} \tilde{U}_i(t) = \sum_{i=1}^{n} \frac{C_i(t)}{G_i(t)}\{dN_i(t) - d\mu(t)\} = 0, \tag{7.3}$$

where $G_i(t) = P(C_i \geq t \mid \mathcal{H}_i(t^-))$ denotes the censoring distribution which is assumed to depend only on the event history up to time t. Equation (7.3) will be valid under conditional independence as specified through $G_i(t)$. In order to apply the method, $G_i(t)$ must be estimated. Let $\widehat{G}_i(t)$ denote

this estimate. The censoring distribution may be estimated using parametric, semi-parametric or non-parametric methods. Correct inference, however, relies on the ability to specify the censoring distribution correctly. Based on the estimated censoring distribution, an inverse probability of censoring weighted estimator of the rate function can be suggested as

$$d\widehat{\tilde{\mu}}(t) = \frac{\sum_{i=1}^{n} C_i(t)\, dN_i(t)/\widehat{G}_i(t)}{\sum_{i=1}^{n} C_i(t)/\widehat{G}_i(t)},$$

where $\widehat{\tilde{\mu}}(t) = \int_0^t d\widehat{\tilde{\mu}}(t)$ denotes a weighted Nelson-Aalen estimate. Analogous derivations apply for the extensions of several other models (see, e.g., Cook and Lawless (2007) or Ghosh and Lin (2002)).

The proportional means regression model suggested by Ghosh and Lin (2002) utilises IPCW to account for terminal events (or more accurately, that there is no observed censoring time for dead individuals). In order to estimate these weights, the censoring distribution should be estimated. If it is believed that no covariates impact the probability of being censored, a Kaplan-Meier estimate of $\widehat{G}(t)$ may be utilised. If censoring is believed to depend on some observed baseline covariates, the censoring probability may be estimated through a Cox model for the censoring hazard. Hence, it may be reasonable to assume that the hazard of being censored can be formulated in the following way,

$$\lambda^C(t \mid A = a, L = l) = \lambda_0^C(t) \exp(\beta_a a + \beta_l l),$$

where $\lambda_0^C(t)$ denotes the baseline censoring hazard, β_a expresses the effect of treatment and β_l expresses the effect of another baseline covariate. Then, the estimated probability of censored for $A = a$ and $L = l$ is given by,

$$\begin{aligned}\widehat{G}(t \mid A = a, L = l) &= \exp\left(-\int_0^t \widehat{\lambda}_0^C(u) \exp(\widehat{\beta}_a a + \widehat{\beta}_l l)\, du\right) \\ &= \exp\left(-\widehat{\Lambda}_0^C(t) \exp(\widehat{\beta}_a a + \widehat{\beta}_l l)\right).\end{aligned}$$

It would be reasonable to choose the Ghosh and Lin model with weights computed through an assumption of independence between recurrent events and censoring (e.g., using Kaplan-Meier to estimate $\widehat{G}(t)$) as a primary analysis for recurrent events in the presence of terminal events. Other censoring weights may be explored as pre-specified sensitivity analyses.

Non-parametric estimators of marginal mean or rate functions may also be updated using IPCW. This holds for the relevant estimators with or without terminal events. The Ghosh and Lin model may also be adjusted to account for potential dependent censoring by updating the estimating equations using IPCW.

Pseudo-observations

The pseudo-observation approach will only provide valid inference if the underlying non-parametric estimators are approximately unbiased (see Chapter 5). Moreover, well-behaved large sample results of estimates derived from generalised estimating equations based on pseudo-observations relies on additional regularity conditions on the underlying non-parametric estimators as discussed in Overgaard, Parner, et al. (2017). In particular, Overgaard, Parner, et al. (2019) highlight that regression analysis for survival probabilities based on pseudo-observations may biased if based on a Kaplan-Meier estimator under dependent censoring. Instead, they study the large sample behaviour

of inverse probability weighted non-parametric estimators of survival probabilities and cumulative incidences instead. Hence, updating the suggested non-parametric estimators to other estimators which can handle dependent censoring, i.e., approximately unbiased again, would (hopefully) ensure that the method still suffices. Large sample behaviour should however be explored on a case-to-case basis as discussed in Chapter 5. Andersen and Perme (2010) as well as Binder et al. (2014) discuss extensions of methods based on pseudo-observations for survival probabilities and cumulative incidences which adjusts for dependent censoring.

Suppose that the interest is to model the expected number of recurrent events using a regression model based on pseudo-observations. Assume first that there are no competing risks. As noted, when there is dependent censoring, the Nelson-Aalen estimator will be biased. Instead, another estimator is suggested by Cook and Lawless (2007) which utilises inverse probability of censoring weights. Cook, Lawless, Lakhal-Chaieb, et al. (2009) discuss various extensions of non-parametric estimators to accommodate dependent censoring for estimation of the marginal rate function for recurrent event with terminal events. Pseudo-observations may be based on inverse probability of censoring weighted versions of either estimator.

Chapter 8

Summary of manuscripts

During the PhD, a total of three manuscripts have been submitted. Manuscripts I and II have been published in peer-reviewed journals. Manuscript III has been submitted and is under revision. Manuscript IV is in preparation. The final pages in the thesis include the manuscripts. Software has been created in relation to Manuscripts II, IIx, and III. Vignettes which provide additional details and examples of the `R` functions as they appear on `GitHub` have been included after each manuscript. Table 8.1 provides an overview of the manuscripts. In summary, the manuscripts contain the following,

Manuscript I discusses challenges in analysing recurrent events with or without competing risks in randomised controlled trials through a case study.

Manuscript II introduces a new method for analysing recurrent events and terminal events. simultaneously using bivariate pseudo-observations.

Manuscript IIx is an `R`-package for computation of pseudo-observations for recurrent events.

Manuscript III proposes a procedure for sample size estimation for a primary recurrent event endpoint with non-negligible terminal events in a randomised controlled trial.

Manuscript IV introduces appropriate methods to handle dependent censoring when using marginal models for recurrent events with or without terminal events.

The discussion in Chapter 9 provides a further summary of the results.

Manuscript I

Manuscript I provides an overview of methods used to analyse recurrent events with or without competing deaths. Here, the challenges in obtaining a clinically relevant treatment effect on recurrent events from a randomised controlled trial based on various models are discussed. The issues are exemplified through two endpoints from the large randomised controlled trial, LEADER (Marso et al. (2016)). The LEADER trial was sponsored by Novo Nordisk and investigated the effects of the drug liraglutide versus placebo on cardiovascular outcomes. Recurrent myocardial infarction and recurrent MACE (composite myocardial infarction, stroke, and cardiovascular death) were explored. For recurrent myocardial infarction, all-cause death acts as a competing risk. Whereas, for the recurrent composite endpoint, non-cardiovascular death acts as a competing risk, but cardiovascular

Manuscript	Title	Status
I	*Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER*	Published in *Pharmaceutical Statistics*
II	*Bivariate pseudo-observations for recurrent event analysis with terminal events*	Published in *Lifetime Data Analysis*
IIx	`R`-package 'recurrentpseudo': *Creates Pseudo-Observations and Analysis for Recurrent Event Data*	Published on `CRAN`
III	*Simulation-based sample size calculations for recurrent events with competing risks*	Under revision for *Pharmaceutical Statistics*
IV	*Marginal models for recurrent events under covariate dependent censoring*	In preparation

Table 8.1: Overview of manuscripts and their status.

death is also included in the recurrent event process.

An extension of the model suggested by Mao and Lin (2016) was proposed. This enables the modelling of the expected number of recurrent composite MACE events with competing non-cardiovascular death. This would be an appropriate model when modelling such data as both cardiovascular and non-cardiovascular death should be handled with care in estimation. Composite recurrent events which includes cardiovascular death are often explored in exploratory or confirmatory analyses without proper attention to how this should be addressed in analyses. Naïve analyses of recurrent composite endpoints will overestimate the marginal mean estimates and regression parameters may be biased.

Differences in estimands for various analysis methods are unclear for the current applications within recurrent events with terminal events. Better practice and understanding of estimands is key for further develop the statistical methodology within recurrent events. Manuscript I distinguish between intensity and marginal models for the analysis of recurrent events with or without competing risks. We recommend marginal models for the analysis of treatment effects in randomised controlled trials. This is motivated by the fact that the marginal models rely on fewer assumptions with simpler treatment effects as compared to intensity-based models. Marginal models, especially the proportional means model, provide a modelling framework for analysing recurrent events which ensures that a clinically relevant and understandable treatment effect can be estimated.

Manuscript II(x)

Manuscript II proposes a new method for analysing recurrent events and deaths simultaneously. The approach is based on bivariate pseudo-observations formulated for the expected number of recurrent events and the survival probability at the same time. Regression models can be based on the bivariate pseudo-observations and inference can be drawn on the effect of covariates on both recurrent events and deaths at the same time. In the presence of terminal events, it is important to consider not only the mean number of events. This is due to the fact that an effective way of reducing the number of recurrent events is by, e.g., administering a treatment which kills individuals.

The regression model is a generalised linear model where the link functions may be chosen

to ensure that the parameter of interest can be extracted from the model, e.g., a mean ratio for recurrent events and a risk ratio for terminal events. Large sample normality of the estimators is shown both using simulation and theoretical considerations. In the situation with a single binary treatment variable, this ensures that inference can be based on bivariate large sample normality. Hypothesis tests constructed using this bivariate normality may be more powerful than the univariate counterparts if a sequential testing procedure is used. Variance estimates based on the sandwich variance estimator, as suggested in the paper, are shown to be conservative in some situations using bootstrap and theory. The bivariate pseudo-observation method is illustrated using two examples: the classical bladder cancer data and recurrent myocardial infarction data from the randomised Novo Nordisk's trial LEADER (Byar (1980), Marso et al. (2016)). A three-dimensional extension of the bivariate model is suggested for analysis of the LEADER data. This model estimates expected number of myocardial infarction events, the cumulative incidence of cardiovascular death and the cumulative incidence of non-cardiovascular death simultaneously given covariates based on pseudo-observations. This allows the user to separate covariate effects, e.g., treatment, associated with cardiovascular versus non-cardiovascular death.

The bivariate pseudo-observation method is implemented in the `R`-package `recurrentpseudo` which is published on `CRAN`. Manuscript IIx contains the package documentation. The `R`-package also contains functions for computing one-dimensional and three-dimensional pseudo-observations. The one-dimensional pseudo-observations are based on the expected number of recurrent events. The three-dimensional pseudo-observations are based on the expected number of recurrent events, the cumulative incidence of death for cause 1 and the cumulative incidence of death for cause 2.

In EMA's Qualification Opinion (2020b) they state that *"The CHMP could also envisage the option to provide an analysis which delivers separate estimates which appropriately summarise the expected effect of the treatment on the (annual event rate for the) recurrent event while alive, and the effect on the terminal event for confirmatory decision making. These estimates should be unbiased from a statistical perspective and assumptions of the model should be transparent. Assessment of the treatment effect on mortality would have to precede the assessment of the treatment effect on the recurrent event."* The bivariate methods provide separate inference for both recurrent events and terminal events. Moreover, the assessment of treatment effects on mortality may be done prior to the assessment of treatment effects on recurrent events using the sequential testing procedure which utilises the bivariate large sample normality. This may be more powerful than testing each component using large sample univariate distributions.

Manuscript III

Manuscript III suggests a simulation-based procedure for conducting power estimation for analysing recurrent events with competing risks. Marginal models for recurrent events and competing deaths are imposed. Recurrent events are assumed to adhere to a proportional means model given treatment (Ghosh and Lin (2002)). Competing deaths are assumed to follow Cox's proportional hazards model given treatment (Cox (1972)). The input parameters in the procedure are quantities that can be extracted based on information which would usually be available in literature. Such as the expected number of recurrent events and the survival probability at one or more times in the reference group as well as the treatment effect on both expected number of recurrent events and the survival probability. Uniform accrual during a period as well as a fixed study length can be accommodated.

Two examples are used to illustrate the procedure with a varying rate of recurrent events and death. Potential power gains of the recurrent event model compared to a model for first events are explored. These examples indicate that the power gains will be highly dependent on the underlying

recurrent event rate. Fritsch et al. (2021) considered how to perform power calculations for recurrent event endpoints with competing deaths based on simulating data from intensities using frailty models. As discussed in Chapter 3 (Section 3.3), adherence to an intensity model does not ensure that a known marginal model is true. Fritsch et al. do not describe the estimand of interest and fit various marginal and conditional models to data simulated using frailty models, which makes interpretation hard. The suggested procedure ensures that the marginal models are correctly specified and that these can be targeted in the simulation.

As a downside, simulation of data using the procedure introduces independence between the recurrent event and death processes. This may not be a realistic if more recurrent events lead to a higher probability of death. This was explored further in a simulation study. The simulation study indicated that even though such a dependence is present it does not necessarily have a large impact on power.

The simulation-based sample size calculation procedure is a valuable tool for designing randomised controlled trials with recurrent events and competing risks. Software for the procedure is available in `R` from the `GitHub`-repository `simpowerrecurrent`[1].

Manuscript IV

Manuscript IV discusses marginal models for recurrent events, with or without terminal events, that accommodates covariate dependent censoring. The paper focuses on proportional rates or means models for recurrent events as suggested by Lin et al. (2000) and Ghosh and Lin (2002). Inverse probability of censoring weights based on the censoring distribution given baseline covariates are utilised in the estimating equations to ensure unbiased inference from the proportional means model suggested by Ghosh and Lin (2002). A regression model based on pseudo-observations of the marginal mean obtained through a weighted Nelson-Aalen estimator is suggested. This is an extension of the regression model for recurrent events based on pseudo-observations of the marginal mean discussed by Andersen, Angst, et al. (2019) and proposed in Manuscript II (Furberg, Andersen, et al. (2023)). Large sample behaviour and theoretical properties of the pseudo-observation approach based on the weighted estimator should be explored more in future work. Overgaard, Parner, et al. (2019) discuss large sample properties of pseudo-observation models for survival probabilities and cumulative incidences under covariate dependent censoring.

Covariate dependent censoring is introduced in a simulation study through a Cox model and a non-proportional piece-wise constant censoring model given covariates. The behaviour of the models are explored based on the dependent censoring. Results remain unbiased if the censoring distribution is correctly specified prior to being used as inverse censoring weights. The methods are illustrated using an application to hospital readmissions for patients with colorectal cancer (González et al. (2005)). Further work on application is planned prior to manuscript submission using recurrent event data which exhibits covariate dependent censoring from a randomised controlled trial from Novo Nordisk. Potentially, the LEADER data could be used with focus on recurrent cardiovascular events occurring while on-treatment focusing on differential treatment adherence for US versus non-US patients but this is to be explored in more detail (Marso et al. (2016)).

The suggested methods will be valuable as sensitivity analyses that explore the assumption of completely independent censoring in a marginal proportional means model for recurrent events. Censoring could be believed to depend on treatment alone or in combination with other baseline covariates. The primary focus of the paper is a treatment comparison based on randomised treatment,

[1] `https://github.com/JulieKFurberg/simpowerrecurrent`

i.e. the effect of being randomised to a given treatment. If another treatment effect is targeted, e.g., the effect of taking the drugs as prescribed, the concepts of causality become important. Thus, depending on the chosen estimand, the appropriate approach may include methods from causal inference.

Chapter 9

Conclusion and perspectives

This thesis has investigated, developed and implemented statistical methods with clinically interpretable treatment effect measures for the analysis of recurrent event endpoints with and without presence of terminal events.

9.1 Contributions

This thesis has contributed with the following:

- Development and characterisation of statistical methods for recurrent event endpoints, particularly in the presence of terminal events. A special emphasis is placed on methods that are used in or applicable for statistical analysis plans for Novo Nordisk's randomised controlled trials.
- Discussion (benefits and shortcomings) as well as recommendations of statistical methods for recurrent event endpoints in randomised controlled trials with a focus of clinically interpretable treatment effects, i.e. estimands.
- Implementation of statistical methods for recurrent events in statistical software that is made publicly available.

These contributions have been made through the research conducted in relation to the manuscripts included in this thesis. Chapter 8 provides detailed summaries of each manuscript. The following serves as an overall summary of the contributions of each manuscript,

Manuscript I *Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER*: This manuscript discusses current methods for the analysis of recurrent events. The interpretation of the methods are discussed with focus on extracting a clinically relevant treatment effect on recurrent events in the presence of terminals events using case studies from a randomised controlled trial. We recommend focusing on marginal models for the analysis of recurrent events. Moreover, a semi-parametric regression model for a composite recurrent event including some death component, but with other competing death causes, is proposed.

Manuscript II *Bivariate pseudo-observations for recurrent event analysis with terminal events*: This manuscript proposes a simultaneous regression model for recurrent and terminal events based on pseudo-observations of the marginal mean and survival probability. Large sample

behaviour of the estimates derived from the procedure is discussed using simulation and theoretical derivations. Testing procedures based on the large sample bivariate normality in the situation with a binary treatment covariate are also discussed.

Manuscript IIx `R`-package 'recurrentpseudo': *Creates Pseudo-Observations and Analysis for Recurrent Event Data*: This manuscript contains the documentation for the `R`-package 'recurrentpseudo'. This is an `R` implementation of the methods discussed in Manuscript II. This package computes one-dimensional (marginal mean), two-dimensional (marginal mean and survival probability), or three-dimensional (marginal mean and cumulative incidences for two death causes) pseudo-observations for recurrent events. This package also provides an interface for the subsequent modelling of the pseudo-observations using generalised estimating equations.

Manuscript III *Simulation-based sample size calculations for recurrent events with competing risks*: This manuscript suggests a simulation-based approach for conducting sample estimation for recurrent events with terminal events in the setting of a randomised controlled trial. The simulation procedure is based on a proportional means model for recurrent events and a proportional hazards model for the terminal event in line with the recommendation of marginal models. It relies on marginal quantities that usually are obtainable from historical data.

Manuscript IV *Marginal models for recurrent events under covariate dependent censoring*: This manuscript discusses current marginal models for recurrent events, with or without terminal events, which may accommodate covariate dependent censoring. Moreover, the manuscript suggests a regression model for recurrent events based on pseudo-observations of the marginal mean computed using inverse censoring weights which adjust for covariate dependent censoring. The existing and new methods are explored using simulation and through two data examples. Further work on this manuscript is intended prior to submission.

In relation to Manuscripts II(x) and III, `R` functions and vignettes with examples are available on `GitHub`. These vignettes are included as they appear on `GitHub` after the relevant manuscripts.

9.2 Perspectives and further work

A key objective of this thesis is to recommend clinically relevant treatment effects, estimands, for analysing recurrent event endpoints with or without competing terminal events. Relevant estimands for recurrent events are discussed in Chapter 4. Bühler et al. (2023) and we argue that multi-state models provide a valuable framework for specifying estimands for randomised controlled trials. Manuscript I provides an overview of existing methods to obtain treatment effects of recurrent events in the presence of competing risks. Interpretation of the treatment effect from each model is discussed using clinical data from the LEADER trial on recurrent myocardial infarction and a composite endpoint (recurrent myocardial infarction, stroke, and cardiovascular death). We argue that conditional hazard models for recurrent events with competing risks can be hard to correctly specify. Thus, it would be easier and more safe to impose marginal models as the dependence between recurrent events can be left unspecified. Frailty models allow for a specification of an explicit dependence structure but unfortunately the resulting treatment effect has a subject specific interpretation. Again, marginal models ensure population-based treatment effects. We recommend focusing on mean ratios for treatment comparisons. This estimand can be obtained by utilizing a proportional means model. Without competing risks, this treatment effect is the anti-logarithm of the parameter from the regression model suggested by Lin et al. (2000). In the presence of competing

risks, the anti-logarithm of the regression parameter from the model suggested by Ghosh and Lin (2002) may be used. Both models estimate the expected number of recurrent events by time t per treatment. If the proportionality assumption from either model is questionable, non-parametric estimates or models based on pseudo-observations may be utilized to estimate the expected number of recurrent events at a given time. The expected number of recurrent events has an easy interpretation from a clinical, patient, regulatory, and statistical perspective. Basing inference on such information would be a benefit since it is easy to understand. This would allow general practitioners to prescribe treatments, hopefully, with a larger knowledge of *what* treatment effect is expected. Andersen and Keiding (2012) suggest three useful principles for the analysis of multi-state models with competing risks: do not condition on the future; do not regard individuals at risk after they have died; and stick to this world. These useful principles have been considered in relation to the recommendation of marginal models for the analysis of recurrent events. Other estimands were discussed in Section 4.4. Of relevance could be to further explore marginal treatment effects obtained from the while-alive estimand, $\mu(t)/\varepsilon(t)$, using pseudo-observations.

Marginal models for recurrent events ensure a clinical interpretation of treatments irrespective of the potential impact of terminal events. If there are competing deaths, the marginal expected number of recurrent events cannot be considered alone. The impact of treatment on mortality needs to be considered alongside the number of recurrent events, since a high mortality rate in one treatment group can directly affect how many recurrent events are observed. This property is partially remedied by focusing on composite recurrent events and death. However, modelling the composite process will only provide inference on the combined effect and the components would need to be analysed separately to obtain individual effects. A marginal proportional means model for recurrent events and death may be fitted using the model suggested by Mao and Lin (2016). A proportional means model for recurrent events and only some death component, with other competing causes of death, may be fitted using the model suggested in Manuscript I. Sample size estimation for both types of composite recurrent event processes and the suggested marginal models is a topic for further research.

As argued, since recurrent events can be prevented by transitioning to a terminal state, it is vital to consider both treatment effects on number of recurrent events and death. The nature of this issue is dual, as treatment may influence each component differently. Manuscript II proposes a bivariate regression model based on pseudo-observations for expected number of events and survival probability simultaneously. This method has the benefit of estimating treatment effects on both components jointly. Moreover, the correlation between the components given covariates can be extracted. Generalised linear models based on pseudo-observations can utilise many link functions and explore time-varying treatment effects. This allows for a rich set of statistical models. Statistical software for estimation of the bivariate pseudo-observations and subsequent regression modelling is implemented in an `R`-package: `recurrentpseudo`. Manuscript IIx covers the content of the `recurrentpseudo` package. An `R` vignette describing the content of the package with examples has been included after Manuscript IIx. EMA has expressed a desire to obtain separate treatment effects on recurrent events and terminal events (see EMA (2020b)). This is possible using the bivariate pseudo-observation method. Moreover, sequential tests which test the components one at a time may be specified which utilises the large sample bivariate distribution.

An understanding of how to design randomised trials with recurrent event endpoints is required to adopt these endpoints into practice. We suggest to base inference on recurrent events on marginal models, especially the proportional means model is of interest. Suppose that a proportional means model is chosen as the primary or confirmatory analysis method for the recurrent event endpoint.

Claiming treatment effects on recurrent events requires the primary or confirmatory endpoint to be formally included in the testing hierarchy. Inclusion ensures that the type I error rate is controlled. Sample size estimation for recurrent event endpoints is a necessary tool in order to secure type I error control and to ensure sufficiently high power. Manuscript III proposes a simulation-based approach for conducting sample size estimation for recurrent events with competing risks. This procedure allows the user to specify quantities that are feasible to suggest based on historical data. Software and an `R` vignette with examples for the simulation procedure is available on `GitHub`. The vignette has been included after Manuscript III. This approach will enable Novo Nordisk, and other sponsors, to make an informed choice when deciding to focus on first or recurrent events when designing clinical trials. Vital for these considerations is the size of both the recurrent and terminal event rates.

Chapter 7 introduces the concept of causality. For randomised trials, it is usually assumed that a causal treatment effect can be obtained directly by design. Causal effects might be found using observational data if certain assumptions can be made. A causal treatment effect may also be difficult to obtain if there is dependent censoring in a randomised controlled trial. If the dependent censoring can be explained by baseline covariates, a causal treatment effect may again be secured. Marginal models that investigate the standard assumption of independent censoring would be of interest as sensitivity analyses for recurrent events or to estimate on-treatment estimands. Manuscript IV discusses ways to explore this assumption; either using a proportional means model with weights estimated by a specific model for the censoring distribution given important baseline covariates or based on pseudo-observations where the estimators utilise inverse censoring weights. Such analyses would be useful to pre-specify as sensitivity analyses for primary or confirmatory recurrent event endpoints. More work on Manuscript IV is intended before submission. Zhao et al. (2014) suggested a tipping-point type analysis for time-to-event data based on multiple imputations. A tipping-point analysis for recurrent events with competing risks is a topic for further research. The introduced concepts of causality are important if other treatment effects than the effect of being randomised are targeted. An example would be the treatment effect which would be obtained by taking a drug as intended.

With the work of this thesis, Novo Nordisk should be able to be front runners for recommending and applying recurrent event methods to their randomised controlled trials. Hopefully, dialogue with the regulatory agencies will drive methodological focus within the biostatistical applications of Novo Nordisk and other sponsors. Specifically, the field would benefit from a discussion of relevant estimands for recurrent events with or without terminal events. Recommendations for statistical practice for recurrent events in randomised trials has been discussed from birth to death in this thesis.

Bibliography

Aalen, O (1978). "Nonparametric inference for a family of counting processes". In: *The Annals of Statistics*, pp. 701–726.

Aalen, OO, RJ Cook, and K Røysland (2015). "Does Cox analysis of a randomized survival study yield a causal treatment effect?" In: *Lifetime Data Analysis* 21, pp. 579–593.

Aalen, OO and S Johansen (1978). "An empirical transition matrix for non-homogeneous Markov chains based on censored observations". In: *Scandinavian Journal of Statistics*, pp. 141–150.

Akacha, M et al. (2018). *Request for CHMP Qualification Opinion: Clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses*. URL: https://www.ema.europa.eu/en/documents/other/qualification-opinion-treatment-effect-measures-when-using-recurrent-event-endpoints-applicants_en.pdf (visited on 12/14/2020).

Amorim, LDAF and J Cai (2015). "Modelling recurrent events: a tutorial for analysis in epidemiology". In: *International Journal of Epidemiology* 44.1, pp. 324–333.

Andersen, PK, J Angst, and H Ravn (2019). "Modeling marginal features in studies of recurrent events in the presence of a terminal event". In: *Lifetime Data Analysis* 25, pp. 681–695.

Andersen, PK, Ø Borgan, et al. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics.

Andersen, PK, RB Geskus, et al. (2012). "Competing risks in epidemiology: possibilities and pitfalls". In: *International Journal of Epidemiology* 41.3, pp. 861–870.

Andersen, PK and RD Gill (1982). "Cox's Regression Model for Counting Processes: A Large Sample Study". In: *The Annals of Statistics* 10.4, pp. 1100–1120.

Andersen, PK and N Keiding (2012). "Interpretability and importance of functionals in competing risks and multistate models". In: *Statistics in Medicine* 31.11-12, pp. 1074–1088.

Andersen, PK and JP Klein (2007). "Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies". In: *Scandinavian Journal of Statistics* 34.1, pp. 3–16.

Andersen, PK, JP Klein, and S Rosthøj (2003). "Generalised linear models for correlated pseudo-observations, with applications to multi-state models". In: *Biometrika* 90, pp. 15–27.

Andersen, PK and MP Perme (2010). "Pseudo-observations in survival analysis". In: *Statistical Methods in Medical Research* 19, pp. 71–99.

Anker, SD and JJV McMurray (2012). "Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials?" In: *European Heart Journal* 33.22, pp. 2764–2765.

Anker, SD, S Schroeder, et al. (2016). "Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency". In: *European Journal of Heart Failure* 18.5, pp. 482–489.

Balan, TA and H Putter (2020). "A tutorial on frailty models". In: *Statistical Methods in Medical Research* 29.11, pp. 3424–3454.

Binder, N, TA Gerds, and PK Andersen (2014). "Pseudo-observations for competing risks with covariate dependent censoring". In: *Lifetime Data Analysis* 20.2, pp. 303–315.

Bretz, F et al. (2009). "A graphical approach to sequentially rejective multiple test procedures". In: *Statistics in Medicine* 28.4, pp. 586–604.

Bühler, A, RJ Cook, and JF Lawless (2023). "Multistate Models as a Framework for Estimand Specification in Clinical Trials of Complex Processes". In: *Statistics in Medicine*.

Byar, D (1980). *The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa*. Springer, pp. 363–370.

Claggett, B et al. (2018). "Comparison of time-to-first event and recurrent-event methods in randomized clinical trials". In: *Circulation* 138.6, pp. 570–577.

Cook, RJ and JF Lawless (1997). "Marginal Analysis of Recurrent Events and A Terminating Event". In: *Statistics in Medicine* 16, pp. 911–924.

— (2018). *Multistate models for the Analysis of Life History Data*. Chapman and Hall/CRC.

— (2007). *The Statistical Analysis of Recurrent Events*. 1st ed. Springer, New York.

Cook, RJ, JF Lawless, L Lakhal-Chaieb, et al. (2009). "Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone". In: *Journal of the American Statistical Association* 104.485, pp. 60–75.

Cook, RJ, JF Lawless, and K-A Lee (2003). "Cumulative processes related to event histories". In: *SORT: Statistics and Operations Research Transactions* 27.1, pp. 0013–0030.

Cox, DR (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, pp. 187–220.

Daniel, R, J Zhang, and D Farewell (2021). "Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets". In: *Biometrical Journal* 63.3, pp. 528–557.

Datta, S and GA Satten (2001). "Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models". In: *Statistics & probability letters* 55.4, pp. 403–411.

European Medicines Agency (EMA) (2020a). *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (Step 5)*. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf (visited on 03/15/2023).

— (1998). *ICH Topic E 9 Statistical Principles for Clinical Trials*. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf (visited on 02/02/2023).

— (2020b). *Qualification opinion of clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses*. URL: https://www.ema.europa.eu/en/documents/other/qualification-opinion-clinically-interpretable-treatment-effect-measures-based-recurrent-event_en.pdf (visited on 12/14/2020).

— (2016). *Reflection paper on assessment of cardiovascular safety profile of medicinal products*. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-assessment-cardiovascular-safety-profile-medicinal-products_en.pdf (visited on 01/18/2023).

Fine, JP and RJ Gray (1999). "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American Statistical Association* 94.446, pp. 496–509.

Food and Drug Administration (FDA) (2021). *E9(R1) Stastical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials*. URL: https://www.fda.gov/media/148473/download (visited on 01/31/2023).

— (2019). *Treatment for Heart Failure: Endpoints for Drug Development. Guidance for Industry. Draft Guidance*. URL: https://www.fda.gov/media/128372/download (visited on 01/31/2023).

— (2020). *Type 2 Diabetes Mellitus: Evaluating the Safety of New Drugs for Improving Glycemic Control. Guidance for Industry*. URL: https://www.fda.gov/media/135936/download (visited on 02/02/2023).

Fritsch, A et al. (2021). "Efficiency Comparison of Analysis Methods for Recurrent Event and Time-to-First Event Endpoints in the Presence of Terminal Events—Application to Clinical Trials in Chronic Heart Failure". In: *Statistics in Biopharmaceutical Research*, pp. 1–12.

Furberg, JK, PK Andersen, et al. (2023). "Bivariate pseudo-observations for recurrent event analysis with terminal events". In: *Lifetime Data Analysis* 29, pp. 256–287.

Furberg, JK, S Rasmussen, et al. (2022). "Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER". In: *Pharmaceutical Statistics* 21.1, pp. 241–267.

Ghosh, D and DY Lin (2002). "Marginal regression models for recurrent and terminal events". In: *Statistica Sinica* 12, pp. 663–688.

— (2000). "Nonparametric analysis of Recurrent Events and Death". In: *Biometrics* 56, pp. 554–562.

Glasziou, PP, RJ Simes, and RD Gelber (1990). "Quality adjusted survival analysis". In: *Statistics in Medicine* 9.11, pp. 1259–1276.

González, JR et al. (2005). "Sex differences in hospital readmission among colorectal cancer patients". In: *Journal of Epidemiology & Community Health* 59.6, pp. 506–511.

Greenland, S, J Pearl, and JM Robins (1999). "Confounding and collapsibility in causal inference". In: *Statistical Science* 14.1, pp. 29–46.

Hernán, MA (2010). "The hazards of hazard ratios". In: *Epidemiology (Cambridge, Mass.)* 21.1, p. 13.

Hernán, MA and JM Robins (2020). *Causal inference: What If*. URL: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2022/12/hernanrobins_WhatIf_20dec22.pdf.

Hougaard, P (2000). *Analysis of Multivariate Survival Data*. Springer.

— (2022). "Choice of time scale for analysis of recurrent events data". In: *Lifetime Data Analysis* 28.4, pp. 700–722.

Jacobsen, M and T Martinussen (2016). "A note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations". In: *Scandinavian Journal of Statistics* 43, pp. 845–862.

Janvin, M et al. (2022). *Causal inference with recurrent and competing events*. URL: https://arxiv.org/abs/2202.08500.

Kaplan, EL and P Meier (1958). "Nonparametric estimation from incomplete observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481.

Lawless, JF and C Nadeau (1995). "Some simple robust methods for the analysis of recurrent events". In: *Technometrics* 37.2, pp. 158–168.

Li, QH and SW Lagakos (1997). "Use of the Wei–Lin–Weissfeld method for the analysis of a recurring and a terminating event". In: *Statistics in Medicine* 16.8, pp. 925–940.

Liang, K-Y and ST Zeger (1986). "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1, pp. 13–22.

Lin, DY et al. (2000). "Semiparametric regression for the mean and rate functions of recurrent events". In: *Journal of the Royal Statistical Society* 62.4, pp. 711–730.

Liu, L, RA Wolfe, and X Huang (2004). "Shared Frailty Models for Recurrent Events and a Terminal Event". In: *Biometrics* 60, pp. 747–756.

Mao, L (2022). "Nonparametric inference of general while-alive estimands for recurrent events". In: *Biometrics*.

Mao, L and DY Lin (2016). "Semiparametric regression for the weighted composite endpoint of recurrent and terminal events". In: *Biostatistics* 17, pp. 390–403.

Marso, SP et al. (2016). "Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes". In: *New England Journal of Medicine* 375.4, pp. 311–322.

Martinussen, T and TH Scheike (2006). *Dynamic Regression Models for Survival Data*. Vol. 1. Springer.

Martinussen, T and S Vansteelandt (2013). "On collapsibility and confounding bias in Cox and Aalen regression models". In: *Lifetime Data Analysis* 19.3, pp. 279–296.

McMurray, JJV et al. (2014). "Angiotension-Neprilysin Inhibition versus Enalapril in Heart Failure". In: *The New England Journal of Medicine* 371.11.

Metcalfe, C and SG Thompson (2007). "Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome?" In: *Statistical Methods in Medical Research* 16.2, pp. 103–122.

Mogensen, UM et al. (2018). "Effect of sacubitril/valsartan on recurrent events in the Prospective comparison of ARNI with ACEI to Determine Impact on Global Mortality and morbidity in Heart Failure trial (PARADIGM-HF)". In: *European Journal of Heart Failure*.

Mütze, T et al. (2019a). "Group sequential designs for negative binomial outcomes". In: *Statistical Methods in Medical Research* 28.8, pp. 2326–2347.

— (2019b). "Group sequential designs with robust semiparametric recurrent event models". In: *Statistical Methods in Medical Research* 28.8, pp. 2385–2403.

Nelson, W (1969). "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1.1, pp. 27–52.

— (1972). "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14.4, pp. 945–966.

Nielsen, GG et al. (1992). "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models". In: *Scandinavian Journal of Statistics* 19.1, pp. 25–43.

Oakes, D (2016). "On the win-ratio statistic in clinical trials with multiple types of event". In: *Biometrika* 103.3, pp. 742–745.

Overgaard, M (2019a). *Counting processes in p-variation with application to recurrent events*. URL: https://arxiv.org/pdf/1903.04296.pdf.

— (2019b). "State occupation probabilities in non-Markov models". In: *Mathematical Methods of Statistics* 28, pp. 279–290.

Overgaard, M, ET Parner, and J Pedersen (2017). "Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations". In: *The Annals of Statistics* 45.5, pp. 1988–2015.

— (2019). "Pseudo-observations under covariate-dependent censoring". In: *Journal of Statistical Planning and Inference* 202, pp. 112–122.

Pavlič, K, T Martinussen, and PK Andersen (2019). "Goodness of fit tests for estimating equations based on pseudo-observations". In: *Lifetime Data Analysis* 25, pp. 189–205.

Pocock, SJ et al. (2012). "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities". In: *European Heart Journal* 33.2, pp. 176–182.

Prentice, RL, BJ Williams, and AV Peterson (1981). "On the regression analysis of multivariate failure time data". In: *Biometrika* 68, pp. 373–79.

Robins, JM (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9-12, pp. 1393–1512.

— (1993). "Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers". In: *Proceedings of the Biopharmaceutical Section, American Statistical Association*. Vol. 24. 3. San Francisco CA, p. 3.

Robins, JM and A Rotnitzky (1992). "Recovery of information and adjustment for dependent censoring using surrogate markers". In: *AIDS Epidemiology: Methodological Issues*, pp. 297–331.

Rogers, JK et al. (2014). "Analysing recurrent hospitalisations in heart failure: a review of statistical methodology, with application to CHARM-Preserved". In: *European Journal of Heart Failure* 16.1, pp. 33–40.

Rubin, DB (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, p. 688.

Schaubel, DE, D Zeng, and J Cai (2006). "A semiparametric additive rates model for recurrent event data". In: *Lifetime Data Analysis* 12.4, pp. 389–406.

Schmidli, H, JH Roger, and M Akacha (2021). "Estimands for recurrent event endpoints in the presence of a terminal event". In: *Statistics in Biopharmaceutical Research*, pp. 1–11.

Solomon, SD et al. (2019). "Angiotensin–Neprilysin Inhibition in Heart Failure with Preserved Ejection Fraction". In: *The New England Journal of Medicine* 381.17, pp. 1609–1620.

Splawa-Neyman, J, DM Dabrowska, and TP Speed (1923 [1990]). "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." In: *Statistical Science*, pp. 465–472.

Stensrud, MJ et al. (2022). "Conditional separable effects". In: *Journal of the American Statistical Association*, pp. 1–13.

Su, C-L, RW Platt, and J-F Plante (2022). "Causal inference for recurrent event data using pseudo-observations". In: *Biostatistics* 23.1, pp. 189–206.

Sun, X, J Ding, and L Sun (2020). "A semiparametric additive rates model for the weighted composite endpoint of recurrent and terminal events". In: *Lifetime Data Analysis* 26.3, pp. 471–492.

Vaupel, JW, KW Manton, and E Stallard (1979). "The impact of heterogeneity in individual frailty on the dynamic of mortality". In: *Demography* 16.3, pp. 439–454.

Wei, J, T Mütze, et al. (2022). "Properties of two while-alive estimands for recurrent events and their potential estimators". In: *Statistics in Biopharmaceutical Research*, pp. 1–11.

Wei, LJ, DY Lin, and L Weissfeld (1989). "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions". In: *Journal of the American Statistical Association* 84 (408), pp. 1065–1073.

Zeng, D and J Cai (2010). "A semiparametric additive rate model for recurrent events with an informative terminal event". In: *Biometrika* 97.3, pp. 699–712.

Zhao, Y et al. (2014). "A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring". In: *Journal of Biopharmaceutical Statistics* 24.2, pp. 229–253.

Manuscript I

Title *Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER*

Authors Julie K. Furberg, Søren Rasmussen, Per K. Andersen and Henrik Ravn

Details Published in Pharmaceutical Statistics (January 2022)

Received: 2 February 2021 | Revised: 7 June 2021 | Accepted: 16 August 2021

DOI: 10.1002/pst.2167

MAIN PAPER

WILEY

Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER

Julie Kjærulff Furberg[1,2] | **Søren Rasmussen**[2] | **Per Kragh Andersen**[1] | **Henrik Ravn**[2]

[1]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

[2]Biostatistics GLP-1 and CV 1, Novo Nordisk, Bagsværd, Denmark

Correspondence
Julie Kjærulff Furberg, Novo Nordisk A/S, Vandtårnsvej 114, 2860 Søborg, Denmark.
Email: jukf@novonordisk.com

Funding information
Innovationsfonden; Novo Nordisk

Abstract

Analysis of recurrent events is becoming increasingly popular for understanding treatment effects in randomised controlled trials. The analysis of recurrent events can improve efficiency and capture disease burden compared to standard time-to-first event analyses. However, the added knowledge about the multi-state process comes at the cost of modelling complexity. High mortality rates can complicate matters even more. A case study using data from a randomised controlled trial, LEADER, is presented to highlight interpretation of common methods as well as potential pitfalls when analysing recurrent events in the presence of a competing risk. The presented methods either target features of the underlying intensity functions or marginal traits of a multi-state process which includes terminal events or not. In particular, approaches to handle death as a part of an event and as a competing risk are discussed. A new method targeting the marginal mean function for a composite endpoint, which includes both death as a component and as a competing risk, will be introduced. Finally, recommendations for how to capture meaningful treatment effects in randomised controlled trials when analysing recurrent and terminal events will be made.

KEYWORDS

competing risks, randomised controlled trials, recurrent events, treatment effects

1 | INTRODUCTION

The analysis of time-to-event outcomes is well-known for quantifying treatment effects for randomised controlled trials. A common choice is to analyse the waiting times until the events of interest using a Cox regression model with treatment as a covariate, for instance analysing the time-to-first heart failure or time-to-first stroke.[1] Recurrent events are events, of the same type, that can happen several times for an individual during the course of their life, for example, hospitalisation or stroke. The analysis of recurrent events has become more popular within the last years, especially within the field of cardiology.[2] This owes to the fact that many classical time-to-event analyses of, for example, heart failure only focus on the first events, despite the recurrent event structure of the data. Using all events, and conducting recurrent event analyses, can help capture disease burden, aid understanding of treatment effects and utilise all data compared to the time-to-first event analyses. However, the analysis of recurrent events can be more complicated in

terms of interpretation and conducting analyses as these analyses target a more complex process as opposed to first events. Recently, results from the PARAGON-HF randomised controlled trial were published, where the primary endpoint was a recurrent composite endpoint consisting of hospitalisation for heart failure or death from cardiovascular disease.[3] Two other randomised controlled trials, CHARM-Preserved and PARADIGM-HF, have also been republished with a focus on recurrent event methods for summarising the results as an alternative to the primary manuscripts focusing on first events.[4–7]

Several models and methods have been put forward for the analysis of recurrent events. Classical and more recent methods will be presented in this paper. The focus will be on the interpretation of treatment effects for these models when utilised to analyse randomised controlled trials, illustrated by a case study. The methods will be applied to data from the LEADER trial, a randomised controlled trial, that focused on comparing cardiovascular events between placebo and an active treatment in a type 2 diabetic population with cardiovascular risk factors.[8]

Mortality rates can be high in controlled trials, especially in trials which include a sick population. This can introduce issues for the inference on recurrent events as a terminal event will preclude further recurrences. These issues are similar to problems with competing risks encountered in classical time-to-event analysis. A potential correlation between recurrent and terminal events can also be present, and treatments may influence mortality. Moreover, the effect of treatment on mortality may also be an important scientific question. A popular approach for managing this problem is to consider composite recurrent events, consisting of both, for example, heart failure and death. In this setting, some of the recurrent event process terminates the process itself, and it is not clear which implications this can have on estimation and interpretation. We will introduce a new method for modelling the marginal mean function for composite endpoints which includes a death component but also are subject to competing risks. No recommendations from health authorities exist for the analysis of recurrent events, however clinical relevance of, for example, recurrent heart failure has been acknowledged by EMA.[9] They also emphasise issues with analysis and interpretation of recurrent events in the presence of terminal events. Thus, they suggest seeking out scientific advice if a recurrent event endpoint is considered as part of a primary endpoint. Recently, a qualification opinion was submitted to EMA with a focus on treatment effect measures for recurrent event endpoints.[10] Overall, EMA acknowledged the gain of recurrent event methods as opposed to time-to-first event methods when mortality rates are low. In the situation with both recurrent event and terminal events, interpretation and benefits are less clear.[11] The present article will have a larger focus on marginal models compared to the qualification opinion, which especially focuses on intensity models.

As suggested by Cook and Lawless, we will distinguish between *intensity models and marginal models* for the analysis of recurrent events.[12,13] Within each of these, we will elaborate on methods explicitly accommodating competing risks from death or not. This will be formulated using multi-state models. Each method will be described and applied to two types of recurrent event endpoints from the LEADER trial: a recurrent myocardial infarction endpoint and a recurrent composite endpoint consisting of myocardial infarction, stroke and cardiovascular death. Recently, data from the LEADER trial was re-analysed considering recurrent events instead of first events.[14] That article mainly has a clinical focus, whereas the emphasis in the present article will be on statistical methodology.

1.1 | Multi-state models for life history data

Life history data captures information about events collected during a period of a lifetime of an individual.[15] Due to the period of collection, this data are typically right-censored. The fact that some individuals will experience the event of interest during this period, and others will not, characterises the right-censoring. The counting process notation provides a mathematical framework for life history data, both applicable for time-to-first event and recurrent event data, see Andersen et al.[16]

A multi-state process in continuous time is a stochastic process $Z(t)$, $t \in T$ with a finite state space $S = \{0,...,K\}$. Here $T = [0,\tau]$, $\tau < \infty$ is a time interval and $Z(t)$ represents the occupied state at time t. The transition intensity function between states is defined as

$$\lambda_{kl}(t|\mathcal{H}(t^-)) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t+\Delta t^-) = l | Z(t^-) = k, \mathcal{H}(t^-))}{\Delta t} \quad (1)$$

for the states $k,l \in S$ with $k \neq l$. Here t^- denotes a time that is infinitesimally smaller than time t. Moreover, $\mathcal{H}(t) = \{Z(s), 0 \leq s \leq t\}$ and $\mathcal{H}(t^-)$ represents the state occupation history over $[0,t)$.[15]

A **Markov** multi-state process is a special case of the general multi-state process that additionally satisfies that $\lambda_{kl}(t|\mathcal{H}(t^-)) = \lambda_{kl}(t)$. This implies that the dependence on the history is given only through the current state. This corresponds to so-called *total time models*: evaluating the time since, for example, randomisation.

For a **semi-Markov** multi-state process, it holds that $\lambda_{kl}(t|\mathcal{H}(t^-)) = h_{kl}(B(t))$, where $h_{kl}(t)$ denotes the transition intensity function and $B(t)$ denotes the time since entry into the current state k. This corresponds to so-called *gap time models*: modelling the duration of the stay in a given state. The interpretation of treatment effects in the analysis of gap times can be difficult for randomised trials, because subjects may be differently selected for the gap times following the first event. If treatment has an effect on the first gap time, a randomised comparison will be questionable on the following gap times.[13] Thus, we will not go into further details with gap time models in this article.

This article focuses on non-parametric and semi-parametric methods for quantifying treatments effects. Many of the semi-parametric approaches, assuming a non-parametric baseline function, could be extended to parametric counterparts. But the semi-parametric models impose less assumptions than the parametric models, and the non-parametric models less than the semi-parametric models. Moreover, this article focuses on multiplicative models as opposed to additive models, as they are most common in the literature and applications. Additive models do however have nice properties, which eases the strict assumptions about, for example, proportional hazards. Model checking will not be elaborated further in this article since the focus will be on the interpretation of treatment effects for the different methods. For any practical application, model checking is important and relevant.

1.2 | Overview of LEADER data

LEADER (Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results) was a double-blind randomised controlled trial investigating the cardiovascular effects of liraglutide, a glucagon like peptide-1 (GLP-1) receptor agonist approved for treatment of type 2 diabetes, investigated versus placebo when added to standard of care in a population with type 2 diabetes and a high cardiovascular risk.[8] A total of 9340 subjects were randomised 1:1 to receive either liraglutide or placebo. The primary composite endpoint was a three-component major cardiovascular adverse events (3-p MACE) endpoint consisting of; non-fatal stroke, non-fatal myocardial infarction or cardiovascular (CV) death. The primary analysis was a time-to-event analysis using a Cox regression modelling of the time to first 3-p MACE with treatment as a covariate. The median follow-up time was 3.8 years.

Cardiovascular adverse events of interest were adjudicated using an event adjudication committee to independently and blindly evaluate the events. Table 1 includes an overview of the number of subjects experiencing the first primary 3-p MACE endpoint and the distribution of subjects experiencing death, MI or stroke per treatment. In the primary LEADER publication, treatment with liraglutide affected the death rate in a positive manner (HR of 0.85, 95% CI: [0.74; 0.97]), which was driven by a difference in time-to-first CV death, as analysed using a Cox regression model with treatment as a covariate.[8] The results for CV death (HR of 0.78, 95% CI: [0.66; 0.93]) was in favour of liraglutide, with no significant difference between treatments on non-CV death (HR of 0.95, 95% CI: [0.77; 1.18]). The same

TABLE 1 Key numbers from the LEADER trial on adjudicated first events per subject

Outcome	Liraglutide ($n = 4668$) *n*, (%), *r*	Placebo ($n = 4672$) *n*, (%), *r*
Primary comp. event (3-p MACE)	608, (13.0%), 3.4	694, (14.9%), 3.9
(First of: non-fatal stroke, non-fatal MI, or CV death)		
Death from any cause	381, (8.2%), 2.1	447, (9.6%), 2.5
Cardiovascular death	219, (4.7%), 1.2	278, (6.0%), 1.6
Non-cardiovascular death	162, (3.5%), 0.9	169, (3.6%), 1.0
First myocardial infarction (fatal and non-fatal)	292, (6.3%), 1.6	339, (7.3%), 1.9
First stroke (fatal and non-fatal)	173, (3.7%), 1.0	199, (4.3%), 1.1

Note: *n* denotes number of randomised subjects. % denotes percentage of total number of randomised subjects. *r* denotes the incidence rate (number of events per 100 patient years).

analysis of myocardial infarction (not a pre-specified analysis) indicated a benefit of liraglutide compared to placebo (HR of 0.85, 95% CI: [0.73; 1.00]). For stroke, there was not a significant difference between the treatments when analysing time-to-first events using the same model (also not a pre-specified analysis) (HR of 0.86, 95% CI: [0.71; 1.06]), but the point estimate has direction in favour of liraglutide. Please be reminded that the LEADER trial was not powered to detect differences between the individual 3-p MACE components. Note, that there were fewer numbers of stroke events compared to events of myocardial infarction. For this case study, we will focus on myocardial infarction, stroke and cardiovascular (and non-cardiovascular) death which all were adjudicated. Specifically, we are focusing on

- Recurrent myocardial infarction (MI). This will be denoted *recurrent MI*.
- Recurrent composite endpoint consisting of myocardial infarction, stroke and cardiovascular death. This will be denoted *recurrent 3-p MACE*. Note that this endpoint contains all fatal or non-fatal stroke and myocardial infarction episodes.

In case of either a myocardial infarction or stroke being fatal, this has been coded as an event occurring at the calendar day of the MI or stroke and then a CV death occurring on the following day. Note, that this definition of 3-p MACE is different from the definition of 3-p MACE used in the recent LEADER article on recurrent events, using 3-p MACE defined as CV death, non-fatal stroke or non-fatal MI.[14] The recurrent MI and 3-p MACE endpoints are illustrated in Table 2. For both endpoints, there are few recurrences beyond five events. When comparing the number of recurrent events as opposed to first events, there are considerably more recurrent events. For recurrent MI, there is a total of 359 (292) and 421 (339) recurrent (first) events for liraglutide and placebo, respectively. Using first events for MI here corresponds to using $292/359 = 0.81$ (liraglutide) and $339/421 = 0.81$ (placebo) of the total number of events. For recurrent 3-p MACE, using first events corresponds to using $608/768 = 0.79$ (liraglutide) and $694/923 = 0.75$ (placebo) of the total number of events. This underlines the number of events that can be gained by using recurrent events as opposed to first events.

These two recurrent event processes illustrate two different kinds of processes. Recurrent MI is a "natural" recurrent event process that might be affected by competing risks in the form of death. Recurrent 3-p MACE is a less "natural" recurrent event process, since CV death is a part of composite endpoint and is considered an event. However, if a CV death occurs, it is not possible to have any subsequent recurrent events.

TABLE 2 Overview of recurrent MI and 3-p MACE episodes for the LEADER trial

	Number of patients with at least *e* events			
	Recurrent MI		**Recurrent 3-p MACE**	
Number of events, *e*	**Liraglutide ($n = 4668$)**	**Placebo ($n = 4672$)**	**Liraglutide ($n = 4668$)**	**Placebo ($n = 4672$)**
1	292	339	608	694
2	46	59	119	163
3	13	14	25	47
4	5	4	11	13
5	1	2	3	3
6	1	1	1	1
7	1	1	1	1
8	0	1	0	1
Total events	359	421	768	923
No events	4376	4333	4060	3978
Censoring before any event	4047	3960	3923	3845
(Competing) Death before any event	329	373	137	133

Note: *n* denotes number of randomised subjects.

Table 1 shows that there are more cardiovascular deaths with placebo than with liraglutide. There is a smaller difference between the treatments on the number of non-cardiovascular deaths. Due to the magnitude of these deaths, assuming that there is non-negligible mortality will be wrong for both endpoints. Moreover, we would expect that an estimated treatment effect for the recurrent MI endpoint to be more affected by the presence of non-negligible death in LEADER, as both CV death and non-CV death act as competing risks. When assuming negligible mortality, results on recurrent 3-p MACE will also be affected, but less so with respect to the difference between the treatments, depending on how CV and non-CV death are handled in these analyses. Table 3 displays the distribution of the event types for the recurrent 3-p MACE endpoint. Figure 1 shows the cumulative incidences estimated using the Aalen-Johansen estimator per treatment and per death cause; either cardiovascular, non-cardiovascular death or all-cause death. Please note that summing the cumulative incidences for CV and non-CV death is equivalent to considering the cumulative incidence for all-cause death. Small differences between treatments in the cumulative incidences estimates for non-CV death are noted. For CV death, there is a notable difference between the two groups from around 15 months after randomisation and beyond.

2 | TARGETS OF ESTIMATION

We are interested in quantifying a clinically relevant treatment effect in terms of the recurrent event process, especially in the context of randomised controlled trials. Different measures of treatment differences (or ratios) are possible.

Two types of multi-state models will be investigated in this article. The first is a recurrent event process with negligible or no mortality as illustrated in Figure 2. The second is a recurrent event process with non-negligible mortality as illustrated in Figure 3. For both these models, there is no "gap" between the at-risk periods. This is to be understood in the sense that an individual is immediately under risk of a new event once an event has occurred. This is not

TABLE 3 Overview of distribution of recurrent 3-p MACE types per treatment

	Number of events	
Number of events	**Liraglutide**	**Placebo**
Cardiovascular death	219	278
Myocardial infarction (fatal and non-fatal)	359	421
Stroke (fatal and non-fatal)	190	224
Total	768	923

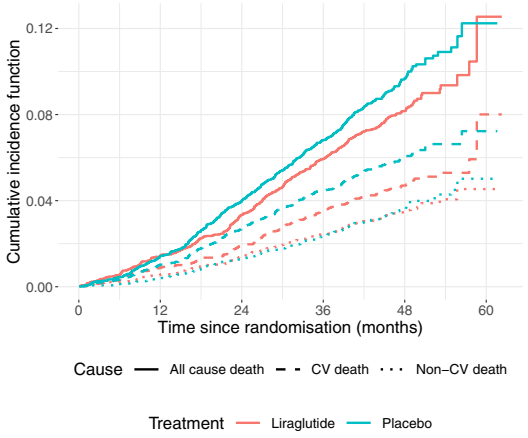


FIGURE 1 Aalen-Johansen estimates of the cumulative incidence functions per treatment and per death cause

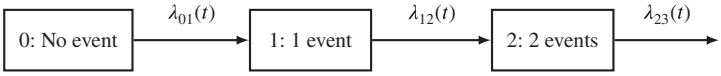


FIGURE 2 A recurrent events process with no "gaps" between at-risk period. Mortality is negligible

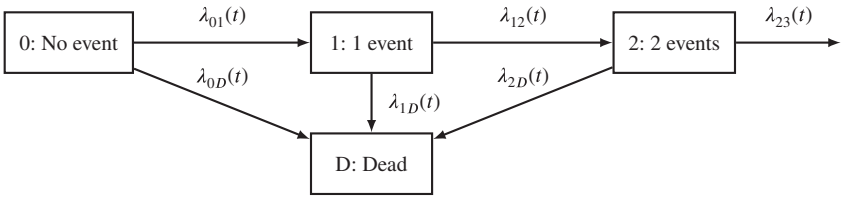


FIGURE 3 A recurrent events process with a terminal event and no "gaps" between at-risk periods. Mortality is non-negligible

TABLE 4 Example of structure of recurrent event data in counting process style. ID denotes the subject identifier. Status denotes the event indicator at the stop time (0: Censoring, 1: Event, 2: Death)

ID	Start time	Stop time	Status	Treatment
1	0	10	1	Placebo
1	10	200	1	Placebo
1	200	350	0	Placebo
2	0	1050	2	Active
3	0	400	1	Active
3	400	1500	2	Active
⋮	⋮	⋮	⋮	⋮

necessarily the clinical reality as one may experience a period where one is no longer under risk of the event. Recurrent hospitalisation could be an example, where individuals are not prone to being hospitalised when already in the hospital. These considerations could also apply to the clinical setting of the LEADER data. However, since only the onset date of the cardiovascular events where adjudicated (and not the stop date), there is no information on the duration of the events from the adjudication committee. In the data, some subjects did also experience, for example, MI events quite close to each other. For instance, one subject had two MI's on two consecutive days. That recurrent events were experienced so closely in time also supports the choice of multi-state models with no "gaps" between at-risk periods.

As we shall see, there is not a single "model for recurrent events." Rather, there are a variety of models targeting different parameters under various assumptions. Naturally, the choice of model will have an effect on the estimates and their final interpretation. Models were fitted in `R`, primarily using the `survival` and `frailtypack` packages, with usage of the `coxph` and `frailtyPenal` functions.[17–20] The data sets have been structured in counting process style exemplified in Table 4. These models can also be implemented in SAS mainly through usage of proc phreg.[21]

Multi-state models provide a general framework for studying stochastic processes. A special case of these are the recurrent event processes. For the processes that we will consider, the notation can be specified even more. For a single recurrent event process, we denote the (recurrent) event times by $T_1, T_2, \ldots$ for event 1, event 2 and so forth. These events are ordered in the sense that $T_1 < T_2 < \ldots$, for example, individuals are experiencing event 1 prior to event 2. The associated counting process, $N(t)$, counts the cumulative number of events at time t for $t \in [0, \tau]$. The event intensity function gives the instantaneous probability of an event in a very small time interval,

$$\lambda(t|\mathcal{H}(t^-)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1|\mathcal{H}(t^-))}{\Delta t} \tag{2}$$

where $\Delta N(t) = N(t+\Delta t^-) - N(t^-)$ denotes the number of events in time $[t, t+\Delta t)$, and $\mathcal{H}(t) = \{N(s) : 0 \leq s \leq t\}$ denotes the history of the process. This will characterise the recurrent event part of the multi-state models specified in Figures 2 and 3. Note, that the subscript denoting which transition is considered is omitted here, since we are considering the situation where $\lambda_{01}(t) = \lambda_{12}(t) = ...$ Alongside this, we denote the survival time by D.

3 | INTENSITY MODELS

Intensity models specify the full distributional nature of the considered multi-state model. Thus it requires a specification of the transition intensities between each of the possible states. Transition intensities and probabilities are conditional on the past as seen from Equations (1) and (2). Hence, model validity and securing no model misspecification quickly becomes key. The intensity models provide a valuable tool for understanding process dynamics.

If having one event makes you more prone to experiencing a second event and so forth, it is unrealistic to assume a common event intensity. In order to remedy such a model misspecification introduced by the correlation between the event probabilities, one solution is to condition on the past, for example, by the number of previous events. However, if interested in quantifying a treatment effect for a randomised controlled trial, this approach should be used with care. This is due to the bias that can be introduced by conditioning on the past, but post-randomisation, in terms of, for example, number of previous events, as the number of subsequent events can be influenced by the treatment. Thus, Cook and Lawless argue that intensity models that condition on the number of previous events are not optimal when analysing randomised controlled trials.[13] Due to this fact, no post-baseline covariates are included in the considered models. For simplicity and in line with the primary manuscript of LEADER, all considered models only include a single covariate, namely the treatment indicator. Other baseline covariates could be included without loss of generality. Let x denote the binary treatment indicator ($x = 0$ for placebo, $x = 1$ for liraglutide).

Several intensity models have been proposed for modelling recurrent event processes. Many of these are semi-parametric models which resemble extensions of the Cox regression model known from classical time-to-event analysis. Some of these semi-parametric approaches carry strong assumptions but have a simple structure.

3.1 | Negligible mortality

The first focus will be on models modelling a multi-state model like the one in Figure 2. For these intensity models, the focus is on specifying the structure of intensity functions for each transition. The case study LEADER data contains non-negligible deaths, corresponding to the reality envisioned in Figure 3. Thus, when analysing recurrent events under the influence of competing risks, this implies that the cause-specific hazard functions $\lambda_{01}(t), \lambda_{12}(t), ...$ in Figure 3 are modelled in the models to be presented. In practice, this can be done by treating the competing death causes as censorings formally due to the factorisation of the likelihood.[22] For the recurrent MI endpoint this implies that both CV and non-CV death are treated as censorings at the time of death in the model fitting. For the recurrent 3-p MACE endpoint, this implies that non-CV death is treated as censorings at the time of death (CV death is a part of the event here) in the model fitting.

3.1.1 | The Andersen–Gill (AG) model

The Andersen–Gill model in its simplest form is an example of a Markov model for recurrent events, suggested as an extension of the classical Cox model for time-to-event outcomes.[23] Here the event intensity function is given by

$$\lambda(t|x) = \lambda_0(t)\exp(\beta x) \tag{3}$$

in this context β is the one-dimensional regression coefficient for the treatment effect, and $\lambda_0(t)$ is the unspecified baseline intensity function. The general model formulation allows for p-dimensional, possibly time-varying, covariates. This model assumes that the event intensity depends over time on the baseline event intensity and that there is a constant effect of treatment. The transition probability of moving from 0 to 1 event is the same as when moving from 1 to 2 events given treatment at the same time t. This assumption is questionable if you have a process where the occurrence

of the event itself will change the rate of further occurrences of the event. For example, if having one stroke, makes you more prone to subsequent strokes, then this model will not hold. Moreover, you are also assuming that the treatment has the same multiplicative effect for all transitions. Finally, this method assumes independence between events, which again may be a questionable assumption. It is possible to fit this model using robust standard errors to ease this assumption. Another remedy to the likely violation of independence could be to condition on the number of previous events, but as mentioned earlier, this will create a potential bias in estimating a treatment effect. Note, that an AG model with a constant baseline intensity function corresponds to a Poisson model using the logarithm of the observation time as offset.

Estimates of the regression coefficients of this model for recurrent MI and 3-p MACE are summarised in Tables 5 and 6. See Table 7 for details on what the estimates mean for each model. The estimated hazard ratios indicate that liraglutide treatment is associated with a lower rate of both recurrent MI $\left(\exp\left(\hat{\beta}\right)=0.849\right)$ and recurrent 3-p MACE $\left(\exp\left(\hat{\beta}\right)=0.828\right)$ compared to placebo. However, note again, that the presented analysis uses all strokes and MI episodes for 3-p MACE as opposed to only non-fatal episodes presented in the primary LEADER publication on first events and the recent LEADER publication on recurrent events.[8,14] The results are still in line with the primary manuscript for LEADER, where the time-to-first event analyses indicated cardiovascular benefits of liraglutide compared to placebo on the primary composite endpoint (non-fatal stroke, non-fatal MI or CV death) and MI. The results from the analyses of time-to-first MI and 3-p MACE analysed using a Cox model with treatment as a covariate are also included in Tables 5 and 6.

3.1.2 | The Prentice–Williams–Peterson (PWP) model

The Prentice–Williams–Peterson model is an example of another Markov model.[24] The event intensity function is given by

$$\lambda(t|x)=\lambda_{k,k+1}(t)\exp\left(\beta_{k,k+1}x\right) \tag{4}$$

where $\lambda_{k,k+1}(t)$ denotes the unspecified baseline intensity function and the $\beta_{k,k+1}$ denotes the effect of treatment for the transition between states k and $k+1$. For this model, the baseline intensities are allowed to vary with each transition. This is equivalent to allowing a time dependent stratum per transition. Here, each transition is associated with a regression coefficient for the treatment effect, $\beta_{k,k+1}$. Thus, $\beta_{k,k+1}$ is a conditional estimate that measures the effect of treatment on the individual transitions. It is possible to choose an overall coefficient $\beta=\beta_{k,k+1}$ across all strata, resulting in the following model

$$\lambda(t|x)=\lambda_{k,k+1}(t)\exp(\beta x) \tag{5}$$

which provides an overall estimate of the treatment effect, β, on the recurrent event process. Note that this estimate is conditional on the previous number of events through the stratification. This model in turn assumes that the treatment has the same constant impact on the different baseline intensities. Thus, the assumption is that the treatment will secure the same, for example, rate decrease (or increase) from 2 to 3 events as from 1 to 2 events, which is proportional to the baseline intensity for a given strata. Note, that individuals are not under risk of the k'th event until the $k-1$'th is experienced. The risk set contains all subjects that are at risk of having the recurrent event. That is, all subjects who were under observation at time t and able to experience event k (required that all prior events are experienced). The property of the risk sets can however also break randomisation, since it is not possible to make a randomised comparison of treatments beyond the first event. This is analogous to the issue with conditioning on the previous number of events in the AG model. If treatment affects the occurrence of the first event, fewer treated subjects will be at risk for the second event. Consequently, some of the treatment effect may be removed through the stratification.

Due to the low number of recurrences beyond five recurrent events for both the MI endpoint and the 3-p MACE composite endpoint, the model in Equation (4) is considered up to and including transition five, that is, from $0\rightarrow 1$, $1\rightarrow 2$, $2\rightarrow 3$, $3\rightarrow 4$, and $4\rightarrow 5$. Estimates of the regression coefficients of these models for recurrent MI and 3-p MACE are summarised in Tables 5 and 6. The PWP model with a common effect is denoted PWP 2 in the table, and the model with an effect per transition is denoted PWP 1. For recurrent MI, the direction of the point estimates indicates that

TABLE 5 Overview of model results for recurrent MI. Placebo is the reference group. The time-to-first event Cox model has treatment as a covariate. PWP 1 and PWP 2 denotes the Prentice–Williams–Peterson model with a treatment effect per transition and an overall treatment effect, respectively. WLW 1 and WLW 2 denotes the Wei–Lin–Weissfeld model with an treatment effect for each event number and an overall treatment effect, respectively

Model	$\widehat{\theta}$	$\mathbf{se}\left(\widehat{\theta}\right)$	$\exp\left(\widehat{\theta}\right)$	**95% CI for** $\exp(\theta)$
Cox	$\widehat{\beta} = -0.159$	0.080	0.853	[0.729; 0.997]
(Time-to-first)				
AG	$\widehat{\beta} = -0.164$	0.072	0.849	[0.737; 0.977]
(Non-par.)				
AG	$\widehat{\beta} = -0.164$	0.072	0.849	[0.737; 0.977]
(Piece. const.)				
PWP 1	$\widehat{\beta}_{01} = -0.159$	0.080	0.853	[0.729; 0.997]
	$\widehat{\beta}_{12} = -0.047$	0.197	0.954	[0.649; 1.404]
	$\widehat{\beta}_{23} = -0.023$	0.400	0.977	[0.446; 2.139]
	$\widehat{\beta}_{34} = 0.629$	0.737	1.875	[0.442; 7.951]
	$\widehat{\beta}_{45} = -0.429$	1.230	0.651	[0.058; 7.257]
PWP 2	$\widehat{\beta} = -0.130$	0.072	0.878	[0.762; 1.011]
Frailty	$\widehat{\beta} = -0.177$	0.088	0.838	[0.705; 0.995]
(Non-par.)	$\widehat{\phi} = 5.687$	Not available from `coxph`[a]		
Frailty	$\widehat{\beta} = -0.177$	0.088	0.838	[0.705; 0.995]
(Piece. const.)	$\widehat{\phi} = 5.694$	Not available from `coxph`[b]		
Joint frailty	Convergence failed			
(Non-par.)				
Joint frailty	$\widehat{\beta} = -0.186$	0.068	0.830	[0.727; 0.949]
(Piece. const.)	$\widehat{\alpha} = -0.211$	0.078	0.809	[0.694; 0.944]
	$\widehat{\phi} = 0.896$	0.031		
	$\widehat{\gamma} = 1.860$	0.115		
WLW 1	$\widehat{\beta}_1 = -0.159$	0.080	0.853	[0.729; 0.997]
	$\widehat{\beta}_2 = -0.255$	0.197	0.775	[0.527; 1.139]
	$\widehat{\beta}_3 = -0.084$	0.384	0.920	[0.433; 1.953]
	$\widehat{\beta}_4 = 0.219$	0.671	1.245	[0.334; 4.639]
	$\widehat{\beta}_5 = -0.701$	1.224	0.496	[0.045; 5.461]
WLW 2	$\widehat{\beta} = -0.167$	0.088	0.846	[0.711; 1.006]
LWYY	$\widehat{\beta} = -0.164$	0.088	0.849	[0.714; 1.009]
GL	$\widehat{\beta} = -0.159$	0.088	0.853	[0.718; 1.013]

Abbreviations: AG, Andersen–Gill; GL, Ghosh–Lin; LWYY, Lin–Wei–Yang–Ying; PWP, Prentice–Williams–Peterson; WLW, Wei–Lin–Weissfeld.
[a]A frailty model (non-par. baseline hazard fitted using `frailtypack`): $\widehat{\theta} = \left(\widehat{\beta}, \widehat{\phi}\right) = (-0.177, 5.614)$ $\left(\mathrm{se}\left(\widehat{\theta}\right)\right) = (0.088, 0.590)$.
[b]A frailty model (piece-wise constant baseline hazard with 5 equidistant cuts fitted using `frailtypack`): $\theta = \left(\widehat{\beta}, \widehat{\phi}\right) = (-0.177, 5.598)$ $\left(\mathrm{se}\left(\widehat{\theta}\right)\right) = (0.088, 0.609)$.

liraglutide is more beneficial than placebo, except for transition $3 \to 4$. This is due to the slight imbalance and the diminished risk sets present for event number four as seen in Table 2: 5 events with liraglutide and 4 events with placebo. Here, this leads to the interpretation that liraglutide seems to increase the intensity of going from 3 to 4 events, whereas decreasing the intensity of all other transition compared to placebo. The low number of events and the

TABLE 6 Overview of model results for recurrent 3-p MACE (stroke, MI or CV death). Placebo is the reference group. The time-to-first event Cox model has treatment as a covariate. PWP 1 and PWP 2 denotes the Prentice–Williams–Peterson model with a treatment effect per transition and an overall treatment effect, respectively. WLW 1 and WLW 2 denotes the Wei–Lin–Weissfeld model with an treatment effect for each event number and an overall treatment effect, respectively. ML 1 denotes the Mao and Lin model that accommodates CV death, but does not address the competing non-CV death. ML 2 denotes the Mao and Lin type model that accommodates both CV and non-CV death. Both ML methods had severity $c = (1,1,1)$

Model	$\widehat{\theta}$	$\mathbf{se}\left(\widehat{\theta}\right)$	$\exp\left(\widehat{\theta}\right)$	**95% CI for** $\exp(\theta)$
Cox	$\widehat{\beta} = -0.142$	0.056	0.868	[0.778; 0.968]
(Time-to-first)				
AG	$\widehat{\beta} = -0.189$	0.049	0.828	[0.752; 0.911]
(Non-par.)				
AG	$\widehat{\beta} = -0.189$	0.049	0.828	[0.752; 0.911]
(Piece. const.)				
PWP 1	$\widehat{\beta}_{01} = -0.142$	0.056	0.868	[0.778; 0.968]
	$\widehat{\beta}_{12} = -0.218$	0.121	0.804	[0.635; 1.020]
	$\widehat{\beta}_{23} = -0.418$	0.250	0.659	[0.403; 1.076]
	$\widehat{\beta}_{34} = 0.547$	0.463	1.727	[0.698; 4.276]
	$\widehat{\beta}_{45} = 0.029$	1.008	1.029	[0.143; 7.417]
PWP 2	$\widehat{\beta} = -0.157$	0.049	0.855	[0.777; 0.941]
Frailty	$\widehat{\beta} = 0.232$	0.067	0.793	[0.696; 0.904]
(Non-par.)	$\widehat{\phi} = 4.395$	Not available from `coxph`[a]		
Frailty	$\widehat{\beta} = -0.232$	0.067	0.793	[0.696; 0.904]
(Piece. const.)	$\widehat{\phi} = 4.403$	Not available from `coxph`[b]		
Joint frailty	Convergence failed			
(Non-par.)				
Joint frailty	$\widehat{\beta} = -0.207$	0.058	0.813	[0.725; 0.911]
(Piece. const.)	$\widehat{\alpha} = -0.083$	0.108	0.921	[0.746; 1.137]
	$\widehat{\phi} = 0.951$	0.029		
	$\widehat{\gamma} = 1.398$	0.149		
WLW 1	$\widehat{\beta}_1 = -0.142$	0.056	0.868	[0.778; 0.968]
	$\widehat{\beta}_2 = -0.323$	0.121	0.724	[0.572; 0.917]
	$\widehat{\beta}_3 = -0.639$	0.247	0.528	[0.325; 0.857]
	$\widehat{\beta}_4 = -0.175$	0.409	0.839	[0.376; 1.871]
	$\widehat{\beta}_5 = -0.026$	0.809	0.975	[0.200; 4.756]
WLW 2	$\widehat{\beta} = -0.192$	0.062	0.825	[0.731; 0.931]
LWYY	$\widehat{\beta} = -0.189$	0.060	0.828	[0.735; 0.931]
GL	$\widehat{\beta} = -0.190$	0.060	0.827	[0.735; 0.931]
ML 1	$\widehat{\beta} = -0.183$	0.059	0.833	[0.742; 0.935]
ML 2	$\widehat{\beta} = -0.183$	0.059	0.832	[0.741; 0.934]

Abbreviations: AG, Andersen–Gill; GL, Ghosh–Lin; LWYY, Lin–Wei–Yang–Ying; ML, Mao–Lin; PWP, Prentice–Williams–Peterson; WLW, Wei–Lin–Weissfeld.

[a]A frailty model (non-par. baseline hazard fitted using `frailtypack`): $\widehat{\theta} = \left(\widehat{\beta}, \widehat{\phi}\right) = (-0.232, 4.400)$ $\left(\mathrm{se}\left(\widehat{\theta}\right)\right) = (0.067, 0.315)$.

[b]A frailty model (piece-wise constant baseline hazard with 10 equidistant cuts fitted using `frailtypack`): $\widehat{\theta} = \left(\widehat{\beta}, \widehat{\phi}\right) = (-0.231, 4.387)$ $\left(\mathrm{se}\left(\widehat{\theta}\right)\right) = (0.066, 0.308)$.

TABLE 7 Overview of methods for analysing recurrent events with terminal events

Method	Model	Target of estimation	Terminal events
Andersen-Gill	Intensity Semi-parametric	Cause-specific hazard ratio	Terminal events correctly handled by technically censoring at time of death
Prentice–Williams–Peterson	Intensity Semi-parametric	Cause-specific (stratified) hazard ratios	Terminal events correctly handled by technically censoring at time of death
Frailty	Intensity Semi-parametric	Cause-specific (conditional) hazard ratio	Terminal events likely handled incorrectly, technically done by censoring at time of death, due to plausible violation of the assumption of independent censoring given frailty
Joint frailty	Intensity Semi-parametric	Cause-specific (conditional) hazard ratios	Recurrent event and terminal event process correctly modelled simultaneously
Wei–Lin–Weissfeld	Marginal Semi-parametric	(Marginal) hazard ratio	Terminal events incorrectly handled by technically censoring at time of death. Estimates relate to a situation where it is not possible to die
Lawless and Nadeau	Marginal Non-parametric	Mean functions	Terminal events incorrectly handled by technically censoring at time of death. Estimates relate to a situation where it is not possible to die
Lawless and Nadeau, Lin–Wei–Yang–Ying	Marginal Semi-parametric	Mean ratio	Terminal events incorrectly handled by technically censoring at time of death. Estimates relate to a situation where it is not possible to die
Cook and Lawless, Ghosh and Lin	Marginal Non-parametric	Mean functions	Terminal events correctly handled by adjustment for death through modelling of survival distribution
Ghosh and Lin	Marginal Semi-parametric	Mean ratio	Terminal events correctly handled by adjustment for death using IPCW or IPSW
Mao and Lin	Marginal Semi-parametric	Mean ratio	Terminal events correctly handled when as defined as part of the composite endpoint. Adjustment done using IPCW
Adjusted Mao and Lin	Marginal Semi-parametric	Mean ratio	Terminal events correctly handled when as defined as part of the composite endpoint or terminal competing events. Adjustment done using IPCW

reduction of the risk sets are reflected in the confidence intervals. A statement like "it reduces the rate of getting 1-3 events, but not the fourth event" does not ring that nice. The picture is clearer for recurrent 3-p MACE, however it still does not allow an easy interpretation of what the effect of treatment is, since the estimates are per transition. The major caveat of this model is that randomisation is not preserved past the first transition due to the stratification. If the treatment is effective, some of the treatment effect will be removed through the stratification. This makes the PWP model less obvious to use to summarise a treatment effect on recurrent events for randomised controlled trials.

3.1.3 | Frailty model

Another common model is the frailty model which can be used for analysing clustered survival data. The frailty terminology was first introduced by Vaupel.[25] A random term is present in this multiplicative model in order to account for possible positive correlation of events within a subject. In the survival context, this random term is known as a frailty, as certain subjects may be more "frail" to experiencing the event of interest. The event intensity function is given by

$$\lambda(t|x,\nu) = \nu\lambda_0(t)\exp(\beta x) \tag{6}$$

here β and $\lambda_0(t)$ are the regression coefficient for the treatment effect and the unspecified baseline intensity function given frailty, respectively. The subject specific frailty is denoted ν. A common choice is to choose a Gamma distributed frailty, with mean equal to 1 and a variance of ϕ. This is what we will consider, although other choices, for example, a log-normal frailty are possible. This model is sometimes called a shared frailty model, referring to that the frailty term is shared within a subject, in the sense that it affects all transitions in the same (multiplicative) way.

Frailty models are subject specific models. Hence, hazard ratios from a frailty model will have a subject specific interpretation, in the sense that the estimates are hazard ratios for a given frailty (ν) in Equation (6). It is assumed that the treatment effect, β, is the same for all frailty levels. Keeping the frailty constant, $\exp(\beta)$ is interpreted as the increase (or decrease) in momentarily probability of having an event when considering liraglutide treatment as opposed to placebo treatment. This interpretation is nicer if one wishes to compare treatments within a cluster, for example, subjects. However, in the setting of summarising a randomised controlled trial this is likely not an interpretation of interest: You would like an interpretation on a population level rather than an interpretation on an individual level. Note that fitting a frailty model with a constant baseline intensity corresponds to fitting a negative binomial model using the logarithm of the observation time as offset.

Parameter estimates are available in Tables 5 and 6. For recurrent MI, the estimated regression coefficient is $\hat{\beta} = -0.177$ with an estimated $\hat{\phi} = 5.69$. For recurrent 3-p MACE, it is $\hat{\beta} = -0.232$ with an estimated $\hat{\phi} = 4.39$. Note, that the estimates for the frailty models are larger on an absolute scale than the corresponding estimates from the AG model. This is a result of the attenuation of the marginal estimates compared to the conditional estimates. Again, this is due to the interpretation of the estimates from the two types of models; estimates from the frailty model describe individual hazard ratios, whereas estimates from, for example, the AG model describe population-based hazard ratios.[26,27] Note, that standard error estimates of $\hat{\beta}$ from R obtained using `coxph` does not take the variability of $\hat{\phi}$ into account, so these should be interpreted with care.[27,28] Estimation of $\hat{\phi}$ using `frailtypack` allows for variance estimation of $\hat{\phi}$.

An underlying assumption of this model, is that conditional on the frailty, censoring is independent.[29] However, when there is non-negligible mortality as for both considered endpoints in the LEADER data, this assumption will most likely not hold. If being more frail to experience recurrent events as MI or 3-p MACE, also makes you more frail to experience death, this conditional independence clearly does not hold. Thus, this model should be interpreted with caution. The limitation of this model is the subject specific interpretation of hazard ratio estimates, and the potential violation of the independent censoring assumption when considering recurrent event data with competing deaths.

3.2 | Non-negligible mortality

As an alternative to the previous models, it is possible to fully specify the intensities in the multi-state model envisioned in Figure 3. That is, both fully specifying the cause-specific intensities for recurrent events, $\lambda_{01}(t), \lambda_{12}(t), ...$, and the cause-specific intensities for death, $\lambda_{0D}(t), \lambda_{1D}(t), ...$

As discussed in the previous section, the frailty model may break down due to dependent censoring in the setting with competing risks. The next model, the joint frailty model, has been proposed as an alternative, which fully specifies the intensities for both recurrent events and death.

3.2.1 | Joint frailty model

Liu, Wolfe and Huang suggested a joint frailty model for modelling the intensity function for recurrent events and a terminal event simultaneously using a common frailty ν.[30] The model specifies that the intensity function for the recurrent events, $\lambda(t)$, and the intensity function for survival, $\lambda^D(t)$, is given by

$$\lambda(t|x,\nu) = \nu \exp(\beta x)\lambda_0(t)$$

$$\lambda^D(t|x,\nu) = \nu^\gamma \exp(\alpha x)\lambda_0^D(t)$$

in the special case of only a single covariate, treatment, affecting both the recurrent event process and the death process. The baseline intensity functions are given by $\lambda_0(t)$ (recurrence) and $\lambda_0^D(t)$ (death). Here, ν denotes the common

frailty term assumed to be Gamma distributed with unit mean and variance ϕ. The regression coefficients, β and α, quantify the effect of treatment on the recurrent event and death process, respectively. The γ parameter expresses the relationship between the recurrent event process and the death process given frailty. For example, if $\gamma > 0$, a higher frailty will result in a higher rate of both recurrent events and death given frailty.

Liu, Wolfe and Huang suggest using the EM-algorithm for estimation as the density has no closed form solution.[30] Rondeau et al suggested to use penalised maximum likelihood for estimation.[31] This is implemented in R in the software package `frailtypack`.[20] Fitting this joint model involves selecting smoothing parameters, one for the recurrent events and one for the survival, up front. The authors suggest facilitating this by first fitting two models, a frailty model for the recurrent event process and a Cox model for the death process, and choose smoothing parameters by cross validation. These estimates are then to be supplied to the model fit for the joint model. Moreover, the method involves a non-trivial selection of number of knots for the splines approximating the baseline intensities. Here, the authors suggest increasing the number of knots from 7 until the graphs of the baseline hazards remain unchanged.

For recurrent MI, $\lambda(t|x,\nu)$ models recurrent MI episodes and $\lambda^D(t|x,\nu)$ models all deaths. For recurrent 3-p MACE, $\lambda(t|x,\nu)$ models recurrent 3-p MACE episodes and $\lambda^D(t|x,\nu)$ models competing non-CV deaths. The joint frailty model with non-parametric baseline intensities for these data sets failed to converge using the `frailtypack` package. This might be due to the size of the data sets (roughly 10,100 records for the MI data set and 10,500 for the 3-p MACE data set). When trying to fit the model on a smaller subset of the data sets, say first 1000 rows, the model did converge. However, it could also be due to the non-trivial choices of parameters to be made. The joint frailty model with non-parametric baseline intensities can be preferable in terms of minimising the number of assumptions. However, the flexibility comes at the cost of selecting several tuning parameters. Due to the failed convergence of the non-parametric version of the joint frailty model, a parametric model with a piece-wise constant baseline hazard was also considered. The piece-wise constant baseline hazard model is also implemented in `frailtypack`. For comparability, we also fitted an AG model and a frailty model using a piece-wise constant baseline hazard function. The intervals were decided using five or ten equally large intervals over the period of follow-up time, for MI and 3-p MACE respectively. The results are displayed in Tables 5 and 6. For the AG and frailty model, there were almost no difference between the estimates obtained using either a non-parametric estimate of the baseline hazard function or the piece-wise constant version for both endpoints. Thus, it should be fair to compare the results from the fitted joint frailty model with piece-wise constant hazard functions for both the recurrent events and the deaths to the results from the AG and frailty model. The results from the joint frailty models differ from those obtained for the frailty model, which may indicate the degree to which the dependent censoring affects the frailty model.

For recurrent MI, the estimated coefficients indicate that liraglutide reduces the rate of MI recurrences $\left(\hat{\beta} = -0.186\right)$ and the rate of death $(\hat{\alpha} = -0.211)$ compared to placebo. The effect on the recurrences is larger than the effect seen for the frailty model $\left(\hat{\beta} = -0.177\right)$. The estimated $\hat{\gamma} = 1.860$ indicates that experiencing myocardial infarction makes you more prone to dying given treatment and frailty. The recurrence rate seems to vary less among patients $\left(\hat{\phi} = 0.896\right)$ for the joint model compared to the frailty models $\left(\hat{\phi} = 5.69\right)$. This could indicate that the deaths accounts for a large amount of the heterogeneity. For recurrent 3-p MACE, the estimated coefficients indicate that liraglutide reduces the rate of 3-p MACE recurrences $\left(\hat{\beta} = -0.207\right)$ and the rate of death $(\hat{\alpha} = -0.083)$ compared to placebo. The estimated effect on the recurrences is slightly smaller than the effect observed for the frailty model $\left(\hat{\beta} = -0.232\right)$. The effect of treatment on death for 3-p MACE is smaller than for MI, which is not surprising, since CV death acts as event and only non-CV death is modelled for the death process for 3-p MACE. A positive $\hat{\gamma} = 1.398$ is observed, indicating that an increased rate in recurrences of 3-p MACE is positively associated with an increased rate of dying of non-CV death causes. The variance of the frailty term is seen to be smaller for the joint frailty model $\left(\hat{\phi} = 0.951\right)$ compared to the frailty model $\left(\hat{\phi} = 4.40\right)$. Again, this may indicate that non-CV death is accounting for a lot of the heterogeneity.

The drawback of the joint frailty model is that the interpretation of the hazard ratio estimates has some of the same disadvantages as the frailty model described in Section 3.1.3. Namely, that the estimates will have a subject specific interpretation. However, since the recurrent event process and the death process are modelled simultaneously, the joint frailty model does not suffer the same probable issues regarding informative censoring as the frailty model.

4 | MARGINAL MODELS

Marginal models provide an alternative to the intensity models. Instead of fully specifying the intensity functions in a multi-state model, these focus on marginal features of the stochastic process. Marginal approaches specify a partial model targeting a specific attribute, which typically will not be conditional on the previous event history. Marginal parameters in a multi-state

model for recurrent events include; time (from 0) to entry into a given state, expected number of events $(E(N(t)))$, state occupation probabilities $(P(Z(t)=l))$ and average time spent in a state $\left(\int_0^\tau P(Z(t)=l)\,dt\right)$.[15]

4.1 | Negligible mortality

The models in this section adhere to the situation in Figure 2. Since the LEADER data contains deaths, a consequence of the below models is that deaths are treated as censorings. We shall see that this censoring is problematic when targeting a marginal parameter.

4.1.1 | Wei–Lin–Weissfeld (WLW) model

The Wei–Lin–Weissfeld model is a marginal model that describes the marginal distributions of waiting times until event one, event two and so forth.[32] The intensity function for the l'th event is given by:

$$\lambda_l(t|x) = \lambda_{0l}(t)\exp(\beta_l x) \tag{7}$$

here β_l and $\lambda_{0l}(t)$ express the effect of treatment and the unspecified baseline intensity function. All subjects are under risk of the l'th event from randomisation. This model ignores the ordering of events as you are under risk of getting the third event even before the second (or even first event is experienced). Thus, the risk set is artificially inflated here, since all subjects are not under risk of getting all possible events from randomisation. As noted by Metcalfe and Thompson, this structure implies that the effects on event $k-1$ will be "carried over" to event k, as a delayed waiting time for the first event also will imply a delay in the waiting time for a subsequent event.[33] As all subjects are at risk for experiencing the k'th event (except if they die prior to the k'th event), the WLW model has been praised for preserving randomisation as opposed to the PWP model.

It is possible to estimate a single treatment effect for this model, by stratifying for the event number and estimating a fixed treatment effect across all. This model is formulated as

$$\lambda_l(t|x) = \lambda_{0l}(t)\exp(\beta x) \tag{8}$$

β denotes the overall effect of treatment. As pointed out by Li and Lagakos,[34] if one targets a marginal hazard, and then censors for deaths the independent censoring assumption in the WLW model is infringed. In turn, the estimated

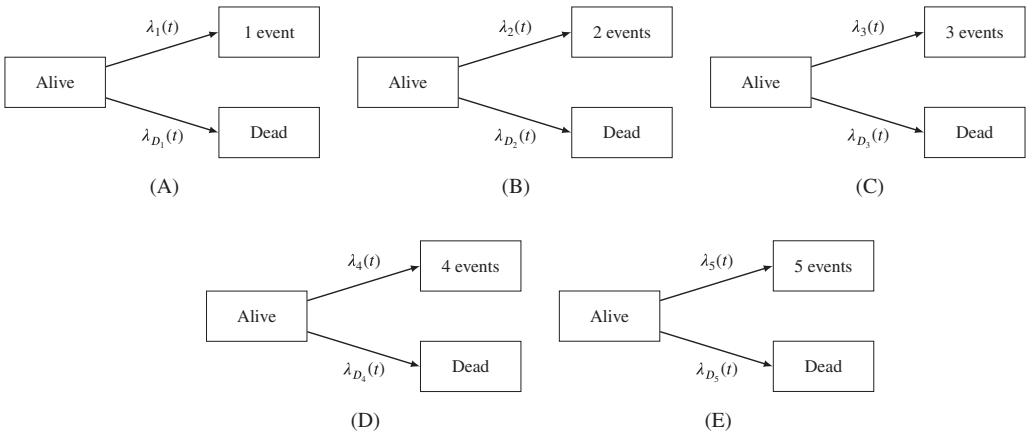


FIGURE 4 The five WLW competing risks models; (A) is the model for the first event, (B) is the model for the second event, (C) is the model for the third event, (D) is the model for the fourth event, (E) is the model for the fifth event

marginal hazard refers to situation where it is not possible to die. Instead, Li and Lagakos suggest considering a marginal WLW model for composite death and recurrent event. One may also perceive the model as describing a marginal cause-specific hazard, which correctly takes death into account as a competing event, see Figure 4. When fitting the WLW model, note that the data structure is slightly different from the counting style process structure, see Table 4, applicable for fitting, for example, the AG model. You need to create a new data set, in which all individuals are under risk of the different events, say up to five events, already from randomisation.

Due to the low number of recurrences beyond five recurrent events for both the recurrent MI endpoint and the 3-p MACE endpoint, the models are considered up to and including five events. Figure 4 displays the five relevant competing risk models, where $\lambda_{D_j}(t), j = 1,...,5$ is the cause-specific hazard of death without event number j. The results of fitting these models are presented in Tables 5 and 6. WLW 1 corresponds to the model in Equation (7), and WLW 2 corresponds to the model in Equation (8). For both models, the standard error estimates are robust. Note, that the estimates of $\hat{\beta}_1$ for the WLW 1 model coincide with the estimates of $\hat{\beta}_{01}$ from the PWP 1 model, which again corresponds to the estimates from the time-to-first event Cox model. We see that the overall direction of the estimates from the WLW 1 models and the PWP 1 models agree for at least the first three events. However, the enlargement of the treatment effects is also seen as well as the inflated precision, which is due to the constructed risk sets. When comparing PWP 2 and WLW 2 models, the estimates are smaller for the WLW models for both endpoints, however with larger standard errors.

Due to the uncompelling nature of the risk sets and their impact, the WLW model would rarely be of interest if modelling a recurrent event process in the context of a randomised trial. Moreover, with non-negligible mortality the estimated marginal hazard refers to a hypothetical population in which it is not possible to die or a marginal cause-specific hazard. All in all, the WLW is not recommendable for analysing recurrent events to capture a treatment effect in randomised controlled trials, neither in the presence of terminal events or not.

4.1.2 | Lawless and Nadeau

Lawless and Nadeau suggested to consider a non-parametric estimate of the marginal mean function $\mu(t) = E(N(t))$.[35] This is the expected number of events per subject by time t. It turns out that $\mu(t)$ can be estimated using the Nelson-Aalen estimator, $\hat{\mu}(t) = \int_0^t \frac{1}{Y(u)} dN(u)$, where $Y(t)$ denotes the number of subjects at risk at time t. Especially, this can be applied to a two-sample setting, where the marginal mean function is estimated per treatment, which we denote $\hat{\mu}_0(t)$ and $\hat{\mu}_1(t)$ for placebo and liraglutide, respectively.

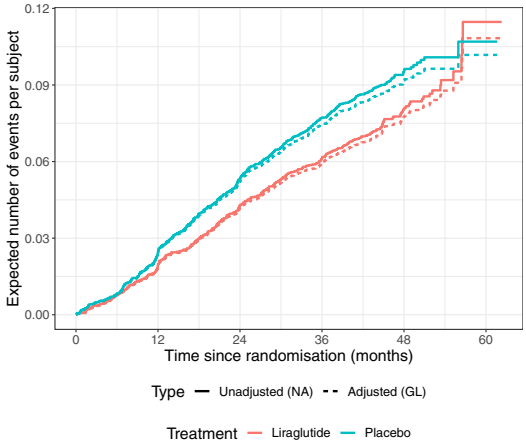


FIGURE 5 Non-parametric marginal mean function estimates per treatment for recurrent MI. "Unadjusted (NA)" refers to the Nelson-Aalen estimates and "Adjusted (GL)" refers to the Ghosh–Lin estimates of $\mu(t)$, respectively

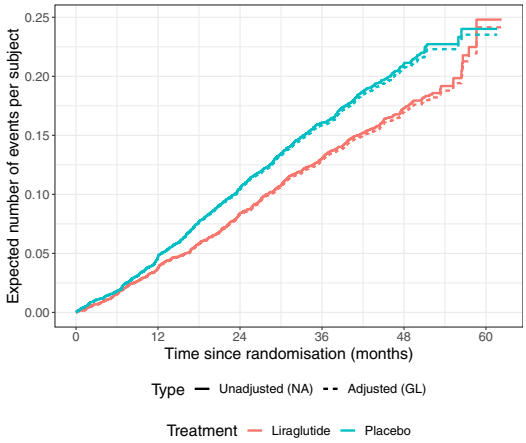


FIGURE 6 Non-parametric marginal mean function estimates per treatment for recurrent 3-p MACE. "Unadjusted (NA)" refers to the Nelson-Aalen estimates and "Adjusted (GL)" refers to the Ghosh–Lin estimates of $\mu(t)$, respectively

When censoring for death (non-CV death for 3-p MACE), the Nelson-Aalen estimates of the expected number of events per subject per treatment can be computed easily using standard software. The estimates for recurrent MI and 3-p MACE are shown in Figures 5 and 6. The expected number of both MI and 3-p MACE events per subject are larger for placebo than liraglutide, with a noticeable difference between the curves after around 1 year. Here we are addressing what happens in a population where no one dies and are still under risk of a recurrent event after death, due to the censorings for death. This might be reasonable if you have little or no deaths in your data, but this is not meaningful with a substantial amount of death as for the LEADER data.

4.1.3 | Lin–Wei–Yang–Ying (LWYY) model

A semi-parametric regression model for the marginal mean was suggested by Lawless and Nadeau and Lin, Wei, Yang and Ying.[35,36] We consider the model that focuses on the marginal mean function, $\mu(t|x) = E(N(t)|x)$. The model is given by

$$\mu(t|x) = \mu_0(t)\exp(\beta x)$$

here β and $\mu_0(t)$ express the regression coefficient for treatment and the baseline mean function, respectively. This is a semi-parametric model, with β expressing the parametric term and $\mu_0(t)$ the non-parametric term. Since events are correlated, a robust sandwich type estimator is used for estimating the variance of β. The LWYY model does not require a full specification of the correlation structure of the entire multi-state process.

The estimated mean functions from the LWYY models stated above are available in Figures 7 and 8. The regression estimates for the LWYY models are displayed in Tables 5 and 6. Fitting an AG model with robust standard errors would correspond to fitting a LWYY model, however with an interpretation on an intensity level (rather than a marginal mean level). The estimating equations for β leave you at the exact same place for the two models. That is why the point estimates for $\hat{\beta}$ for the LWYY model coincide with those of the AG model. However, standard error estimates differ, since the LWYY model does not assume independence between events, as opposed to the equivalent AG model. Hence, standard errors are slightly larger for the LWYY model as compared to the AG model, indicating a dependence between events.

If the assumption of the multiplicative form of the marginal mean function for LWYY model is questionable, for example, if the effect of treatment changes over time, it might be preferable to model $\mu(t)$ non-parametrically, using the approach outlined by Lawless and Nadeau (see previous section). A two-sample test for comparing the two treatments is also available in this non-parametric set-up.

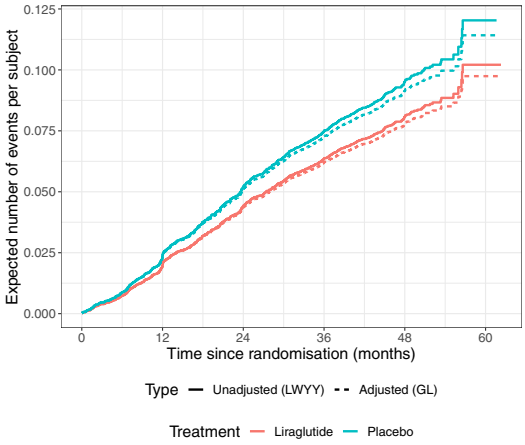


FIGURE 7 Model-based marginal mean function estimates per treatment for recurrent MI. "Unadjusted (LWYY)" refers to the Lin–Wei–Yang–Ying estimate and "Adjusted (GL)" refers to the Ghosh and Lin estimate of $\mu(t|x)$, respectively

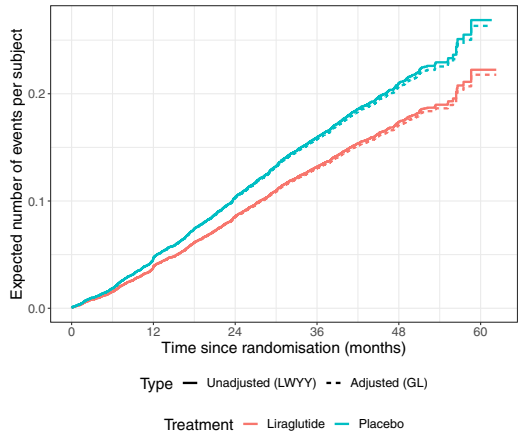


FIGURE 8 Model-based marginal mean function estimates per treatment for recurrent 3-p MACE. "Unadjusted (LWYY)" refers to the Lin–Wei–Yang–Ying estimate and "Adjusted (GL)" refers to the Ghosh and Lin estimate of $\mu(t|x)$, respectively

An assumption of the LWYY model is that the censoring mechanism is independent. This assumption is questionable when censoring for death as dying implies that no subsequent events can occur. Moreover, it implies that the estimated mean functions address the expected number of event per subject over time in a population where it is not possible to die. As for the non-parametric model suggested by Lawless and Nadeau (see previous section), the LWYY model will not be reasonable to use when analysing recurrent events data with deaths in the magnitude as seen in the LEADER data.

4.2 | Non-negligible mortality

The models in this section describe the reality in Figure 3, where it is both possible to experience recurrent events and competing deaths. The following marginal models explicitly handles competing deaths.

4.2.1 | Ghosh and Lin (GL)

This approach was first suggested by Cook and Lawless in 1997, and further developed by Ghosh and Lin in 2000.[37,38] They suggest a non-parametric method for estimating the marginal expected number of recurrent events per subject up to time t, namely $\mu(t) = E(N(t))$, acknowledging that no recurrent events can happen after death. It holds that

$$\mu(t) = \int_0^t S(u^-)\, dR(u)$$

where $S(t) = P(D > t)$ and $dR(t) = E\{dN(t) | D \geq t\}$. Then $\mu(t)$ can be estimated by

$$\hat{\mu}(t) = \int_0^t \hat{S}(u)\, d\hat{R}(u)$$

where $\hat{S}(t)$ is the Kaplan–Meier estimate of S and $\hat{R}(t)$ is the Nelson-Aalen estimate of R. This can be estimated per treatment group, $\hat{\mu}_0(t)$ and $\hat{\mu}_1(t)$.

With no mortality, $\hat{\mu}(t) = \hat{R}(t)$. However, with mortality, it holds that $\hat{\mu}(t) < \hat{R}(t)$. Thus, the Nelson-Aalen estimator, when treating deaths as censoring, overestimates the mean function. This is the same type of issue known from competing risk: When treating deaths as censorings, you are assuming that dead individuals are still under risk having an event and the mean will be overestimated.

The estimates of $\hat{\mu}_0(t)$ and $\hat{\mu}_1(t)$ obtained from the above is illustrated in Figures 5 and 6. These are plotted alongside the Nelson-Aalen estimates of $\mu(t)$ (censoring for competing deaths). As expected, since there is non-negligible mortality, the Nelson-Aalen estimates are upwards biased as they are not appropriately correcting for the deaths that are occurring. The impact of this adjustment is larger for recurrent MI compared to recurrent 3-p MACE, as the CV death component acts as an event for the composite endpoint. The magnitude of this bias depends on the data at hand. For the 3-p MACE endpoint, a fully correct adjustment should also correct for the CV death. This will be discussed later.

4.2.2 | Ghosh and Lin (GL) model

Ghosh and Lin suggested to consider a regression model for the marginal mean function acknowledging the fact that no recurrences can happen after death.[39] This is an extension of the LWYY model allowing for non-negligible mortality. Mortality is adjusted for using either an inverse probability of censoring weights (IPCW) or an inverse probability of survival weights (IPSW). We consider the following model for the mean function

$$\mu(t|x) = \mu_0(t)\exp(\beta x)$$

here β and $\mu_0(t)$ express the regression coefficient for treatment and the baseline mean function, respectively. This model specification coincides with that of the LWYY model formulation above. However, the estimating equations are updated as to re-weight the data from subjects that have died, either using IPCW or IPSW. While it may be reasonable to assume that the censoring times are independent of covariates, the same most likely does not hold for the survival distribution.[39] Thus, IPCW may be implemented by either imposing a semi-parametric or a non-parametric model for the censoring distribution. Whereas, IPSW can be implemented using a semi-parametric model for the survival distribution. We have focused on the IPCW approach, with weights estimated by estimating the censoring distribution using Kaplan–Meier (unconditional on any covariates). This bears resemblance to the Fine and Gray model also utilising IPCW.[40] We will denote this by the GL model.

The estimated mean functions are shown from Figures 7 and 8. In these Figures, the estimates obtained from LWYY model (censoring for competing deaths) and the GL model above are displayed together. As expected, the LWYY models are overestimating the expected number of events per subjects when censoring for deaths. The estimates for the regression coefficients are given in Tables 5 and 6. For recurrent MI, the estimated regression coefficient for the GL model $\left(\hat{\beta} = -0.159, 95\%\ \text{CI}: [0.718; 1.013]\right)$ was slightly larger compared to the LWYY model estimates $\left(\hat{\beta} = -0.164, 95\%\ \text{CI}: [0.714; 1.009]\right)$, but with similar standard errors. From Table 1, it is seen that more deaths occur in the placebo group $(447\,(9.6\%))$ compared to the liraglutide group $(381\ (8.2\%))$. This is apparent in the GL estimates

versus the LWYY estimates. For the GL model, the difference between the treatments in $\hat{\beta}$ is smaller, since the placebo subjects will have less at risk time compared to liraglutide subjects due to the excess deaths. For recurrent 3-p MACE, the estimated regression coefficients and standard errors for the GL and LWYY model are more or less identical. This is not surprising due to the very equal distribution of non-CV deaths between liraglutide $(162\,(3.5\%))$ and placebo $(169\,(3.6\%))$ as seen in Table 1. Thus, even though the mean functions are overestimated using the LWYY versus the GL model (see Figure 8), the difference between the two treatments expressed through the regression coefficient is fairly similar for 3-p MACE.

4.2.3 | Mao and Lin (ML) model

Mao and Lin suggests modelling the mean function for a severity weighted composite endpoint, where death is a part of the composite endpoint.[41] Thus, we only use this method for modelling the recurrent composite 3-p MACE endpoint. Here it is assumed that there is K different types of events, including the terminal event. For each event type, $N_k(t)$ for $k = 1,...,K$, denotes the cumulative number of the k'th event type experienced by time t. Each event type is assigned a relative severity, c_k, for the severity weighted process $N(t) = \sum_{k=1}^{K} c_k N_k(t)$. It is assumed that the marginal mean function satisfies that

$$\mu(t|x) = \mu_0(t)\exp(\beta x)$$

with β denoting the treatment effect and $\mu_0(t)$ the baseline mean function. Since the terminal event is a part of the composite event, and occurrence of this precludes any further events from happening, the estimating equations are updated to re-weight observations from patients that die. This is done using IPCW, as suggested by Ghosh and Lin.[39] The estimating equations are the same as those for Ghosh and Lin's model, however now with respect to the severity weighted counting process. Again, the IPCW weights can be estimated either under an assumption of independence between censoring times and covariates, corresponding to estimating the censoring distribution non-parametrically or under a conditional independence assumption, corresponding to using a semi-parametric model. We have considered the non-parametric version, where the censoring distribution was estimated using Kaplan–Meier, ensuring comparability between this and the implemented GL method. For comparability with the previous estimates from the LWYY and GL models, it will be most relevant to consider equal severity of $c = (c_1, c_2, c_3) = (1,1,1)$. A different choice of severity will model the expected number of events for a different process, namely the severity weighted counting process. We will denote this by the ML model.

We are interested in modelling the recurrent composite 3-p MACE endpoint using the method suggested by Mao and Lin. However, their article considers a situation where all-cause death is a component of the composite endpoint. Our composite endpoint only includes CV death as a component, and does not include non-CV death, which will act as a competing risk. Firstly, we have applied the methods of Mao and Lin, adjusting for the CV death part of the composite event, to the recurrent 3-p MACE data with equal severity of 1, and censoring for non-CV deaths. Secondly, we have extended the method to additionally adjust for non-CV death in the calculation of the IPCW weights, and applied it to the recurrent 3-p MACE data with severity $c = (1,1,1)$. This corresponds to a mixture of a ML and GL approach. Intuitively, the first approach should provide an improvement of the results obtained from fitting the LWYY model to 3-p MACE. Here, non-CV death is censored, but CV death is adjusted for. The second approach should provide an improvement of fitting the GL model to 3-p MACE and only adjusting for non-CV death as a competing event. Here, the adjustment should correct for both the CV death event that is a part of the recurrent event process and for the non-CV death acting as a competing event. Theoretical details on this method are available in Appendix.

The results are available in Figure 9. The regression coefficient estimates are available in Table 6. The impact of adjusting for CV-death on the composite 3-p MACE endpoint using the ML method is seen in Figure 9. Clearly and as expected, not adjusting for a terminating event in the composite recurrent event process overestimates the marginal mean function. Moreover, this adjustment affects the estimated regression coefficients due to the uneven distribution of CV deaths among placebo and liraglutide, as seen in Table 6 under "ML 1". The difference between the two treatments becomes smaller using the ML 1 approach compared to the LWYY method, when subjects are still under risk of a recurrent event after experiencing a CV death. The second approach, adjusting for both CV death and non-CV death using a slight modification of the ML method is shown in Figure 9. Here, the GL method overestimates the mean function as the CV deaths in the composite event is not accounted for. The regression coefficient estimates for this approach are

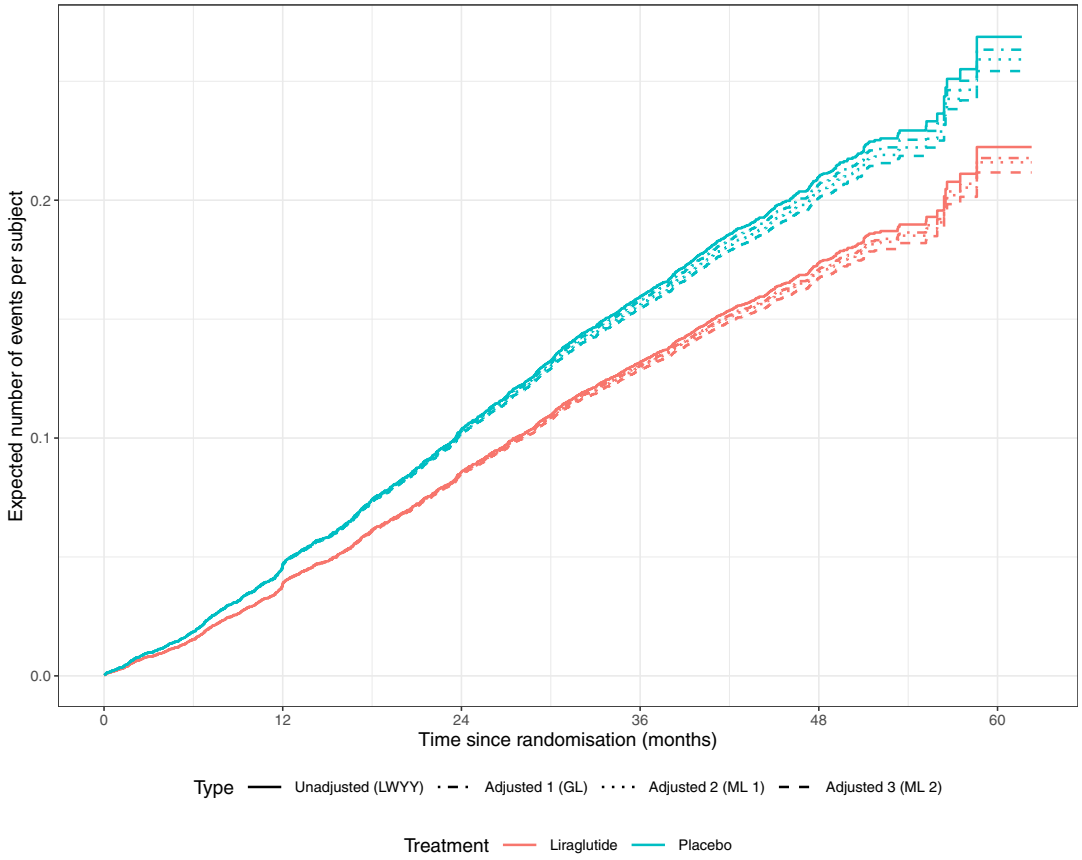


FIGURE 9 Model-based marginal mean function estimates per treatment for recurrent 3-p MACE. "Unadjusted (LWYY)", "Adjusted 1 (GL)", "Adjusted 2 (ML 1)" and "Adjusted 3 (ML 2)" refers to the estimates of $\mu(t|x)$ using the Lin–Wei–Yang–Ying, Ghosh–Lin, Mao and Lin (adjusted for CV death alone) and the updated Mao and Lin (adjusted for both non-CV and CV death) models, respectively

seen in Table 6 under "ML 2". These are very comparable with the estimates seen for the first "ML 1" approach, as the adjustment for non-CV death does not affect the difference between treatments as measured through the regression coefficient, due to the balanced distribution of non-CV death in each treatment group. It does, however, make a difference compared to both the GL (and LWYY) methods. The second approach would be more correct in adjusting for the different kinds of death that can occur in relation to endpoints like this. The relative differences between the methods are not that big (as seen in Figure 9), but this will depend on the considered data set.

5 | DISCUSSION

This article presents different current non-parametric and semi-parametric models for the analysis of recurrent events, but it does not provide an exhaustive list of all available methods. The focus has been to emphasise relevant approaches and potential pitfalls for analysing randomised controlled trials and quantifying a meaningful treatment effect on recurrent events. Several large cardiovascular outcome and heart failure trials only consider the first events, and analyse time-to-first events. This is done even though the endpoint of interest is recurrent in nature, for example, heart failure hospitalisation. This entails a loss of information, since data past the first event is not used for the analysis. Moreover, characterising the effect of treatment on the entire course of disease progression could yield valuable information. Also,

if treatment behaves in a similar way on recurrent events as for first events, modelling recurrent events can imply a gain in efficiency and an increase in power. This paper highlights that the analysis of recurrent events requires careful consideration. This is not surprising as one now tries to capture the behaviour of a more complicated multi-state model as opposed to a simple and classical time-to-event multi-state model.

The complexity that is introduced in modelling recurrent events in the presence of competing risks should not be taken lightly, and this requires an understanding of *what* is being estimated. Table 7 provides an overview of the different presented models summarising their properties. Depending on the framework, each method has potential issues and benefits. As a variety of methods exist, it is vital to understand what is being estimated in each, in order to select a relevant approach and to understand how to interpret results. Data content and model understanding is key for drawing correct inference: What are you estimating? The focus on estimation complements the recent ICH guideline on estimands, where it is stated that; "An estimand is a precise description of the treatment effect reflecting the clinical question posed a given clinical trial object".[42] It is underlined that it is vital to specify intercurrent events, which are events happening post randomisation that may affect the interpretation of treatment effects, this could for instance be death. Clearly, deaths are intercurrent events that can influence the interpretation of treatment effects when analysing recurrent events and should thus be handled in an appropriate manner.

In the context of analysing recurrent events in the presence of non-negligible mortality for quantifying a treatment effect in a randomised controlled trial, the following considerations are highlighted for each of the presented methods. The AG model adjusted for treatment alone will likely be "too" simple a model for application as the independence assumption will be unrealistic. As a remedy to this, one could consider adjusting for the previous number of events in the AG model. This will be inappropriate as the treatment effect will be underestimated if recurrent events are correlated and the treatment is effective. The PWP model is also inappropriate as the treatment effect will be underestimated in a similar fashion, however here through stratification. The WLW model preserves randomisation, but introduces unrealistic risk sets from a clinical perspective. Also, the interpretation of the marginal model falls short in the presence of competing deaths. The frailty model estimates a subject specific treatment effect, which is rarely of interest in a randomised controlled trial. Moreover, the model may break down in the presence of competing risks due to dependent censoring. The joint frailty model targets both recurrent events and deaths at once, and the censoring assumption is likely no longer violated. But, being a frailty model, the treatment estimates still have a subject specific interpretation. The marginal LWYY model does not need to model the dependence between events as opposed to the intensity models. The GL model provides an improvement of the LWYY model, adjusting the risk sets and estimation to reflect deaths. The model still targets number of events, which can be problematic if a treatment has an effect on death, and in turn reduces the number of events. The ML model further addresses this, as deaths are considered a part of the severity weighted event itself. However, it can be difficult to decide which severity to use. In summary, each method has their benefits but surely also their own set of issues. From a clinical perspective, both recurrent events and deaths are vital for understanding treatment effects, and it is suitable to consider modelling approaches targeting both at once. Cook and Lawless advocate for using marginal methods for analysing recurrent events in randomised trials, since intensity models are conditional on past history and subject to model misspecification.[13] For application to recurrent event data with competing deaths, we believe that the marginal methods suggested by Ghosh and Lin, Mao and Lin or a combination provide understandable quantities that adequately addresses competing deaths.[39,41] These methods come with the caveat that a treatment effective in killing individuals will also effectively reduce the number of events. The problem with censoring for deaths when targeting marginal parameters in a recurrent event setting is analogous to the issue with considering Kaplan–Meier estimates, where competing causes have been censored, in a competing risk setting.[22] Here, 1 – Kaplan–Meier estimates will overestimate the true failure probabilities, as the Nelson-Aalen estimates will overestimate the marginal mean function, if competing deaths are censored. The intensity models do not suffer from this issue as they are focusing on what happens in an instantaneous moment, and as such, censoring for competing death causes does not affect inference. Consequently, Hengelbrock et al suggest to present both cause-specific hazard ratio estimates from intensity models and non-parametric mean frequency estimates per Ghosh and Lin's method (accounting for competing deaths) to analyse recurrent safety data with competing death causes.[43] We believe that a marginal regression model that simultaneously targets the recurrent event and terminal event at once would be of value. This is a topic for further research.This case study has presented that in the presence of non-negligible death, many marginal approaches for modelling means of recurrent events, should accommodate competing deaths. The impact of not adjusting for terminal events will depend on the analysed data. Several authors have published results using marginal methods applied to recurrent event data in randomised trials. For PARADIGM, a recurrent composite

heart failure hospitalisation and CV death endpoint and a recurrent heart failure hospitalisation endpoint were analysed using a LWYY model.[6] The same was done for PARAGON-HF.[3] For CHARM-Preserved, a recurrent composite heart failure hospitalisation and CV death was analysed using a AG model with robust variance estimation, however where "rate ratios" are reported (as opposed to "hazard ratios").[4] Note, that the LWYY model will not be appropriate if there are competing deaths, which is the case for these trials. The marginal mean function will be overestimated if applied naively to data with competing deaths, where the deaths are censored. Results from the REDUCE-IT clinical trial has also been published using the WLW model to applied to composite endpoints (which included cardiovascular death).[44]

An important aspect presented in this article is that defining a composite endpoint which consists not only of a recurrent event of interest, for example, MI, but also death, for example, CV death, does not salvage all issues related to estimation and mortality for recurrent event models. Firstly, this does not adequately adjust the risk sets for those subjects that experience a CV death event, since they are assumed to still be under risk of a subsequent recurrent event. Secondly, if a competing event, say non-CV death, plays a role, this should also be adequately adjusted for. The impact of not adequately adjusting for both aspects was presented in this article. Composite recurrent endpoints with a death component are becoming quite popular for the analysis of recurrent events, however care should be taken. Especially, it should be noted that it is not appropriate to conduct recurrent event analysis targeting a marginal parameter on a composite endpoint with a death component, due to the implication on the risk sets. With an imbalance in deaths, this will create a bias in parameter estimates. The exploratory analysis for PARADIGM-HF trial was a LWYY model adjusted for treatment and region applied to a composite of total heart failure (HF) hospitalisation and CV death treating non-CV deaths as censorings.[6] This exactly carries two of the aforementioned issues: The model does not acknowledge that you can die from either CV death or non-CV death. Moreover, there seems to be an imbalance regarding CV deaths for the two treatment groups, thus implying a potential direct bias on the mean ratio by not adequately adjusting for CV deaths.

The win ratio approach for analysing composite endpoints suggested by Pocock has gained focus in the last years within the cardiovascular community.[45] It was applied to the CHARM-Preserved data to analyse recurrent heart failure hospitalisation and CV death.[4] This method addresses a potential ranking of components for a composite endpoint, for example, CV death is more severe than hospitalisation. However, this method provides more of a ranked based approach for analysing a time-to-event endpoints, rather than a focus on capturing a recurrent event process. Moreover, the estimated win-ratio parameter (number of "winners" on treatment A divided by number of "losers" on treatment A) depends on the distribution of follow-up time, as also noted by Oakes.[46] This might not be the best property, especially when considering treatment with an effect on mortality. Finally, this method requires some matching of individuals based on risk scores, which could be problematic in practice.

As mentioned, a paper focusing on recurrent events in LEADER was published within the recent years.[14] Many of the issues discussed in the present article on the inference of treatment effects in a recurrent event setting can be found there and can be subject to critique. However, and unfortunately these analyses also present methods that are being applied in practice at the moment for randomised controlled trials. When comparing the time-to-first event analysis with the recurrent event analysis applied to the two endpoints in Tables 5 and 6, we can see the potential benefit of considering recurrent event methods. The estimated parameter (hazard ratio, mean ratio) coincides when comparing the Cox and the GL model for recurrent MI, but the GL method has a narrower confidence interval. The estimates from the Cox and ML 2 models differ, in the sense that the point estimates are larger with the ML 2 model, but the confidence intervals are also slightly wider. Note that the interpretation of the estimates also differs; hazard ratios versus mean ratios. Nevertheless, this indicates that efficiency may be gained by focusing on recurrent event methods as opposed to time-to-first event methods.

ACKNOWLEDGMENTS

The clinical trial LEADER was sponsored by Novo Nordisk A/S. This research was conducted as a part of the PhD education of JK Furberg from which she received funding from Novo Nordisk A/S. H Ravn and S Rasmussen are employed at Novo Nordisk A/S.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

De-identified individual participant data, study protocol and redacted Clinical Study Report will be available according to Novo Nordisk data sharing commitments. The data will be made available permanently after research completion and approval of product and product use in both EU and US. Data will be shared with bona fide researchers submitting a research proposal requesting access to data and for use as approved by the Independent Review Board according to the IRB Charter (see novonordisk-trials.com). Access request proposal form and the access criteria can be found at novonordisk-trials.com. The data will be made available on a specialised SAS data platform.

ORCID

Julie Kjærulff Furberg https://orcid.org/0000-0001-8785-1462

REFERENCES

1. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B (Methodol)*. 1972;34(2):187-220.
2. Anker SD, McMurray JJV. Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials? *Eur J Heart J*. 2012;33:2764-2765.
3. Solomon SD, McMurray JJ, Anand IS, et al. Angiotensin–Neprilysin inhibition in heart failure with preserved ejection fraction. *N Engl J Med*. 2019;381(17):1609-1620.
4. Rogers JK, Pocock SJ, McMurray JJV, et al. Analysing recurrent hospitalisations in heart failure: a review of statistical methodology, with application to CHARM-preserved. *Eur J Heart Fail*. 2014;16(1):33-40.
5. Yusuf S, Pfeffer MA, Swedberg K, et al. Effects of candesartan in patients with chronic heart failure and preserved leftventricular ejection fraction: the CHARM-preserved trial. *Lancet*. 2003;362:777-781.
6. Mogensen UM, Gong J, Jhund PS, et al. Effect of sacubitril/valsartan on recurrent events in the prospective comparison of ARNI with ACEI to determine impact on global mortality and morbidity in heart failure trial (PARADIGM-HF). *Eur J Heart Fail*. 2018;20(4):760-768.
7. McMurray JJV, Packer M, Desai AS, et al. Angiotension-Neprilysin inhibition versus Enalapril in heart failure. *N Engl J Med*. 2014;371(11).993-1004.
8. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. *N Engl J Med*. 2016;375(4):311-322.
9. EMA Committee for Medicinal Products for Human Use. *Guideline on Clinical Investigation of Medicinal Products for the Treatment of Chronic Heart Failure. Online*: European Medicines Agengy (EMA). Committee for Medicinal Products for Human Use (CHMP); 2017.
10. Akacha M, Binkowitz B, Bretz F, et al. Request for CHMP Qualification Opinion: Clinically Interpretable Treatment Effect Measures Based on Recurrent Event Endpoints that Allow for Efficient Statistical Analyses. 2018.
11. Committee for Medicinal Products for Human Use (CHMP). *Qualification Opinion of Clinically Interpretable Treatment Effect Measures Based on Recurrent Event Endpoints That Allow for Efficient Statistical Analyses*. European Medicines Agengy (EMA). Committee for Medicinal Products for Human Use (CHMP); 2020.
12. Cook RJ, Lawless JF. Analysis of repeated events. *Stat Methods Med Res*. 2002;11:141-166.
13. Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events*. 1st ed. Springer; 2007.
14. Verma S, Bain BC, Buse JB, et al. Occurrence of first and recurrent major adverse cardiovascular EventsWith Liraglutide treatment among patients with type 2 diabetes and high risk of cardiovascular events. *JAMA Cardiol*. 2019;4(12):1214-1220.
15. Cook RJ, Lawless JF. *Multistate Models for the Analysis of Life History Data*. 1st ed. CRC Press; 2018.
16. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. 1st ed. Springer Series in Statistics; 1993.
17. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020.
18. Therneau TM. A Package for Survival Analysis in R. *R package version 3.1–11*. R; 2020.
19. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer; 2000.
20. Rondeau V, Mazroui Y, Gonzalez JR. Frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw*. 2012;47(4):1-28.
21. SAS Institute Inc. *SAS/STAT® 13.2 User's Guide The PHREG Procedure*; SAS Institute; 2014.
22. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*. 2012;41(3):861-870.
23. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10(4):1100-1120.
24. Prentice R, Williams B, Peterson A. On the regression analysis of multivariate failure time data. *Biometrika*. 1981;68:373-379.
25. Vaupel JW, Manton KW, Stallard E. The impact of heterogeneity in individual frailty on the dynamic of mortality. *Demography*. 1979;16(3):439-454.
26. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. 2nd ed. Oxford Statistical Science Series; 2002.
27. Martinussen T, Scheike T. *Dynamic Regression Models for Survival Data*. Springer. 1 ed. 2006.
28. Balan TA, Putter H. A tutorial on frailty models. *Stat Methods Med Res*. 2020;29(11):3424-3454.

29. Nielsen GG, Gill RD, Andersen PK, Sørensen TIA. A counting process approach to maximum likelihood estimation in frailty models. *Scand J Stat*. 1992;19(1):25-43.
30. Lui L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics*. 2004;60:747-756.
31. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*. 2007;8:708-721.
32. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc*. 1989;84:1065-1073.
33. Metcalfe C, Thompson SG. Wei, Lin andWeissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Stat Methods Med Res*. 2007;16:103-122.
34. Li QH, Lagakos SW. Use of theWei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Stat Med*. 1998;16:925-940.
35. Lawless JF, Nadeau JC. Some simple robust methods for the analysis of recurrent events. *Dent Tech*. 1995;37:158-168.
36. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc*. 2000;62(4):711-730.
37. Cook R, Lawless JF. Marginal analysis of recurrent events and a terminating event. *Stat Med*. 1997;16:911-924.
38. Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. *Biometrics*. 2000;56:554-562.
39. Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. *Stat Sin*. 2002;12:663-688.
40. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
41. Mao L, Lin DY. Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. *Biostatistics*. 2016;17:390-403.
42. ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials (E9 (R1))*; European Medicines Agengy (EMA). Committee for Medicinal Products for Human Use (CHMP); 2019.
43. Hengelbrock J, Gillhaus J, Kloss S, Leverkusb F. Safety data from randomized controlled trials: applying models for recurrent events. *Pharm Stat*. 2016;15(4):315-323.
44. Bhatt DL, Steg PG, Miller M, et al. Effects of icosapent ethyl on total ischemic events from REDUCE-IT. *J Am Coll Cardiol*. 2019;73(22):2791-2802.
45. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33:176-182.
46. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. *Biostatistics*. 2016;103(3):742-745.

How to cite this article: Furberg JK, Rasmussen S, Andersen PK, Ravn H. Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER. *Pharmaceutical Statistics*. 2022;21(1):241-267. doi:10.1002/pst.2167

APPENDIX: A THEORETICAL RESULTS FOR MARGINAL MODEL FOR COMPOSITE EVENTS WITH A DEATH COMPONENT AND OTHER COMPETING DEATHS

This appendix section contains the theoretical results for the approach outlined in Section 4.2.3 on the combination of the methods suggested by Ghosh and Lin and Mao and Lin.[39,41] The derivations in the following largely resembles the steps in these articles.

We are interested in modelling a recurrent composite event process where a terminal event is a part of the recurrent event process, for example, CV death. Moreover, there is competing risks for dying of other causes, for example, non-CV death. Figures A1 and A2 visualise the processes of interest.

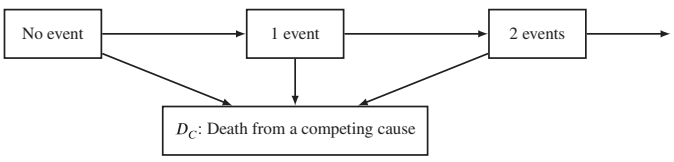


FIGURE A1 A recurrent event process with a competing terminal event

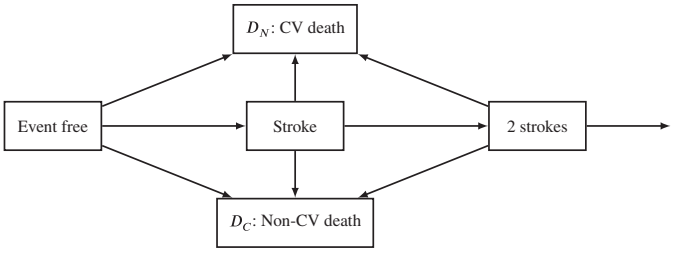


FIGURE A2 Example of a composite recurrent event process with CV death as a component and non-CV death as a competing risk

For $k = 1,...,K$, let $N_k^*(t)$ denote the cumulative number of the k'th event type by time t. A severity, c_k, is assigned to the k'th event type according to the relative severity. One of the composite event types is a terminal event of some type. We define the severity weighted sum of the K counting processes: $N^*(t) = \sum_{k=1}^{K} c_k N_k^*(t)$. We specify the following regression model for the marginal rate function, acknowledging that no recurrences can occur after death.

$$E\Big\{dN^*(t)|Z(t)\Big\} = d\mu(t) = e^{\beta^T Z(t)} d\mu_0(t) \tag{A1}$$

where $Z(t)$ denotes a p-dimensional, possibly external time-varying, covariate vector which has a multiplicative effect on the marginal baseline rate function $d\mu_0(t)$. $\mu_0(t)$ is an arbitrary increasing function. With no time-varying covariates, Equation (A1) reduces to the proportional means model: $E\Big\{N^*(t)|Z\Big\} = \mu(t) = e^{\beta^T Z}\mu_0(t)$, where $\mu(t)$ denotes the marginal mean function and $\mu_0(t)$ denotes the baseline mean function.

Let D_N denote the time to the terminal event defined as part of the composite recurrent event process. Let D_C denote the time to the competing terminal event. In practice, $N^*(t)$, D_C and D_N are subject to right censoring. Let $D = D_C \wedge D_N$. Let C denote the censoring time, which is assumed to be independent of $N^*(t)$, D_C and D_N given $Z(t)$. Let $N(t) = N^*(t \wedge C)$, $X = D \wedge C$, and $\delta = I(D \le C)$. The observed data consists of $\{N_i(t), Z_i(t), X_i, \delta_i; t \le X_i\}$ for subject $i = 1,...,n$.

To make valid inference for model (A1), we need to exclude the dependent censoring introduced by death from the "at-risk" indicators from the estimating equations from the classical LWYY model.[36] In order to do so, a subject remains at-risk until independent censoring occurs (even though a subject dies prior to the independent censoring). So one needs to consider the at-risk indicators given by $I(C \ge t)$ instead of $I(X \ge t)$. Replacing $Y(t)$ with $I(C \ge t)$ in the estimating equations yields

$$\sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{\sum_I^n I(C_j \ge t) Z_j(t) e^{\beta^T Z_j(t)}}{\sum_{j=1}^n I(C_j \ge t) e^{\beta^T Z_j(t)}} \right\} \mathrm{d}N_i(t)$$

However, C is not fully observed. An inverse probability of censoring weighting (IPCW) technique can be employed for estimating $I(C \ge t)$. Let the weights at time t be given by

$$w(t) = \frac{I(C \ge D \wedge t) G(t|Z)}{G(D \wedge t|Z)} = \frac{I(C \ge D \wedge t) G(t|Z)}{G(X \wedge t|Z)}$$

which can be estimated by

$$\hat{w}(t) = \frac{I(C \ge D \wedge t) \hat{G}(t|Z)}{\hat{G}(D \wedge t|Z)}$$

where $\hat{G}(D\wedge t|Z)$ is the estimator of $G(D\wedge t|Z)$, either estimated using a proportional hazards model or with a Kaplan–Meier estimate. Hence $I(C_j \geq t)$ will be replaced by

$$\hat{w}_j(t) = \frac{I(C_j \geq D_j \wedge t)\hat{G}(t|Z_j)}{\hat{G}(D_j \wedge t|Z_j)}$$

since $E(w_j(t)) = G(t) = E(I(C_j \geq t))$. This leads to the following estimating equation for β

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{\sum_{j=1}^{n} \hat{w}_j(t) Z_j(t) e^{\beta^T Z_j(t)}}{\sum_{j=1}^{n} \hat{w}_j(t) e^{\beta^T Z_j(t)}} \right\} \mathrm{d}N_i(t)$$

Let $\hat{\beta}$ be the solution to $U(\beta) = 0$. Correspondingly, the baseline mean function $\mu_0(t)$ can be estimated by the Breslow type estimator

$$\hat{\mu}_0(t) = \sum_{i=1}^{n} \int_0^{t} \frac{\mathrm{d}N_i(u)}{\sum_{j=1}^{n} \hat{w}_j(u) e^{\hat{\beta}^T Z_j(u)}}$$

If $\hat{w}_j(t)$ is estimated using a Kaplan–Meier estimator (assuming that the censoring is independent of the covariates), it holds that

Theorem 1. *The estimator $\hat{\beta}$ is consistent and it holds that $\hat{\beta} \xrightarrow{a.s.} \beta$. Moreover, $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a zero-mean random normal variables with a covariance matrix that can be consistently estimated by $\hat{\Gamma} = \hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$. Here*

$$\hat{A} = -n^{-1}\frac{\partial U(\hat{\beta})}{\partial \beta} = n^{-1}\sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \frac{\sum_{j=1}^{n} \hat{w}_j(t) Z_j(t) e^{\hat{\beta}^T Z_j(t)}}{\sum_{j=1}^{n} \hat{w}_j(t) e^{\hat{\beta}^T Z_j(t)}} \right\}^{\bigotimes 2} \cdot \frac{\hat{w}_j(t) e^{\hat{\beta}^T Z_j(t)}}{\sum_{j=1}^{n} \hat{w}_j(t) e^{\hat{\beta}^T Z_j(t)}} \mathrm{d}N_i(t),$$

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} (\hat{\eta}_i + \hat{\psi}_i)^{\bigotimes 2},$$

$$\hat{\eta}_i = \int_0^{\tau} \left\{ Z_i(t) - \frac{\sum_{j=1}^{n} \hat{w}_j(t) Z_j(t) e^{\hat{\beta}^T Z_j(t)}}{\sum_{j=1}^{n} \hat{w}_j(t) e^{\hat{\beta}^T Z_j(t)}} \right\} d\hat{M}_i(t),$$

$$\hat{\psi}_i = \int_0^{\tau} \frac{\hat{q}(t)}{\sum_{j=1}^{n} Y_j(t)} d\hat{M}_i^C(t),$$

$$\hat{q}(t) = -n^{-1}\sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(u) - \frac{\sum_{j=1}^{n} \hat{w}_j(u) Z_j(u) e^{\hat{\beta}^T Z_j(u)}}{\sum_{j=1}^{n} \hat{w}_j(u) e^{\hat{\beta}^T Z_j(u)}} \right\} I(u \geq t \geq X_i) d\hat{M}_i(u),$$

$$\hat{M}_i(t) = \int_0^{t} \hat{w}_i(u) \left\{ dN_i(u) - e^{\hat{\beta} Z_i(u)} d\hat{\mu}_0(u) \right\},$$

$$\hat{M}_i^C(t) = dN_i^C(t) - \int_0^{t} Y_i(u) d\hat{\Lambda}_0(u),$$

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^C(u)}{\sum_{j=1}^{n} Y_j(u)},$$

$$N_i^C(t) = I(X_i \le t, \delta_i = 0), \quad Y_i(t) = I(X_i \ge t)$$

Proof. The structure of the estimating equation for β coincides with the estimating equations available in Ghosh and Lin and Mao and Lin.[39,41] For details on the asymptotic derivations for these estimates see section 2.5 in Ghosh and Lin.[39]

Manuscript II

Title *Bivariate pseudo-observations for recurrent event analysis with terminal events*

Authors Julie K. Furberg, Per K. Andersen, Sofie Korn, Morten Overgaard and Henrik Ravn

Details Published in Lifetime Data Analysis (April 2023)

Lifetime Data Analysis (2023) 29:256–287
https://doi.org/10.1007/s10985-021-09533-5

Bivariate pseudo-observations for recurrent event analysis with terminal events

Julie K. Furberg[1] · Per K. Andersen[2] · Sofie Korn[3] · Morten Overgaard[4] · Henrik Ravn[1]

Received: 25 March 2021 / Accepted: 4 September 2021 / Published online: 5 November 2021

Abstract

The analysis of recurrent events in the presence of terminal events requires special attention. Several approaches have been suggested for such analyses either using intensity models or marginal models. When analysing treatment effects on recurrent events in controlled trials, special attention should be paid to competing deaths and their impact on interpretation. This paper proposes a method that formulates a marginal model for recurrent events and terminal events simultaneously. Estimation is based on pseudo-observations for both the expected number of events and survival probabilities. Various relevant hypothesis tests in the framework are explored. Theoretical derivations and simulation studies are conducted to investigate the behaviour of the method. The *method* is applied to two real data examples. The bivariate marginal pseudo-observation model carries the strength of a two-dimensional modelling procedure and performs well in comparison with available models. Finally, an extension to a three-dimensional model, which decomposes the terminal event per death cause, is proposed and exemplified.

Keywords Recurrent events · Terminal events · Pseudo-observations · Simultaneous model · Multi-state model

Julie K. Furberg
jukf@novonordisk.com

1 Biostatistics GLP-1 and CV 1, Novo Nordisk A/S, Vandtårnsvej 114, Søborg, Denmark

2 Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

3 Biostatistics 1, LEO Pharma A/S, Ballerup, Denmark

4 Research unit for Biostatistics, Department of Public Health, Aarhus University, Aarhus, Denmark

1 Introduction

In many cardiovascular trials, the events of interest are events that can happen several times for an individual. Examples include recurrent myocardial infarction or recurrent hospitalisation for heart failure. Often, the trial population will have a high mortality rate. Moreover, experiencing these severe recurrent events can increase the risk of dying. Further, death will preclude any subsequent recurrent events. For such situations, it will be desirable to gain knowledge not only about the recurrence of events but also mortality.

Various methods and models have been suggested for analysing recurrent events in the presence of a competing terminal event. Non-parametric and semi-parametric marginal regression models for estimating the mean number of recurrent events per subject in the presence of a competing risk have been considered by Cook and Lawless (1997) and Ghosh and Lin (2000, 2002). Liu et al. (2004) have proposed a joint frailty regression model that specifies both an intensity function for recurrent events and an intensity function for death through a subject level shared frailty term. Cause-specific hazard functions for either recurrent events or a terminal event can be targeted using standard models such as the Andersen–Gill model (Andersen and Gill 1982) or the Cox model (Cox 1972), respectively. Figure 1 visualizes a relevant multi-state model for recurrent events with mortality as a terminal event. Andersen et al. (2019) discuss the modelling of marginal features of such a multi-state model.

The present paper proposes a joint marginal regression model targeting a two-dimensional marginal parameter for both recurrence and death: the expected number of events per subject and the survival probability. The estimation of the parameters in the marginal model is based on pseudo-observations as suggested by Andersen et al. (2003). Large sample behaviour will be explored both based on theoretical derivations and simulation. Moreover, will the performance of the method be investigated against other models using simulation. Suitable hypothesis tests based on the model will be explored with a special focus on data from randomised controlled trials. A bootstrap experiment is conducted to assess variance estimation in the proposed model. The model will be applied to two data examples: a well-known data set on bladder cancer recurrences (Byar 1980) and data from a large cardiovascular outcomes trial, LEADER (Marso et al. 2016).

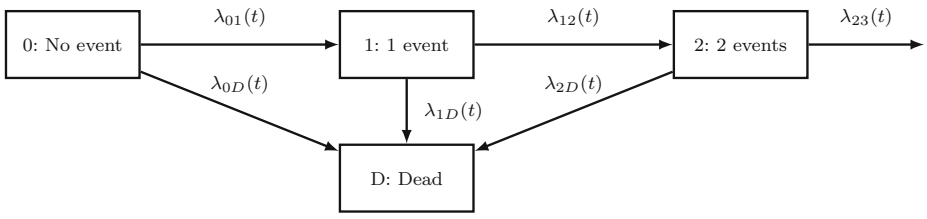


Fig. 1 A recurrent event process with a terminal event

2 Model and mathematical framework

Let D^* denote the survival time and let $N^*(t)$ denote the number of recurrent events by time t. As no recurrent events can be experienced beyond death, $N^*(t)$ remains constant after D^*. Let C denote the censoring time, which is assumed to be independent of both D^* and $N^*(t)$. Due to censoring, neither D^* nor $N^*(t)$ is fully observed. The data consists of $X = \{N(\cdot), D, \delta, Z\}$, where $N(t) = N^*(t \wedge C)$, $D = D^* \wedge C$, $\delta = I(D^* \leq C)$, and Z denotes p baseline covariates. Additionally, it is assumed that C is independent of Z. For each individual, $i = 1, \ldots, n$, we observe $X_i = \{N_i(\cdot), D_i, \delta_i, Z_i\}$, which is assumed to be independent replicates of $X = \{N(\cdot), D, \delta, Z\}$. In the present setting, the marginal parameters of interest are the marginal mean function and the survival probability,

$$\mu(t) = E(N^*(t)) = \int_0^t S(u^-) dR(u),$$
$$S(t) = P(D^* > t),$$

where $dR(t) = E(dN^*(t) \mid D^* \geq t)$. We will model $(\mu(t), S(t))$ conditional on covariates Z simultaneously using a regression model applied to pseudo-observations. Due to right-censoring, both D^* and $N^*(t)$ are incompletely observed and require special attention during analysis. In the absence of censoring, standard regression models could target both survival probabilities and expected number of events. The framework of pseudo-observations is a method for analysing event history data using generalised linear models. The basic idea behind pseudo-observations is as follows. Let $W = (N^*(\cdot), D^*)$ and let θ be the parameter of interest which has the form $\theta = E(f(W))$. The pseudo-observation for $f(W)$ for individual i, $\hat{\theta}_i$, is given by

$$\hat{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i},$$

for a sufficiently well-behaved estimator $\hat{\theta}$. Here, $\hat{\theta}^{-i}$ denotes the estimate obtained by applying $\hat{\theta}$ to the data set where individual i is eliminated. For further details on pseudo-observations and their usage, see Andersen et al. (2003) or Andersen and Perme (2010). Subsequently, the incompletely observed $f(W_i) = f(N_i^*(\cdot), D_i^*)$ is replaced by $\hat{\theta}_i$ for all individuals. Then, $\hat{\theta}_i$ may be used as the outcome in a generalised linear model with some link function g,

$$g(E(f(W) \mid Z)) = \xi^T Z.$$

The pseudo-observations are computed for all individuals, censored or not, and used as responses in the above generalised linear model. The bivariate marginal pseudo-observation model considers

$$\theta = \begin{pmatrix} \mu(t) \\ S(t) \end{pmatrix} = \begin{pmatrix} E(N^*(t)) \\ E(I(D^* > t)) \end{pmatrix}.$$

Regression models based on pseudo-observations for $\mu(t)$ and $S(t)$ separately have been explored in Andersen et al. (2019). In the presence of competing deaths, Cook and Lawless (1997) and Ghosh and Lin (2000), suggested to use the estimator of $\mu(t)$ given by

$$\hat{\mu}(t) = \int_0^t \hat{S}(u^-)\, d\hat{R}(u), \tag{1}$$

where $\hat{R}(t)$ denotes the Nelson–Aalen estimator of $R(t)$ and $\hat{S}(t)$ denotes the Kaplan–Meier estimator of $S(t)$. Alternatively, $\hat{\mu}(t)$ can be expressed using the inverse probability of censoring (IPCW) form for which

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \frac{1}{\hat{K}(u^-)}\, dN_i(u), \tag{2}$$

where $\hat{K}(t)$ is a Kaplan–Meier estimate of $K(t) = P(C > t)$. The estimator $\hat{\mu}(t)$ is approximately unbiased when the censoring times are independent of the recurrent events and terminal event processes (Ghosh and Lin 2000). Moreover, the Kaplan–Meier estimator $\hat{S}(t)$ is also approximately unbiased under independent censoring (Andersen et al. 1993). As an estimator for θ we will consider,

$$\hat{\theta} = \begin{pmatrix} \hat{\mu}(t) \\ \hat{S}(t) \end{pmatrix}.$$

For a given time $t \in [0, \tau]$, the pseudo-observation for subject i $(i = 1, \ldots, n)$ is given by

$$\hat{\theta}_i = \begin{pmatrix} \hat{\mu}_i(t) \\ \hat{S}_i(t) \end{pmatrix} = \begin{pmatrix} n\hat{\mu}(t) - (n-1)\hat{\mu}^{-i}(t) \\ n\hat{S}(t) - (n-1)\hat{S}^{-i}(t) \end{pmatrix},$$

where $\hat{\mu}(t)$ denotes the estimate of the marginal mean function based on the entire sample and $\hat{\mu}^{-i}(t)$ denotes the estimate based on leaving out subject i's observations. Moreover, $\hat{S}(t)$ denotes the estimate of the survival probability based on the entire sample and $\hat{S}^{-i}(t)$ denotes the estimate based on leaving out subject i's observations. Regardless of whether an individual was censored, died or experienced a recurrent event, a pseudo-observation is computed at time t. For a given individual, the behaviour of $\hat{\theta}_i$ as a function of t depends on the data for subject i. The $\hat{\mu}_i(t)$ component is exemplified for four individuals in Fig. 5 using data from a bladder cancer study. The $\hat{S}_i(t)$ component is exemplified for different individuals in Andersen and Perme (2010). The pseudo-observations may take on a variety of different values and are for instance not necessarily positive, but, with little or no censoring up to time t, the pseudo-observations will resemble the true observations, see e.g. Fig. 7. We will compute pseudo-observations at times $t_1, \ldots, t_k$ for both $\hat{\mu}_i(\cdot)$ and $\hat{S}_i(\cdot)$. The same times, $t_1, \ldots, t_k$, are used for computing $\hat{\mu}_i(\cdot)$ and $\hat{S}_i(\cdot)$. The model can be extended

to considering different time points for $\hat{\mu}_i(\cdot)$ and $\hat{S}_i(\cdot)$ but for ease of notation we have chosen to use the same time points. This implies that

$$\hat{\theta}_i = \left(\begin{pmatrix} \hat{\mu}_i(t_l) \\ \hat{S}_i(t_l) \end{pmatrix}, \, l = 1, \ldots, k \right)$$

is $(2k \times 1)$-dimensional for individual i. We consider a proportional means model where

$$\mu(t \mid Z) = \mu_0(t) \exp(\beta^T Z).$$

and a proportional hazards model where

$$S(t \mid Z) = \exp(-\Lambda_0(t) \exp(\gamma^T Z)).$$

This implies that

$$g \begin{pmatrix} \mu(t \mid Z) \\ S(t \mid Z) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta^T Z \\ \log(\Lambda_0(t)) + \gamma^T Z \end{pmatrix}, \tag{3}$$

which defines a generalised linear model in terms of a vector of model parameters ξ that consist of β, γ, and the intercepts $\log(\mu_0(t_l))$ and $\log(\Lambda_0(t_l))$ for $l = 1, \ldots, k$ with link function $g(x, y) = (\log(x), \text{cloglog}(y)) = (\log(x), \log(-\log(y)))$.

The model parameters ξ will be estimated using generalised estimating equations (GEE) (Liang and Zeger 1986) by doing a regression of $\hat{\theta}_i$ on Z_i with mean function

$$m_i(\xi) = m(\xi; Z_i) = \left(g^{-1} \begin{pmatrix} \log(\mu_0(t_l)) + \beta^T Z_i \\ \log(\Lambda_0(t_l)) + \gamma^T Z_i \end{pmatrix}, \, l = 1, \ldots, k \right).$$

Then the estimating equations are given by

$$U(\xi) = \sum_i \left(\frac{\partial m_i}{\partial \xi} \right)^T V_i^{-1} (\hat{\theta}_i - m_i(\xi)) = \sum_i U_i(\xi) = 0, \tag{4}$$

where V_i is a $2k \times 2k$ working covariance matrix for $\hat{\theta}_i$. Let $\hat{\xi}$ denote the solution to $U(\xi) = 0$. Work by Graw et al. (2009), Jacobsen and Martinussen (2016), Overgaard et al. (2017) and Overgaard (2019) indicate that $\hat{\beta}$ and $\hat{\gamma}$ are marginally asymptotically normally distributed. We will argue that $(\hat{\beta}, \hat{\gamma})$ has a large sample bivariate normal distribution using simulation and theoretical derivations. Theoretical details are included in Appendix A.1. Thus, we can assume that $\hat{\xi}$ is asymptotically normal with mean ξ and a covariance matrix which may be estimated by

$$\hat{\Sigma} = I(\hat{\xi})^{-1} \widehat{\text{var}}(U(\xi)) I(\hat{\xi})^{-1}, \tag{5}$$

where

$$I(\xi) = \sum_i \left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1} \left(\frac{\partial m_i}{\partial \xi}\right), \quad \widehat{\text{var}}(U(\xi)) = \sum_i U_i(\hat{\xi}) U_i(\hat{\xi})^T,$$

where $\hat{\Sigma}$ is $2(p+k) \times 2(p+k)$-dimensional. As a working covariance matrix, we will use the identity matrix, known as the working independence covariance matrix. Note, that this is not the same as assuming independence for θ_i within subject, as the sandwich covariance estimator in Eq. (5) accounts for dependence within a subject. Theoretical results indicate that the sandwich variance estimator in Eq. (5) may tend to be slightly conservative (Jacobsen and Martinussen 2016; Overgaard et al. 2017; Overgaard 2019). The theoretical results in Appendix A.1 also indicate that the sandwich covariance estimator leads to conservative variance estimates.

The proposed two-dimensional modelling approach could be extended to other high-dimensional models. For instance, if one is particularly interested in a certain death cause, the pseudo-observations could be based on cumulative incidences for those causes in an analogous manner. The example based on LEADER data will exemplify this by considering a model for the parameter

$$\theta = \begin{pmatrix} \mu(t) \\ C_1(t) \\ C_2(t) \end{pmatrix} = \begin{pmatrix} E(N^*(t)) \\ E(I(D^* \leq t, \Delta = 1)) \\ E(I(D^* \leq t, \Delta = 2)) \end{pmatrix}, \tag{6}$$

where $\Delta = \{1, 2\}$ represents a cause-of-death indicator. Thus, $C_1(t)$ and $C_2(t)$ are the cumulative incidences for causes 1 and 2, respectively. In this example, we will use the link functions, $h(x, y, z) = (\log(x), \text{cloglog}(1-y), \text{cloglog}(1-z))$. Other link functions than these and those chosen for Eq. (3) could be utilized. However, using the cloglog-link ensures that the estimated probabilities lie between 0 and 1. Whereas, using the log-link ensures that the estimated expected number of events is positive.

2.1 Hypothesis testing

Assume that we are considering a simple setting where Z is *one-dimensional*. Here, we are interested in the effect of treatment, β and γ, on recurrent events and death, respectively. Let $\hat{\beta}$ and $\hat{\gamma}$ denote the estimates of β and γ derived from the described estimation procedure. *We will argue that*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \overset{as}{\sim} \mathcal{N}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right), \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}. \tag{7}$$

We will denote the corresponding correlation matrix by Ω, with

$$\Omega = \begin{pmatrix} 1 & \omega_{12} \\ \omega_{12} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & 1 \end{pmatrix}.$$

Under Eq. (7), we wish to create testing procedures for evaluating the effect of treatment on the recurrent events and death. Along the lines of Wei et al. (1989) and Ghosh and Lin (2002), we consider the following testing approaches.

Local tests

The following local tests are based on the marginal distributions of $\hat{\beta}$ and $\hat{\gamma}$ and evaluates the treatment effect separately. The two hypotheses and the testing procedures are specified below.

1. $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$. Reject H_0 if $P(|U| > |\tilde{\beta}|) \leq \alpha$, where $\tilde{\beta} = \frac{\hat{\beta}}{\sqrt{\text{var}(\hat{\beta})}}$ and $U \sim \mathcal{N}(0, 1)$.
2. $H_0 : \gamma = 0$ versus $H_a : \gamma \neq 0$. Reject H_0 if $P(|U| > |\tilde{\gamma}|) \leq \alpha$, where $\tilde{\gamma} = \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}$ and $U \sim \mathcal{N}(0, 1)$.

Global test

This global test is based on the bivariate distribution of $(\hat{\beta}, \hat{\gamma})$ and evaluates the treatment effect simultaneously. The hypothesis and the testing procedure are specified below.

1. $H_0 : \beta = \gamma = 0$ versus $H_a : \beta \neq 0$ or $\gamma \neq 0$. Let

$$T_{global} = \begin{pmatrix} \hat{\beta} & \hat{\gamma} \end{pmatrix} \hat{\Sigma}^{-1} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}$$

   H_0 is rejected if $P(Y > T_{global}) \leq \alpha$, where Y is χ^2-distributed with 2 degrees of freedom.

Sequential tests

The sequential tests are based on the bivariate distribution of $(\hat{\beta}, \hat{\gamma})$ first, subsequently the marginal distribution, and evaluates the treatment effect in a sequential manner. The two hypotheses and the testing procedures are specified below.

1. Order the values of $\tilde{\beta}$ and $\tilde{\gamma}$ according to absolute size. Assume that $|\tilde{\beta}| < |\tilde{\gamma}|$ without loss of generality
2. Let $H_l : \eta = 0$ be the ordered hypotheses for $\eta = \{\beta, \gamma\}$ with $l = \{1, 2\}$ representing the order
3. H_1 is rejected if $P(|\max(W_1, W_2)| > |\tilde{\beta}|) \leq \alpha$ where (W_1, W_2) is a bivariate standard normal variable with unit variance and correlation given by the correlation between $\tilde{\beta}$ and $\tilde{\gamma}$
4. If H_1 is rejected, H_2 is tested. H_2 is rejected if $P(|W_2| > |\tilde{\gamma}|) \leq \alpha$

This multiple testing procedure maintains the overall type I error at α (Ghosh and Lin 2000).

Alternatively, a hierarchical closed testing procedure that evaluates non-inferiority with respect to mortality and subsequently superiority with respect to recurrent events could also be of interest. This would correspond to the following testing procedure,

1. $H_0 : \gamma \geq \delta$ versus $H_a : \gamma < \delta$, where δ is the non-inferiority margin,
2. $H_0 : \beta \leq 0$ versus $H_a : \beta > 0$.

The second null hypothesis is only investigated if the first null has been rejected. This maintains the overall type I error at α.

3 Simulation studies

In order to examine the behaviour of the proposed model, simulation studies were conducted.

3.1 Data generation

The simulation scheme largely follows the elegant simulation procedure used in Ghosh and Lin (2002). We assume that we have a total sample size of n. Moreover, we assume that we have a single *one-dimensional* covariate, namely a binary treatment indicator for which $Z \sim \text{Bin}(1, p)$ with $p = 0.5$. We wish to simulate data according to

$$\begin{aligned} \log\left(\mu(t \mid Z)\right) &= \beta_0 + \beta Z, \\ \log\left(-\log\left(S(t \mid Z)\right)\right) &= \gamma_0 + \gamma Z, \end{aligned}$$

where β_0 and γ_0 denote the intercepts on log and cloglog scale, respectively. In this notation, β_0 and γ_0 are k-dimensional. It holds that,

$$\mu(t \mid Z) = \int_0^t S(u^- \mid Z)\, dR(u \mid Z),$$

where $S(t \mid Z) = P(D^* > t \mid Z)$ and $dR(t \mid Z) = E(dN^*(t) \mid D^* \geq t, Z)$ (Cook and Lawless 1997; Ghosh and Lin 2000). We consider the following data generating frailty models

$$\lambda^D(t \mid Z, \nu) = \nu \exp(\gamma_D Z)\lambda_0^D, \tag{8}$$

$$dR(t \mid Z, \nu) = \nu S_0(t \mid \nu)^{1-\exp(\gamma_D Z)} \exp(\beta Z)\, dt, \tag{9}$$

where $S_0(t \mid \nu) = \exp\left(-\int_0^t \lambda^D(u \mid Z = 0, \nu)\, du\right)$ and $\nu > 0$ is a frailty term that is assumed to be generated from a positive stable distribution with Laplace transform $\exp(-\nu^\rho)$, $\rho \in (0, 1]$. This frailty term introduces dependence between the recurrent events and the terminal event. The gap times between two successive events and the survival time will have a Kendall's tau correlation coefficient of $1 - \rho$. If $\rho = 1$, there is no association between successive recurrent events and deaths. If $\rho < 1$, there is a positive association between successive recurrent events and death and the magnitude

of the association is determined by ρ. Integrating Eqs. (8) and (9) with respect to ν implies that,

$$\mu(t \mid Z) = \int_0^\infty \mu(t \mid Z, \nu) f(\nu)\, d\nu = \mu_0(t) \exp(\beta Z),$$

where $f(\cdot)$ denotes the density of ν and

$$\mu_0(t) = \frac{1}{\lambda_0^D} \left(\exp\left(- \left(\lambda_0^D t \right)^\rho \right) + 1 \right).$$

Then considering a log-link function implies that

$$\log\left(\mu(t \mid Z)\right) = \log\left(\exp\left(\beta Z\right) \mu_0(t)\right) = \log(\mu_0(t)) + \beta Z,$$

where $\beta_0 = \log(\mu_0(t))$. Moreover, it holds that

$$S(t \mid Z) = \int_0^\infty S(t \mid Z, \nu) f(\nu)\, d\nu = \exp\left(- \exp(\gamma_D \rho Z) \left(\lambda_0^D t \right)^\rho \right).$$

Thus, considering a cloglog-link function implies that,

$$\log\left(-\log\left(S(t \mid Z)\right)\right) = \log\left(\left(\lambda_0^D t \right)^\rho \right) + \gamma Z,$$

where $\gamma_0 = \log\left(\left(\lambda_0^D t \right)^\rho \right)$ and $\gamma = \gamma_D \rho$. We will let $n = \{100, 500, 1000\}$, $\lambda_0^D = \{0.25, 1.5\}$, $\beta = \{0.5\}$, $\gamma_D = \{-0.2, 0, 0.2\}$, and $\rho = \{0.75, 1\}$. Censoring times are uniform on $[0, 5]$, independently of $N^*(\cdot)$ and D^*. At time 1 with $\rho = 1$, this corresponds to a baseline survival probability of $S(1 \mid Z = 0) = \exp(-0.25 \cdot 1) = 0.78$ and $S(1 \mid Z = 0) = \exp(-1.5 \cdot 1) = 0.22$ for $\lambda_0^D = \{0.25, 1.5\}$, respectively.

3.2 Choice of k and t_k

The proposed bivariate marginal model based on pseudo-observations requires a choice of number of time points, k, and time points, $t = (t_1, \ldots, t_k)$. It is possible to compute pseudo-observations at all observed event times, but this can quickly become computationally burdensome. Moreover, it is not ensured that the asymptotic properties of the estimation procedure will hold for computation at all event times as the theoretical results have been shown for pseudo-observations computed at a finite number of times (Overgaard et al. 2017; Overgaard 2019). We suggest selecting t based on one of three approaches: fixed time points, percentiles of event and death times or equidistantly chosen time points based on event and death times. For analysing randomised controlled trials, it will be natural to consider fixed time points, due to typical requirements of pre-specification. There could for instance be a clinical interest in comparing survival and expected number of recurrent events two years after randomisation.

We have compared the performance of the three approaches by simulation for $k = \{1, 2, 3, 4, 5, 10\}$ time points. We let $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 0.75)$. Other parameter choices provide similar results. The results are shown in Fig. 2. All three approaches are unbiased in estimating both β and γ for all k. The standard errors appear to be decreasing as a function of k and appear to stabilize after $k = 3$. This suggests that little is gained in terms of precision by including more than three time points. Also, the standard errors appear smaller for the equidistant and fixed time points approaches. Note, that the fixed time point approach is sensitive to the particular choice of t. The choice of relevant approach will depend on the application.

3.3 Investigation of normality assumption

In this section, we will investigate the assumption of large sample normality of $(\hat{\beta}, \hat{\gamma})$ estimated using the bivariate marginal pseudo-observation model by simulation. We will restrict our attention to $k = 1$ computed at $t = 2$ based on $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 0.75)$. A total of 1000 simulations were performed. Figure 3 displays eight plots that investigate the performance of the estimation using the bivariate marginal pseudo-observation model. The scatter plot displays the $\hat{\beta}$ estimates versus the $\hat{\gamma}$ estimates. The true β and $\gamma = \gamma_D \cdot \rho$ values are indicated by a dashed line. Contour lines have been added for the bivariate normal distribution with mean (β, γ) and variance given by the average of the estimated variances across simulations. The normalised plot has transformed the original scatter plot points using the bivariate normal distribution. The contour lines correspond to a bivariate standard normal distribution. Moreover, the number of points within each contour have been counted and compared to the total number of 1000 simulated points. This count is compared to the cumulative distribution function of the bivariate standard normal distribution. Both the simple and normalised scatter plot indicate that $(\hat{\beta}, \hat{\gamma})$ follows the relevant bivariate normal distribution.

Marginal normality of $\hat{\beta}$ and $\hat{\gamma}$ has been investigated using QQ plots. The theoretical quantiles of the expected distributions have been compared to the sample quantiles. The QQ plots indicate that the samples quantiles are aligned with the theoretical quantiles from the marginal distributions. Moreover, the predicted values of $\hat{\mu}(2)$ and $\hat{S}(2)$ have been compared to the true curves. Finally, histograms of the predicted values of $\hat{\mu}(2)$ and $\hat{S}(2)$ per treatment group alongside the true values have been displayed.

All visual examinations of the simulations are in line with the assumed bivariate normality of $(\hat{\beta}, \hat{\gamma})$. Similar results were seen for other parameter choices and when increasing k (see Appendix B). Appendix A.1 contains theoretical derivations that support the bivariate normality of $(\hat{\beta}, \hat{\gamma})$.

3.4 Comparison with other models

For the following, the performance of the marginal two-dimensional pseudo-observation model has been compared to other regression models for $\mu(t)$ and $S(t)$. For estimation of $\mu(t)$, the pseudo-observation model has been compared to that of Lin et al. (2000), and Ghosh and Lin (2002). The model of Ghosh and Lin has been fitted

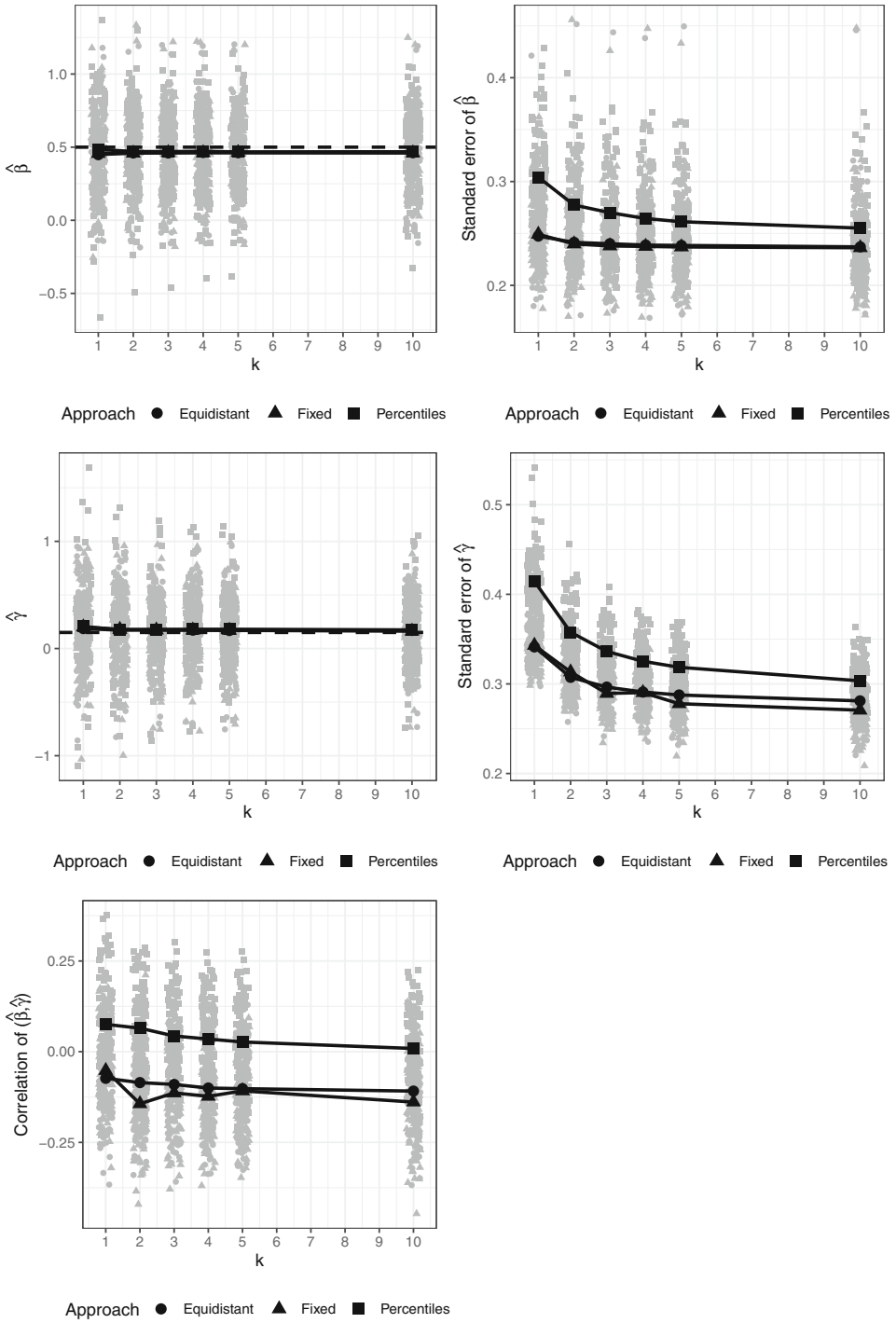


Fig. 2 Estimates obtained from the bivariate marginal pseudo-observation model using time points selected as either equidistantly, fixed times or percentiles for different k's based on 100 simulated data sets per k. The average value of the estimates are colored black and connected with lines for each choice and number of time points

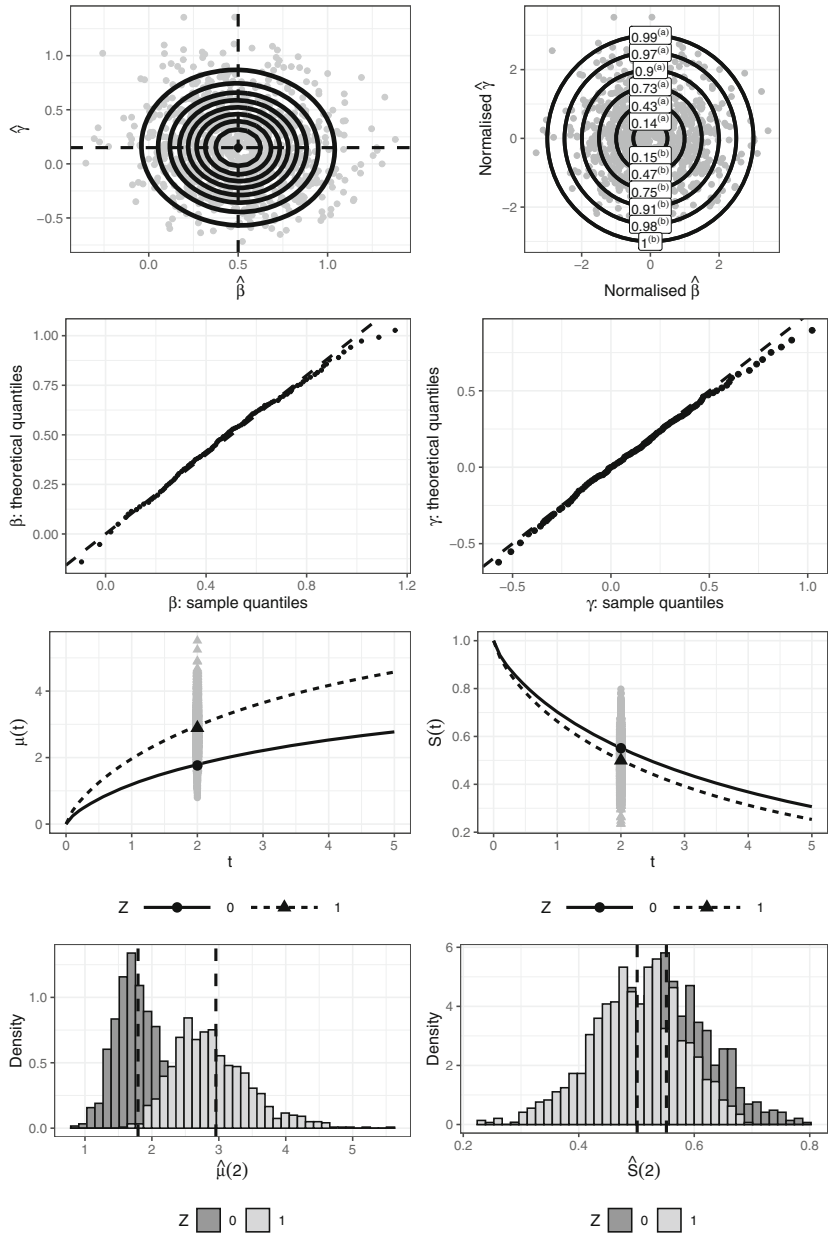


Fig. 3 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets with $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 0.75)$. *First row:* left hand side: scatter plot of $\hat{\beta}$ versus $\hat{\gamma}$ estimates, Right hand side: Normalised scatter plot compared to a bivariate standard normal distribution. *Second row:* left hand side: QQ-plot of $\hat{\beta}$ estimates, Right hand side: QQ-plot of $\hat{\gamma}$ estimates. *Third row:* left hand side: true $\mu(t)$ curve, predicted $\hat{\mu}(2)$ values (grey) and average of predicted $\hat{\mu}(2)$ values (black). Right hand side: True $S(t)$ curve, predicted $\hat{S}(2)$ values (grey) and average of predicted $\hat{S}(2)$ values (black). *Fourth row:* left hand side: histogram of predicted $\hat{\mu}(2)$ values with true $\mu(2)$ values indicated by a dashed line. Histogram of predicted $\hat{S}(2)$ values with true $S(2)$ values indicated by a dashed line. [a]Number of points in circle divided by number of simulated points. [b]Cumulative distribution function, $P(X, Y) <$ circle (Color figure online)

using inverse probability of censoring weights (IPCW). These models *also* assume that,

$$\mu(t \mid Z) = \mu_0(t) \exp(\beta Z).$$

For estimation of $S(t)$, the model has been compared to the regression model of Cox (1972). This model *also* assumes that,

$$\lambda^D(t \mid Z) = \lambda_0^D(t) \exp(\gamma Z).$$

The pseudo-observation model presented in the comparisons are based on pseudo-observations computed at three time points, $t = (1, 2, 3)$. Tables 1 and 2 show the results from the simulation studies on $S(t)$ and $\mu(t)$, respectively. The regression model of Lin, Wei, Yang and Ying (LWYY) does not accommodate competing deaths in estimation of $\mu(t)$. As a result, the estimated parameters in this model relate to a world where it is not possible to die. However, when the treatment does not affect death ($\gamma_D = 0$), the LWYY model still performs well in spite of the occurring deaths. When $\gamma_D \neq 0$, the LWYY model becomes biased, and the direction of bias depends on the sign of γ_D. The regression model of Ghosh and Lin (GL) has been applied to estimate $\mu(t)$ and performs well in most parameter settings. The bivariate marginal regression model based on pseudo-observations (Pseudo) performs similarly to the GL model in all scenarios. When ρ decreases, making the dependence between recurrent events and deaths larger, both the GL and Pseudo models have difficulties in estimating the treatment effect on recurrent events, β, in an unbiased manner.

For estimation of $S(t)$, the bivariate marginal regression model based on pseudo-observations has been compared to a Cox model. The two models are performing similarly in terms of bias and precision for all parameter choices.

3.5 Bootstrap experiment

Recent developments within the theoretical field of pseudo-observations *allude* to the fact that variance estimation done using a generalised estimating equation (GEE) framework with a working independence covariance matrix and using the sandwich variance estimator will overestimate the variance in some scenarios (Jacobsen and Martinussen 2016; Overgaard et al. 2017; Overgaard 2019). Thus, conservative tests are to be expected when fitting the model using this approach (see Sect. 2 and Eq. (5)). However, using this sandwich covariance matrix makes estimation simpler, and as noted in Jacobsen and Martinussen (2016), Overgaard et al. (2017), and Overgaard (2019), this will result in a negligible variance bias in most cases. In order to further investigate the large sample behaviour of the proposed *method*, a bootstrap experiment is carried out to investigate the distributional nature of $(\hat{\beta}, \hat{\gamma})$. This will enable assessment of potential bias in estimating both parameter estimates and their variance estimates. The bootstrap experiment is carried out in the following way,

1. A single data set is simulated according to the above simulation scheme with parameters $(n, \lambda_0^D, \beta, \gamma_D, \rho)$.

Table 1 Comparison of regression models for $S(t)$ based on 1000 simulations

$(\lambda_0^D, \beta, \gamma_D, \rho)$	n	Cox			Pseudo		
		$\text{Bias}_{\hat{\gamma}}$	$\text{SD}_{\hat{\gamma}}$	$\text{Mean}_{\text{SE}(\hat{\gamma})}$	$\text{Bias}_{\hat{\gamma}}$	$\text{SD}_{\hat{\gamma}}$	$\text{Mean}_{\text{SE}(\hat{\gamma})}$
(0.25, 0.5, 0, 1)	100	0.006	0.320	0.313	0.001	0.329	0.319
	500	0.004	0.139	0.137	0.006	0.144	0.140
	1000	0.008	0.097	0.097	0.007	0.100	0.099
(1.5, 0.5, 0, 1)	100	−0.009	0.226	0.220	−0.004	0.206	0.198
	500	0.000	0.099	0.097	−0.001	0.089	0.087
	1000	0.001	0.067	0.068	0.001	0.062	0.062
(1.5, 0.5, 0, 0.75)	100	−0.005	0.228	0.222	−0.003	0.228	0.220
	500	−0.002	0.097	0.098	−0.003	0.097	0.097
	1000	0.002	0.070	0.069	0.002	0.069	0.069
(0.25, 0.5, 0.2, 1)	100	−0.005	0.297	0.302	−0.016	0.302	0.306
	500	0.001	0.135	0.133	−0.010	0.139	0.135
	1000	0.003	0.093	0.094	−0.006	0.093	0.095
(0.25, 0.5, −0.2, 1)	100	−0.002	0.315	0.325	0.002	0.321	0.334
	500	−0.003	0.139	0.142	0.005	0.145	0.146
	1000	0.000	0.100	0.100	0.007	0.102	0.103
(0.25, 0.5, 0.2, 0.75)	100	−0.012	0.305	0.293	−0.016	0.309	0.298
	500	−0.004	0.129	0.129	−0.008	0.131	0.131
	1000	0.002	0.089	0.091	−0.004	0.092	0.092
(0.25, 0.5, −0.2, 0.75)	100	−0.001	0.305	0.308	0.001	0.318	0.315
	500	−0.004	0.140	0.135	−0.001	0.141	0.138
	1000	0.003	0.098	0.095	0.007	0.099	0.098

Note that $\gamma = \gamma_D \rho$. $\text{Bias}_{\hat{\gamma}} = \text{Mean}_{\hat{\gamma}} - \gamma$ and $\text{SD}_{\hat{\gamma}}$ denotes the bias and the standard deviation of the $\hat{\gamma}$ estimates, respectively. $\text{Mean}_{\text{SE}(\hat{\gamma})}$ denotes the mean of the estimated standard error for $\hat{\gamma}$

Cox Cox model, *Pseudo* pseudo-observation regression model

Table 2 Comparison of regression models for $\mu(t)$ based on 1000 simulations

$(\lambda_0^D, \beta, \gamma_D, \rho)$	n	LWYY			GL			Pseudo		
		$\text{Bias}_{\hat{\beta}}$	$\text{SD}_{\hat{\beta}}$	$\text{Mean}_{\text{SE}(\hat{\beta})}$	$\text{Bias}_{\hat{\beta}}$	$\text{SD}_{\hat{\beta}}$	$\text{Mean}_{\text{SE}(\hat{\beta})}$	$\text{Bias}_{\hat{\beta}}$	$\text{SD}_{\hat{\beta}}$	$\text{Mean}_{\text{SE}(\hat{\beta})}$
(0.25, 0.5, 0, 1)	100	0.002	0.140	0.136	0.000	0.178	0.175	0.002	0.173	0.173
	500	−0.002	0.063	0.061	−0.004	0.081	0.079	−0.003	0.078	0.077
	1000	−0.003	0.044	0.043	−0.007	0.057	0.056	−0.006	0.055	0.054
(1.5, 0.5, 0, 1)	100	0.003	0.249	0.239	0.011	0.316	0.303	0.009	0.308	0.300
	500	0.002	0.110	0.106	0.001	0.139	0.135	0.002	0.136	0.133
	1000	−0.002	0.075	0.075	−0.003	0.094	0.095	−0.003	0.092	0.094
(1.5, 0.5, 0, 0.75)	100	0.012	0.306	0.282	0.018	0.324	0.309	0.017	0.321	0.308
	500	−0.007	0.129	0.127	−0.004	0.136	0.138	−0.004	0.134	0.137
	1000	−0.007	0.089	0.089	−0.009	0.098	0.097	−0.008	0.097	0.097
(0.25, 0.5, 0.2, 1)	100	0.084	0.139	0.137	0.013	0.183	0.180	0.016	0.176	0.176
	500	0.071	0.062	0.061	−0.005	0.083	0.081	−0.003	0.081	0.078
	1000	0.071	0.045	0.043	−0.005	0.060	0.057	−0.006	0.057	0.055
(0.25, 0.5, −0.2, 1)	100	−0.066	0.142	0.136	−0.006	0.174	0.170	−0.009	0.167	0.169
	500	−0.064	0.061	0.061	−0.001	0.076	0.077	−0.001	0.073	0.076
	1000	−0.062	0.045	0.043	−0.001	0.057	0.054	−0.001	0.055	0.054
(0.25, 0.5, 0.2, 0.75)	100	0.024	0.253	0.232	−0.026	0.248	0.231	−0.030	0.251	0.239
	500	0.039	0.110	0.108	−0.017	0.109	0.108	−0.018	0.113	0.111
	1000	0.042	0.075	0.077	−0.016	0.074	0.076	−0.017	0.076	0.079
(0.25, 0.5, −0.2, 0.75)	100	−0.055	0.227	0.222	−0.007	0.228	0.219	−0.008	0.235	0.228
	500	−0.054	0.104	0.101	−0.004	0.104	0.100	−0.004	0.108	0.104
	1000	−0.058	0.071	0.072	−0.010	0.072	0.071	−0.010	0.075	0.074

$\text{Bias}_{\hat{\beta}} = \text{Mean}_{\hat{\beta}} - \beta$ and $\text{SD}_{\hat{\beta}}$ denotes the bias and the standard deviation of the $\hat{\beta}$ estimates, respectively. $\text{Mean}_{\text{SE}(\hat{\beta})}$ denotes the mean of the estimated standard error for $\hat{\beta}$
LWYY Lin, Wei, Yang and Ying model, *GL* Ghosh and Lin model, *Pseudo* Pseudo-observation regression model

2. Based on this data set, pseudo-observations of $(\mu(t), S(t))$ are computed and a bivariate marginal model is fitted based on $(\hat{\mu}_i(t), \hat{S}_i(t))$ using GEE as described in Sect. 2. This estimation uses a working independence covariance matrix assumption. Here, we have chosen $k = 1$ and $t = 2$. The estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\Sigma}$ are saved.
3. In order to assess this estimation, the simulated data set is sampled with replacement. That is, replacement for each individual, since the individuals are independent and identically distributed replicates. This resampled data set is then fed into the estimation procedure a total of B times, and the estimates of $\hat{\beta}_{boot,b}$, $\hat{\gamma}_{boot,b}$, $\hat{\Sigma}_{boot,b}$ and $\hat{\Omega}_{boot,b}$ for $b = 1, \ldots, B$ are saved. Here, the calculation of the pseudo-observations is a part of the bootstrap step.
4. Based on the B bootstrap samples, the following is computed,

$$\hat{\beta}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{boot,b}, \quad \hat{\gamma}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_{boot,b},$$

$$\hat{\sigma}_{11,boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_{boot,b} - \hat{\beta}_{boot})^2, \quad \hat{\sigma}_{22,boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\gamma}_{boot,b} - \hat{\gamma}_{boot})^2,$$

$$\hat{\sigma}_{12,boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_{boot,b} - \hat{\beta}_{boot})(\hat{\gamma}_{boot,b} - \hat{\gamma}_{boot}),$$

$$\hat{\omega}_{12,boot} = \frac{\hat{\sigma}_{12,boot}}{\sqrt{\hat{\sigma}_{11,boot}}\sqrt{\hat{\sigma}_{22,boot}}}.$$

The results of the bootstrap experiment are displayed in Table 3. For all parameter settings, the pairs $(\hat{\beta}, \hat{\beta}_{boot})$ and $(\hat{\gamma}, \hat{\gamma}_{boot})$ were close to each other. All in all, the bootstrap variances are very close to the variance estimates from the bivariate marginal pseudo-observation model for all of the chosen settings. This indicates that the model is performing well in terms of correctly estimating the variance using the assumed sandwich covariance matrix [see Eq. (5)]. The model is seen to be slightly conservative, as expected, when looking at other more extreme parameter setting such as when $\beta = 2$. This is also supported by the formula for the asymptotic variance of $\sqrt{n}(\hat{\xi} - \xi)$ as seen in Appendix A.1.

4 Applications

The suggested model will be applied to two studies, a bladder cancer study and a cardiovascular trial, LEADER.

4.1 Bladder cancer

A clinical cancer trial was conducted by the Veterans Administration Cooperative Urological Research Group (Byar 1980). A total of 118 patients with stage I bladder cancer

Table 3 Results from bootstrap experiment on the bivariate marginal regression model based on pseudo-observations

$(\lambda_0^D, \beta, \gamma_D, \rho)$	n	$\sqrt{\hat{\sigma}_{11}}$	$\sqrt{\hat{\sigma}_{11,boot}}$	$\sqrt{\hat{\sigma}_{22}}$	$\sqrt{\hat{\sigma}_{22,boot}}$	$\hat{\omega}_{12}$	$\hat{\omega}_{12,boot}$
(0.25, 0.5, 0, 1)	100	0.215	0.225	0.335	0.344	−0.363	−0.316
	500	0.082	0.080	0.158	0.160	−0.446	−0.449
	1000	0.057	0.057	0.117	0.115	−0.447	−0.418
(0.25, 0.5, 0.2, 1)	100	0.194	0.205	0.347	0.366	−0.312	−0.312
	500	0.080	0.081	0.158	0.160	−0.409	−0.420
	1000	0.060	0.060	0.113	0.116	−0.404	−0.419
(0.25, 0.5, −0.2, 1)	100	0.185	0.191	0.375	0.380	−0.506	−0.521
	500	0.081	0.081	0.177	0.180	−0.394	−0.418
	1000	0.057	0.057	0.123	0.124	−0.362	−0.360
(0.25, 0.5, 0.2, 0.75)	100	0.313	0.307	0.349	0.368	−0.056	−0.048
	500	0.124	0.125	0.152	0.159	−0.015	0.016
	1000	0.083	0.082	0.104	0.106	−0.056	−0.052
(0.25, 0.5, −0.2, 0.75)	100	0.222	0.220	0.405	0.442	−0.233	−0.220
	500	0.103	0.104	0.152	0.156	−0.010	−0.077
	1000	0.080	0.089	0.110	0.116	0.000	−0.001
(0.25, 2, 0, 1)	100	0.220	0.181	0.388	0.395	−0.384	−0.361
	500	0.102	0.084	0.158	0.162	−0.303	−0.376
	1000	0.078	0.062	0.115	0.115	−0.302	−0.320
(0.25, 2, 1, 1)	100	0.248	0.202	0.353	0.360	−0.127	−0.271
	500	0.125	0.101	0.151	0.153	−0.221	−0.367
	1000	0.092	0.071	0.105	0.106	−0.240	−0.444
(0.25, 2, −1, 1)	100	0.264	0.252	0.493	0.540	−0.223	−0.150
	500	0.102	0.078	0.231	0.235	−0.267	−0.227
	1000	0.075	0.061	0.145	0.145	−0.270	−0.250

The results are based on $B = 1000$. For each parameter setting, a single data set is simulated. Here, $\sqrt{\hat{\sigma}_{11}}$, $\sqrt{\hat{\sigma}_{22}}$ and $\hat{\omega}_{12}$ denotes the estimates from the bivariate pseudo-observation model fit based on the single simulated data. The bootstrap estimates, $\sqrt{\hat{\sigma}_{11,boot}}$, $\sqrt{\hat{\sigma}_{22,boot}}$ and $\hat{\omega}_{12,boot}$, are computed as described in the text

were randomised to receive placebo, pyridoxine or thiotepa. Following randomisation, the patients were examined for occurrences of superficial bladder tumours (recurrent events) and any deaths were registered. We will focus on the comparison of placebo and thiotepa (86 patients in total). Across both treatment groups, the median follow-up time was 30 months.

Figure 4 displays the Kaplan–Meier estimates, $\hat{S}(t)$, and the marginal mean estimates computed using Eq. (1), $\hat{\mu}(t)$, per treatment group. Pseudo-observations of the marginal mean function for all times were computed for four select individuals, see Fig. 5. The pseudo-observations, $\hat{\mu}_i(t)$, jump at event times. Prior to any events, censoring or death, $\hat{\mu}_i(t)$ decreases over time. After death, $\hat{\mu}_i(t)$ decreases over time and after censoring, $\hat{\mu}_i(t)$ increases over time.

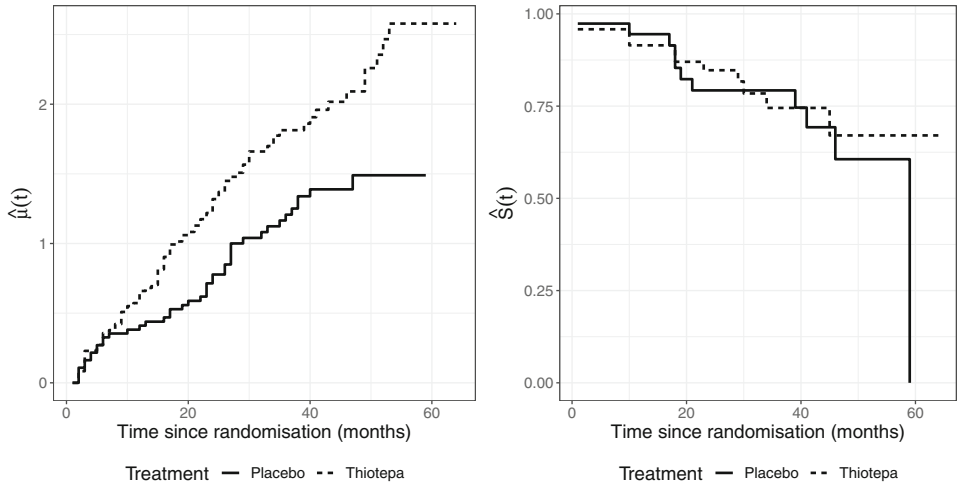


Fig. 4 Non-parametric estimates computed on the bladder cancer data. **Left hand side:** Marginal mean estimates for recurrent bladder cancer, $\hat{\mu}(t)$, per treatment group computed using equation (1). **Right hand side:** Kaplan-Meier estimates for any death, $\hat{S}(t)$, per treatment group

The bivariate marginal pseudo-observation model was applied to the bladder cancer data. The model was applied to $k = 1$ with $t = (30)$ and $k = 3$ with $t = (20, 30, 40)$. Parameter estimates from the models are available in Table 4. Model predictions are available in Table 5. In comparison, parameter estimates from the model of Ghosh and Lin applied to recurrent bladder cancer results in $\hat{\beta} = -0.406$ with $\text{SE}(\hat{\beta}) = 0.286$. Similarly, parameter estimates from a Cox model applied to death results in $\hat{\gamma} = 0.281$ with $\text{SE}(\hat{\gamma}) = 0.430$.

Based on the parameter estimates from the pseudo-observation model with $k = 3$, we have performed the hypotheses tests suggested in Sect. 2.1. For the local tests,

$$\tilde{\beta} = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}_{11}}} = -1.533, \quad \tilde{\gamma} = \frac{\hat{\gamma}}{\sqrt{\hat{\sigma}_{22}}} = 0.139.$$

As we have two-sided alternatives, the corresponding p-values are computed as,

$$P(|U| > |\tilde{\beta}|) = 2 \cdot P(U < \tilde{\beta}) = 0.125,$$
$$P(|U| > |\tilde{\gamma}|) = 2 \cdot P(U > \tilde{\gamma}) = 0.889,$$

where U denotes an univariate standard normal variable. Based on the p-values from the local tests, thiotepa treatment does not seem to be significantly different from placebo on either recurrent events or death at a 0.05 significance level. For the global test, under Eq. (7), it holds that,

$$T_{global} = \begin{pmatrix} -0.427 & 0.066 \end{pmatrix} \hat{\Sigma}^{-1} \begin{pmatrix} -0.427 \\ 0.066 \end{pmatrix} = 2.387.$$

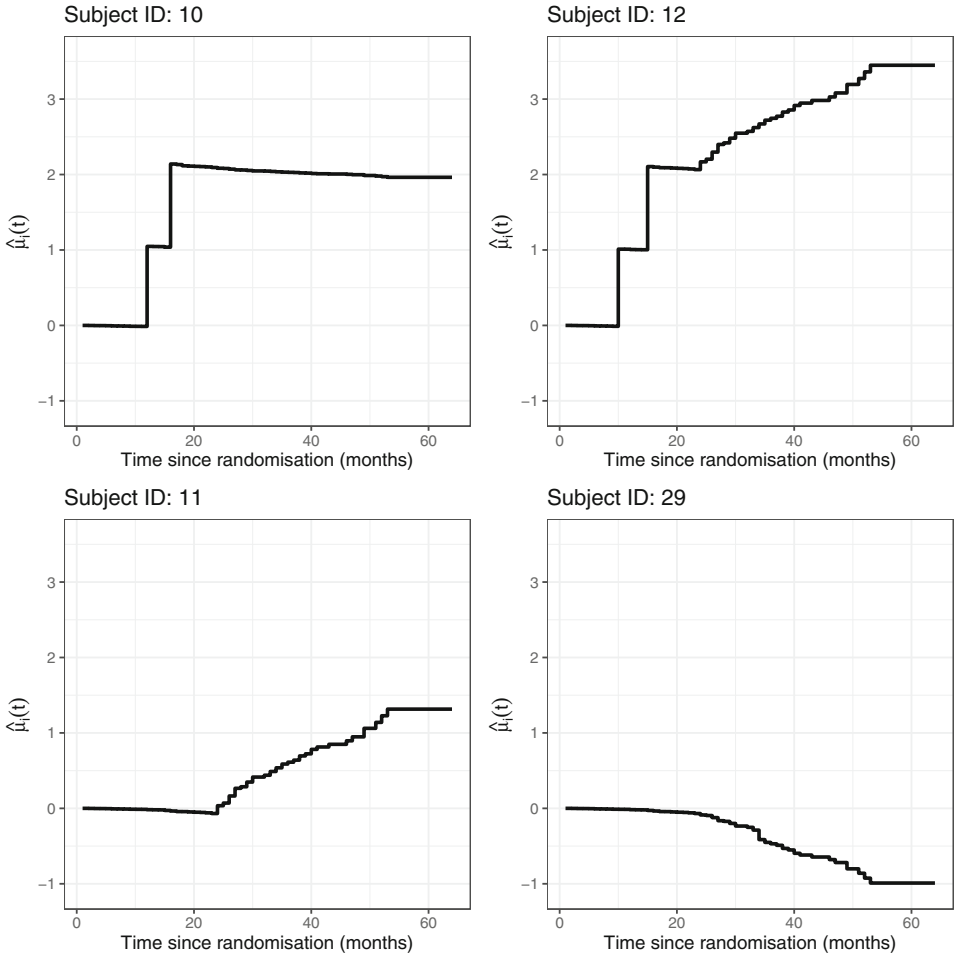


Fig. 5 Pseudo-observations, $\hat{\mu}_i(t)$, over time for four subjects from the bladder cancer data. Subject 10 has two bladder cancer events at times 12 and 16, and dies at time 18. Subject 12 also has two bladder cancer events at time 10 and 15 and is censored at time 23. Subject 11 is censored at time 23, and subject 29 dies at time 34, without previous bladder cancer

For which,

$$P(Y > T_{global}) = 0.303,$$

where Y is $\chi^2(2)$-distributed. Again, there is no evidence to indicate that the effect of thiotepa treatment on recurrences or death is different from placebo treatment. For the sequential test,

$$P(|\max(W_1, W_2)| > |\tilde{\beta}|) = 2 \cdot P(\max(W_1, W_2) < \tilde{\beta}) = 0.008,$$
$$P(|W_2| > |\tilde{\gamma}|) = 2 \cdot P(W_2 > \tilde{\gamma}) = 0.889,$$

Table 4 Parameter estimates from the bivariate marginal pseudo-observation model based on the bladder cancer data

Model	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$	$\sqrt{\hat{\sigma}_{11}}$	$\sqrt{\hat{\sigma}_{22}}$	$\hat{\omega}_{12}$
$k = 1, t = (30)$	−0.464	−0.048	0.075	0.261	−0.001	0.275	0.511	−0.011
$k = 3, t = (20, 30, 40)$	−0.427	0.066	0.078	0.228	0.005	0.279	0.477	0.036

Table 5 Predictions from the bivariate marginal pseudo-observation model based on the bladder cancer data

Model		$\hat{\mu}(t_l \mid Z = 0)$	$\hat{\mu}(t_l \mid Z = 1)$	$\hat{S}(t_l \mid Z = 0)$	$\hat{S}(t_l \mid Z = 1)$
$k = 1$	$t_1 = 30$	1.666	1.048	0.785	0.794
	$t_1 = 20$	1.037	0.693	0.853	0.847
$k = 3$	$t_2 = 30$	1.640	1.096	0.792	0.784
	$t_3 = 40$	1.966	1.313	0.749	0.740

Here Z denotes the binary treatment variable, where $Z = 0$ denotes placebo and $Z = 1$ denotes thiotepa

where (W_1, W_2) is defined as in Section 2.1. The tests indicates that thiotepa treatment is effective in reducing recurrent events compared to placebo. However, there does not seem to be a difference between the treatments on mortality.

4.2 LEADER

LEADER (Liraglutide Effect and Action in Diabetes: Evaluation of cardiovascular outcome Results) (Marso et al. 2016) was a double-blind randomized controlled trial investigating the cardiovascular effects of liraglutide, a glucagon like peptide-1 (GLP-1) receptor agonist approved for treatment of type 2 diabetes, investigated versus placebo when added to standard of care in a population with type 2 diabetes and a high cardiovascular risk. A total of 9340 subjects were randomised 1:1 to receive either liraglutide or placebo. The primary composite endpoint was a three-component major cardiovascular adverse events (3-p MACE) endpoint consisting of; non-fatal stroke, non-fatal myocardial infarction (MI) or cardiovascular (CV) death. The primary analysis was a time-to-event analysis using a Cox regression model considering the time to first 3-p MACE with treatment as a covariate. The median follow-up time was 3.8 years. For this application, we will focus on recurrent myocardial infarction, which consists of both fatal and non-fatal events. If a myocardial infarction was fatal, this has been coded as an MI event occurring on a given calendar day, and then a cardiovascular death on the subsequent calendar day.

Figure 6 displays the $\hat{S}(t)$ and $\hat{\mu}(t)$ per treatment based on LEADER data. The bivariate marginal pseudo-observation model, with $k = 1$ and $t = (30)$ and with $k = 3$ and $t = (20, 30, 40)$, was applied to the LEADER data. Parameter estimates from the models are available in Table 6. Model predictions are available in Table 7. In comparison, parameter estimates from the model of Ghosh and Lin applied to LEADER data results in $\hat{\beta} = -0.159$ with $\text{SE}(\hat{\beta}) = 0.088$. As a result of the different estimation procedures, the estimates are slightly different. Similarly, parameter estimates from a Cox model applied to death results in $\hat{\gamma} = -0.166$ with $\text{SE}(\hat{\gamma}) = 0.070$.

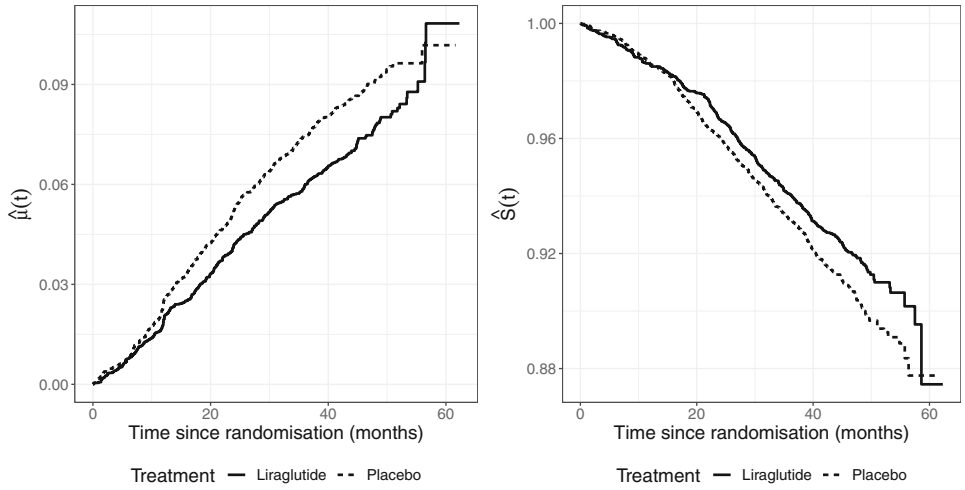


Fig. 6 Non-parametric estimates computed on the LEADER data. *Left hand side:* marginal mean estimates for recurrent myocardial infarction, $\hat{\mu}(t)$, per treatment group computed using Eq. (1). *Right hand side:* Kaplan–Meier estimates for any death, $\hat{S}(t)$, per treatment group

Based on the parameter estimates from the pseudo-observation model with $k = 3$, we have performed the hypotheses tests suggested in Sect. 2.1. For the local tests,

$$\tilde{\beta} = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}_{11}}} = -2.235, \quad \tilde{\gamma} = \frac{\hat{\gamma}}{\sqrt{\hat{\sigma}_{22}}} = -1.988.$$

Then,

$$P(|U| > |\tilde{\beta}|) = 2 \cdot P(U < \tilde{\beta}) = 0.025,$$
$$P(|U| > |\tilde{\gamma}|) = 2 \cdot P(U < \tilde{\gamma}) = 0.047.$$

Thus, there is evidence to indicate that liraglutide treatment is effective in reducing the number of recurrences and mortality compared to placebo. The evidence is stronger for the treatment effect on recurrent events. For the global test, under Eq. (7), it holds that,

$$T_{global} = \begin{pmatrix} -0.218 & -0.163 \end{pmatrix} \hat{\Sigma}^{-1} \begin{pmatrix} -0.218 \\ -0.163 \end{pmatrix} = 8.138.$$

For which, $P(Y > T_{global}) = 0.017$. Hence there is evidence indicating that liraglutide treatment has a significant effect on either recurrent events or death compared to placebo at a 0.05 level. For the sequential test,

$$P(|\max(W_1, W_2)| > |\tilde{\beta}|) = 2 \cdot P(\max(W_1, W_2) < \tilde{\beta}) < 0.001,$$
$$P(|W_2| > |\tilde{\gamma}|) = 2 \cdot P(W_2 < \tilde{\gamma}) = 0.047.$$

Table 6 Parameter estimates from the bivariate marginal pseudo-observation model based on the LEADER data

Model	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$	$\sqrt{\hat{\sigma}_{11}}$	$\sqrt{\hat{\sigma}_{22}}$	$\hat{\omega}_{12}$
$k = 1, t = (30)$	−0.212	−0.157	0.010	0.009	0.001	0.101	0.093	0.093
$k = 3, t = (20, 30, 40)$	−0.218	−0.163	0.010	0.007	0.001	0.097	0.082	0.100

Table 7 Predictions from the bivariate marginal pseudo-observation model based on the LEADER data

Model		$\hat{\mu}(t_l \mid Z = 0)$	$\hat{\mu}(t_l \mid Z = 1)$	$\hat{S}(t_l \mid Z = 0)$	$\hat{S}(t_l \mid Z = 1)$
$k = 1$	$t_1 = 30$	0.064	0.052	0.946	0.953
	$t_1 = 20$	0.042	0.033	0.970	0.975
$k = 3$	$t_2 = 30$	0.064	0.052	0.946	0.953
	$t_3 = 40$	0.081	0.065	0.921	0.932

Here Z denotes the binary treatment variable, where $Z = 0$ denotes placebo and $Z = 1$ denotes liraglutide

The sequential test results again substantiates that liraglutide treatment is effective on both recurrences and death compared to placebo.

The three-dimensional model mentioned in Sect. 2 and Eq. (6) is applied to the LEADER data. Here, the two death causes of interest are cardiovascular (CV) death and non-CV death. Thus, the model simultaneously models the marginal expected number of recurrent MI events, the cumulative incidence of CV death and the cumulative incidence of non-CV death conditional on treatment. The pseudo-observations of $\hat{\mu}_i(t)$, $\hat{C}_{1i}(t)$ and $\hat{C}_{2i}(t)$ for an individual is shown in Fig. 7. This subject had several recurrent MI events and died of a CV death. The majority of all censorings occur after 40 months, with only 27 out of 9340 subjects censored prior to 40 months. The model applied to $t = 30$ results in

$$\hat{\xi} = \begin{pmatrix} -0.212 \\ -0.320 \\ 0.093 \end{pmatrix}, \ \hat{\Sigma} = \begin{pmatrix} 0.010 & 0.001 & 0.000 \\ 0.001 & 0.014 & 0.000 \\ 0.000 & 0.000 & 0.023 \end{pmatrix}, \ \hat{\Omega} = \begin{pmatrix} 1 & 0.108 & 0.013 \\ 0.108 & 1 & -0.025 \\ 0.013 & -0.025 & 1 \end{pmatrix}$$

for $\mu(t)$, $C_1(t)$, $C_2(t)$, respectively. Here, $C_1(t)$ and $C_2(t)$ denote the cumulative incidence for CV death and non-CV death, respectively. The estimates indicate that liraglutide is beneficial in terms of reducing the recurrent MI events and CV mortality compared to placebo. There does not seem to be a difference between the treatments in terms of non-CV death. Moreover, we see that the estimated correlation is largest between the treatment effect on recurrent MI and CV death. When applying the model to $t = (20, 30, 40)$, the estimates are

$$\hat{\xi} = \begin{pmatrix} -0.218 \\ -0.292 \\ 0.034 \end{pmatrix}, \ \hat{\Sigma} = \begin{pmatrix} 0.010 & 0.001 & 0.000 \\ 0.001 & 0.011 & 0.000 \\ 0.000 & 0.000 & 0.018 \end{pmatrix}, \ \hat{\Omega} = \begin{pmatrix} 1 & 0.117 & 0.012 \\ 0.117 & 1 & -0.032 \\ 0.012 & -0.032 & 1 \end{pmatrix}.$$

Note, that the estimates for $\mu(t)$ in the three-dimensional model coincides with the estimates for $\mu(t)$ from the two-dimensional model. In comparison, fitting a Fine &

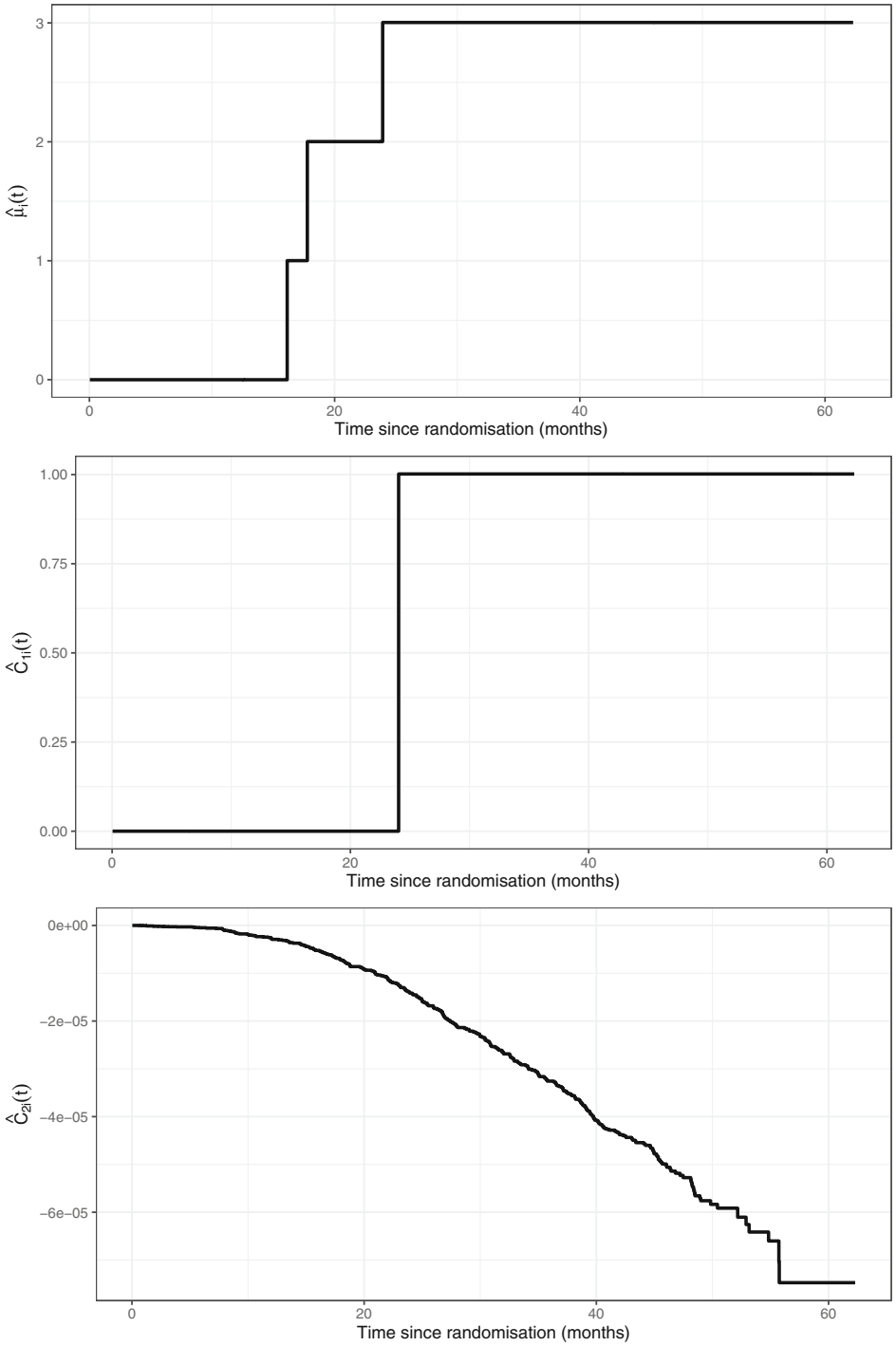


Fig. 7 Pseudo-observations of $\hat{\mu}_i(t)$, $\hat{C}_{1i}(t)$ and $\hat{C}_{2i}(t)$ for all times for a given individual based on LEADER data. The subject experienced three myocardial infarctions at 16, 18 and 24 months after randomisation and died of a cardiovascular death 24 months after randomisation

Gray model to each death cause results in $\hat{\xi}_2 = -0.243$ (SE$(\hat{\xi}_2) = 0.090$) for CV death and $\hat{\xi}_3 = -0.039$ (SE$(\hat{\xi}_3) = 0.110$) for non-CV death.

5 Discussion

The proposed bivariate marginal regression model based on pseudo-observations performs similarly to the regression models proposed by Ghosh and Lin (2002) and Cox (1972), see Sect. 3.4. The suggested model, however, carries the strength of a two-dimensional modelling approach with joint covariance estimation. The model requires independent censoring for computing both marginal mean estimates and survival estimates in an unbiased manner. Andersen and Perme (2010) and Binder et al. (2014) address dependent censoring when computing pseudo-observations for survival probabilities or cumulative incidences using stratification and an IPCW approach allowing the censoring distribution to depend on certain covariates using e.g. a Cox model. Similar approaches can likely be used to accommodate dependent censoring in the estimation of $\mu(t)$ [see Eq. (2)], but extending the bivariate model to address potential dependent censoring is a topic for further research.

The simulations focus on a single binary covariate, analogous to a treatment indicator. But the proposed model *can be extended* to include other baseline covariates of interest *as done in the general model formulation*. For assessing the fit of the two-dimensional marginal model based on pseudo-observations, pseudo-residuals explored by Andersen and Perme (2010) or cumulative sums of pseudo-observations considered in Pavlič et al. (2019) may be of use. Non-proportional hazards in terms of a covariate effect may be investigated using a model based on pseudo-observations that includes an interaction term between the covariate and the time points that are used in the computation. Hence, any non-constant effect of treatment on either recurrent events or survival over the chosen time points may be investigated.

The joint frailty model targets intensity functions for recurrent events and death simultaneously, conditional on a subject level frailty (Liu et al. 2004). The bivariate marginal regression model based on pseudo-observations also targets the recurrent event process and the death process jointly. This marginal model estimates a mean ratio and a hazard ratio parameter for the recurrent event and terminal event process, respectively. Whereas, the joint frailty model estimates hazard ratios conditional on a subject level frailty. Hazard ratios can be harder to understand than mean ratio parameters. Moreover, the parameters in frailty type models are characterized by having subject specific interpretations, which may not be desirable if aiming for population specific estimates. Finally, the joint frailty model has a stronger parametric assumption about the relationship between recurrent events and death than the bivariate marginal model. Another strength of the pseudo-observation model is the flexibility in the choice of link functions which can make interpretation of the final model easier. This allows the choice of several different effect parameters compared to the joint frailty model. It may be desirable to consider an id-link or a log-link for the cumulative incidences which would correspond to risk differences or ratios. Restricted mean survival times could be considered as an alternative to cumulative incidences in the bivariate marginal model based on pseudo-observations.

Theoretical properties and large sample behaviour of the proposed model has been explored through simulation and in theoretical details in Appendix A.1. It has been shown that estimation of $\hat{\beta}$ and $\hat{\gamma}$ separately using the suggested approach ensures large sample normality (Overgaard et al. 2017; Overgaard 2019). Our theoretical results and the presented simulation studies support that the same holds for the joint estimation of $(\hat{\beta}, \hat{\gamma})$.

The bivariate marginal pseudo-observation model may be extended to more high-dimensional marginal modelling problems. This is exemplified using the LEADER data and the three-dimensional model for recurrent myocardial infarction, cardiovascular death and non-cardiovascular death. This model can guide understanding of how treatment affects each of the components and how these are related. Moreover, as the sequential tests indicate, power may be gained by utilizing the multivariate distribution. Often, these high-dimensional issues are analysed using marginal models for each component. The generalisation of the bivariate marginal pseudo-observation model targets high-dimensional marginal parameters in a simple manner.

Funding This research was carried out as part of xxx's Ph.D. education. For the PhD education, she received funding from Novo Nordisk A/S and Innovation Fund Denmark. xxx is supported by the Novo Nordisk Foundation Grant NNF17OC0028276.

Code availability R code for the bivariate marginal pseudo-observation model is available upon request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

A Appendix

A.1 Theoretical details on bivariate normality of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$

According to Overgaard et al. (2017), the pseudo-observation approach of this paper produces consistent and asymptotically normal parameter estimates under essentially two conditions. One condition is that the estimate $\hat{\theta}$ of $\theta = E(f(W))$ can be seen as a functional, ϕ, of the empirical distribution, F_n, in a Banach space setting such that ϕ is two times (Fréchet) differentiable with a Lipschitz continuous second order derivative and such that $\|F_n\|$ converges at a certain rate. This condition ensures that the close approximation of the pseudo-observation $\hat{\theta}_i = \theta + \dot{\theta}(X_i) + \frac{1}{n-1}\sum_{j\neq i}\ddot{\theta}(X_i, X_j) + o_P(n^{-\frac{1}{2}})$ (uniformly in i) in terms of the estimator's first and second order influence functions, $\dot{\theta}$ and $\ddot{\theta}$, holds. This, in turn, implies that the less close approximation $\hat{\theta}_i = \theta + \dot{\theta}(X_i) + o_P(1)$ also holds. The other condition is therefore that $E(\dot{\theta}(X) \mid Z) = E(f(W) \mid Z) - \theta$, which means that the pseudo-observations carry the right information and ensures that the estimating equation is unbiased under the model. The result of Overgaard et al. (2017) is formulated for one-dimensional pseudo-observations, but generalizes to multi-dimensional outcomes. In a multi-dimensional setting, the requirements then need to hold for each outcome separately.

For pseudo-observations of the Kaplan–Meier estimate $\hat{S}(t_l)$, the conditions above hold under assumption of positivity, i.e. $P(C > t_l) > 0$, and completely independent censoring, i.e. that C is independent of (D^*, Z), as described by Overgaard et al. (2017) based on the work of Graw et al. (2009) and Jacobsen and Martinussen (2016). For pseudo-observations of $\hat{\mu}(t_l)$, the conditions were established by Overgaard (2019), see Example 8, under similar assumptions of positivity, completely independent censoring, here that C is independent of (N^*, D^*, Z), and additionally the assumption that $N^*(t_l)$ has a little more than finite fourth moment.

The result of Overgaard et al. (2017) is that, under regularity conditions, estimates, $\hat{\xi} = \hat{\xi}_n$, exist that solve (4) with high probability for large n such that

$$\sqrt{n}(\hat{\xi}_n - \xi)$$

is asymptotically normal with mean 0 and variance

$$M^{-1}\Psi M^{-1},$$

where

$$M = E\left(\left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1} \frac{\partial m_i}{\partial \xi}\right)$$

and

$$\Psi = \mathrm{Var}\left(\left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1}(\theta + \dot{\theta}(X_i) - m(\xi; Z_i)) + h(X_i)\right)$$

with

$$h(x) = E\left(\left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1}\ddot{\theta}(x, X_i)\right).$$

In summary, the suggested pseudo-observation approach produces consistent and asymptotically normal parameter estimates under the assumptions

1. positivity, $P(C > t_k) > 0$,
2. completely independent censoring, i.e. C is independent of (N^*, D^*, Z),
3. a little more than finite fourth moment of $N^*(t_k)$.

It is worth noting that the suggested estimate of Ψ can be expected to consistently estimate $\mathrm{Var}\left(\left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1}(\theta + \dot{\theta}(X_i) - m(\xi; Z_i))\right)$ but not $\mathrm{Var}\left(\left(\frac{\partial m_i}{\partial \xi}\right)^T V_i^{-1}(\theta + \dot{\theta}(X_i) - m(\xi; Z_i)) + h(X_i)\right)$. In other words, any contribution from the second order terms of h are not included and so the estimate, and thereby the standard errors of the sandwich variance estimator, can be expected to be biased.

B Plots from simulation of bivariate normality of $(\hat{\beta}, \hat{\gamma})$

This appendix displays additional plots visualizing the bivariate normal distribution of $(\hat{\beta}, \hat{\gamma})$ for different parameter settings and k.

$(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 1)$ and $t = 2$

See Appendix Fig. 8.

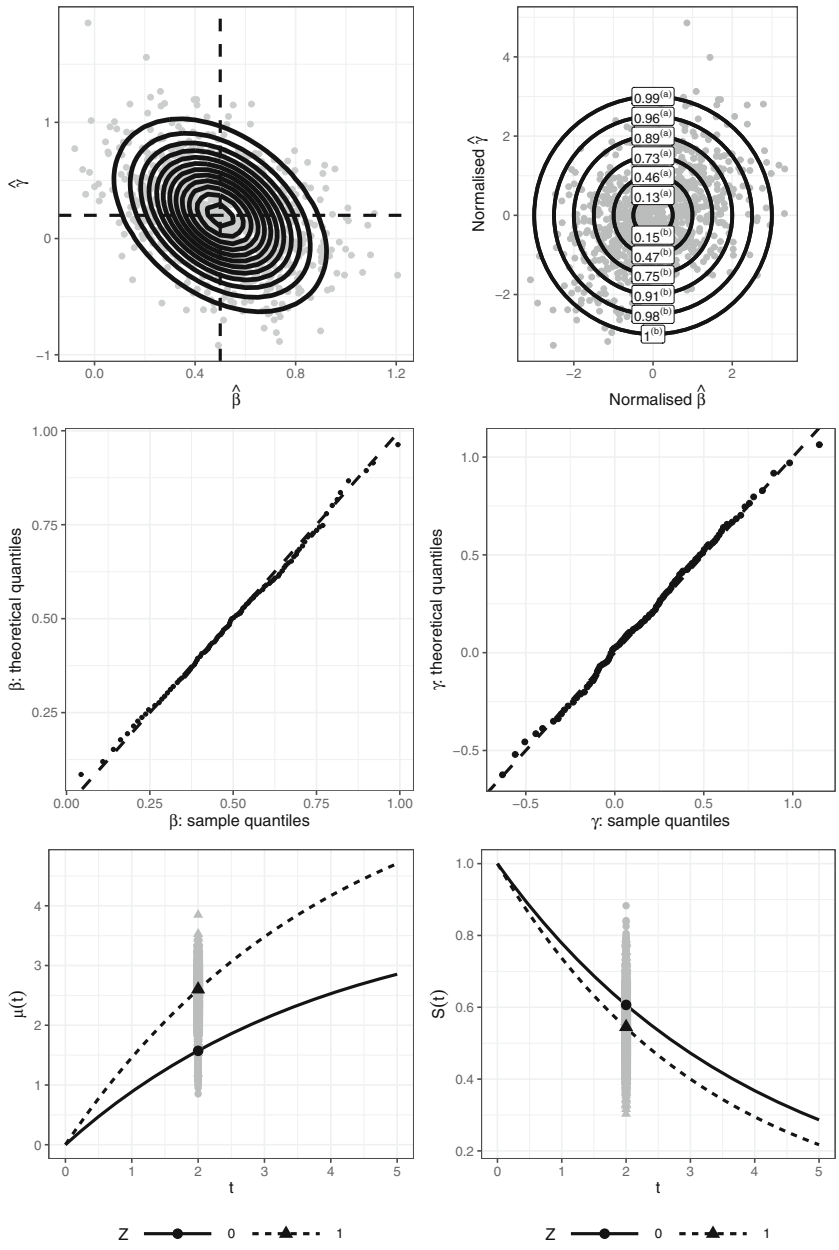


Fig. 8 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets and $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 1)$. The pseudo-observations are computed based on $k = 1$ with $t = 2$.

$(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, -0.2, 1)$ and $t = 2$

See Appendix Fig. 9.

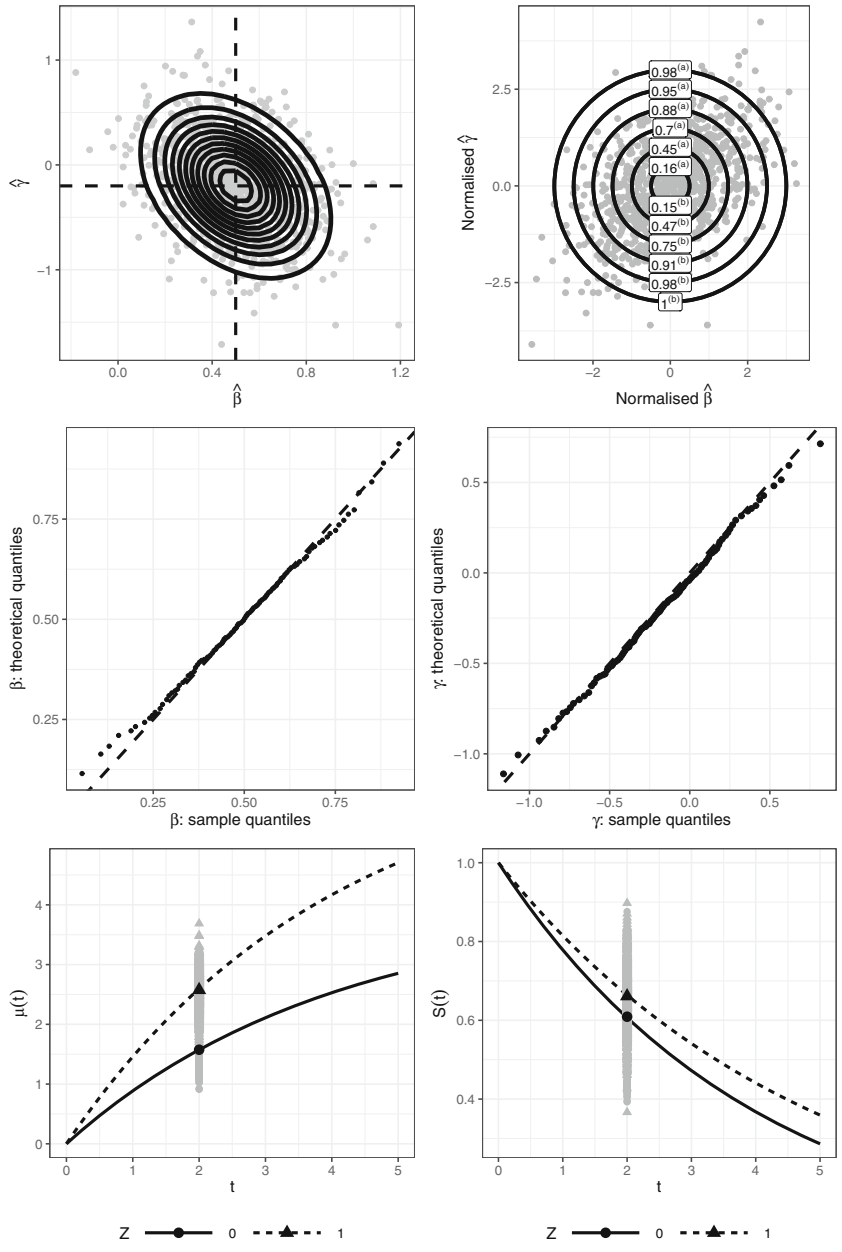


Fig. 9 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets and $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, -0.2, 1)$. The pseudo-observations are computed based on $k = 1$ with $t = 2$

$(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 0.75)$ and $t = (1, 2, 3)$

See Appendix Fig. 10.

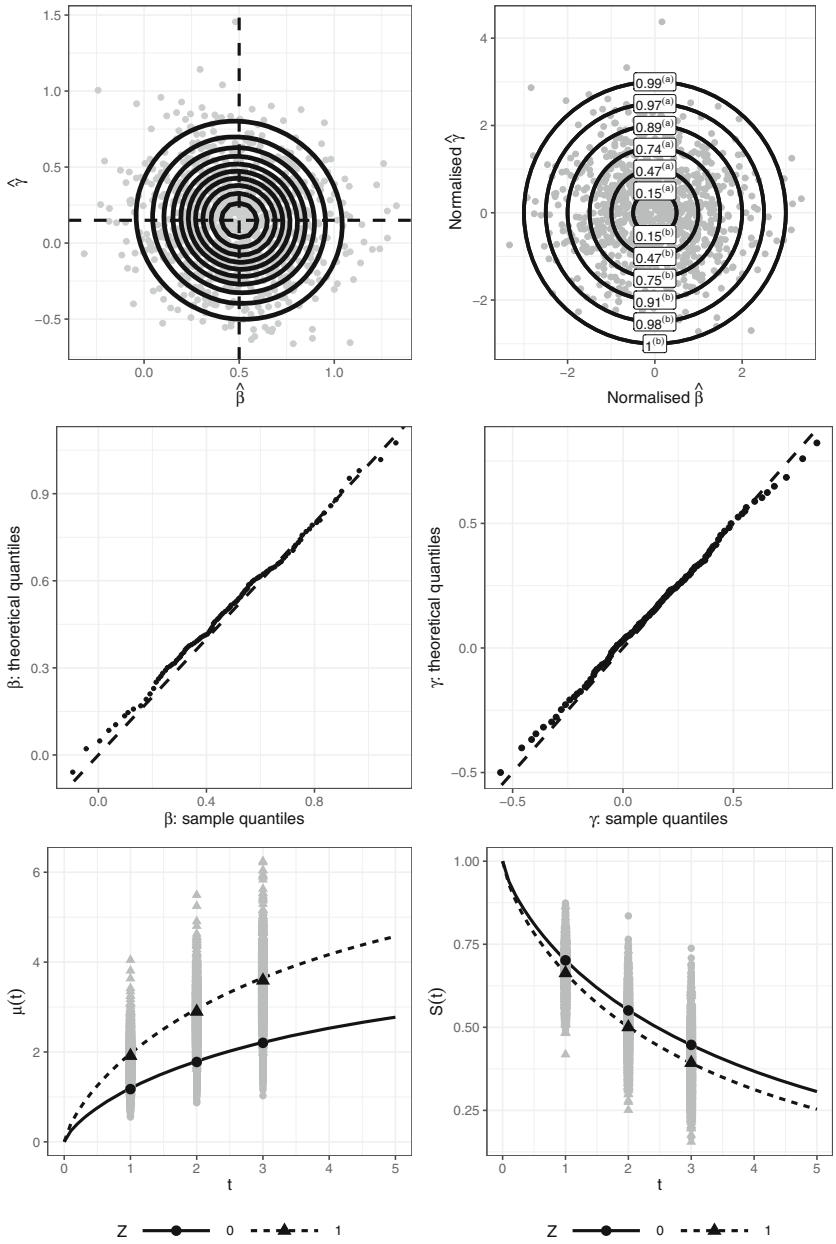


Fig. 10 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets and $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 0.75)$. The pseudo-observations are computed based on $k = 3$ with $t = (1, 2, 3)$

$(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 1)$ and $t = (1, 2, 3)$

See Appendix Fig. 11.

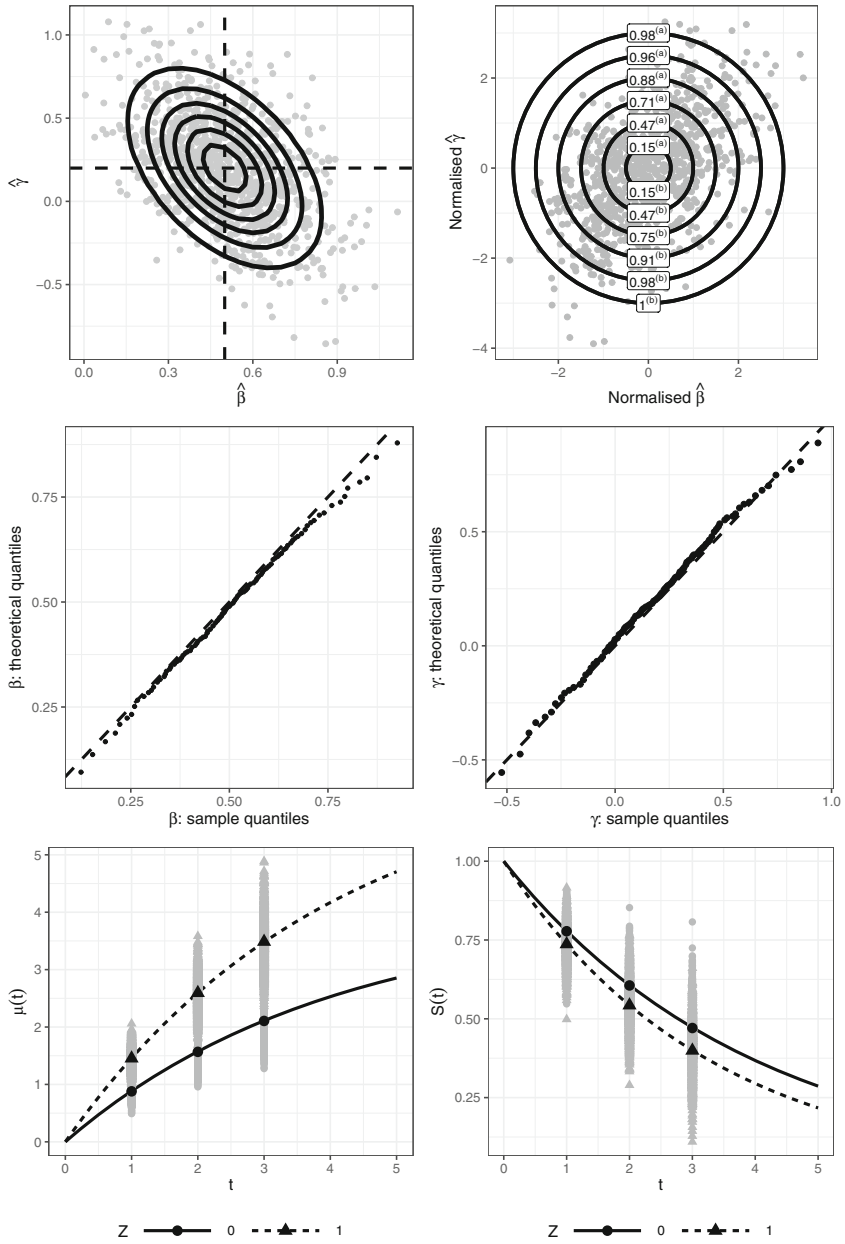


Fig. 11 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets and $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, 0.2, 1)$. The pseudo-observations are computed based on $k = 3$ with $t = (1, 2, 3)$

$(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, -0.2, 1)$ and $t = (1, 2, 3)$

See Appendix Fig. 12.

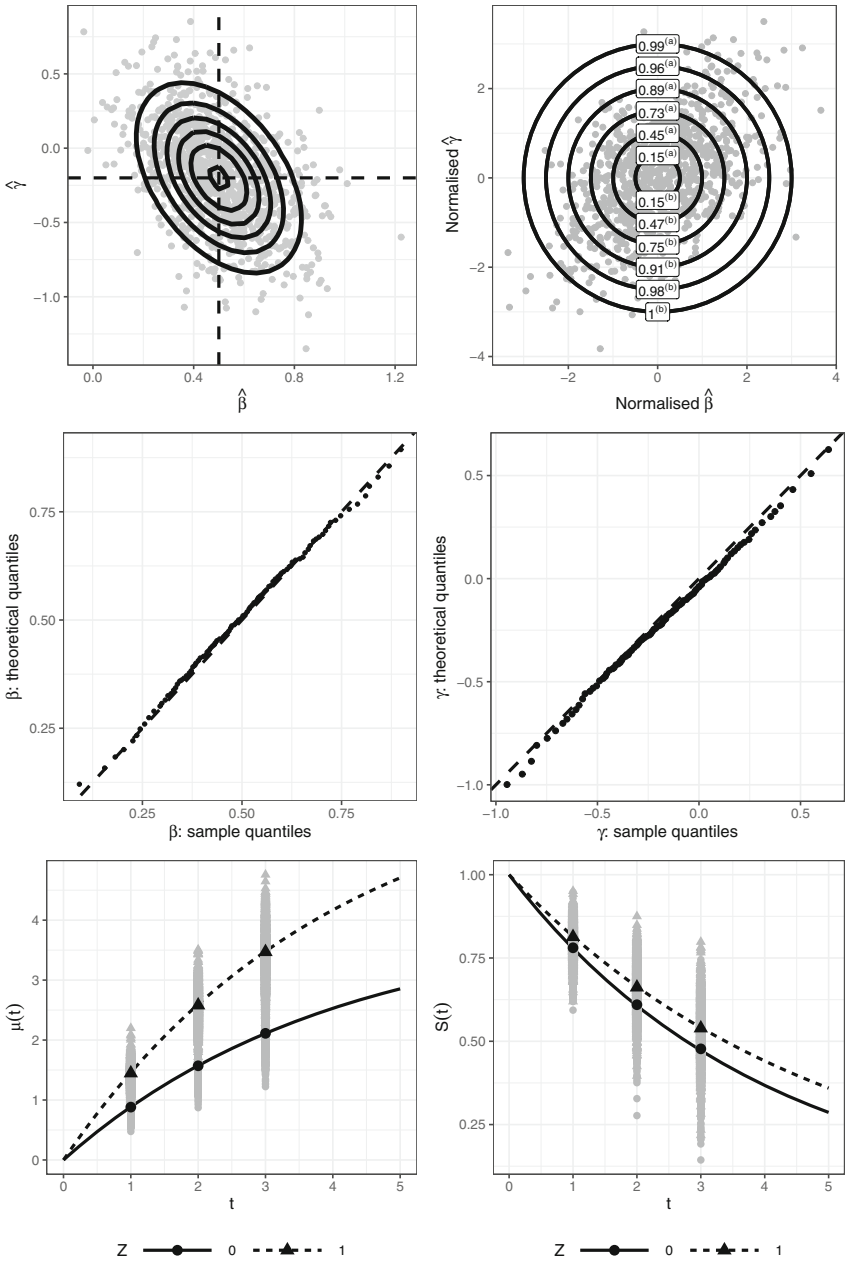


Fig. 12 Plots investigating normality assumption of $(\hat{\beta}, \hat{\gamma})$ using 1000 simulated data sets and $(n, \lambda_0^D, \beta, \gamma_D, \rho) = (100, 0.25, 0.5, -0.2, 1)$. The pseudo-observations are computed based on $k = 3$ with $t = (1, 2, 3)$

References

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10(4):1100–1120
Andersen PK, Perme MP (2010) Pseudo-observations in survival analysis. Stat Methods Med Res 19:71–99
Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer series in statistics. Springer
Andersen PK, Klein JP, Rosthøj S (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90:15–27
Andersen PK, Angst J, Ravn H (2019) Modeling marginal features in studies of recurrent events in the presence of a terminal event. Lifetime Data Anal 25(4):681–695
Binder N, Gerds TA, Andersen PK (2014) Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Anal 20(2):303–315
Byar D (1980) The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa. Springer
Cook R, Lawless JF (1997) Marginal analysis of recurrent events and a terminating event. Stat Med 16:911–924
Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B (Methodol) 34(2):187–220
Ghosh D, Lin D (2000) Nonparametric analysis of recurrent events and death. Biometrics 56:554–562
Ghosh D, Lin D (2002) Marginal regression models for recurrent and terminal events. Stat Sin 12:663–688
Graw F, Gerds TA, Schumacher M (2009) On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal 15:241–255
Jacobsen M, Martinussen T (2016) A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. Scand J Stat 43:845–862
Liang KY, Zeger ST (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22
Lin D, Wei L, Yang I, Ying Z (2000) Semiparametric regression for the mean and rate functions of recurrent events. J R Stat Soc 62(4):711–730
Liu L, Wolfe RA, Huang X (2004) Shared frailty models for recurrent events and a terminal event. Biometrics 60:747–756
Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JFE, Nauck MA, Nissen SE, Pocock S, Poulter NR, Ravn LS, Steinberg WM, Stockner M et al (2016) Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med 375(4):311–322
Overgaard M (2019) Counting processes in p-variation with application to recurrent events. https://arxiv.org/pdf/1903.04296.pdf
Overgaard M, Parner ET, Pedersen J (2017) Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. Ann Stat 45(5):1988–2015
Pavlič K, Martinussen T, Andersen PK (2019) Goodness of fit tests for estimating equations based on pseudo-observations. Lifetime Data Anal 25:189–205
Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc 84:1065–1073

Manuscript IIx

Title R-package 'recurrentpseudo': *Creates Pseudo-Observations and Analysis for Recurrent Event Data*

Authors Julie K. Furberg

Details Published on CRAN (September 2022)

URL https://cran.r-project.org/web/packages/recurrentpseudo/index.html

Package 'recurrentpseudo'

September 19, 2022

Title Creates Pseudo-Observations and Analysis for Recurrent Event Data

Version 1.0.0

Description Computation of one-, two- and three-dimensional pseudo-observations based on recurrent events and terminal events. Generalised linear models are fitted using generalised estimating equations. Technical details on the bivariate procedure can be found in ``Bivariate pseudo-observations for recurrent event analysis with terminal events'' (Furberg et al., 2021) <doi:10.1007/s10985-021-09533-5>.

License GPL (>= 2)

URL https://github.com/JulieKFurberg/recurrentpseudo

Encoding UTF-8

RoxygenNote 7.2.1

Imports survival, geepack, stringr, prodlim

Depends magrittr, dplyr

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation no

Author Julie Kjærulff Furberg [aut, cre] (<https://orcid.org/0000-0001-8785-1462>)

Maintainer Julie Kjærulff Furberg <julie_furberg@hotmail.com>

Repository CRAN

Date/Publication 2022-09-19 14:16:16 UTC

R topics documented:

`pseudo.geefit` *Function that makes GEE model fit for recurrent pseudo-observations*

Description

This function fits a GEE model based on pseudo-observations of the marginal mean function, and the survival probability or cumulative incidences of two death causes as returned by `pseudo.onedim()` (marginal mean function), or `pseudo.twodim()` (marginal mean function and survival probability), or `pseudo.threedim()` (marginal mean function and cumulative incidences of death causes 1 and 2)

Usage

```
pseudo.geefit(pseudodata, covar_names)
```

Arguments

`pseudodata`	Data set containing pseudo-observations. Expecting output from pseudo.twodim()
`covar_names`	Vector with covariate names to be found in "pseudodata". E.g. covar_names = c("Z", "Z1")

Value

An object of class `pseudo.geefit`.

- `xi` contains the estimated model parameters
- `sigma` contains the estimated variance matrix corresponding to `xi`

References

Furberg, J.K., Andersen, P.K., Korn, S. et al. Bivariate pseudo-observations for recurrent event analysis with terminal events. Lifetime Data Anal (2021). https://doi.org/10.1007/s10985-021-09533-5

Examples

```
# Bladder cancer data from survival package
require(survival)

# Make a three level status variable
bladder1$status3 <- ifelse(bladder1$status %in% c(2, 3), 2, bladder1$status)

# Add one extra day for the two patients with start=stop=0
# subset(bladder1, stop <= start)
bladder1[bladder1$id == 1, "stop"] <- 1
bladder1[bladder1$id == 49, "stop"] <- 1
```

```
# Restrict the data to placebo and thiotepa
bladdersub <- subset(bladder1, treatment %in% c("placebo", "thiotepa"))

# Make treatment variable two-level factor
bladdersub$Z <- as.factor(ifelse(bladdersub$treatment == "placebo", 0, 1))
levels(bladdersub$Z) <- c("placebo", "thiotepa")

head(bladdersub)

# Two-dimensional (bivariate pseudo-obs) model fit

# Computation of pseudo-observations
pseudo_bladder_2d <- pseudo.twodim(tstart = bladdersub$start,
                                   tstop = bladdersub$stop,
                                   status = bladdersub$status3,
                                   id = bladdersub$id,
                                   covar_names = "Z",
                                   tk = c(30),
                                   data = bladdersub)

# Data in wide format
head(pseudo_bladder_2d$outdata)

# Data in long format
head(pseudo_bladder_2d$outdata_long)

# GEE fit
fit_bladder_2d <- pseudo.geefit(pseudodata = pseudo_bladder_2d,
                                covar_names = c("Z"))
fit_bladder_2d
```

`pseudo.onedim` *Function that computes univariate pseudo-observations*

Description

This function computes univariate pseudo-observations of the marginal mean function (in the presence of terminal events)

Usage

```
pseudo.onedim(tstart, tstop, status, covar_names, id, tk, data)
```

Arguments

`tstart`	Start time - expecting counting process notation
`tstop`	Stop time - expecting counting process notation
`status`	Status variable (0 = censoring, 1 = recurrent event, 2 = death)
`covar_names`	Vector containing names of covariates intended for further analysis

id	ID variable for subject
tk	Vector of time points to calculate pseudo-observations at
data	Data set which contains variables of interest

Value

An object of class pseudo.onedim.

- outdata contains the semi-wide version of the computed pseudo-observations (one row per time, tk, per id).
- outdata_long contains the long version of the computed pseudo-observations (one row per observation, several per id).
- indata contains the input data which the pseudo-observations are based on.
- ts vector with time points used for computation of pseudo-observations.
- k number of time points used for computation of pseudo-observations (length(ts)).

References

Furberg, J.K., Andersen, P.K., Korn, S. et al. Bivariate pseudo-observations for recurrent event analysis with terminal events. Lifetime Data Anal (2021). https://doi.org/10.1007/s10985-021-09533-5

Examples

```
# Example: Bladder cancer data from survival package
require(survival)

# Make a three level status variable
bladder1$status3 <- ifelse(bladder1$status %in% c(2, 3), 2, bladder1$status)

# Add one extra day for the two patients with start=stop=0
# subset(bladder1, stop <= start)
bladder1[bladder1$id == 1, "stop"] <- 1
bladder1[bladder1$id == 49, "stop"] <- 1

# Restrict the data to placebo and thiotepa
bladdersub <- subset(bladder1, treatment %in% c("placebo", "thiotepa"))

# Make treatment variable two-level factor
bladdersub$Z <- as.factor(ifelse(bladdersub$treatment == "placebo", 0, 1))
levels(bladdersub$Z) <- c("placebo", "thiotepa")
head(bladdersub)

# Pseudo observations
pseudo_bladder_1d <- pseudo.onedim(tstart = bladdersub$start,
                                   tstop = bladdersub$stop,
                                   status = bladdersub$status3,
                                   id = bladdersub$id,
                                   covar_names = "Z",
                                   tk = c(30),
```

```
                                    data = bladdersub)
head(pseudo_bladder_1d$outdata)

# GEE fit
fit_bladder_1d <- pseudo.geefit(pseudodata = pseudo_bladder_1d,
                                covar_names = c("Z"))
fit_bladder_1d
```

`pseudo.threedim` *Function that computes 3-dim pseudo-observations*

Description

This function computes 3-dimensional pseudo-observations of the marginal mean function (in the presence of terminal events) and cumulative incidences of death causes 1 and 2

Usage

```
pseudo.threedim(tstart, tstop, status, covar_names, id, tk, data, deathtype)
```

Arguments

`tstart`	Start time - expecting counting process notation
`tstop`	Stop time - expecting counting process notation
`status`	Status variable (0 = censoring, 1 = recurrent event, 2 = death)
`covar_names`	Vector containing names of covariates intended for further analysis
`id`	ID variable for subject
`tk`	Vector of time points to calculate pseudo-observations at
`data`	Data set which contains variables of interest
`deathtype`	Type of death (cause 1 or cause 2)

Value

An object of class `pseudo.threedim`.

- `outdata` contains the semi-wide version of the computed pseudo-observations (one row per time, tk, per id).
- `outdata_long` contains the long version of the computed pseudo-observations (one row per observation, several per id).
- `indata` contains the input data which the pseudo-observations are based on.
- `ts` vector with time points used for computation of pseudo-observations.
- `k` number of time points used for computation of pseudo-observations (length(ts)).

References

Furberg, J.K., Andersen, P.K., Korn, S. et al. Bivariate pseudo-observations for recurrent event analysis with terminal events. Lifetime Data Anal (2021). https://doi.org/10.1007/s10985-021-09533-5

Examples

```
# Example: Bladder cancer data from survival package
require(survival)

# Make a three level status variable
bladder1$status3 <- ifelse(bladder1$status %in% c(2, 3), 2, bladder1$status)

# Add one extra day for the two patients with start=stop=0
# subset(bladder1, stop <= start)
bladder1[bladder1$id == 1, "stop"] <- 1
bladder1[bladder1$id == 49, "stop"] <- 1

# Restrict the data to placebo and thiotepa
bladdersub <- subset(bladder1, treatment %in% c("placebo", "thiotepa"))

# Make treatment variable two-level factor
bladdersub$Z <- as.factor(ifelse(bladdersub$treatment == "placebo", 0, 1))
levels(bladdersub$Z) <- c("placebo", "thiotepa")
head(bladdersub)

# Add deathtype variable to bladder data
# Deathtype = 1 (bladder disease death), deathtype = 2 (other death reason)
bladdersub$deathtype <- with(bladdersub, ifelse(status == 2, 1, ifelse(status == 3, 2, 0)))
table(bladdersub$deathtype, bladdersub$status)

# Pseudo-observations
pseudo_bladder_3d <- pseudo.threedim(tstart = bladdersub$start,
                                     tstop = bladdersub$stop,
                                     status = bladdersub$status3,
                                     id = bladdersub$id,
                                     deathtype = bladdersub$deathtype,
                                     covar_names = "Z",
                                     tk = c(30),
                                     data = bladdersub)
pseudo_bladder_3d

# GEE fit
fit_bladder_3d <- pseudo.geefit(pseudodata = pseudo_bladder_3d,
                                covar_names = c("Z"))
fit_bladder_3d
```

`pseudo.twodim` *Function that computes bivariate pseudo-observations*

Description

This function computes bivariate pseudo-observations of the marginal mean function (in the presence of terminal events) and the survival probability

Usage

```
pseudo.twodim(tstart, tstop, status, covar_names, id, tk, data)
```

Arguments

tstart	Start time - expecting counting process notation
tstop	Stop time - expecting counting process notation
status	Status variable (0 = censoring, 1 = recurrent event, 2 = death)
covar_names	Vector containing names of covariates intended for further analysis
id	ID variable for subject
tk	Vector of time points to calculate pseudo-observations at
data	Data set which contains variables of interest

Value

An object of class pseudo.twodim.

- outdata contains the semi-wide version of the computed pseudo-observations (one row per time, tk, per id).
- outdata_long contains the long version of the computed pseudo-observations (one row per observation, several per id).
- indata contains the input data which the pseudo-observations are based on.
- ts vector with time points used for computation of pseudo-observations.
- k number of time points used for computation of pseudo-observations (length(ts)).

References

Furberg, J.K., Andersen, P.K., Korn, S. et al. Bivariate pseudo-observations for recurrent event analysis with terminal events. Lifetime Data Anal (2021). https://doi.org/10.1007/s10985-021-09533-5

Examples

```
# Example: Bladder cancer data from survival package
require(survival)

# Make a three level status variable
bladder1$status3 <- ifelse(bladder1$status %in% c(2, 3), 2, bladder1$status)

# Add one extra day for the two patients with start=stop=0
# subset(bladder1, stop <= start)
bladder1[bladder1$id == 1, "stop"] <- 1
```

```
bladder1[bladder1$id == 49, "stop"] <- 1

# Restrict the data to placebo and thiotepa
bladdersub <- subset(bladder1, treatment %in% c("placebo", "thiotepa"))

# Make treatment variable two-level factor
bladdersub$Z <- as.factor(ifelse(bladdersub$treatment == "placebo", 0, 1))
levels(bladdersub$Z) <- c("placebo", "thiotepa")
head(bladdersub)

# Pseudo observations
pseudo_bladder_2d <- pseudo.twodim(tstart = bladdersub$start,
                                   tstop = bladdersub$stop,
                                   status = bladdersub$status3,
                                   id = bladdersub$id,
                                   covar_names = "Z",
                                   tk = c(30),
                                   data = bladdersub)
head(pseudo_bladder_2d$outdata)

# GEE fit
fit_bladder_2d <- pseudo.geefit(pseudodata = pseudo_bladder_2d,
                                covar_names = c("Z"))
fit_bladder_2d
```

Index

Vignette for recurrentpseudo

Title Vignette for R-package 'recurrentpseudo': *recurrentpseudo: An R package for analysing recurrent events in the presence of terminal events using pseudo-observations*

Authors Julie K. Furberg

URL `https://github.com/JulieKFurberg/recurrentpseudo`

recurrentpseudo: An R package for analysing recurrent events in the presence of terminal events using pseudo-observations

Introduction

This package computes pseudo-observations for recurrent event data in the presence of terminal events. Three versions exist: One-dimensional, two-dimensional or three-dimensional pseudo-observations.

Following the computation of pseudo-observations, the marginal mean function, survival probability and/or cumulative incidences can be modelled using generalised estimating equations.

See Furberg et al. (*Bivariate pseudo-observations for recurrent event analysis with terminal events (2023)*) for technical details on the procedure.

Package and main functions

```
# Development version

#require(devtools)
#devtools::install_github("JulieKFurberg/recurrentpseudo", force = TRUE)

require(recurrentpseudo)


# Main functions

#?pseudo.onedim
#?pseudo.twodim
#?pseudo.threedim
#?pseudo.geefit
```

Notation

Let D^* denote the survival time and let $N^*(t)$ denote the number of recurrent events by time t. Let C denote the time of censoring. Due to right-censoring, the data consists of $X = \{N(\cdot), D, \delta, Z\}$ where $N(t) = N^*(t \wedge C)$, $D = D^* \wedge C$, $\delta = I\left(D^* \leq C\right)$ and Z denotes p baseline covariates.

We observe $X_i = \{N_i(\cdot), D_i, \delta_i, Z_i\}$ for each individual $i = 1, \ldots, n$.

We consider the marginal mean function, $\mu(t)$, given by

$$\mu(t) = E(N^*(t)) = \int_0^t S(u^-)\, dR(u), \quad dR(t) = E(dN^*(t) \mid D^* \geq t)$$

and the survival probability, $S(t)$, given by

$$S(t) = P(D^* > t).$$

Moreover, we consider the cumulative incidences for death causes 1, $C_1(t)$, and 2, $C_2(t)$

$$C_1(t) = E(I(D^* \leq t, \Delta = 1)), \quad C_2(t) = E(I(D^* \leq t, \Delta = 2))$$

where $\Delta = \{1, 2\}$ represents a cause-of-death indicator.

Introduction to pseudo-observations

The following section serves as a fast introduction to pseudo-observations, which the methods of this package is based on.

For more detailed information, please see

- Andersen and Perme (*Pseudo-observations in survival analysis (2010)*) or
- Andersen, Klein and Rosthøj (*Generalised linear models for correlated pseudo-observations, with applications to multi-state models (2003)*)

We wish to formulate a model for

$$\theta = E(f(X))$$

where $X = X_1, \ldots, X_n$ denotes a vector of survival times (or other survival data) for n individuals and f denotes some function. An example would be $\theta = E(I(D^* > t)) = P(D^* > t)$.

Assume that a sufficiently nice estimator $\hat{\theta}$ of θ exists. For a fixed time, $t \in [0, \tau]$, the pseudo-observation for the i'th individual at t is given by

$$\hat{\theta}_i(t) = n \cdot \hat{\theta}(t) - (n-1) \cdot \hat{\theta}^{-i}(t)$$

where $\hat{\theta}(t)$ denotes the estimate based on the total data set, and $\hat{\theta}^{-i}(t)$ denotes the estimate based on the same data set but omitting observations from individual i.

Since the survival times are subject to right-censoring, standard inference on survival data is adjusted to accommodate this, e.g. in likelihood estimation.

However, since all subjects has a valid pseudo-observation, $\hat{\theta}_i(t)$, at one or more times, these can be used as an outcome variable in a generalised linear model. Note, that this is regardless of the whether a subject is alive, censored or died at time t.

Assume that g denotes a link function, then we wish to fit

$$g(E(f(X) \mid Z)) = \xi^T Z.$$

Following, $f(X)$ is replaced by $\hat{\theta}_i(\cdot)$ in the model fit.

The model parameters, ξ, are estimated using generalised estimating equations (GEE), see Liang and Zeger (*Longitudinal data analysis using generalized linear models (1986)*).

The GEE procedure accommodates the fact that each individual can have several (pseudo-)observations.

One-dimensional pseudo-observations

The one-dimensional pseudo-observations model is based on the parameter $\theta = \mu(t)$, which is estimated by

$$\hat{\theta} = \hat{\mu}(t) = \int_0^t \hat{S}(u^-)\, d\hat{R}(u),$$

where $\hat{S}(t)$ denotes the Kaplan-Meier estimator of $S(t)$ and $\hat{R}(t)$ denotes the Nelson-Aalen estimator of $R(t)$.

We assume that

$$\log(\mu(t \mid Z)) = \log(\mu_0(t)) + \beta^T Z.$$

Two-dimensional pseudo-observations

The two-dimensional pseudo-observations model is based on the parameter $\theta = (\mu(t), S(t))$, which is estimated by

$$\hat{\theta} = \begin{pmatrix} \hat{\mu}(t) \\ \hat{S}(t) \end{pmatrix}.$$

We assume that

$$\begin{pmatrix} \log(\mu(t \mid Z)) \\ \text{cloglog}\left(S(t \mid Z)\right) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta^T Z \\ \log(\Lambda_0(t)) + \gamma^T Z \end{pmatrix}.$$

Three-dimensional pseudo-observations

The three-dimensional pseudo-observations model is based on the parameter $\theta = (\mu(t), C_1(t), C_2(t))$, which is estimated by

$$\hat{\theta} = \begin{pmatrix} \hat{\mu}(t) \\ \hat{C}_1(t) \\ \hat{C}_2(t) \end{pmatrix}$$

where $\hat{C}_1(t)$ and $\hat{C}_2(t)$ are the Aalen-Johansen estimates of the cumulative incidences for causes 1, $C_1(t)$, and 2, $C_2(t)$, respectively.

We assume that

$$\begin{pmatrix} \log(\mu(t \mid Z)) \\ \text{cloglog}\left(1 - C_1(t \mid Z)\right) \\ \text{cloglog}\left(1 - C_2(t \mid Z)\right) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta^T Z \\ \log(\Lambda_{10}(t)) + {\gamma_1}^T Z \\ \log(\Lambda_{20}(t)) + {\gamma_2}^T Z \end{pmatrix}.$$

Example - Bladder cancer data from survival package

The functions in `recurrentpseudo` will be exemplified using the well-known bladder cancer data from the survival package. This data set considers data from a clinical cancer trial conducted by the Veterans Administration Cooperative Urological Research Group (Byar: *The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa* (1980)) Here, 118 patients with stage I bladder cancer were randomised to receive placebo, pyridoxine or thiotepa. After randomisation, information on occurrences of superficial bladder tumours and any deaths were collected.

We focus on the comparison between placebo and thiotepa ($n = 86$ in total). We model recurrent bladder tumours, and adjust for death (cause 1: bladder cancer disease death, cause 2: other causes).

One-, two- and three-dimensional pseudo-observations are computed based on a single time point, $t = 30$ months.

For the comparison between placebo and thiotepa on recurrent bladder tumours, the effect measure of interest is the mean ratio $\exp(\beta)$.

```
# Example: Bladder cancer data from survival package
require(survival)
#> Indlæser krævet pakke: survival
```

```r
# Make a three level status variable
bladder1$status3 <- ifelse(bladder1$status %in% c(2, 3), 2, bladder1$status)

# Add one extra day for the two patients with start=stop=0
# subset(bladder1, stop <= start)
bladder1[bladder1$id == 1, "stop"] <- 1
bladder1[bladder1$id == 49, "stop"] <- 1

# Restrict the data to placebo and thiotepa
bladdersub <- subset(bladder1, treatment %in% c("placebo", "thiotepa"))

# Make treatment variable two-level factor
bladdersub$Z <- as.factor(ifelse(bladdersub$treatment == "placebo", 0, 1))
levels(bladdersub$Z) <- c("placebo", "thiotepa")
head(bladdersub)
#>   id treatment number size recur start stop status rtumor rsize enum status3
#> 1  1   placebo      1    1     0     0    1      3      .     .    1       2
#> 2  2   placebo      1    3     0     0    1      3      .     .    1       2
#> 3  3   placebo      2    1     0     0    4      0      .     .    1       0
#> 4  4   placebo      1    1     0     0    7      0      .     .    1       0
#> 5  5   placebo      5    1     0     0   10      3      .     .    1       2
#> 6  6   placebo      4    1     1     0    6      1      1     1    1       1
#>         Z
#> 1 placebo
#> 2 placebo
#> 3 placebo
#> 4 placebo
#> 5 placebo
#> 6 placebo
```

We fit the univariate pseudo-observation model using the binary treatment indicator as covariate, i.e. we model

$$\log(\mu(t \mid Z)) = \log(\mu_0(t)) + \beta Z$$

One-dimensional pseudo-observations and GEE fit can be computed using the following code,

```r
# Pseudo observations at t = 30
pseudo_bladder_1d <- pseudo.onedim(tstart = bladdersub$start,
                                   tstop = bladdersub$stop,
                                   status = bladdersub$status3,
                                   id = bladdersub$id,
                                   covar_names = "Z",
                                   tk = c(30),
                                   data = bladdersub)
head(pseudo_bladder_1d$outdata)
#>              mu k ts id       Z
#> 1 -0.0004269178 1 30  1 placebo
#> 2 -0.0004269178 1 30  2 placebo
#> 3  1.2359654463 1 30  3 placebo
#> 4  1.0739859010 1 30  4 placebo
#> 5 -0.0958639918 1 30  5 placebo
#> 6  1.0122441163 1 30  6 placebo

# GEE fit
fit_bladder_1d <- pseudo.geefit(pseudodata = pseudo_bladder_1d,
                                covar_names = c("Z"))
fit_bladder_1d
```

```
#> $xi
#>
#> (Intercept)  0.5590869
#> Zthiotepa   -0.4359054
#>
#> $sigma
#>             (Intercept)   Zthiotepa
#> (Intercept)  0.02662095 -0.02662095
#> Zthiotepa   -0.02662095  0.07934314
#>
#> attr(,"class")
#> [1] "pseudo.geefit"

# Treatment differences
xi_diff_1d <- as.matrix(c(fit_bladder_1d$xi[2]), ncol = 1)

mslabels <- c("treat, mu")
rownames(xi_diff_1d) <- mslabels
colnames(xi_diff_1d) <- ""
xi_diff_1d
#>
#> treat, mu -0.4359054

# Variance matrix for differences
sigma_diff_1d <- matrix(c(fit_bladder_1d$sigma[2,2]),
                        ncol = 1, nrow = 1,
                        byrow = T)

rownames(sigma_diff_1d) <- colnames(sigma_diff_1d) <- mslabels
sigma_diff_1d
#>            treat, mu
#> treat, mu 0.07934314
```

Thus, the estimated mean ratio is $\exp(\hat{\beta}) = 0.6466789$ (standard error and confidence intervals can be found using the Delta method).

Alternatively, the bivariate pseudo-observation model using the binary treatment indicator as covariate can be fitted, i.e.

$$\begin{pmatrix} \log(\mu(t \mid Z)) \\ \text{cloglog}\left(S(t \mid Z)\right) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta Z \\ \log(\Lambda_0(t)) + \gamma Z \end{pmatrix}$$

Two-dimensional pseudo-observations and GEE fit can be computed using the following code

```
# Pseudo observations at t = 30
pseudo_bladder_2d <- pseudo.twodim(tstart = bladdersub$start,
                                   tstop = bladdersub$stop,
                                   status = bladdersub$status3,
                                   id = bladdersub$id,
                                   covar_names = "Z",
                                   tk = c(30),
                                   data = bladdersub)
head(pseudo_bladder_2d$outdata)
#>              mu          surv k ts id       Z
#> 1 -0.0004269178  1.421085e-14 1 30  1 placebo
#> 2 -0.0004269178  1.421085e-14 1 30  2 placebo
#> 3  1.2359654463  8.170875e-01 1 30  3 placebo
#> 4  1.0739859010  8.170875e-01 1 30  4 placebo
```

```
#> 5 -0.0958639918 -5.305763e-02 1 30  5 placebo
#> 6  1.0122441163 -5.305763e-02 1 30  6 placebo

# GEE fit
fit_bladder_2d <- pseudo.geefit(pseudodata = pseudo_bladder_2d,
                                covar_names = c("Z"))
fit_bladder_2d
#> $xi
#>
#> esttypemu              0.55908687
#> esttypemu:Zthiotepa   -0.43590539
#> esttypesurv           -1.41652478
#> esttypesurv:Zthiotepa -0.04800778
#>
#> $sigma
#>                          esttypemu esttypemu:Zthiotepa  esttypesurv
#> esttypemu              0.026620952        -0.026620952 -0.003481085
#> esttypemu:Zthiotepa   -0.026620952         0.079343139  0.003481085
#> esttypesurv           -0.003481085         0.003481085  0.123251791
#> esttypesurv:Zthiotepa  0.003481085         0.002758847 -0.123251791
#>                       esttypesurv:Zthiotepa
#> esttypemu                      0.003481085
#> esttypemu:Zthiotepa            0.002758847
#> esttypesurv                   -0.123251791
#> esttypesurv:Zthiotepa          0.260915569
#>
#> attr(,"class")
#> [1] "pseudo.geefit"

# Treatment differences
xi_diff_2d <- as.matrix(c(fit_bladder_2d$xi[2],
                          fit_bladder_2d$xi[4]), ncol = 1)

mslabels <- c("treat, mu", "treat, surv")
rownames(xi_diff_2d) <- mslabels
colnames(xi_diff_2d) <- ""
xi_diff_2d
#>
#> treat, mu   -0.43590539
#> treat, surv -0.04800778


# Variance matrix for differences
sigma_diff_2d <- matrix(c(fit_bladder_2d$sigma[2,2],
                          fit_bladder_2d$sigma[2,4],
                          fit_bladder_2d$sigma[2,4],
                          fit_bladder_2d$sigma[4,4]),
                          ncol = 2, nrow = 2,
                          byrow = T)

rownames(sigma_diff_2d) <- colnames(sigma_diff_2d) <- mslabels
sigma_diff_2d
#>               treat, mu treat, surv
#> treat, mu   0.079343139 0.002758847
#> treat, surv 0.002758847 0.260915569
```

Finally, one could fit the three-dimensional pseudo-observation model to the bladder cancer data.

Three-dimensional pseudo-observations and GEE fit can be computed using the following code

```
# Add deathtype variable to bladder data
# Deathtype = 1 (bladder disease death), deathtype = 2 (other death reason)
bladdersub$deathtype <- with(bladdersub, ifelse(status == 2, 1, ifelse(status == 3, 2, 0)))
table(bladdersub$deathtype, bladdersub$status)
#>
#>       0   1   2   3
#>   0  55 132   0   0
#>   1   0   0   2   0
#>   2   0   0   0  20

# Pseudo-observations
pseudo_bladder_3d <- pseudo.threedim(tstart = bladdersub$start,
                                     tstop = bladdersub$stop,
                                     status = bladdersub$status3,
                                     id = bladdersub$id,
                                     deathtype = bladdersub$deathtype,
                                     covar_names = "Z",
                                     tk = c(30),
                                     data = bladdersub)
head(pseudo_bladder_3d$outdata_long)
#>   k ts id esttype             y       Z
#> 1 1 30  1      mu -4.269178e-04 placebo
#> 2 1 30  1    surv  1.421085e-14 placebo
#> 3 1 30  1    cif1  0.000000e+00 placebo
#> 4 1 30  1    cif2  1.000000e+00 placebo
#> 5 1 30  2      mu -4.269178e-04 placebo
#> 6 1 30  2    surv  1.421085e-14 placebo

# GEE fit
fit_bladder_3d <- pseudo.geefit(pseudodata = pseudo_bladder_3d,
                                covar_names = c("Z"))
fit_bladder_3d
#> $xi
#>
#> esttypemu              0.5590869
#> esttypemu:Zthiotepa   -0.4359054
#> esttypecif1           -3.7618319
#> esttypecif1:Zthiotepa  0.2930357
#> esttypecif2           -1.5431978
#> esttypecif2:Zthiotepa -0.1005109
#>
#> $sigma
#>                          esttypemu esttypemu:Zthiotepa esttypecif1
#> esttypemu              0.026620952        -0.026620952  0.01663610
#> esttypemu:Zthiotepa   -0.026620952         0.079343139 -0.01663610
#> esttypecif1            0.016636098        -0.016636098  1.07839851
#> esttypecif1:Zthiotepa -0.016636098         0.013359996 -1.07839851
#> esttypecif2           -0.006027688         0.006027688 -0.02642283
#> esttypecif2:Zthiotepa  0.006027688         0.001779996  0.02642283
#>                       esttypecif1:Zthiotepa  esttypecif2 esttypecif2:Zthiotepa
#> esttypemu                       -0.01663610 -0.006027688           0.006027688
#> esttypemu:Zthiotepa              0.01336000  0.006027688           0.001779996
#> esttypecif1                     -1.07839851 -0.026422825           0.026422825
#> esttypecif1:Zthiotepa            2.01305239  0.026422825          -0.057715255
#> esttypecif2                      0.02642283  0.138167379          -0.138167379
```

```
#> esttypecif2:Zthiotepa           -0.05771525 -0.138167379           0.299045959
#>
#> attr(,"class")
#> [1] "pseudo.geefit"

# Treatment differences
xi_diff_3d <- as.matrix(c(fit_bladder_3d$xi[2],
                          fit_bladder_3d$xi[4],
                          fit_bladder_3d$xi[6]), ncol = 1)

mslabels <- c("treat, mu", "treat, cif1", "treat, cif2")
rownames(xi_diff_3d) <- mslabels
colnames(xi_diff_3d) <- ""
xi_diff_3d
#>
#> treat, mu   -0.4359054
#> treat, cif1  0.2930357
#> treat, cif2 -0.1005109


# Variance matrix for differences
sigma_diff_3d <- matrix(c(fit_bladder_3d$sigma[2,2],
                          fit_bladder_3d$sigma[2,4],
                          fit_bladder_3d$sigma[2,6],

                          fit_bladder_3d$sigma[2,4],
                          fit_bladder_3d$sigma[4,4],
                          fit_bladder_3d$sigma[4,6],

                          fit_bladder_3d$sigma[2,6],
                          fit_bladder_3d$sigma[4,6],
                          fit_bladder_3d$sigma[6,6]

                          ),
                        ncol = 3, nrow = 3,
                        byrow = T)

rownames(sigma_diff_3d) <- colnames(sigma_diff_3d) <- mslabels
sigma_diff_3d
#>               treat, mu treat, cif1  treat, cif2
#> treat, mu   0.079343139  0.01336000  0.001779996
#> treat, cif1 0.013359996  2.01305239 -0.057715255
#> treat, cif2 0.001779996 -0.05771525  0.299045959
```

We can compare the three model fits. Note, that the μ components match each other.

```
# Compare - should match for mu elements
xi_diff_1d
#>
#> treat, mu -0.4359054
xi_diff_2d
#>
#> treat, mu   -0.43590539
#> treat, surv -0.04800778
xi_diff_3d
#>
#> treat, mu   -0.4359054
```

```
#> treat, cif1  0.2930357
#> treat, cif2 -0.1005109

sigma_diff_1d
#>            treat, mu
#> treat, mu 0.07934314
sigma_diff_2d
#>               treat, mu treat, surv
#> treat, mu   0.079343139 0.002758847
#> treat, surv 0.002758847 0.260915569
sigma_diff_3d
#>               treat, mu treat, cif1  treat, cif2
#> treat, mu   0.079343139  0.01336000  0.001779996
#> treat, cif1 0.013359996  2.01305239 -0.057715255
#> treat, cif2 0.001779996 -0.05771525  0.299045959
```

More covariates

Assume that we wish to add extra baseline covariates to the model fit. For the sake of illustration, we have simulated a continuous covariate, Z_2, and a categorical covariate, Z_3. The covariate Z_1 corresponds to the binary treatment covariate ($Z = 1$ is thiotepa and $Z = 0$ is placebo). In order to make estimation for these models possible, the pseudo-observations are calculated at three time points, namely $t = 20, 30, 40$ months.

For the one-dimensional model for μ it holds that,

$$\log(\mu(t \mid Z)) = \log(\mu_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3.$$

This can be fitted using the below code,

```
require(dplyr)

## One-dim
# A binary variable, Z1_
# A continuous variable, Z2_
# A categorical variable, Z3_
set.seed(0308)

bladdersub <- as.data.frame(
  bladdersub %>% group_by(id) %>%
  mutate(Z1_ = Z,
         Z2_ = rnorm(1, mean = 3, sd = 1),
         Z3_ = sample(x = c("A", "B", "C"),
                      size = 1, replace = TRUE,
                      prob = c(1/4, 1/2, 1/4))
         ))
# head(bladdersub, 20)

# Make pseudo obs at more timepoints (more data)
# Pseudo observations at t = 20, 30, 40
pseudo_bladder_1d_3t <- pseudo.onedim(tstart = bladdersub$start,
                                      tstop = bladdersub$stop,
                                      status = bladdersub$status3,
                                      id = bladdersub$id,
                                      covar_names = c("Z1_", "Z2_", "Z3_"),
                                      tk = c(20, 30, 40),
                                      data = bladdersub)
```

```
fit1 <- pseudo.geefit(pseudodata = pseudo_bladder_1d_3t,
                      covar_names = c("Z1_", "Z2_", "Z3_"))

fit1$xi
#>
#> (Intercept)  0.39412273
#> Ztime30      0.42336922
#> Ztime40      0.59966454
#> Z1_thiotepa -0.29479824
#> Z2_         -0.08253287
#> Z3_B         0.01965615
#> Z3_C        -0.38332247
fit1$sigma
#>              (Intercept)       Ztime30       Ztime40  Z1_thiotepa           Z2_
#> (Intercept)  0.136673486 -0.0061702979 -0.0109943095 -0.036100986 -0.0247560186
#> Ztime30     -0.006170298  0.0046658382  0.0061013009  0.001167246  0.0002746085
#> Ztime40     -0.010994310  0.0061013009  0.0100442253  0.005504053 -0.0004015364
#> Z1_thiotepa -0.036100986  0.0011672458  0.0055040527  0.090063224 -0.0078557067
#> Z2_         -0.024756019  0.0002746085 -0.0004015364 -0.007855707  0.0109918233
#> Z3_B        -0.055441417  0.0005892506  0.0051883496  0.037829933 -0.0035769385
#> Z3_C        -0.041748863  0.0075014896  0.0141048652  0.023020550 -0.0082026411
#>                      Z3_B         Z3_C
#> (Intercept) -0.0554414168 -0.041748863
#> Ztime30      0.0005892506  0.007501490
#> Ztime40      0.0051883496  0.014104865
#> Z1_thiotepa  0.0378299330  0.023020550
#> Z2_         -0.0035769385 -0.008202641
#> Z3_B         0.0869004239  0.050773663
#> Z3_C         0.0507736625  0.200303239

fit1$xi[4]
#> [1] -0.2947982
```

Or for two-dimensional pseudo-observations, it holds that

$$\begin{pmatrix} \log(\mu(t \mid Z)) \\ \operatorname{cloglog}\left(S(t \mid Z)\right) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \\ \log(\Lambda_0(t)) + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 \end{pmatrix}.$$

Or for three-dimensional pseudo-observations, it holds that

$$\begin{pmatrix} \log(\mu(t \mid Z)) \\ \operatorname{cloglog}\left(1 - C_1(t \mid Z)\right) \\ \operatorname{cloglog}\left(1 - C_2(t \mid Z)\right) \end{pmatrix} = \begin{pmatrix} \log(\mu_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \\ \log(\Lambda_{10}(t)) + \gamma_{11} Z_1 + \gamma_{12} Z_2 + \gamma_{13} Z_3 \\ \log(\Lambda_{20}(t)) + \gamma_{21} Z_1 + \gamma_{22} Z_2 + \gamma_{23} Z_3 \end{pmatrix}.$$

These two models are fitted using the below code,

```
## Two-dim
# Pseudo observations at t = 20, 30, 40
pseudo_bladder_2d_3t <- pseudo.twodim(tstart = bladdersub$start,
                                      tstop = bladdersub$stop,
                                      status = bladdersub$status3,
                                      id = bladdersub$id,
                                      covar_names = c("Z1_", "Z2_", "Z3_"),
                                      tk = c(20, 30, 40),
                                      data = bladdersub)
```

```r
fit2 <- pseudo.geefit(pseudodata = pseudo_bladder_2d_3t,
                      covar_names = c("Z1_", "Z2_", "Z3_"))

# fit2$xi
# fit2$sigma


## Three-dim
pseudo_bladder_3d_3t <- pseudo.threedim(tstart = bladdersub$start,
                                        tstop = bladdersub$stop,
                                        status = bladdersub$status3,
                                        id = bladdersub$id,
                                        covar_names = c("Z1_", "Z2_", "Z3_"),
                                        deathtype = bladdersub$deathtype,
                                        tk = c(20, 30, 40),
                                        data = bladdersub)

fit3 <- pseudo.geefit(pseudodata = pseudo_bladder_3d_3t,
                      covar_names = c("Z1_", "Z2_", "Z3_"))

# fit3$xi
# fit3$sigma

## Compare for mu
fit1$xi[4]
#> [1] -0.2947982
fit2$xi[4]
#> [1] -0.2947982
fit3$xi[4]
#> [1] -0.2948043
```

Citation

To cite the `recurrentpseudo` package please use the following reference,

> Julie K. Furberg, Per K. Andersen, Sofie Korn, Morten Overgaard, Henrik Ravn: Bivariate pseudo-observations for recurrent event analysis with terminal events (Lifetime Data Analysis, 2023)

Manuscript III

Title *Simulation-based sample size calculations for recurrent events with competing risks*

Authors Julie K. Furberg, Per K. Andersen, Thomas Scheike and Henrik Ravn

Details Under revision for Pharmaceutical Statistics (March 2023)

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

MAIN PAPER

Simulation-based sample size calculations for recurrent events with competing risks

Julie Kjærulff Furberg*[1,2] | Per Kragh Andersen[1] | Thomas Scheike[1] | Henrik Ravn[2,3]

[1]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark
[2]Biostatistics OSCD & Outcomes 1, Novo Nordisk, Bagsværd, Denmark
[3]Methods, Innovation & Outreach (MIO), Novo Nordisk, Bagsværd, Denmark

Correspondence
*Julie Kjærulff Furberg, Novo Nordisk A/S, Vandtårnsvej 114, 2860 Søborg, Denmark.
Email: jukf@novonordisk.com

Abstract

In randomised controlled trials, the outcome of interest could be recurrent events, such as hospitalisations for heart failure. If mortality rates are non-negligible, both recurrent events and competing terminal events need to be addressed when formulating the estimand and statistical analysis is no longer trivial. In order to design future trials with primary recurrent event endpoints with competing risks, it is necessary to be able to perform power calculations to determine sample sizes. This paper introduces a simulation-based approach for power estimation based on a proportional means model for recurrent events and a proportional hazards model for terminal events. The simulation procedure is presented along with a discussion of what the user needs to specify to use the approach. Data from a randomised controlled trial, LEADER, is used as the basis for generating data for a future trial. Finally, potential power gains of recurrent event methods as opposed to first event methods are discussed.

KEYWORDS:
Sample size calculation, recurrent events, competing risks, treatment effects, randomised controlled trials

1 | INTRODUCTION

The application of methods from life history analysis is common and encountered in various fields of medical research, for instance in clinical trials within diseases, such as cancer, heart failure and diabetes. For randomised controlled trials (RCTs), the aim could be to compare two treatments and their effect on a survival outcome, e.g. an analysis of time-to-death comparing two treatments. Within recent years, the analysis of recurrent events has gained more focus. Recurrent events, as opposed to terminal events, are events that can happen several times for an individual over the course of their life, e.g. hospitalisation for heart failure. Methods and tools for sample size calculation for time-to-event outcomes are available using standard software, however, the same does not hold for recurrent event analyses. Several estimands may be targeted in estimation depending on the scientific question of interest.[?] [1] This, in turn, impacts the feasibility of conducting a sample size calculation in the recurrent event setting. At times, the recurrent event process of interest can be stopped by the occurrence of a terminal event (see Figure 1). When such a transition to a final state is possible, attention should be paid to the proposed analysis of recurrent events, similar to the extra care that needs to be taken when conducting an analysis of time-to-first events in the presence of competing risks.[2]

For this paper, the goal is to estimate a sample size for an RCT to detect a difference between two treatments on a recurrent event endpoint in the presence of competing deaths. Here, the aim is to compare two treatments in terms of their effect on the expected number of events. This can be achieved by modelling the marginal mean function, $\mu(t)$ and the estimand considered in this paper concentrates on the ratio of the expected number of recurrent events between two treatments for all t, a parameter

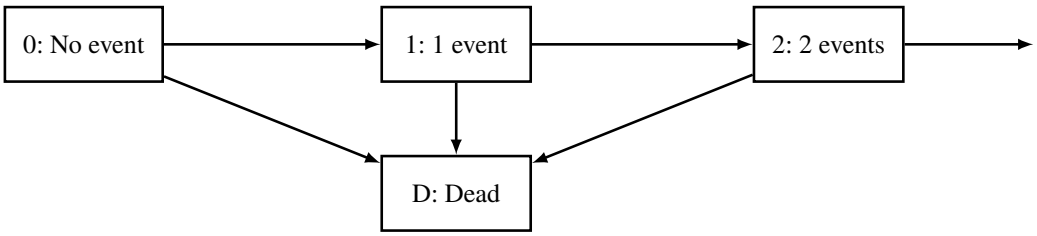


FIGURE 1 A recurrent event process with a terminal event.

that may be estimated using the proportional means model suggested by Ghosh and Lin.[3] Moreover, a Cox model is assumed for the hazard of death.[4] A large number of trials fulfilling criteria specified by the user will be simulated according to these models in order to estimate the power at a given significance level for a specific sample size. For a fixed desired power, e.g. 80%, an approximate sample size can also be found using this approach. Data is simulated in flexible ways allowing for various shapes of baseline marginal mean function for recurrent events and baseline cumulative hazards for death. The method is illustrated using several applications, including a real life application using data from the large randomised trial, LEADER, as the basis for designing a future trial.[5] For clinical trials with complex designs or assumptions, simulation-based power estimation is a important tool. For instance, simulation-based power estimation of complex group-sequential design trials with survival outcomes is an option in the R-package `rpact`.[6]

Recently, there has been an increased focus on applying recurrent event methods when the target is in fact recurrent, e.g. heart failure.[7,8] This would ensure better use of data and improve the understanding of the total disease burden. Potential power gains when applying recurrent event methods as opposed to time-to-first-event methods will also be explored using simulation in this paper.

2 | MATHEMATICAL SETTING AND ESTIMAND

In the present setting, the desire is to model the expected number of events in the presence of terminal events when both are subject to right-censoring (see Figure 1). Let D^* denote the survival time and let $N^*(t)$ denote the number of recurrent events by time t. After death, D^*, has occurred, $N^*(t)$ has no further jumps since no additional recurrent events can happen. Let C denote the censoring time. Neither D^* nor $N^*(t)$ are fully observed due to right-censoring, and we observe $N(t) = N^*(t \wedge C)$, $D = D^* \wedge C$ and $\delta = I(D^* \leq C)$. It is assumed that there is a binary treatment variable, Z. Hence, the data consists of $X = \{N(\cdot), D, \delta, Z\}$. For each subject $i = 1, \ldots, n$, we observe $X_i = \{N_i(\cdot), D_i, \delta_i, Z_i\}$ which are independent and identically distributed replicates of X. It is assumed that C is independent of $N^*(\cdot)$, D^* and Z.

In order to draw inference, the expected number of events in situations where competing terminal events can occur is modelled. This is done by considering the marginal mean function, $\mu(t)$, given by

$$\mu(t) = E(N^*(t)) = \int_0^t S(u^-)\, dR(u),$$

where $S(t) = P(D^* > t)$ and $dR(u) = E(dN^*(t) \mid D^* \geq t)$. Consider, now, a comparison of the expected number of events, $\mu_k(t^*)$, at time t^* between each treatment group, $k = \{0, 1\}$. This could be a ratio $\frac{\mu_1(t^*)}{\mu_0(t^*)}$ or a difference $\mu_1(t^*) - \mu_0(t^*)$ or a log-ratio, $\log\left(\frac{\mu_1(t^*)}{\mu_0(t^*)}\right)$. For this application, focus will be placed on the log-mean ratio. Let

$$\kappa(t^*) = \log\left(\frac{\mu_1(t^*)}{\mu_0(t^*)}\right) = \log(\mu_1(t^*)) - \log(\mu_0(t^*))$$

Assume that the interest is to focus on the hypotheses,

$$H_0 : \kappa(t^*) = \kappa_0, \quad H_a : \kappa(t^*) \neq \kappa_0. \tag{1}$$

Without loss of generality, we assume that $\kappa_0 = 0$, but a general $\kappa_0 > 0$ could easily be accommodated. A simple model for $\mu(\cdot)$ can be formulated by considering the semi-parametric proportional means model of Ghosh and Lin [3],

$$\mu(t \mid Z) = \mu_0(t) \exp(\beta Z), \tag{2}$$

where $\mu_0(t)$ denotes the unspecified baseline mean function, β the effect of treatment Z. The effect of treatment, β, is assumed to be constant for all t. In this setting, β is estimated using inverse probability of censoring weights (IPCW) to account for the competing deaths. When there are no competing deaths, this is equivalent to the estimating equations of Lin, Wang, Ying, Yang. [9] Under equation (2), $\kappa(t^*)$ simplifies to

$$\kappa(t^*) = \kappa = \log\left(\frac{\mu_0(t^*)\exp(\beta)}{\mu_0(t^*)}\right) = \log\left(\exp(\beta)\right) = \beta$$

for all t^*. Note, that under this model, it is not necessary to choose t^* due to the structure that is imposed on $\mu(\cdot)$. But if we wish to calculate power under other models, e.g. using non-parametric estimates of $\mu(t)$, it would typically require a choice of t^*.**?** [10] For the purpose of this paper, we will assume that the model in equation (2) holds. Hence, an estimate that targets the estimand of interest can be extracted directly when fitting the Ghosh and Lin model. Thus, in line with equation (1) we shall now consider the following hypotheses,

$$H_0 : \beta = 0, \quad H_a : \beta \neq 0. \tag{3}$$

The desire is thus to approximate a minimum sample size if one were to design a future trial where the primary analysis was a Ghosh and Lin analysis of a recurrent event endpoint in the presence of terminal events. Furthermore, it is assumed that the cumulative hazard of death given treatment, $\Lambda^D(t \mid Z)$, follows Cox's proportional hazards model, such that, [4]

$$\Lambda^D(t \mid Z) = \Lambda_0^D(t) \exp(\gamma Z), \tag{4}$$

where $\Lambda_0^D(t)$ denotes the unspecified baseline cumulative hazard of death and γ is the effect of treatment, Z. The cumulative hazard of being censored is denoted $\Lambda^C(t)$, which is assumed not to depend on treatment, Z. We will mainly focus on the situation with exponential censoring times, which implies that $\Lambda^C(t) = ct$ for a constant c. Furthermore, the addition of administrative censoring at a fixed time is allowed.

3 | SIMULATION OF TRIAL DATA

Simulation is used in order to estimate the minimum sample size (or power) for a future trial targeting the above log-mean ratio. For a given sample size, effect size, the power can be estimated by simulating a large number of trials according to some assumed characteristics. We wish to simulate data under equation (2) such that $\mu(t \mid Z) = \mu_0(t)\exp(\beta Z)$. This is achieved by using the fact that

$$\mu(t \mid Z) = E(N^*(t) \mid Z) = \int_0^t P(D^* \geq u \mid Z)\, E(dN^*(u) \mid Z, D^* \geq u).$$

If the rate of recurrent events at time t while still alive given treatment, $E(dN^*(t) \mid Z, D^* \geq t)$, satisfies the following,

$$E(dN^*(t) \mid Z, D^* \geq t) = \frac{d\mu_0(t)\exp(\beta Z)}{P(D^* \geq t \mid Z)}, \tag{5}$$

the survival probabilities cancel out and the proportional means model in equation (2) is fulfilled, such that

$$\mu(t \mid Z) = \int_0^t d\mu_0(u)\exp(\beta Z) = \mu_0(t)\exp(\beta Z).$$

In Section 6, we will elaborate further on the impact of assuming that the rate in equation (5) holds. Moreover, it holds that

$$P(D^* > t \mid Z) = S(t \mid Z) = \exp(-\Lambda^D(t \mid Z)).$$

If we further assume a proportional hazards model for death, it holds that

$$P(D^* > t \mid Z) = \exp\left(-\Lambda_0^D(t)\exp(\gamma Z)\right). \tag{6}$$

Hence, when $\mu_0(t)$, β, $\Lambda_0^D(t)$, and γ are specified, it is possible to simulate both recurrent events and deaths using equations (5) and (6). The resulting data then adheres to the proportional means model in equation (2) and the proportional hazards model in equation (4). The forms of $\mu_0(t)$ and $\Lambda_0^D(t)$ are allowed to vary freely. First, censoring times and death times are simulated once per subject using $\Lambda^D(t \mid Z)$ and $\Lambda^C(t)$. Next, for each subject, recurrent events are simulated while $t < D \wedge C$ using the rate in equation (5). It is assumed that there is a 1:1 randomisation and that $Z \sim \text{Bin}(1, p)$ with $p = 0.5$.

Censoring is simulated as a mixture of exponential censoring (during the trial) and administrative censoring upon a chosen trial closure. Hence, throughout the trial, the cumulative hazard of censoring is given by $\Lambda^C(t) = \Lambda_0^C(t) = ct$ for a constant c. Furthermore, uniform accrual until a specified trial day can be added.

3.1 | Input parameters for planning a future trial

In order to plan a future trial and perform power estimation using the suggested procedure, a number of quantities needs to be specified:

1. The total sample size, n. It is assumed that the randomisation is 1:1.
2. A set of values for $(t, \mu_0(t))$, i.e. the expected number of events in the reference group at times t.
3. A set of values for $(t, \Lambda_0^D(t))$, i.e. the cumulative hazard of death in the reference group at times t.
4. The censoring rate through the trial, such that $\Lambda^C(t) = ct$. Here, c is supplied.
5. The log-mean ratio, β, i.e. the effect of treatment on recurrent events.
6. The log-hazard ratio, γ, i.e. the effect of treatment on death.
7. Length of enrollment, $[0, \tau_a]$. Uniform accrual until τ_a is assumed.
8. Total length of study duration, τ_{max}. Administrative censoring occurs at τ_{max}.

For $(t, \mu_0(t))$ and $(t, \Lambda_0^D(t))$, either entire trajectories or few time points can be supplied. Piece-wise constant rates are approximated between each t. The linear approximation may be less good if there are only a few time points. In order to perform power estimation, the α-level is also required.

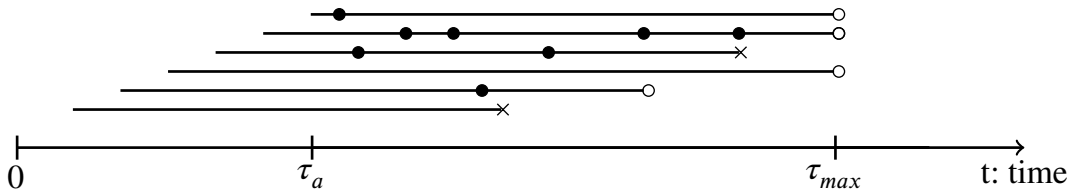


FIGURE 2 Illustration of enrollment and study closure for six individuals. The filled circles, unfilled circles, and the crosses represent recurrent events, censoring and death, respectively. The subjects are uniformly accrued until τ_a. The study had a duration of τ_{max}. The fifth individual was lost-to-follow-up prior to the study closure, τ_{max}. Individual three and six died prior to study closure.

Figure 2 illustrates the consequences of τ_a and τ_{max} for six individuals. These six individuals are uniformly accrued between time 0 and τ_a. A single subject is lost-to-follow-up between accrual and study closure, τ_{max}. All subjects, that did not die prior to study closure, were administratively censored at study closure, τ_{max}. Figure 3 shows examples of three simulated data sets according to calendar time with $(\tau_a, \tau_{max}) = \{(18, \infty), (0, 30), (18, 30)\}$. A single data set with a planned uniform enrollment of 18 months was simulated (left figure). Another data set with no accrual but with a fixed length of 30 months was simulated (middle figure). Finally, both 18 months of accrual and 30 months study length was simulated (right figure). Note, that the situation

with no accrual and a fixed study length, e.g. $(\tau_a, \tau_{max}) = (0, 30)$, also corresponds to the situation with a fixed follow-up time per individual and delayed entry according to accrual. Scenarios of interest would typically include a fixed follow-up time for all individuals or varying follow-up times per individual. These scenarios corresponds to the middle and right figures in Figure 3.

The ingredients are analogous to those required to make a sample size calculation for time-to-event outcomes using a Cox model. However, since modelling the marginal mean requires both information on the recurrent events and death, reference rates and treatment effects should be specified for both recurrent events and death. Similar considerations applies if interested in performing sample size calculations targeting a Fine and Gray model since the cumulative incidences depends on all cause-specific hazards as discussed in Latouche and Porcher.[11]

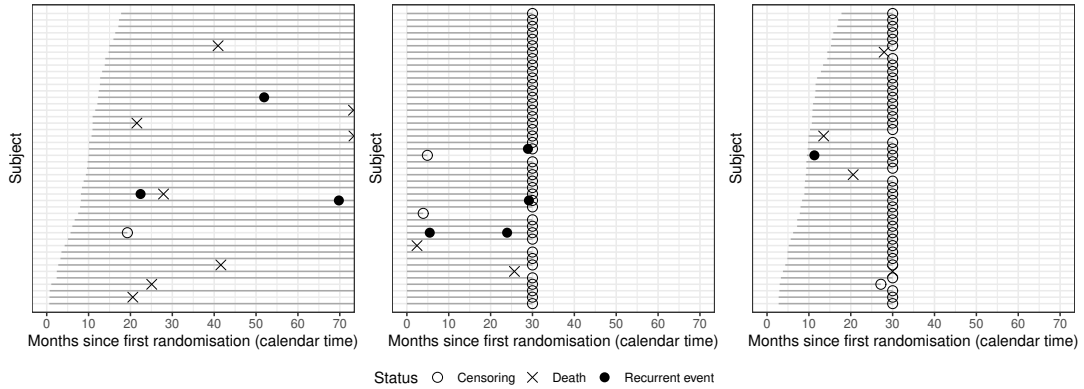


FIGURE 3 Examples of individual life courses for a subset of three simulated data sets. Left: 18 months of uniform accrual and no fixed trial length, $(\tau_a, \tau_{max}) = (18, \infty)$. Middle: No accrual and 30 months of fixed trial duration, $(\tau_a, \tau_{max}) = (0, 30)$. Right: 18 months of uniform accrual and study length of 30 months, $(\tau_a, \tau_{max}) = (18, 30)$.

4 | EXAMPLE OF TRIAL DATA SIMULATION

To illustrate the flexibility of the simulation procedure, a simple setup will be used with linear $\mu_0(t)$ and $\Lambda_0^D(t)$, where

$$\mu_0(t) = 0.1t, \quad \Lambda_0^D(t) = 0.01t, \quad \text{for} \quad t \geq 0.$$

Moreover, we let $\beta = \{-0.2, 0, 0.2\}$ and $\gamma = \{-0.2, 0, 0.2\}$. A $\beta = 0.2$ corresponds to $\exp(\beta) = \exp(0.2) = 1.22$ more recurrent events on average with $Z = 1$ versus $Z = 0$. Figure 4 displays $\mu(t \mid Z)$ and $\Lambda^D(t \mid Z)$ under these choices. It is assumed that $\Lambda^C(t) = 0.01t$. Uniform accrual during 20 days and a fixed study length of 80 days was assumed ($(\tau_a, \tau_{max}) = (20, 80)$). Simulating $n = 1000$ individuals with an even randomisation and $\beta = 0.2$ and $\gamma = 0$ yields the distribution in Table 1. This particular data set results in an estimated $\hat{\beta} = 0.189$ ($\text{se}(\hat{\beta}) = 0.043$). From Figure 5, it is visible that the estimated $\hat{\mu}_0(t)$ from the Ghosh and Lin model based on the simulated data follows the true $\mu_0(t)$ as intended.

Treatment, Z	censoring	recurrent event	death
0	304	2541	200
1	297	3031	199

TABLE 1 Number of events in a single simulated data set for each treatment group with $n = 1000$, $\beta = 0.2$ and $\gamma = 0$.

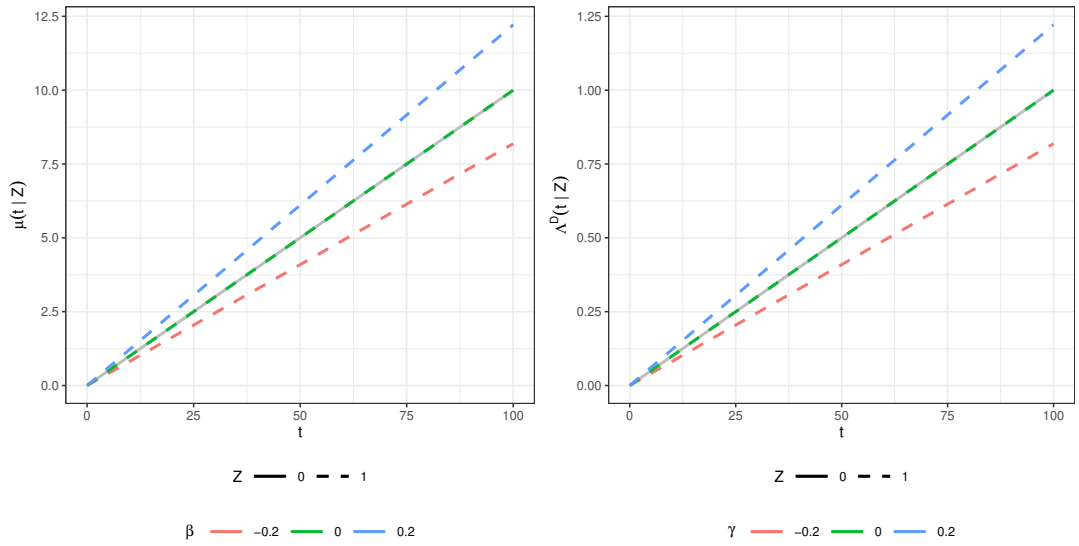


FIGURE 4 Plots of $\mu(t \mid Z) = \mu_0(t)\exp(\beta Z) = 0.1\exp(\beta Z)t$ for $\beta = \{-0.2, 0, 0.2\}$ and $\Lambda^D(t \mid Z) = \Lambda_0^D(t)\exp(\gamma Z) = 0.01\exp(\gamma Z)t$ for $\gamma = \{-0.2, 0, 0.2\}$.

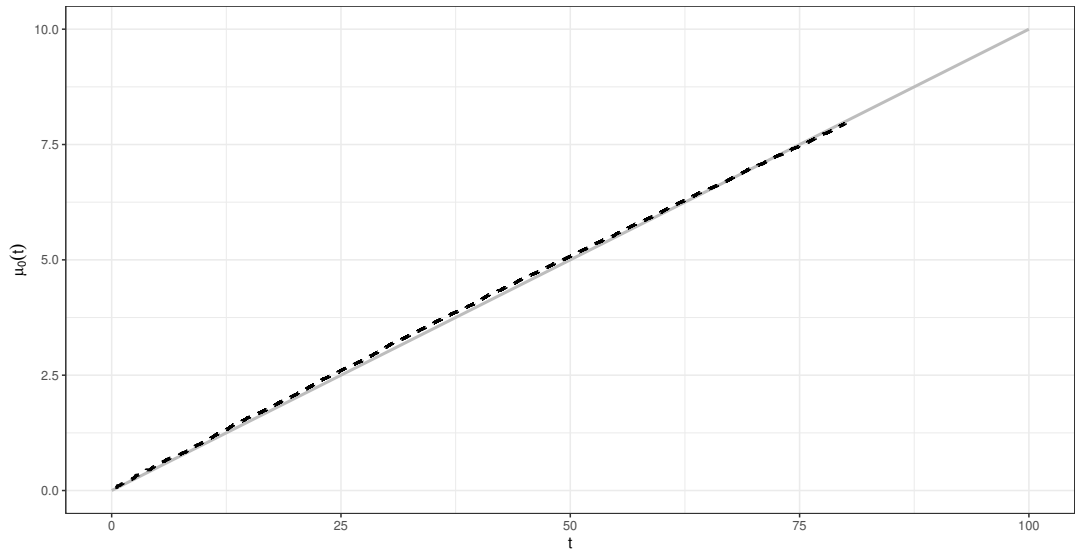


FIGURE 5 The fully drawn grey line represents the true $\mu_0(t) = 0.1t$. The stapled black line is the estimated $\hat{\mu}_0(t)$ from the Ghosh and Lin model applied to the simulated data.

			Recurrent events				First events				
β	γ	n	Total no. of events (per id)	P(rej. H_0), GL (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)	P(rej. H_0), Cox (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)	P(rej. H_0), Log-rank	P(rej. H_0), Gray's
0.2	0	100	554.93 (5.55)	0.319	0.198	0.138	0.153	0.198	0.220	0.151	0.130
		200	1109.55 (5.55)	0.552	0.202	0.097	0.259	0.200	0.155	0.256	0.203
		500	2774.51 (5.55)	0.913	0.202	0.062	0.544	0.199	0.097	0.537	0.399
0.2	0.2	100	553.70 (5.54)	0.280	0.193	0.144	0.184	0.229	0.221	0.185	0.113
		200	1112.66 (5.56)	0.482	0.199	0.101	0.284	0.209	0.155	0.282	0.166
		500	2777.92 (5.56)	0.884	0.199	0.064	0.593	0.214	0.098	0.590	0.315
0.2	-0.2	100	553.32 (5.53)	0.346	0.206	0.134	0.133	0.190	0.220	0.133	0.129
		200	1113.88 (5.57)	0.571	0.202	0.094	0.234	0.192	0.154	0.228	0.238
		500	2778.80 (5.56)	0.919	0.204	0.059	0.495	0.191	0.097	0.488	0.499
-0.2	0	100	453.04 (4.53)	0.296	-0.201	0.144	0.161	-0.213	0.224	0.158	0.141
		200	910.35 (4.55)	0.495	-0.197	0.101	0.255	-0.198	0.157	0.255	0.193
		500	2272.39 (4.54)	0.873	-0.200	0.064	0.549	-0.201	0.099	0.551	0.414
-0.2	0.2	100	452.55 (4.53)	0.257	-0.194	0.149	0.127	-0.174	0.224	0.126	0.125
		200	907.22 (4.54)	0.472	-0.201	0.105	0.227	-0.188	0.157	0.223	0.231
		500	2272.60 (4.55)	0.850	-0.199	0.067	0.431	-0.178	0.099	0.429	0.460
-0.2	-0.2	100	453.95 (4.54)	0.310	-0.199	0.139	0.168	-0.211	0.223	0.166	0.114
		200	908.50 (4.54)	0.523	-0.199	0.098	0.270	-0.209	0.157	0.267	0.152
		500	2276.10 (4.55)	0.894	-0.202	0.062	0.598	-0.217	0.099	0.595	0.338
0	0	100	499.02 (4.99)	0.064	-0.005	0.140	0.051	0.009	0.221	0.050	0.047
		200	995.48 (4.98)	0.053	-0.001	0.099	0.057	0.001	0.155	0.057	0.044
		500	2497.45 (4.99)	0.056	0.002	0.063	0.049	0.001	0.098	0.047	0.056
0	0.2	100	502.41 (5.02)	0.049	-0.003	0.145	0.052	0.013	0.222	0.047	0.055
		200	998.02 (4.99)	0.053	-0.003	0.103	0.052	0.017	0.156	0.050	0.050
		500	2490.16 (4.98)	0.051	-0.002	0.065	0.064	0.015	0.098	0.064	0.051
0	-0.2	100	499.71 (5.00)	0.052	0.001	0.135	0.059	-0.022	0.220	0.056	0.059
		200	1000.18 (5.01)	0.061	-0.004	0.096	0.052	-0.018	0.155	0.051	0.052
		500	2504.58 (5.01)	0.054	-0.001	0.061	0.051	-0.015	0.097	0.051	0.050

TABLE 2 Probability of rejecting the null based on 1000 simulations and various parameter settings. The average of $\hat{\beta}$ and se($\hat{\beta}$) across the 1000 simulations are also displayed. For recurrent events, a Ghosh and Lin model is applied to the recurrent events. For first events, a Cox model is applied to the first recurrent events. Here, a Wald test and estimates are reported. Moreover, a log-rank test comparing the cause-specific cumulative hazards for $Z = 0, 1$ and a Gray's test comparing the cumulative incidences for $Z = 0, 1$ has been computed. The number of recurrent events in total and per subject is averaged across the 1000 simulations.

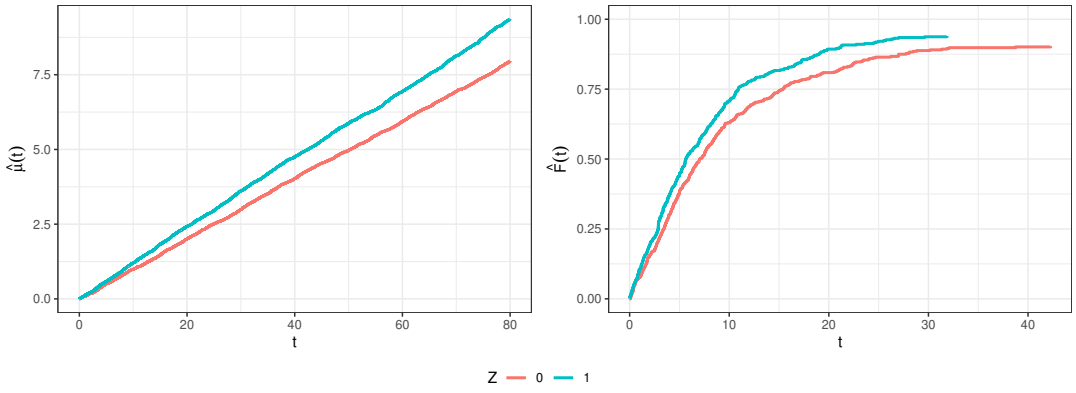


FIGURE 6 Non-parametric estimates of the marginal mean function for recurrent events, $\mu(t)$, and cumulative incidence for a first event, $F(t)$, per treatment based on a single simulated data set with $n = 1000$ and $\beta = 0.2$ and $\gamma = -0.2$.

A total of 1000 simulations were performed for each combination of $\beta = \{0, 0.2, -0.2\}$ with $\gamma = \{0, 0.2, -0.2\}$ and $n = \{100, 200, 500\}$. The results are shown in Table 2 under 'Recurrent events'. In Table A1 in Appendix A, the distributions of recurrent events and deaths per treatment is illustrated for $\beta = -0.2$. When $\beta \neq 0$, the alternative hypothesis is true and the P(reject H_0) corresponds to the approximate power. But when $\beta = 0$, the null is true and the P(reject H_0) corresponds to the approximate type I error rate.

As expected, the power increases as a function of the sample size when $\beta \neq 0$. Moreover, the effect of γ on the power is clear. When treatment increases the cumulative hazard of death ($\gamma > 0$), the power is reduced if $\beta > 0$. Conversely, when treatment decreases the cumulative hazard of death ($\gamma < 0$), the power is increased for $\beta > 0$. If $\beta < 0$, the opposite holds true. For $\beta = 0$, type I error control is maintained.

Potential power gains of recurrent event methods compared to time-to-first-event methods applied to data with a recurrent nature has been explored previously.[12,13] In Table 4 under 'First events' results from analyses of first events have been performed in order to compare the estimated power for recurrent events with first events. A Cox model has been fitted to the first recurrent events. This model estimates the cause-specific hazard of having a first event. The results from a Wald test for the proportional (cause-specific) hazards model has been averaged across the simulations. Analogously, log-rank tests comparing the cause-specific cumulative hazard of a first recurrent event for each treatment has been performed at each simulation. Alternatively, and more similar to modelling $\mu(t)$ for recurrent events, Gray's test was calculated in each simulation to compare the cumulative incidences of having a first event.[14] Please note that the null hypotheses that are explored for the first events are different from the null for recurrent events due to the difference in estimation and modelling. Figure 6 displays the non-parametric estimates per treatment group of the marginal mean function for recurrent events, $\mu(t)$, and cumulative incidence for a first event, $F(t)$, from a single simulated data set. For this simulation, $n = 1000$, $\beta = 0.2$, $\gamma = -0.2$. Since $\beta > 0$, treatment $Z = 1$ has more recurrent events on average compared to $Z = 0$. But since $\gamma < 0$, the cumulative hazard of dying is smaller with $Z = 1$ versus $Z = 0$. Hence, when $\gamma \neq 0$, the results from the Gray's test are preferable to compare to the results for Wald test for recurrent events. This is due to the fact that $\mu(t)$ and $F(t)$ contain information on both recurrent or first events as well as deaths as opposed to the cause-specific counterparts from the log-rank test or the Cox model. Thus, we focus on the power estimated from the Gray's test comparing the cumulative incidences of having an event. For this example, the power gain of focusing on recurrent events as opposed to first events is clear when comparing results from the Wald test for recurrent events with the Gray's test for the first events for all parameter settings.

4.1 | Alternative shapes for $\mu_0(t)$ and $\Lambda_0^D(t)$

The shapes of $\mu_0(t)$ and $\Lambda_0^D(t)$ may be more complex than just linear functions, such as $\mu_0(t) = at$ which was explored in the previous example. Instead, a fast accumulation followed by a slow accumulation of recurrent events, e.g. $\mu_0(t) = t^{0.5}$, could be considered. Oppositely, in some scenarios, first a slow uptake followed by a fast uptake in recurrent events might also be possible, e.g. $\mu_0(t) = \exp(0.03t) - 1$. Similar thoughts applies to the form of $\Lambda_0^D(t)$ which may vary according to the application.

Figure 7 displays these two alternative shapes of $\mu_0(t)$ along with the base example $\mu_0(t) = 0.1t$. For 1000 simulations, let $\Lambda_0^D(t) = 0.01t$, $n = 500$, $\beta = 0.2$, $\gamma = 0$, $(\tau_a, \tau_{max}) = (20, 80)$ and $c = 0.01$. When $\mu_0(t) = t^{0.5}$ the estimated power is 0.993. When $\mu_0(t) = \exp(0.03t) - 1$, the estimated power is 0.703. This can be compared with the linear accumulation, $\mu_0(t) = 0.1t$, in Table 2, where the estimated power was 0.913 for the same setting. For this example, an initial fast accumulation of recurrent events will result in a higher power to detect differences and a slow initial uptake will result in a lower power compared to the linear case. Similar considerations applies for varying the underlying shape of $\Lambda_0^D(t)$.

5 | APPLICATION TO LEADER DATA

The input parameters chosen by the user may advantageously be based on available historic data. In this application, we illustrate how to use the procedure based on such data. The baseline rates in the procedure are based on the observed baseline rates in the previous trial, LEADER. The baseline rates are supplied in three ways, either based on constant rates as estimated from the trial data, piece-wise constant or using non-parametric estimates directly. Moreover, the impact of varying study length and accrual is illustrated.

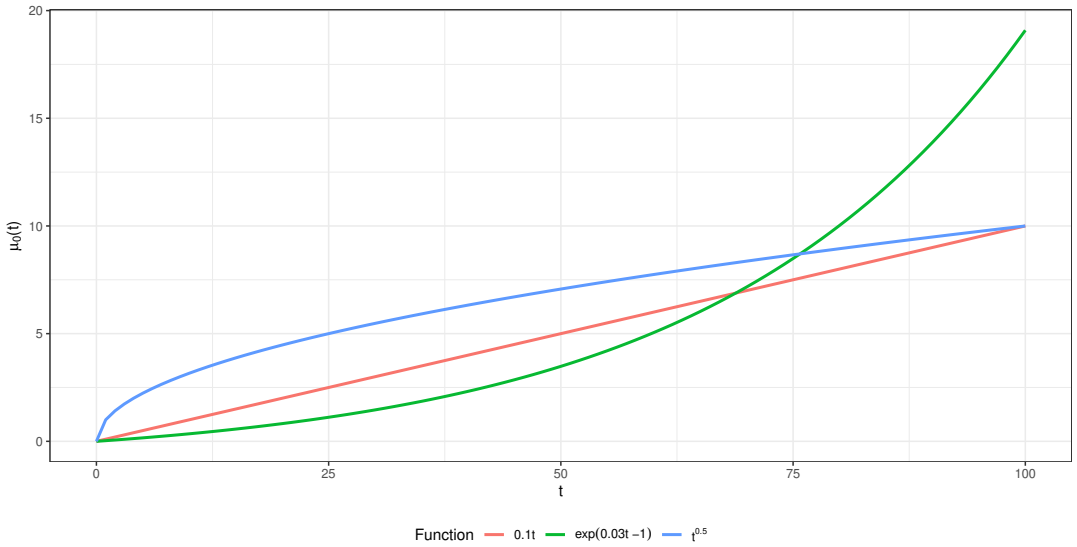


FIGURE 7 Various shapes of $\mu_0(t)$, either $\mu_0(t) = 0.1t$, $\mu_0(t) = t^{0.5}$ or $\mu_0(t) = \exp(0.03t) - 1$.

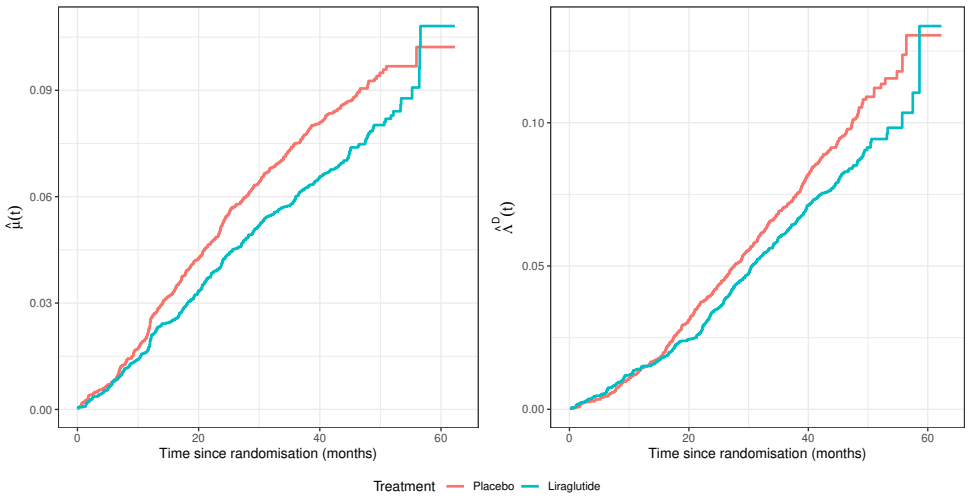


FIGURE 8 Non-parametric estimates computed on the LEADER data. **Left hand side:** Marginal mean estimates for recurrent myocardial infarction, $\hat{\mu}(t)$, per treatment group. **Right hand side:** Cumulative hazard estimates for any death, $\hat{\Lambda}^D(t)$, per treatment group.

LEADER (Liraglutide Effect and Action in Diabetes: Evaluation of cardiovascular outcome Results) was a randomized controlled trial investigating the cardiovascular effects of liraglutide, a treatment for type 2 diabetes, investigated versus placebo when added to standard of care in a population with type 2 diabetes and a high cardiovascular risk.[5] A total of 9340 subjects were randomised 1:1 to receive either liraglutide or placebo. The median follow-up time was 3.8 years. The primary endpoint was a

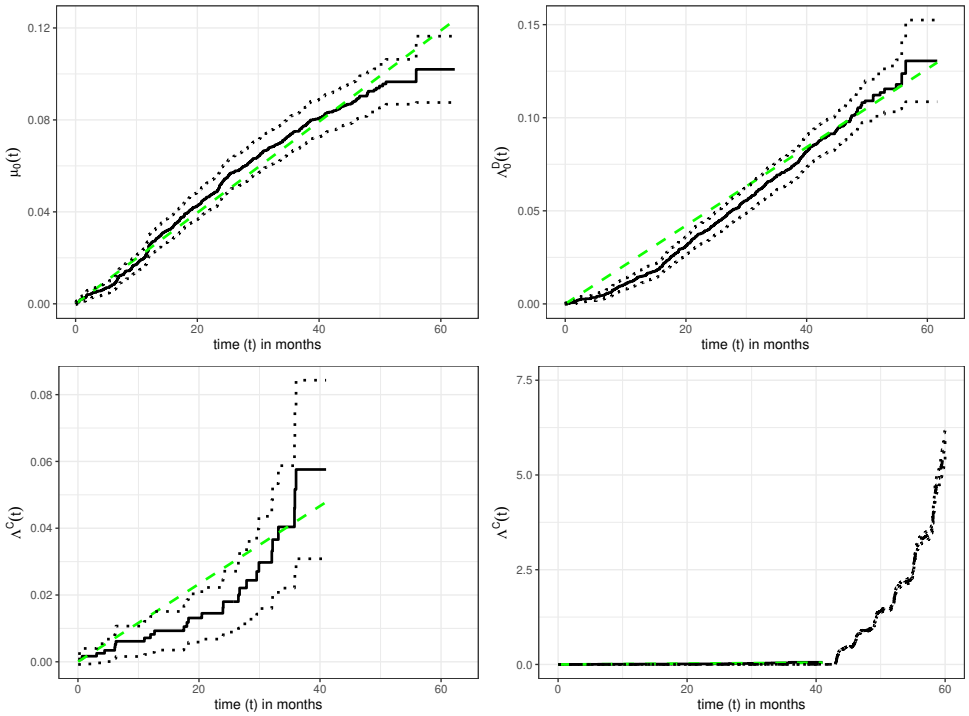


FIGURE 9 Linear approximations of $\mu_0(t)$, $\Lambda_0^D(t)$, and $\Lambda^C(t)$ (stapled, green lines) based on occurrence-exposure rates from LEADER data. Non-parametric estimates of $\hat{\mu}_0(t)$, $\hat{\Lambda}_0^D(t)$ and $\hat{\Lambda}^C(t)$ (fully drawn, black lines) displayed with poine-wise confidence intervals based (dotted, black lines) on LEADER data.

three-component composite major adverse cardiovascular adverse events (MACE) endpoint consisting of time-to-first non-fatal stroke, non-fatal myocardial infarction or cardiovascular death. For this application, the focus is on recurrent myocardial infarction where all cause death acts as terminal events. Figure 8 displays the estimated marginal mean function and cumulative hazard of death for each treatment group using non-parametric estimators for $\mu(t)$, $\hat{\mu}(t)$ (Cook and Lawless, later Ghosh and Lin), and $\Lambda^D(t)$, $\hat{\Lambda}^D(t)$ (Nelson-Aalen).[3 15 16 17] On average, the placebo group has more recurrent myocardial infarction episodes compared to the liraglutide group. Also, mortality is improved with liraglutide versus placebo. A Ghosh and Lin model applied to recurrent myocardial infarction results in,

$$\hat{\beta} = -0.159, \quad \text{se}(\hat{\beta}) = 0.088.$$

This corresponds to a Wald test statistic of -1.807 (p-value of 0.071). Thus, with a total sample size of $n = 9340$ it is not possible to reject the null hypothesis. An estimated $\hat{\beta} = -0.159$ corresponds to $\exp(-0.159) = 0.853$ fewer events on average with liraglutide versus placebo. For all cause death, a Cox model with treatment as a covariate, results in $\hat{\gamma} = -0.166$ with $\text{se}(\hat{\gamma}) = 0.070$.

Now, we assume that we were to design a future trial similar to the LEADER trial. The aim would be to show that there is a significant difference between two treatments on recurrent myocardial infarction as modelled using a Ghosh and Lin model. Hence, the future trial is designed with an active treatment versus a placebo treatment as encountered in LEADER. In line with the previous models, it is assumed that

$$\mu(t \mid Z) = at \exp(\beta Z), \quad \Lambda^D(t \mid Z) = bt \exp(\gamma Z),$$

such that $\mu_0(t) = at$ and $\Lambda_0^D(t) = bt$. Here, a and b is estimated using LEADER data. The monthly rate of recurrent events and deaths are estimated to be $\hat{a} = 0.00198$ and $\hat{b} = 0.00210$, respectively, based on the occurrence-exposure rates (number of events divided by time at risk). Exponential censoring times are simulated according to the observed censoring distribution in LEADER. Data beyond 1250 days (41 months) after randomisation are disregarded to avoid the impact by heavy administrative censoring in the last months of the trial. It is also assumed that $\Lambda_0^C(t) = ct$ during the trial. The monthly rate of censoring is estimated to be $\hat{c} = 0.00117$ using the occurrence-exposure rate before 41 months. To begin with, a fixed trial length of 4 years with no enrollment period is assumed ($\tau_a = 0$ and $\tau_{max} = 4$ years). This corresponds to 4 years of follow-up per individual. Another fixed duration with and without uniform enrollment will be illustrated later. Figure 9 displays the linear approximations of $\mu_0(t)$, $\Lambda_0^D(t)$ and $\Lambda_0^C(t)$ alongside the non-parametric estimates of $\hat{\mu}_0(t)$, $\hat{\Lambda}_0^D(t)$ and $\hat{\Lambda}_0^C(t)$ based on the LEADER data. Here, $\hat{\mu}_0(t)$ is estimated using a non-parametric estimator (see Cook and Lawless, Ghosh and Lin) and $\hat{\Lambda}_0^D(t)$ and $\hat{\Lambda}_0^C(t)$ is estimated using the Nelson-Aalen estimator of the cumulative hazard.[?] [10] [16] The bottom right panel of Figure 9 displays the cumulative hazard of censoring for the entire period including the administrative censoring which begins to occur after 40 months. The linear fits are not perfect but may still be adequate.

	total subjects	censoring	recurrent event	death
Actual data				
placebo	4672	4225	421	447
liraglutide	4668	4287	359	381
Simulated data				
placebo	4698	4276	410	422
liraglutide	4642	4266	344	376

TABLE 3 First panel: Event distribution of recurrent myocardial infarction, censoring and death based on LEADER data. Second panel: Event distribution based on a single data set with $n = 9340$ simulated with similar characteristics as the LEADER data.

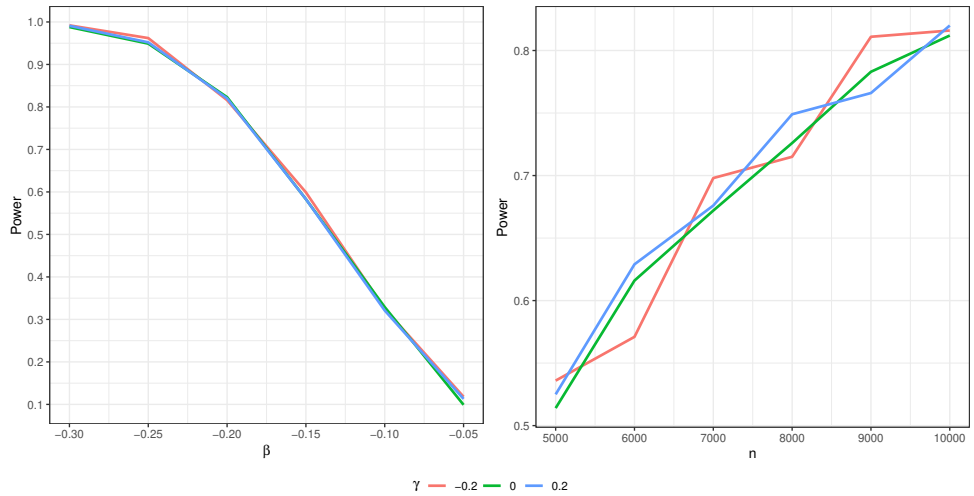


FIGURE 10 Power approximations based on several simulations for a future trial. Placebo data from LEADER is used to estimate $\mu_0(t)$, $\Lambda_0^D(t)$ and $\Lambda^C(t)$. The trial length is assumed to be 4 years.
Left figure: Power for $\beta = \{-0.05, -0.10, -0.15, -0.20, -0.25, -0.30\}$ and $\gamma = \{-0.2, 0, 0.2\}$ for $n = 10000$.
Right figure: Power for $n = \{5000, 6000, 7000, 8000, 9000, 10000\}$ and $\gamma = \{-0.2, 0, 0.2\}$ for $\beta = -0.2$.

			Recurrent events				First events				
β	γ	n	Total no. of events (per id)	P(rej. H_0), GL (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)	P(rej. H_0), Cox (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)	P(rej. H_0), Log-rank	P(rej. H_0), Gray's
-0.2	0	5000	420.72 (0.084)	0.556	-0.205	0.098	0.536	-0.204	0.100	0.536	0.541
		7500	631.09 (0.084)	0.699	-0.203	0.080	0.683	-0.203	0.082	0.683	0.686
		10000	840.80 (0.084)	0.823	-0.201	0.070	0.813	-0.201	0.071	0.813	0.806
-0.2	-0.2	5000	420.50 (0.084)	0.534	-0.201	0.098	0.555	-0.210	0.100	0.555	0.511
		7500	631.35 (0.084)	0.686	-0.196	0.080	0.708	-0.206	0.082	0.708	0.672
		10000	840.88 (0.084)	0.825	-0.200	0.069	0.839	-0.209	0.071	0.839	0.801
-0.2	0.2	5000	420.87 (0.084)	0.541	-0.203	0.098	0.482	-0.193	0.100	0.482	0.522
		7500	630.95 (0.084)	0.698	-0.198	0.080	0.621	-0.188	0.082	0.622	0.670
		10000	840.84 (0.084)	0.829	-0.201	0.070	0.768	-0.189	0.071	0.768	0.814
-0.15	0	5000	440.22 (0.086)	0.337	-0.148	0.097	0.328	-0.147	0.099	0.328	0.329
		7500	644.31 (0.086)	0.477	-0.151	0.079	0.470	-0.152	0.081	0.471	0.471
		10000	860.05 (0.086)	0.603	-0.150	0.069	0.578	-0.150	0.070	0.578	0.584
-0.15	-0.2	5000	431.19 (0.086)	0.320	-0.144	0.097	0.335	-0.152	0.099	0.335	0.305
		7500	645.51 (0.086)	0.464	-0.149	0.079	0.485	-0.157	0.081	0.486	0.436
		10000	860.94 (0.086)	0.591	-0.151	0.069	0.622	-0.160	0.070	0.622	0.581
-0.15	0.2	5000	430.13 (0.086)	0.347	-0.152	0.097	0.294	-0.141	0.099	0.296	0.338
		7500	646.46 (0.086)	0.463	-0.150	0.079	0.393	-0.139	0.081	0.394	0.446
		10000	859.64 (0.086)	0.617	-0.154	0.069	0.519	-0.143	0.070	0.522	0.598
-0.1	0	5000	439.62 (0.088)	0.191	-0.102	0.096	0.196	-0.103	0.098	0.197	0.195
		7500	660.67 (0.088)	0.227	-0.095	0.078	0.215	-0.095	0.080	0.215	0.220
		10000	879.98 (0.088)	0.336	-0.101	0.068	0.311	-0.100	0.069	0.312	0.314
-0.1	-0.2	5000	440.01 (0.088)	0.175	-0.097	0.096	0.196	-0.106	0.098	0.196	0.171
		7500	660.54 (0.088)	0.234	-0.098	0.078	0.258	-0.107	0.080	0.258	0.218
		10000	881.42 (0.088)	0.311	-0.101	0.068	0.350	-0.111	0.069	0.351	0.306
-0.1	0.2	5000	440.29 (0.088)	0.192	-0.105	0.096	0.156	-0.093	0.098	0.157	0.180
		7500	660.18 (0.088)	0.243	-0.101	0.078	0.197	-0.090	0.080	0.197	0.233
		10000	880.71 (0.088)	0.326	-0.100	0.068	0.243	-0.089	0.069	0.243	0.308

TABLE 4 Probability of rejecting the null based on 1000 simulations and various parameter settings. Baseline rates for recurrent events, death and censoring are based on LEADER placebo data. The trial length is assumed to be 4 years.

We let $n = 9340$ and assume that the true values of the parameters were $\beta = -0.2$ and $\gamma = -0.2$. Table 3 shows an overview of a distribution of event types based on data from LEADER and on a single simulated data set with the above characteristics. The simulated data seems to mimic the actual data fairly well in terms of event counts.

Suppose that a power of 80% with a 5% significance level is targeted for the future trial. The power is approximated using 1000 simulations. When $n = 10000$ and $\beta = \gamma = -0.2$ there is a power of 0.825 to detect differences ($\beta \neq 0$). Hence, this would be a sufficient sample size to detect a difference of $\beta = -0.2$ with a 80% power and 5% significance level. This difference corresponds to $\exp(\beta) = 0.82$ fewer events on average with active treatment versus placebo given the remaining assumptions. For $\beta = -0.15$, the power would be 0.591 instead. Thus, this would require a higher sample size than 10000.

Figure 10 displays the approximate power for $\beta = \{-0.05, -0.10, -0.15, -0.20, -0.25, -0.30\}$ and $\gamma = \{-0.2, 0, 0.2\}$ for $n = 10000$. Figure 10 also displays the approximate power for $\beta = -0.2$ and $\gamma = \{-0.2, 0, 0.2\}$ for sample sizes in $n = \{5000, 6000, 7000, 8000, 9000, 10000\}$. Table 4 displays the results of power approximations for 1000 simulations, $\beta = \{-0.1, -0.15, -0.2\}$, $\gamma = \{-0.2, 0, 0.2\}$ and $n = \{5000, 7500, 10000\}$ based on LEADER data. In Table A2 in Appendix A, the distributions of recurrent events and deaths per treatment is illustrated for $\beta = -0.2$. The approximate power using both the Ghosh and Lin model on recurrent myocardial infarction and time-to-first myocardial infarction is presented. On average, the power is higher with the recurrent event method when comparing the average Ghosh-Lin Wald test with the Gray's test on the cumulative incidences. The power gains are not substantial but this example has a very low event rate compared to the previous example in Section 4. This suggests that the power benefits may be quite dependent on *how* recurrent the endpoint of interest is.

5.1 | Alternative shapes for $\mu_0(t)$ and $\Lambda_0^D(t)$

The assumption of a constant rate of recurrent events or death may be too restrictive in practice. Alternatively, piece-wise constant rates could be considered. Piece-wise constant rates of recurrent events and deaths are estimated based on occurrence-exposure rates within 12 month intervals until 60 months after randomisation as displayed in Table 5. Alternatively, $\hat{\mu}_0(t)$ and $\hat{\Lambda}_0^D(t)$ values could be read of at each year. It is assumed that $\mu_0(t) = a_t t$ and $\Lambda_0^D(t) = b_t t$ for t in $[0, 12]$, $(12, 24]$, $(24, 36]$, $(36, 48]$ and $(48, 60]$

months with the monthly rates given in Table 5. The set of data points $(t, \mu_0(t))$ and $(t, \Lambda_0^D(t))$ for $t = 0, 12, 24, 36, 48$ and 60 months are used in the simulation procedure. Figure 11 displays the piece-wise linear baseline functions that are assumed in the simulation alongside the linear baseline functions for the scenario described in the previous section. The piece-wise constant rates seem to catch the form, especially for $\mu_0(t)$, fairly well. Based on 1000 simulations, the approximate power is 0.827 when using the assumed piece-wise constant rates, while keeping all other assumptions fixed. This is comparable to the approximate power based on the constant rates (0.825). Due to the similarity of the constant and piece-wise constant models in Figure 11, it is not surprising that the power estimates using either approach is fairly close.

Interval, t in months	Monthly rate of recurrent events, a_t	Monthly rate of deaths, b_t
[0, 12]	0.00206	0.00120
(12, 24]	0.00237	0.00190
(24, 36]	0.00199	0.00238
(36, 48]	0.00153	0.00231
(48, 60]	0.00103	0.00261

TABLE 5 Estimates of piece-wise constant monthly rates of recurrent events and deaths based on occurrence-exposure rates in the placebo group from LEADER data.

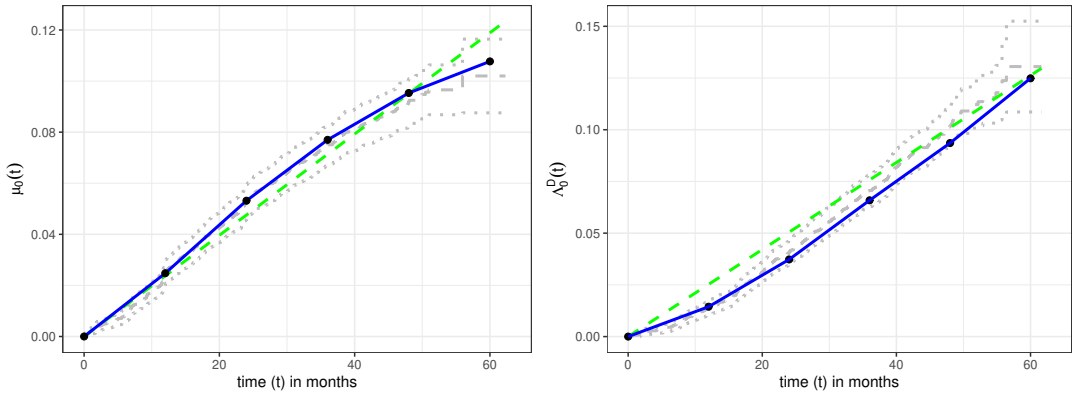


FIGURE 11 Constant (green) and piece-wise constant (blue) rates for $\mu_0(t)$ and $\Lambda_0^D(t)$ based on five data points based on occurrence-exposure rates. Non-parametric $\hat{\mu}_0(t)$ and $\hat{\Lambda}_0^D(t)$ estimates based on LEADER data (grey).

Instead of assuming some underlying shape of the true $\mu_0(t)$ and $\Lambda_0^D(t)$, the baseline estimates from the LEADER data can be used directly in the simulation. This may be too sensitive to the particular data for designing a future trial. While the remaining assumptions are fixed as described earlier, we assume that the true baseline functions equal the baseline estimates from data, i.e. $\mu_0(t) = \hat{\mu}_0(t)$ and $\Lambda_0^D(t) = \hat{\Lambda}_0^D(t)$. Thus, the set of values $(t, \hat{\mu}_0(t))$ and $(t, \hat{\Lambda}_0^D(t))$ is supplied to the procedure. Here, the time points refer to the recurrent event times and death times for $\hat{\mu}_0(t)$ and $\hat{\Lambda}_0^D(t)$. With 1000 simulations for $\beta = \gamma = -0.2$ and $n = 10000$, this results in a power of 0.871. The estimated power using the assumption of constant rates was 0.825 under the same settings. The power using the baseline estimates directly is slightly higher but comparable to the results using either the constant or piece-wise constant rate assumption.

5.2 | Study length and enrollment

The power calculations presented so far for this example assumed a fixed trial length of 4 years and did not take enrollment into account ($(\tau_a, \tau_{max}) = (0, 48)$ months). The impact on power of varying recruitment and study length is explored in this section. Based on the LEADER's trial design, the total study length was planned to be 60 months with 18 months of enrollment. For the last randomised subjects, this implied a follow-up period of 42 months. This is visible in Figure 9, where the last randomised subjects are administratively censored around 42 months after randomisation. The power approximations are modified to accommodate earlier study closure, participant enrollment or both. The underlying assumptions are still the same as for the first example based on constant rates in Section 5. We will explore uniform accrual in $\tau_a = 18$ months and a fixed trial length of $\tau_{max} = 30$ months. A total of 1000 simulations were performed. 18 months of uniform accrual was added in addition to the fixed trial duration of 48 months ($(\tau_a, \tau_{max}) = (18, 48)$), which results in a power of 0.814 versus the power of 0.825 ($(\tau_a, \tau_{max}) = (0, 48)$). Assume now that the study length was 30 months instead of 48 months with no enrollment period ($(\tau_a, \tau_{max}) = (0, 30)$), this would result in a power of 0.638. Including both 18 months of uniform accrual and administrative censoring after 30 months ($(\tau_a, \tau_{max}) = (18, 30)$) results in a power of 0.459.

This displays the flexibility in terms of varying assumptions on study length and enrollment. The procedure could also be expanded to accommodate non-uniform or seasonal accrual.

6 | DEPENDENCE BETWEEN $N^*(\cdot)$ AND D^*

The proposed simulation procedure relies on the assumption that the rate of recurrent events while alive in order has a specific form in order to ensure that the marginal models hold (see Section 3, equation (5)). It turns out, that assuming this particular rate introduces independence between $N^*(\cdot)$ and D^*. Hence, if there is dependence between $N^*(\cdot)$ and D^* this will not be correctly reflected in the data sets simulated according to the proposed procedure. If experiencing more recurrent events increases the probability of dying, regardless of treatment, Z, it indicates that there could be some dependence between $N^*(\cdot)$ and D^*. The following simulation studies has been conducted in order to evaluate the impact on power if there is dependence between $N^*(\cdot)$ and D^*.

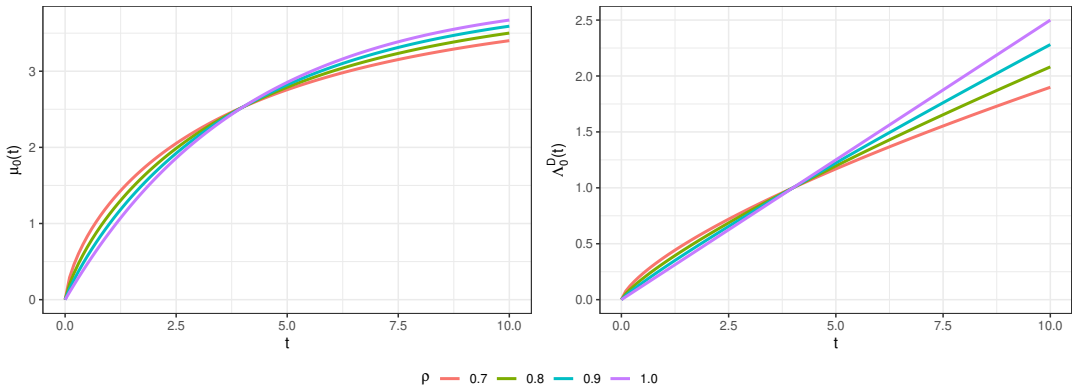


FIGURE 12 Shapes of $\mu_0(t)$ and $\Lambda_0^D(t)$ for $\lambda_0^D = 0.25$ and $\rho = \{0.7, 0.8, 0.9, 1.0\}$.

A simulation procedure similar to the one proposed in Ghosh and Lin has been employed to ensure dependence using frailties.[3] It holds that,

$$\mu(t \mid Z) = \int_0^t S(u^- \mid Z)\, d\, R(u \mid Z),$$

				Simulation 1: Independence between $N^*(\cdot)$ and D^*				Simulation 2: Dependence between $N^*(\cdot)$ and D^*			
ρ	β	γ	n	Total no. of events (per id)	P(reject H_0), GL (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)	Total no. of events (per id)	P(reject H_0), GL (Wald)	Mean of $\hat{\beta}$	Mean of se($\hat{\beta}$)
1	-0.2	0	100	249.14 (2.49)	0.157	-0.189	0.203	252.06 (2.52)	0.182	-0.211	0.203
			500	1244.24 (2.49)	0.588	-0.201	0.091	1263.88 (2.53)	0.607	-0.203	0.091
			1000	2489.96 (2.49)	0.872	-0.204	0.065	2522.26 (2.52)	0.855	-0.196	0.064
		-0.2	100	250.45 (2.50)	0.165	-0.191	0.195	251.26 (2.51)	0.180	-0.194	0.196
			500	1244.59 (2.49)	0.608	-0.196	0.088	1258.62 (2.52)	0.627	-0.199	0.087
			1000	2489.95 (2.49)	0.888	-0.196	0.062	2516.68 (2.52)	0.887	-0.200	0.062
0.9	-0.2	0	100	246.89 (2.47)	0.181	-0.208	0.199	250.84 (2.50)	0.174	-0.200	0.207
			500	1236.35 (2.47)	0.621	-0.203	0.090	1250.05 (2.50)	0.587	-0.204	0.094
			1000	2467.74 (2.47)	0.879	-0.199	0.063	2496.36 (2.50)	0.837	-0.200	0.067
		-0.2	100	247.80 (2.48)	0.178	-0.195	0.191	250.78 (2.51)	0.197	-0.218	0.199
			500	1235.11 (2.47)	0.625	-0.197	0.086	1246.49 (2.49)	0.589	-0.196	0.090
			1000	2472.92 (2.47)	0.903	-0.199	0.061	2499.04 (2.50)	0.880	-0.200	0.063
0.8	-0.2	0	100	246.30 (2.46)	0.192	-0.209	0.194	248.04 (2.48)	0.158	-0.192	0.215
			500	1218.23 (2.44)	0.659	-0.205	0.088	1240.11 (2.48)	0.544	-0.200	0.098
			1000	2437.11 (2.44)	0.900	-0.201	0.062	2471.71 (2.47)	0.845	-0.205	0.069
		-0.2	100	244.54 (2.45)	0.178	-0.200	0.189	247.01 (2.47)	0.182	-0.198	0.204
			500	1223.99 (2.45)	0.628	-0.195	0.084	1238.32 (2.48)	0.585	-0.198	0.093
			1000	2447.83 (2.45)	0.902	-0.194	0.060	2476.02 (2.48)	0.871	-0.203	0.066
0.7	-0.2	0	100	242.14 (2.42)	0.204	-0.212	0.192	243.49 (2.43)	0.177	-0.203	0.224
			500	1208.23 (2.42)	0.654	-0.202	0.086	1224.92 (2.45)	0.507	-0.204	0.102
			1000	2416.57 (2.42)	0.909	-0.203	0.061	2456.95 (2.46)	0.803	-0.200	0.072
		-0.2	100	241.04 (2.41)	0.198	-0.196	0.186	245.05 (2.45)	0.165	-0.199	0.211
			500	1206.32 (2.41)	0.646	-0.195	0.083	1224.11 (2.45)	0.553	-0.203	0.096
			1000	2416.02 (2.42)	0.910	-0.195	0.059	2450.65 (2.45)	0.839	-0.204	0.068

TABLE 6 Probability of rejecting the null based on 1000 simulations and various parameter settings. For simulation 2, data is simulated using the described frailty models. For simulation 1, data is simulated according to the procedure described in this article, based on the same $\mu_0(t)$ and $\Lambda_0^D(t)$ as for simulation 2. For each simulation, a Ghosh-Lin model has been applied to the data in order to decide whether to reject the null. The number of recurrent events in total and per subject is averaged across the 1000 simulations.

where $S(t \mid Z) = P(D^* > t \mid Z)$ and $dR(t \mid Z) = E(dN^*(t) \mid D^* \geq t, Z)$.[?] [10] We consider the following data generating frailty models,

$$\lambda^D(t \mid Z, \nu) = \nu \exp(\gamma_D Z)\lambda_0^D, \tag{7}$$

$$dR(t \mid Z, \nu) = \nu S_0(t \mid \nu)^{1-\exp(\gamma_D Z)} \exp(\beta Z)\, dt, \tag{8}$$

where $S_0(t \mid \nu) = \exp\left(-\int_0^t \lambda^D(u \mid Z = 0, \nu)\, du\right)$ and $\nu > 0$ is a frailty term that is assumed to be generated from a positive stable distribution with Laplace transform $\exp(-\nu^\rho)$, $\rho \in (0, 1]$. The frailty term introduces dependence between the recurrent events and the terminal event. The waiting times between two successive recurrent events and the survival time will have a Kendall's tau correlation coefficient of $1 - \rho$. If $\rho < 1$ ($\rho = 1$), there is a positive (no) association between recurrent events and death. Under this data generating model, both the proportional means model for recurrent events and proportional hazards model for death is fulfilled. It holds that,

$$\mu(t \mid Z) = \int_0^\infty \mu(t \mid Z, \nu) f(\nu)\, d\nu = \mu_0(t) \exp(\beta Z),$$

$$S(t \mid Z) = \int_0^\infty S(t \mid Z, \nu) f(\nu)\, d\nu = \exp\left(-\Lambda_0^D(t) \exp(\gamma Z)\right),$$

where $\mu_0(t)$ and $\Lambda_0^D(t)$ is given by using equations (8) and (7) and integrating out, such that

$$\mu_0(t) = \frac{1}{\lambda_0^D}\left(-\exp\left(-\left(\lambda_0^D t\right)^\rho\right) + 1\right), \quad \Lambda_0^D(t) = \left(\lambda_0^D t\right)^\rho, \quad \gamma = \gamma_D \rho.$$

A total of 1000 simulations were performed for $\lambda_0^D = 0.25$, $n = \{100, 500, 1000\}$ with $\beta = -0.2$, $\gamma_D = \{0, -0.2\}$ and $\rho = \{0.7, 0.8, 0.9, 1\}$. The higher ρ, the less dependence between $N^*(\cdot)$ and D^*. Exponential censoring with a rate of 0.1 is assumed.

Additionally, administrative censoring is added at time 10. The choice of ρ and λ_0^D dictates the form of $\mu_0(t)$ and $\Lambda_0^D(t)$. Figure 12 displays the shapes of $\mu_0(t)$ and $\Lambda_0^D(t)$ for $\rho = \{0.7, 0.8, 0.9, 1.0\}$.

First, data is simulated using the outlined procedure, which allows for dependence between $N^*(\cdot)$ and D^*. For each simulation, a Ghosh and Lin model is fitted. Next, an approximate power based on 1000 simulations is estimated. Following this, data which adheres to the same $\mu_0(t)$, $\Lambda_0^D(t)$, censoring pattern, β and γ is simulated according to the procedure suggested in this paper. For each simulated data set, a Ghosh and Lin model is fitted and the overall power is approximated. As mentioned, this procedure introduces independence between $N^*(\cdot)$ and D^*. Hence, we wish to compare the effect on the power between the two methods, where the dependence is increased, as controlled by ρ. The results are displayed in Table 6. With no association between successive recurrent events and death, $\rho = 1$, the results from the two simulation types are similar. Here, the estimated power is comparable when comparing the two simulation types per parameter setting. However, when as ρ is decreased, such that there is an increased dependence between successive recurrent events and death, the results from the two simulation types are more different. Overall, it seems that the power will be slightly overestimated using simulation 1, which introduces independence (in a scenario when this is violated). Thus, care should be taken if there is dependence between $N^*(\cdot)$ and D^* for the specific scenario in which the power procedure is to be applied. Most likely, the estimated power will too high due to the dependence which is not accounted for.

Cox models for time to death with the number of previous events at t^-, $N^*(t^-)$, as a time-varying covariate has been fitted, in order to judge if the dependence between $N^*(\cdot)$ and D^* imposed by equations (7) and (8) is present in the simulated data sets. We focus on $\rho = 0.7$, such that there should be dependence between D^* and $N^*(\cdot)$. Moreover, we let $\beta = -0.2$, $\gamma = 0$ and $n = 1000$. A Cox model added to a single simulated data set according to simulation 2 (dependence imposed) results in an estimated hazard ratio of 1.200 ($\hat{\eta} = 0.182$, $\text{se}(\hat{\eta}) = 0.023$) with a significant Wald test p-value of below 0.0001 for the effect of previous number of events on death (η). Conversely, a Cox model for time to death conditional on the number of previous events applied on a single data set simulated using simulation 1 (independence imposed) results in an estimated hazard ratio of 0.960 ($\hat{\eta} = -0.041$, $\text{se}(\hat{\eta}) = 0.027$) with an insignificant p-value of 0.130. 1000 data sets have been simulated using simulation 2 resulting in an average hazard ratio of 1.222 (average $\hat{\eta} = 0.201$, average of $\text{se}(\hat{\eta}) = 0.020$) with a mean p-value of below 0.0001. Similarly, 1000 data sets simulated according to simulation 1 results in an average estimated hazard ratio of 0.999 (average $\hat{\eta} = -0.001$, average of $\text{se}(\hat{\eta}) = 0.027$) with a mean p-value of 0.500. Thus, it is clear that the imposed dependence between D^* and $N^*(\cdot)$ is not apparent from the data sets simulated using the proposed simulation procedure. Still, the difference between the estimated power using either approach may be small as seen from Table 6.

Kendall's tau correlation coefficient estimated on LEADER data between the first recurrent events and deaths for censored data resulted in a correlation coefficient of 0.015, which corresponds to a $\rho = 0.985$.[18] Hence, for this data, it would not be unrealistic to assume that $\rho = 1$. If the waiting time until death is modelled conditional on the number of recurrent events at time t^-, $N^*(t^-)$, using a Cox model based on the LEADER data, the hazard ratio estimate for the effect of the previous number of events is 2.071 ($\hat{\eta} = 0.728$, $\text{se}(\hat{\eta}) = 0.072$). This effect is significant with a Wald test p-value below 0.0001. If the same model is instead applied to the data set which was simulated according to the characteristics in the LEADER data (corresponding to the data behind Table 3) using the procedure proposed in this paper, the estimated hazard ratio is 1.168 ($\hat{\eta} = 0.155$, $\text{se}(\hat{\eta}) = 0.171$) with a Wald test p-value of 0.363. If we simulate 1000 data sets with the same characteristics, the mean hazard ratio was 0.998 (average $\hat{\eta} = -0.002$, average $\text{se}(\hat{\eta}) = 0.171$) with an average Wald test p-value of 0.500. This indicates that there is some dependence between D^* and $N^*(\cdot)$ in the LEADER data which is not correctly reflected in the data sets simulated according to the proposed procedure.

7 | DISCUSSION

Cook and Lawless distinguish between intensity and marginal models for the analysis of recurrent events.[19,20] Typically, analysis either targets marginal or conditional features of the recurrent event process. In this paper, we have focused on marginal features. This has the benefit that the needed quantities can be obtained from simple and marginal quantities, such as the reference recurrent event rate, the expected log-mean-ratio between two treatments, the survival rate in the reference group, etc. Methods that are based on intensities typically involve the specification of parameters that can be hard to justify, such as frailty parameters. Frailty models target subject specific treatment effects since the parameters are conditional on the frailty. Thus, such models will not be able to directly estimate a population-level summary measure of the treatment effect which an estimand should target according

to ICH E9.[21] This requirement of ensuring a population-level summary measure in the estimand makes the marginal models preferable when focusing on recurrent events with competing deaths.

Wu and Cook suggested closed form solutions for trial design when the intention is to model the processes using Prentice, Williams, Peterson (PWP) type models, which are stratified Andersen-Gill models.[22,23,24] However, as argued by Cook and Lawless and discussed in Furberg et al., if the randomised treatment has an effect on the number of recurrent events, the PWP model stratified according to the number of events may in turn remove some of the treatment effect which is an undesirable feature for analysing recurrent events in a randomised controlled trial.[19,1]

Fritsch et al. considered power calculations for recurrent event endpoint in the presence of terminal events with a comparison to time-to-event analysis.[12] Their simulations are based on a joint frailty model approach, where non-compatible models are applied afterwards. As the authors note, this makes it hard to compare power across methods and type I error control is no longer maintained. Moreover, several of the models are mis-specified if there is a treatment effect on death. Finally, an overall estimand discussion is missing from the paper and it is unclear what the estimand of interest is when modelling recurrent events with terminal events. The approach outlined in this paper ensures that identifiable parameters can be inserted in the simulation procedure in a way where the marginal interpretation is kept. Moreover, type I error control and the correct power interpretation is kept using the suggested approach regardless of the impact of treatment on recurrent events and death. Potential power gains has also been outlined using two examples. Whether or not recurrent event methods will provide large benefits in terms of power compared to traditional time-to-event analyses is seen to depend on the underlying recurrent event rate. The example based on LEADER data shows that minor improvements in power can be expected if focusing on recurrent events versus first events when the yearly recurrent event rate is low. The first example in Section 4 had more substantial improvements as well as a high event rate.

The decision on when to close the clinical trial is governed by the selected stopping rules. A variety of stopping rules may be applicable depending on the aim of the trial. For the purpose of this paper, we have only considered the following stopping rules; either when each individual has had a fixed number of months of follow-up, or when the total study length has reached a specific length. Alternatively, the study may be stopped if and only if at least x subjects has had one recurrent event combined with a minimum of follow-up time. Moreover, if the observed drop-out rate or enrollment rate is lower than planned, the trial may be continued according to some pre-specified rules.

The models in this paper rely on the assumptions of proportional means for recurrent events and proportional hazards for deaths for the two treatments groups. This proportionality may be unreasonable on both or either of the two components. The cumulative hazard of death may be specified using alternative models to the Cox model and the simulation procedure would only require an update in how deaths (equation (6)) and recurrent events while alive (equation (5)) are generated. More flexible non-parametric methods can be chosen to estimate $\mu(t)$ and $\Lambda^D(t)$. However, in order to compare two treatments this requires a choice of t^* to e.g. compute the log-mean ratio, $\log(\mu_1(t^*)) - \log(\mu_0(t^*))$. The suggested procedure may be modified to simulate recurrent events and death using non-parametric estimators for $\mu_{0j}(t)$ and $S_j(t)$ for treatment $j = 0, 1$ (Cook and Lawless, Ghosh and Lin, and Kaplan-Meier), which then imposes no proportionality between treatments.[?,10,25] This could be used to judge whether the proportionality assumption is reasonable for both recurrent events and deaths. Subsequently, treatments may be compared at a chosen t^*. This is a topic for further research. Alternatively, AUC under $\mu(t)$ until τ may be considered and compared for each treatment (Claggett et al.), but the interpretation of this measure is less intuitive.[26] The while-alive estimand that focuses on the ratio of the marginal mean function and the restricted mean survival time (RMST), $\mu(t)/RMST(t) = \mu(t)/\int_0^t S(u)\,du$, could also be targeted as another measure. However, it is less clear how to impose specific regression models on this measure.[27,28] A topic for further research, is a regression model for $\mu(t)/RMST(t)$ formulated using pseudo-observations.

The proposed simulation procedure introduces independence between the number of recurrent events and death ($N^*(\cdot)$ and D^*) through the assumed recurrent event rate while alive (see Section 3, equation (5)). If for instance more recurrent events leads to a higher probability of death, irrespective of treatment, there would be dependence between $N^*(\cdot)$ and D^*. However, if there is a treatment effect on both death and recurrent events, it can be hard to disentangle if there is any additional dependence. The bivariate pseudo-observation approach suggested by Furberg et al. which models $(\mu(t), S(t)) = (E(N^*(t)), P(D^* > t))$ simultaneously given covariates could be used to characterize if there is dependence between $N^*(\cdot)$ and D^* given treatment.[29] The simulation studies in Section 6 which simulated data with dependence between $N^*(\cdot)$ and D^* indicated that the estimated

power may be too high if there is unaccounted dependence. This could be relaxed by considering frailty type models to generate data which adheres to the marginal models. However, this would typically involve the choice of additional parameters, e.g. frailty variance. If the frailty models in Section 6 fits the application and future data, λ_0^D and ρ can be estimated directly using baseline estimates of $\mu_0(t)$ and $\Lambda_0^D(t)$, which then subsequently can be used to simulate trial data. A procedure which allows one to specify dependence in a practical way as well as maintaining the correct marginal structure is a topic for future work. If the sample size calculation is to be based on available historical data, the Kendall's tau correlation coefficient for censored may be calculated based on first event and death times to quantify a potential dependence. Kendall's tau based on LEADER data for the first recurrent event and death times is estimated to be approximately 0.015, which indicates that the dependence is low for this example. Still, analyses of time to death conditional on the number of recurrent events at time t^- based on LEADER data seem to capture a stronger dependence than data sets simulated according to LEADER characteristics based on the procedure proposed in this paper. Even though such a dependence may exist, the results from the simulation studies indicate that this does not necessarily have a large impact on the estimated power. For the scenarios we have seen, the impact on power is manageable if there is unaccounted dependence between recurrent events and death.

The simulation-based procedure provides a flexible method for performing power calculations when targeting a marginal mean regression model in the setting with recurrent events and competing risks. The input parameters are various marginal parameters which are possible to obtain from historic data. The procedure is available online based on `R` software alongside a descriptive vignette with examples.

ACKNOWLEDGMENTS

The clinical trial LEADER was sponsored by Novo Nordisk A/S.

Financial disclosure

This research was conducted as a part of the PhD education of JK Furberg from which she received funding from Novo Nordisk A/S. H Ravn is employed at Novo Nordisk A/S.

Conflict of interest

The authors declare no potential conflict of interests.

SOFTWARE AVAILABILITY

An `R`-function simulating data according to the described procedure is available through GitHub in the repository `simpowerrecurrent`, (`https://github.com/JulieKFurberg/simpowerrecurrent`). The `R` function is based on functions from the `mets` `R`-package.[30,31,32]

DATA AVAILABILITY STATEMENT

De-identified individual participant data, study protocol and redacted Clinical Study Report will be available according to Novo Nordisk data sharing commitments. The data will be made available permanently after research completion and approval of product and product use in both EU and US. Data will be shared with *bona fide* researchers submitting a research proposal requesting access to data and for use as approved by the Independent Review Board according to the IRB Charter (see novonordisk-trials.com). Access request proposal form and the access criteria can be found at novonordisk-trials.com. The data will be made available on a specialised SAS data platform.

References

1. Furberg JK, Rasmussen S, Andersen PK, Ravn H. Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in LEADER. *Pharmaceutical Statistics* 2022; 21(1): 241–267.

2. Andersen PK, Geskus RB, De Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* 2012; 41(3): 861–870.

3. Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. *Statistica Sinica* 2002; 12: 663-688.

4. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972; 34(2): 187-220.

5. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *New England Journal of Medicine* 2016; 375(4): 311-322.

6. Wassmer G, Pahlke F. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. 2022. R package version 3.3.1.

7. Roubille F, Huet F, Duflos C. Total Burden of Events: A New Standard Stressing the Positive Impact of Inflammation Modulation. *Journal of the American College of Cardiology* 2020; 76(14): 1671–1673.

8. Everett BM, MacFadyen JG, Thuren T, Libby P, Glynn RJ, Ridker PM. Inhibition of interleukin-1β and reduction in atherothrombotic cardiovascular events in the CANTOS trial. *Journal of the American College of Cardiology* 2020; 76(14): 1660–1670.

9. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society* 2000; 62(4): 711-730.

10. Ghosh D, Lin DY. Nonparametric analysis of Recurrent Events and Death. *Biometrics* 2000; 56: 554-562.

11. Latouche A, Porcher R. Sample size calculations in the presence of competing risks. *Statistics in medicine* 2007; 26(30): 5370–5380.

12. Fritsch A, Schlömer P, Mendolia F, Mütze T, Jahn-Eimermacher A, Consortium REQO. Efficiency Comparison of Analysis Methods for Recurrent Event and Time-to-First Event Endpoints in the Presence of Terminal Events—Application to Clinical Trials in Chronic Heart Failure. *Statistics in Biopharmaceutical Research* 2021: 1–12.

13. Committee for Medicinal Products for Human Use (CHMP) . Qualification opinion of clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses. 2020.

14. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999.

15. Cook R, Lawless JF. Marginal Analysis of Recurrent Events and A Terminating Event. *Statistics in Medicine* 1997; 16: 911-924.

16. Nelson W. Hazard plotting for incomplete failure data. *Journal of Quality Technology* 1969; 1(1): 27–52.

17. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics* 1972; 14(4): 945–966.

18. Hougaard P. *Analysis of multivariate survival data*. 564. Springer. 1 ed. 2000.

19. Cook RJ, Lawless JF. Analysis of repeated events. *Statistical Methods in Medical Research* 2002; 11: 141-166.

20. Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events*. Springer, New York. 1 ed. 2007.

21. ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use . Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (E9 (R1)). 2019.

22. Wu L, Cook RJ. The design of intervention trials involving recurrent and terminal events. *Statistics in Biosciences* 2013; 5(2): 261–285.

23. Prentice R, Williams B, Peterson A. On the regression analysis of multivariate failure time data. *Biometrika* 1981; 68: 373-79.

24. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* 1982; 10(4): 1100-1120.

25. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 1958; 53(282): 457–481.

26. Claggett BL, McCaw ZR, Tian L, et al. Quantifying Treatment Effects in Trials with Multiple Event-Time Outcomes. *NEJM Evidence* 2022: EVIDoa2200047.

27. Wei J, Mütze T, Jahn-Eimermacher A, Roger J. Properties of two while-alive estimands for recurrent events and their potential estimators. *Statistics in Biopharmaceutical Research* 2022: 1–11.

28. Mao L. Nonparametric inference of general while-alive estimands for recurrent events. *Biometrics* 2022.

29. Furberg JK, Andersen PK, Korn S, Overgaard M, Ravn H. Bivariate pseudo-observations for recurrent event analysis with terminal events. *Lifetime Data Analysis* 2021: 1–32.

30. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2020.

31. Holst KK, Scheike TH, Hjelmborg JB. The Liability Threshold Model for Censored Twin Data. *Computational Statistics and Data Analysis* 2016; 93: 324-335.

32. Scheike TH, Holst KK, Hjelmborg JB. Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Analysis* 2014; 20(2): 210-233.

How to cite this article: Furberg JK., Scheike T. , Andersen PK., and Ravn H. (2022), Simulation-based sample size calculations for recurrent events with competing risks, *Pharmaceutical Statistics*, *xxxxxxxxx*.

APPENDIX

A DISTRIBUTION OF RECURRENT EVENTS AND DEATHS ACCORDING TO TREATMENT

Supplementary to Tables 2 and 4, the total number of recurrent events and per subject averaged across simulations and the number of deaths per treatment group is illustrated for $\beta = -0.2$. For both tables, the results are based on 1000 simulations. For each table, the total number (per id) of recurrent events per treatment, Z, is averaged across the simulations. The number of deaths per treatment, Z, is also averaged across the simulations. Since $\beta = -0.2$, the treatment reduces the number of recurrent events. When $\gamma = 0$, there is no effect of treatment on death. When $\gamma > 0$, treatment is associated with an increased mortality rate. Conversely, when $\gamma < 0$, treatment is associated with a decreased mortality rate. The settings corresponding to Table 2 are illustrated in Table A1. The settings corresponding to Table 4 are illustrated in Table A2.

β	γ	n	Treatment, Z	Total number of events (per id)	Total number of deaths
-0.2	0	100	0	250.83 (5.24)	18.82
			1	204.86 (4.32)	18.68
		200	0	501.66 (5.24)	37.42
			1	409.29 (4.31)	37.64
		500	0	1251.40 (5.23)	93.82
			1	1021.36 (4.31)	93.76
-0.2	-0.2	100	0	250.62 (5.26)	18.74
			1	205.80 (4.35)	16.18
		200	0	496.74 (5.20)	37.58
			1	410.87 (4.35)	32.48
		500	0	1250.51 (5.23)	94.15
			1	1027.18 (4.36)	80.91
-0.2	0.2	100	0	249.79 (5.25)	18.83
			1	206.20 (4.30)	21.80
		200	0	499.30 (5.22)	37.50
			1	410.33 (4.29)	43.31
		500	0	1254.17 (5.24)	93.97
			1	1020.34 (4.28)	107.81

TABLE A1 The additional assumptions behind the simulations are identical to those in Table 2 described in Section 4.

β	γ	n	Treatment, Z	Total number of events (per id)	Total number of deaths
-0.2	0	5000	0	234.36 (0.093)	234.36
			1	189.09 (0.076)	234.29
		7500	0	350.56 (0.093)	347.15
			1	284.35 (0.076)	350.61
		10000	0	463.57 (0.093)	468.90
			1	380.15 (0.076)	469.39
-0.2	-0.2	5000	0	231.74 (0.093)	233.75
			1	190.05 (0.076)	193.47
		7500	0	346.70 (0.092)	350.57
			1	284.79 (0.076)	289.75
		10000	0	464.07 (0.093)	468.76
			1	379.18 (0.076)	386.47
-0.2	0.2	5000	0	231.40 (0.093)	234.28
			1	189.61 (0.076)	283.49
		7500	0	347.73 (0.093)	351.09
			1	284.15 (0.076)	423.61
		10000	0	462.91 (0.093)	466.74
			1	379.70 (0.076)	565.45

TABLE A2 The additional assumptions behind the simulations are identical to those in Table 4 described in Section 5.

Vignette for simpowerrecurrent

Title Vignette for R-package 'simpowerrecurrent': *simpowerrecurrent: Simulation-based sample size calculations for recurrent events with competing deaths*

Authors Julie K. Furberg

URL `https://github.com/JulieKFurberg/simpowerrecurrent`

simpowerrecurrent: Simulation-based sample size calculations for recurrent events with competing deaths

Introduction

Let $N^*(t)$ denote the expected number of recurrent events by time t and let D^* denote the time of death. Let C denote the time of censoring. Due to right-censoring, $N(t) = N^*(t \wedge C)$ and $D = D^* \wedge C$ is observed. Moreover, the censoring indicator is observed, $\delta = I(D^* \leq C)$. There is only a single binary treatment variable, Z. For each of the n subjects, the following is observed $X_i = \{N_i(\cdot), D_i, \delta_i, Z_i\}, i = 1, \ldots, n$. X_i are assumed to be independent and identically distributed replicates of $X = \{N(\cdot), D, \delta, Z\}$. It is assumed that C is independent of Z.

It is assumed that the proportional means model of Ghosh and Lin (2002) holds, such that

$$E(N^*(t) \mid Z) = \mu(t \mid Z) = \mu_0(t) \exp(\beta Z),$$

where $\mu_0(t)$ is the baseline mean function for recurrent events, and β is the effect of treatment, Z, on recurrent events.

Moreover, it is assumed that Cox's proportional hazard model (1972) hold for the terminal events, such that

$$\Lambda^D(t \mid Z) = \Lambda_0^D(t) \exp(\gamma Z),$$

where $\Lambda_0^D(t)$ is the cumulative baseline hazard for death, and γ is the effect of treatment, Z, on death.

In order to simulate a single data set according to these models, the following should be specified

- The total sample size, n. It is assumed that the randomisation is 1:1.
- A set of values for $(t, \mu_0(t))$, i.e. the expected number of events in the reference group at times t.
- A set of values for $(t, \Lambda_0^D(t))$, i.e. the cumulative hazard of death in the reference group at times t.
- The censoring rate through the trial, such that $\Lambda^C(t) = ct$. Here, c is supplied.
- The log-mean ratio, β, i.e. the effect of treatment on recurrent events.
- The log-hazard ratio, γ, i.e. the effect of treatment on death.
- Max enrollment day after randomisation, τ_a. Uniform accrual until τ_a is assumed.
- Total length of study duration, τ_{max}. Administrative censoring occurs at τ_{max}.

In order to perform a power calculation or equivalently estimate a sample size, the following should additionally be specified

- The total number of simulations
- The significance level, α

Download from Github

The functions can be downloaded from GitHub using the below code,

```
#require(devtools)
#devtools::install_github("JulieKFurberg/simpowerrecurrent", force = TRUE)
require(simpowerrecurrent)

# A couple of extra packages
require(ggplot2); require(survival); require(mets); require(dplyr)
```

Simulation of a single data set

The following example displays how to simulate data from the above model using the required input parameters.

Here, we assume that $n = 100$, and that

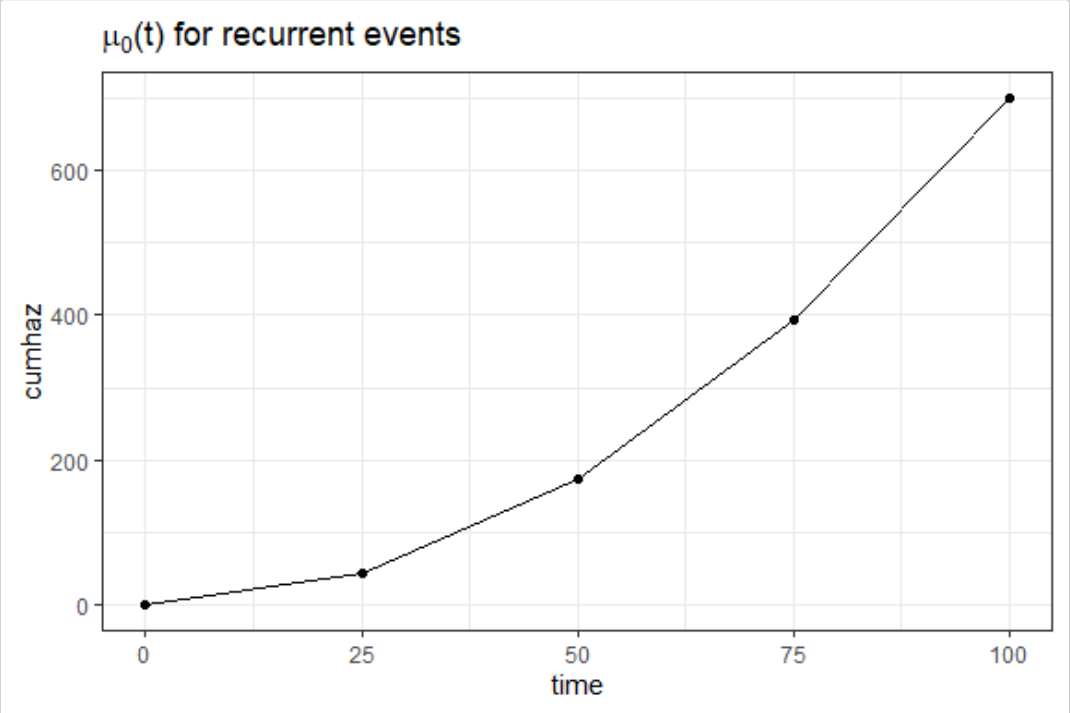
$$\mu(t \mid Z) = 0.07t^2 \exp(-0.1Z), \quad \Lambda^D(t \mid Z) = 0.05t \exp(-0.1Z)$$

Moreover, it is assumed that the cumulative hazard of being censored during the trial is,

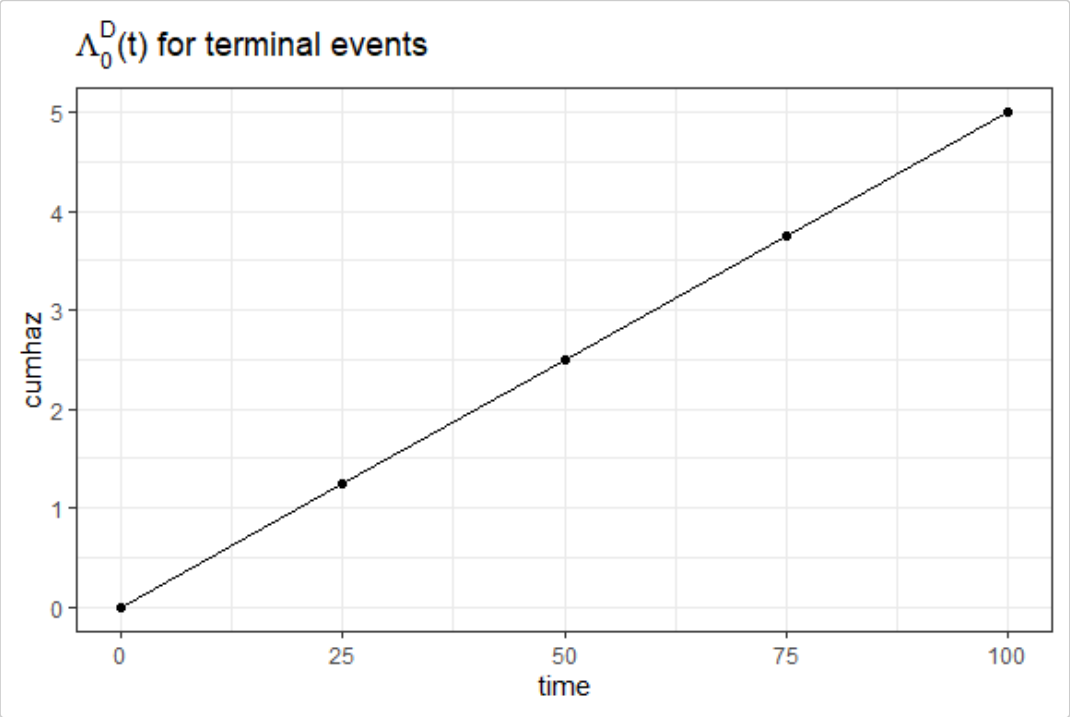
$$\Lambda^C(t) = 0.03t.$$

Furthermore, there is uniform accrual of the 100 subjects during 10 days. The study closed after 30 days from the first enrollment.

```
mu0 <- function(time, a){a  * time^2}
LamD0 <- function(time, b){b * time}

times <- seq(0, 100, by = 25)

cumhaz_mu <- data.frame(time = times,
                        cumhaz = mu0(time = times, a = 0.07))

cumhaz_S <- data.frame(time = times,
                       cumhaz = LamD0(time = times, b = 0.05))
# Visual
ggplot(aes(x = time, y = cumhaz), data = cumhaz_mu) +
  geom_line() + ggtitle(expression(mu[0]*"(t) for recurrent events")) +
  geom_point() + theme_bw()
```



```
ggplot(aes(x = time, y = cumhaz), data = cumhaz_S) +
  geom_line() + ggtitle(expression(Lambda[0]^D*"(t) for terminal events")) +
  geom_point() + theme_bw()
```




```
# Simulating a single data set
set.seed(1234)
sim1 <- simrecurprop(n = 100,
                     beta = -0.1,
                     gamma = -0.1,
                     mu0 = cumhaz_mu,
                     Lam0D = cumhaz_S,
                     crate = 0.03,
                     accrualtime = 10,
                     admincens = 30)

head(sim1)
#>   id    start      stop status Z
#> 1  1 8.329633  8.513353      1 0
#> 2  1 8.513353  8.598851      1 0
#> 3  1 8.598851  8.645844      1 0
#> 4  1 8.645844  9.219162      1 0
#> 5  1 9.219162  9.234392      1 0
#> 6  1 9.234392 10.357293      1 0

# Overview of the data set
with(sim1, table(Z, status))
#>    status
#> Z     0   1   2
```

```
#>   0  25 922  30
#>   1  20 906  25
```

In the output data set, the following variables are included,

- `id`: The subject id
- `start`: The start time of the record for individual i. Counting process style
- `stop`: The stop time of the record for individual i. Counting process style
- `status`: Status at the stopping time for subject i for record j.
- `z`: The binary treatment covariate

Estimation of power for previous example

```
simres1 <- powerest(nsims = 100,
                    n = 100,
                    beta = -0.1,
                    gamma = -0.1,
                    mu0 = cumhaz_mu,
                    Lam0D = cumhaz_S,
                    alpha = 0.05,
                    crate = 0.03,
                    accrualtime = 10,
                    admincens = 30)

head(simres1$resmat)
#>              beta    sebeta reject?       pval
#> [1,] -0.008535714 0.1703983       0 0.96004850
#> [2,] -0.311417513 0.1466871       1 0.03375334
#> [3,]  0.001625712 0.1514896       0 0.99143766
#> [4,] -0.356570420 0.1492128       1 0.01686321
#> [5,] -0.218203859 0.1600075       0 0.17265874
#> [6,] -0.238368925 0.1640195       0 0.14614148

simres1$power
#> [1] 0.09
simres1$betamean
#> [1] -0.04806245
simres1$betasemean
#> [1] 0.1555369
```

The results from fitting a Ghosh and Lin model to each simulated data set is contained in the data.frame, `resmat`, with the columns `beta`, `sebeta`, `reject?` and `pval`. There is one row per simulation in `resmat`. The results correspond to the estimated $\hat{\beta}$ and $\text{se}(\hat{\beta})$ from the Ghosh and Lin model as well as the decision to reject the null or not (alongside a two-sided p-value). Here, the hypotheses of interest are,

$$H_0 : \beta = 0, \quad H_a : \beta \neq 0.$$

The approximate power is contained in `power`. The average of all $\hat{\beta}$ across simulations is contained in `betamean`. The average of the standard errors for $\hat{\beta}$ across simulations is contained in `betasemean`.

Manuscript IV

Title *Marginal models for recurrent events under covariate dependent censoring*

Authors Julie K. Furberg, Per K. Andersen and Henrik Ravn

Details In preparation

Marginal models for recurrent events with covariate dependent censoring

Julie K. Furberg, Per K. Andersen and Henrik Ravn

March 28, 2023

Abstract

Marginal models for recurrent events can rely on an assumption of completely independent censoring. That is, the recurrent event process should be completely independent of the censoring process. In some situations, it would desirable to relax this assumption to conditionally independent censoring given covariates. This would allow censoring to depend on observed baseline covariates. An example could be inference on recurrent events from a randomised controlled trial with or without the impact of competing risks. The initial assumption may be that censoring is independent of treatment which may be relaxed to conditionally independent censoring given treatment and potentially additional baseline covariates. This paper discusses proportional means regression models for recurrent events with or without terminal events adjusted for covariate dependent censoring using inverse censoring weights. Moreover, it proposes to formulate a regression model for the expected number of recurrent events computed using pseudo-observations based on inverse censoring weighted non-parametric estimators. The performance and use of the methods will be illustrated through simulation studies. ***Intention: Furthermore, the methods are applied to two data examples: recurrent hospital admissions among patients with colorectal cancer and recurrent cardiovascular events among patients with diabetes.***

1 Introduction

The statistical analysis of recurrent events is an area of interest for capturing treatment effects from randomised controlled trials. Current practice focuses on the analysis of first events. In light of this, the analysis of recurrent events may bring a better understanding of disease progression, treatment effects and utilization of data. Therefore, pharmaceutical companies and regulatory agencies have expressed an interest in statistical analysis of recurrent events (Akacha et al. (2018), EMA (2020b), FDA (2019) and Rogers et al. (2014)). Consensus on which treatment effect(s) and statistical framework to prefer is still lacking. We suggest recurrent event models which capture a clinically relevant treatment effect.

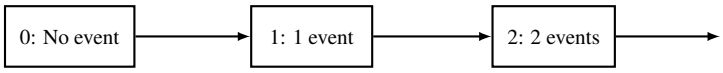


Figure 1: A recurrent event process.

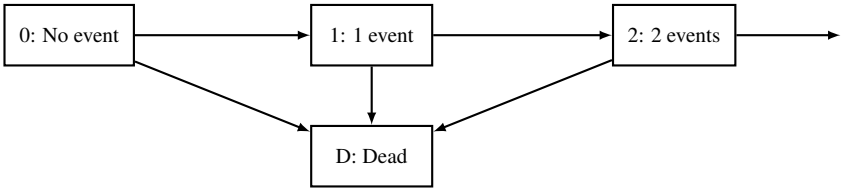


Figure 2: A recurrent event process with a terminal event.

Multi-state models are a powerful framework for describing the underlying processes of interest. Cook and Lawless (2018) provide a detailed overview and introduction to multi-state models. Bühler et al. (2022) discuss their power in describing treatment effects, estimands, for randomised controlled trials. The processes of interest for this paper may be illustrated using two multi-state models; a recurrent event process with and without a terminal event. These multi-state models are illustrated in Figures 1 and 2. With no or negligible mortality, Figure 1 may adequately describe the reality. With high mortality rates, reality is better reflected by Figure 2.

Cook and Lawless (2007) recommend focusing on marginal features of the recurrent event process if the goal is to obtain a simple treatment effect from randomised trials. In this spirit, this paper focuses on the marginal expected number of recurrent events. Non-parametric estimators of this quantity has been put forward (see Lawless and Nadeau (1995), Cook and Lawless (1997), and Ghosh and Lin (2000)). Semi-parametric regression models have also been developed for the situation with or without terminal events (see Lin et al. (2000) and Ghosh and Lin (2002)). In some cases, the aforementioned methods rely on an assumption of independence between the recurrent event and censoring process. The purpose of this paper is to discuss extensions of the existing methods to allow for covariate dependent censoring. This has been discussed earlier by Cook, Lawless, et al. (2009) and Ghosh and Lin (2002). We will discuss these non-parametric estimators and semi-parametric regression models for the marginal mean, both based on inverse censoring weights. Moreover, a method based on pseudo-observations of the marginal mean computed based on weighted non-parametric estimators will be proposed. This is an extension of the method considered in Andersen, Angst, et al. (2019) and Furberg et al. (2023).

Assume that a confirmatory analysis of recurrent events has been specified using, e.g., a proportional means model under an assumption of independent censoring. Then, the suggested methods will be valuable as sensitivity analyses that explore departures from an assumption of completely independent censoring by allowing covariate dependent censoring. Alternatively, marginal models which accommodate conditionally independent censoring given covariates may be chosen as candidates for the primary analyses. This would more closely resemble the intensity-based models that automatically accommodate conditionally independent censoring given covariates as specified as the history of the process.

The structure of the paper is as follows. First, the mathematical framework will be introduced and discussed. Subsequently, the methods will be illustrated using a simulation study. Finally, the methods will be applied to two real data sets.

2 Mathematical framework

Let $N^*(t)$ denote the number of recurrent events by time t. Assume that Z denotes p baseline covariates. The mathematical framework accommodates both the situation with or without terminal events. Let D^* denote the survival time. Without terminal events, D^* never occurs such that $D^* = \infty$. With terminal events, $N^*(t)$ remains constant after D^*. Let C denote the censoring time. Due to right-censoring, both $N^*(t)$ and D^* are incompletely observed. Thus, the observed data consists of $X = \{N(\cdot), D, \Delta, Z\}$ where $N(t) = N^*(t \wedge C)$, $D = D^* \wedge C$ and $\Delta = I(D^* \leq C)$. For each individual, $i = 1, \ldots, n$, we observe the data $X_i = \{N_i(\cdot), D_i, \Delta_i, Z_i\}$ which are independent replicates of X. The parameter of interest is the expected number of recurrent events,

$$\mu(t) = E(N^*(t)) = \int_0^t S(u^-)\, dR(u), \tag{1}$$

where $S(t) = P(D^* > t)$ and $dR(t) = E(dN^*(t) \mid D^* \geq t)$. Without terminal events, equation (1) reduces to $\mu(t) = \int_0^t E(dN^*(t))$.

Non-parametric inference Without terminal events and under independent censoring, Lawless and Nadeau (1995) argue that an approximately unbiased non-parametric estimator of $\mu(t)$ is the Nelson-Aalen estimator (Aalen (1978), Nelson (1969), Nelson (1972)). With terminal events, Cook and Lawless (1997) and Ghosh and Lin (2000) suggest to estimate $\mu(t)$ using the non-parametric plug-in estimator,

$$\widehat{\mu}(t) = \int_0^t \widehat{S}(u^-)\, d\widehat{R}(u), \tag{2}$$

where $\widehat{S}(t)$ denotes the Kaplan-Meier estimator of $S(t)$ and $\widehat{R}(t)$ denotes the Nelson-Aalen estimator of $R(t)$ (Kaplan and Meier (1958)). Without terminal events, $\widehat{\mu}(t) = \int_0^t d\widehat{R}(u) = \widehat{R}(t)$. Both estimators, however, rely an assumption of independent censoring. That is, C is completely independent of $N^*(\cdot)$ such that $E(N^*(t) \mid C \geq t) = E(N^*(t))$.

Cook, Lawless, et al. (2009) suggest extensions to non-parametric estimators for recurrent events that allow for dependent censoring using inverse probability of censoring weights (IPCW). They suggest a non-parametric estimator based on IPCW for the rate and mean function. To that end, let $C_i(t) = I(t \leq C_i)$ denote the censoring process for individual i. Moreover, assume that the censoring probability for individual i only depends on the history until up to time t, $G_i(t) = P(C_i \geq t \mid \mathscr{H}_i(\infty)) = P(C_i \geq t \mid \mathscr{H}_i(t^-))$. Then, the weighted rate function is given by

$$d\widehat{\mu}(t) = \frac{\sum_{i=1}^n C_i(t) dN_i(t)/\widehat{G}_i(t)}{\sum_{i=1}^n C_i(t)/\widehat{G}_i(t)},$$

where $\widehat{G}_i(t)$ denotes an estimate of $G_i(t)$ obtained by specifying a censoring model. The mean function, $\widehat{\mu}(t) = \int_0^t d\widehat{\mu}(u)$, corresponds to a weighted Nelson-Aalen estimate. This solves the following estimating equations,

$$\sum_{i=1}^{n} U_i(t) = \sum_{i=1}^{n} \frac{C_i(t)}{G_i(t)} \{dN_i(t) - d\mu(t)\} = 0.$$

This weighted Nelson-Aalen type estimator may be used to compute the marginal mean function under covariate dependent censoring by specifying the censoring distribution conditional on the covariates. Of note, this implies that the contribution of all individuals with events are re-weighted according to the censoring probability.

Proportional means regression model Lin et al. (2000) suggest a semi-parametric multiplicative regression model for recurrent events. The proportional rates model states that,

$$d\mu(t \mid Z(t)) = E(dN^*(t) \mid Z(t)) = d\mu_0(t) \exp(\beta^T Z(t)), \tag{3}$$

where $d\mu_0(t)$ denotes the unspecified baseline function. With no time-varying covariates, $Z(t) = Z$, and the model simplifies to

$$\mu(t \mid Z) = E(N^*(t) \mid Z) = \mu_0(t) \exp(\beta^T Z). \tag{4}$$

This is known as the proportional means model. We denote this model by the LWYY model. The estimating equations of this model are given by,

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(u) - \bar{Z}(\beta, u)\} \, dN_i(u),$$

with

$$\bar{Z}(\beta, u) = \frac{\sum_{j=1}^{n} Y_i(t) Z_i(t) \exp(\beta^T Z_i(t))}{\sum_{j=1}^{n} Y_i(t) \exp(\beta^T Z_i(t))},$$

where $Y_i(t) = I(C_i \geq t)$ denotes the at-risk process or equivalently the censoring process for individual i ($C_i(t) = I(C_i \geq t)$). The solution to $U(\beta) = 0$ is denoted $\widehat{\beta}$. The censoring mechanism is assumed to be conditionally independent given covariates such that

$$E(dN^*(t) \mid Z(t), C \geq t) = E(dN^*(t) \mid Z(t)), \tag{5}$$

for all $t \geq 0$ (Lin et al. (2000)). Ghosh and Lin (2002) suggested an extension of the proportional means or rates model for recurrent events with terminal events. The model adheres to both equation (3) and equation (4). With terminal events, $N^*(t)$ still records the observed recurrent events but the observation of recurrent events may be prevented by death. We denote it the GL model. The LWYY model targets the expected number of recurrent events in a world without mortality. Whereas, the GL model targets the expected number of recurrent events with mortality. That is, the reason for observing a low number of recurrent events in the presence of terminal events can be due to; a higher mortality rate, a lower rate of recurrent events while alive or a mixture. Thus, in the presence of terminal events, it will be important to consider mortality given covariates in addition to considering the number of recurrent events.

Inference for the GL model can be done using either inverse probability of censoring or survival weights. Technically, the adjustment for mortality is done by acknowledging that the censoring times are unknown for individuals who already died. The estimating equations using IPCW for the Ghosh-Lin model are given by,

$$U^C(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(u) - \bar{Z}^C(\beta, u) \right\} \widehat{w}_i^C(t) \, dN_i(u),$$

with

$$\bar{Z}^C(\beta, u) = \frac{\sum_{j=1}^{n} \widehat{w}_i^C(t) Z_i(t) \exp(\beta^T Z_i(t))}{\sum_{j=1}^{n} \widehat{w}_i^C(t) \exp(\beta^T Z_i(t))},$$

and

$$\widehat{w}_i^C(t) = \frac{I(C_i \geq D_i \wedge t) \widehat{G}(t)}{\widehat{G}(C_i \wedge D_i \wedge t)}.$$

The censoring distribution can be allowed to depend on covariates, $Z(t)$, but it is required that C and D^* are conditionally independent given $Z(\cdot)$. If we believe that $N^*(\cdot)$ and C are completely independent, such that $E(N^*(t) \mid C \geq t) = E(N^*(t))$,

it may be reasonable to base $\widehat{G}(t)$ on a Kaplan-Meier estimate of $G(t)$. Whereas, if $N^*(\cdot)$ and C are conditionally independent given $Z(\cdot)$, we may wish to base $\widehat{G}$ on, e.g., a Cox model, such that the censoring hazard is given by

$$\lambda^C(t \mid Z(t)) = \lambda_0^C(t)\exp(\delta^T Z(t)),$$

where $\lambda_0^C(t)$ denotes the baseline censoring hazard and δ the effect of covariates, $Z(t)$. Accordingly,

$$\widehat{G}(t \mid Z(t)) = \exp\left(-\int_0^t \exp(\widehat{\delta}^T Z(t))\, d\widehat{\Lambda}_0^C(t)\right),$$

where $\widehat{\delta}$ denotes the partial maximum likelihood estimator of δ and $\widehat{\Lambda}_0^C(t)$ denotes the Breslow estimate of the cumulative censoring hazard $\Lambda_0^C(t) = \int_0^t \lambda_0^C(u)\, du$ (Cox (1972) and Breslow (1974)). Then, the weights should be modified to

$$\widehat{w}_i^C(t) = \frac{I(C_i \geq D_i \wedge t)\widehat{G}(t \mid Z_i)}{\widehat{G}(C_i \wedge D_i \wedge t \mid Z_i)}.$$

Inference on recurrent events with dependent censoring is also discussed in Ghosh and Lin (2003).

3 Pseudo-observations

Andersen, Klein, et al. (2003) proposed using pseudo-observations for the analysis of life history data. For an introduction to pseudo-observations within survival analysis, see Andersen and Perme (2010). The parameter of interest for our marginal model is the expected number of recurrent events, that is

$$\theta = \mu(t) = E(N^*(t)).$$

Furberg et al. (2023) propose a bivariate model based on pseudo-observations for analysing recurrent events and terminal events simultaneously. For recurrent events, the parameter of interest is the expected number of recurrent events, θ. Furberg et al. discuss the computation of pseudo-observations with completely independent censoring. We suggest modifying this approach to use weighted non-parametric estimators that can accommodate covariate dependent censoring. We consider the IPCW estimator (see Section 2),

$$\widehat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{1}{\widehat{G}(u^-)}\, dN_i(u).$$

Hence, for a fixed $t \in [0, \tau]$, the pseudo-observation for individual i is given by

$$\widehat{\mu}_i(t) = n\widehat{\mu}(t) - (n-1)\widehat{\mu}_{-i}(t),$$

where $\widehat{\mu}(t)$ is the estimate based on the entire sample and $\widehat{\mu}_{-i}(t)$ is the leave-one-out estimator based on the entire sample where the observations from individual i is omitted. A generalised linear model which models the pseudo-observations conditional on covariates may be formulated as

$$g(\mu(t \mid Z)) = \beta^T Z,$$

where g denotes a link function. The link functions, $g(x) = x$ or $g(x) = \log(x)$, lead to the following two models,

$$\begin{aligned} &\text{Model 1:} \quad \mu(t \mid Z) = \beta^T Z, \\ &\text{Model 2:} \quad \log(\mu(t \mid Z)) = \beta^T Z. \end{aligned}$$

Model 1 targets mean differences and model 2 targets mean ratios. Model 2 corresponds to the proportional means model (the LWYY model with no competing risks and the GL model with competing risks). Model parameters are estimated using generalised estimating equations (GEE) as discussed in Furberg et al. (2023). Pseudo-observations may be computed at more than one time point to model interaction between covariates and time. Dependence within an individual is handled through the specification of the dependence in the GEE framework. Results indicate that little precision is gained by computing pseudo-observations at more than five time points (Furberg et al. (2023)).

Binder et al. (2014) discussed regression models for cumulative incidences based on pseudo-observations adjusted for covariate dependent censoring. In particular, they propose alternative ways of computing the leave-one-out estimator which are less computationally burdensome. In their simulations, computations are faster and bias is negligible when using the faster leave-one-out estimators. Three versions are natural to consider for the computation of the leave-one-out estimator of the marginal mean. For illustration, assume that the censoring distribution follows a Cox model, such that

$$G(t \mid Z_i) = \exp(-\Lambda_0^C(t)\exp(\delta^T Z_i)),$$

where Z_i denotes the baseline covariate(s) for individual i. In particular, the leave-one-out estimators may be computed in three ways. First, the censoring model may be fitted n times resulting in n estimates where individual i is left out, such that we obtain $\widehat{\delta}_{-i}$ and $\widehat{\Lambda}^C_{-i,0}(t)$. Here,

$$\widehat{\mu}_{-i}(t) = \frac{1}{n-1}\int_0^t \sum_{k\neq i} \frac{1}{\exp\left(-\widehat{\Lambda}^C_{-i,0}(u^-)\exp(\widehat{\delta}^T_{-i}Z_k)\right)}\, dN_k(u). \tag{6}$$

Alternatively, the censoring distribution may be fitted a single time which results in $\widehat{\delta}$ and $\widehat{\Lambda}^C_0(t)$. Hence,

$$\widehat{\mu}_{-i}(t) = \frac{1}{n-1}\int_0^t \sum_{k\neq i} \frac{1}{\exp\left(-\widehat{\Lambda}^C_0(u^-)\exp(\widehat{\delta}^T Z_k)\right)}\, dN_k(u). \tag{7}$$

Finally, a combination may be considered where the regression coefficient is estimated based on all subjects but the cumulative baseline hazard of censoring is computed n times. This leads to

$$\widehat{\mu}_{-i}(t) = \frac{1}{n-1}\int_0^t \sum_{k\neq i} \frac{1}{\exp\left(-\widehat{\widehat{\Lambda}}^C_{-i,0}(u^-)\exp(\widehat{\delta}^T Z_k)\right)}\, dN_k(u). \tag{8}$$

where $\widehat{\delta}$ is used in the computation of $\widehat{\widehat{\Lambda}}^C_{-i,0}(u^-)$. We expect the computation in equation (6) to be most expensive, followed by equation (8) and finally equation (7). Analogous considerations apply for considering a different censoring distribution.

Theoretical derivations and large sample properties of the suggested pseudo-observation method is beyond the scope of this article. The asymptotic distribution of the estimator of the regression coefficient should be explored in more detail. Overgaard et al. (2019) discuss large sample properties of pseudo-observation models for survival probabilities and cumulative incidences under covariate dependent censoring.

4 Simulation studies

The suggested methods will be illustrated using simulation. For generality, data is generated from a recurrent event, death and censoring process (see Figure 2). That is, we assume that there is non-negligible mortality. We imagine that the aim is to infer the effect of a single binary treatment Z on the expected number of recurrent events. This inference is complicated by the presence of death and censoring, which both is affected by Z. It is assumed that,

$$\mu(t \mid Z) = E(N^*(t) \mid Z) = \mu_0(t)\exp(\beta Z), \tag{9}$$

$$S(t \mid Z) = P(D^* > t \mid Z) = \exp\left(-\Lambda^D_0(t)\exp(\gamma Z)\right), \tag{10}$$

$$G(t \mid Z) = P(C > t \mid Z) = \exp\left(-\Lambda^C_0(t)\exp(\delta Z)\right). \tag{11}$$

We consider the following data generating frailty models as suggested by Ghosh and Lin (2002),

$$\lambda^D(t \mid Z, \nu) = \nu\exp(\gamma_D Z)\lambda^D_0,$$
$$dR(t \mid Z, \nu) = \nu S_0(t \mid \nu)^{1-\exp(\gamma_D Z)}\exp(\beta Z)\, dt,$$

where $S_0(t \mid \nu) = \exp\left(-\int_0^t \lambda^D(u \mid Z = 0, \nu)\, du\right)$ and $\nu > 0$ is a frailty term that is assumed to be generated from a positive stable distribution with Laplace transform $\exp(-\nu^\rho)$, $\rho \in (0, 1]$. Marginalizing over the random effect leads to,

$$\mu_0(t) = \frac{1}{\lambda^D_0}\left(-\exp\left(-\left(\lambda^D_0 t\right)^\rho\right) + 1\right), \quad \Lambda^D_0(t) = \left(\lambda^D_0 t\right)^\rho, \quad \gamma = \gamma_D\rho.$$

Moreover, it is assumed that $\Lambda^C_0(t) = \lambda^C_0 t$. The true censoring distribution follows a Cox model. Thus, inference from a Ghosh-Lin model where the weights are calculated assuming completely independent censoring (Kaplan-Meier used to estimate $\widehat{G}(t)$) will be biased. If the censoring model is correctly specified as a Cox model, inference from the Ghosh-Lin model with updated weights should again be unbiased. Additional administrative censoring was imposed at time 8. A total of 1000 simulations will be done per parameter setting. We explore $\delta = \{0, 0.2, 0.5\}$ with $n = \{100, 200, 500\}$, $\beta = \{0, 0.2\}$, $\gamma = \{0, 0.2\}$, and $\rho = 1$. It is assumed that $\lambda^C_0 = 0.2$ and $\lambda^D_0 = 0.25$. For these settings, the average number of recurrences is 2-3 recurrent events per individual.

Proportional means regression model For each simulated data set, three Ghosh-Lin type of models have been fitted; Ghosh-Lin model 1 (GL 1) with weights based on a Kaplan-Meier estimate of the censoring distribution $\widehat{G}(t)$, Ghosh-Lin model 2 (GL 2) with weights based on a stratified Cox model with $\widehat{G}_Z(t) = \exp(-\widehat{\Lambda}^C_{0Z}(t))$, and Ghosh-Lin model 3 (GL 3) with weights based on a Cox model with $\widehat{G}(t \mid Z) = \exp(-\widehat{\Lambda}^C_0(t)\exp(\widehat{\beta}Z))$. Models based on weights from GL 2 or 3 are expected to capture the dependent censoring imposed in the simulation procedure.

Table 1 displays the results from the simulation study for the three types of Ghosh-Lin models. The first Ghosh-Lin model (GL 1) is only unbiased while $\delta = 0$ as the censoring distribution is equal for the two treatments. This is captured by estimating the censoring distribution using the Kaplan-Meier estimator. If $\delta > 0$, GL 1 becomes biased whereas the models based on different censoring distributions for the two treatment groups (GL 2 and 3) are still unbiased. Results are similar for GL 2 and GL 3 where the censoring distribution is either assumed to follow a Cox model given treatment or the censoring distribution is estimated separately per treatment group.

					GL 1: $\widehat{G}(t)$ **KM**			**GL 2:** $\widehat{G}_Z(t)$ **Strat. Cox**			**GL 3:** $\widehat{G}(t \mid Z)$ **Cox**		
δ	β	γ	ρ	n	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$
0	0.2	0	1	100	0.200	0.200	0.192	0.202	0.196	0.193	0.202	0.195	0.193
				200	0.195	0.135	0.137	0.194	0.132	0.137	0.195	0.132	0.137
				500	0.201	0.088	0.087	0.201	0.086	0.087	0.201	0.086	0.087
		0.2		100	0.204	0.205	0.201	0.203	0.203	0.201	0.206	0.201	0.201
				200	0.201	0.145	0.143	0.200	0.141	0.143	0.201	0.141	0.143
				500	0.196	0.086	0.091	0.194	0.083	0.091	0.195	0.084	0.091
0.2	0.2	0	1	100	0.174	0.192	0.192	0.199	0.190	0.193	0.199	0.188	0.192
				200	0.172	0.139	0.137	0.198	0.136	0.137	0.198	0.136	0.137
				500	0.178	0.088	0.087	0.204	0.086	0.087	0.204	0.086	0.087
		0.2		100	0.159	0.208	0.203	0.184	0.205	0.203	0.187	0.203	0.203
				200	0.174	0.143	0.144	0.199	0.139	0.144	0.200	0.139	0.144
				500	0.169	0.091	0.091	0.196	0.088	0.091	0.196	0.089	0.091
0.5	0.2	0	1	100	0.125	0.197	0.194	0.187	0.195	0.194	0.192	0.193	0.193
				200	0.133	0.139	0.138	0.199	0.135	0.138	0.201	0.135	0.138
				500	0.125	0.089	0.088	0.195	0.086	0.088	0.194	0.087	0.088
		0.2		100	0.116	0.216	0.203	0.181	0.208	0.203	0.187	0.209	0.203
				200	0.124	0.147	0.145	0.195	0.142	0.144	0.199	0.142	0.145
				500	0.115	0.092	0.092	0.191	0.091	0.093	0.191	0.090	0.093

Table 1: Results from simulation study based on 1000 simulations for Ghosh-Lin model types. The average of the $\widehat{\beta}$ parameter estimates is denoted $E(\widehat{\beta})$. Moreover, $E(\text{se}(\widehat{\beta}))$ denotes the average of the estimated standard errors of the $\widehat{\beta}$ estimates. The standard deviation of the $\widehat{\beta}$ estimates is denoted $SD(\widehat{\beta})$. Censoring times follow a Cox model per equation (11).

				GL 1: $\widehat{G}(t)$ **KM**			**GL 2:** $\widehat{G}_Z(t)$ **Strat. Cox**			**GL 3:** $\widehat{G}(t \mid Z)$ **Cox**		
β	γ	ρ	n	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$
0.2	0	1	100	0.020	0.227	0.212	0.175	0.217	0.206	0.220	0.211	0.206
			200	0.028	0.151	0.152	0.188	0.145	0.148	0.223	0.141	0.148
			500	0.031	0.100	0.097	0.196	0.095	0.095	0.228	0.093	0.095
	0.2		100	−0.007	0.244	0.227	0.165	0.232	0.221	0.223	0.221	0.221
			200	0.004	0.163	0.162	0.182	0.158	0.157	0.230	0.152	0.158
			500	0.009	0.104	0.103	0.195	0.099	0.100	0.236	0.097	0.101

Table 2: Results from simulation study based on 1000 simulations for Ghosh-Lin model types. The average of the $\widehat{\beta}$ parameter estimates is denoted $E(\widehat{\beta})$. Moreover, $E(\text{se}(\widehat{\beta}))$ denotes the average of the estimated standard errors of the $\widehat{\beta}$ estimates. The standard deviation of the $\widehat{\beta}$ estimates is denoted $SD(\widehat{\beta})$. Censoring times follow piece-wise constant hazard ratios, see equation (12).

Instead of adhering to a Cox model, the censoring distribution could follow a piece-wise constant hazard ratios model. To that end, we let

$$\lambda^C_0(t) = 0.1,$$
$$\lambda^C_1(t) = \begin{cases} 0.7, & \text{for} \quad 0 \le t < 2, \\ 0.1, & \text{for} \quad 2 \le t < 5, \\ 0.5, & \text{for} \quad t \ge 5, \end{cases}$$

such that

$$G_Z(t) = G(t \mid Z) = P(C > t \mid Z) = \exp\left(-\lambda^C_Z(t)t\right), \quad \text{for} \quad Z \in \{0,1\}. \tag{12}$$

For such a situation, inverse censoring weights based on a model stratified for Z would be correct (GL 3) since weights based on a Cox model would be mis-specified (GL 2). This has been explored in a simulation study where all other assumptions are kept fixed. Table 2 displays the results. GL 1 is very biased and GL 3 also exhibits some bias. The Ghosh-Lin model based on the stratified censoring model (GL 2) performs best.

Pseudo-observation regression model Alternatively, a model for the marginal mean based on pseudo-observations could have been applied to the simulated data. Thus, three different models based on pseudo-observations have been fitted for each simulated data set. All three models are based on pseudo-observations computed for $\widehat{\mu}(t)$ at time 7 using a log-link function (Model 2). The weighted Nelson-Aalen estimators which are used in the computations of the pseudo-observations varies between the three models. The inverse weights are as for the Ghosh-Lin type models based on; Pseudo-observation model 1 (Pseudo 1) using a Kaplan-Meier estimate for $\widehat{G}(t)$, Pseudo-observation model 2 (Pseudo 2) using a Stratified Cox model with $\widehat{G}_Z(t) = \exp(-\widehat{\Lambda}^C_{0Z}(t))$ and Pseudo-observation model 3 (Pseudo 3) using a Cox model with $\widehat{G}(t \mid Z) = \exp(-\widehat{\Lambda}^C_0(t)\exp(\widehat{\beta}Z))$. To ease the computational burden, the leave-one-estimator suggested in equation (7) has been used for computation of the pseudo-observations.

					Pseudo 1: $\widehat{G}(t)$ **KM**			**Pseudo 2:** $\widehat{G}_Z(t)$ **Strat. Cox**			**Pseudo 3:** $\widehat{G}(t \mid Z)$ **Cox**		
δ	β	γ	ρ	n	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$
0	0.2	0	1	100	0.202	0.245	0.233	0.201	0.237	0.233	0.202	0.230	0.232
				200	0.196	0.167	0.165	0.198	0.164	0.165	0.197	0.161	0.165
		0.2		100	0.192	0.265	0.247	0.192	0.259	0.248	0.190	0.252	0.247
				200	0.202	0.175	0.175	0.203	0.170	0.175	0.199	0.167	0.174
0.2	0.2	0	1	100	0.164	0.269	0.242	0.201	0.256	0.243	0.198	0.249	0.243
				200	0.159	0.176	0.171	0.196	0.171	0.172	0.197	0.169	0.172
		0.2		100	0.154	0.267	0.253	0.196	0.258	0.254	0.189	0.253	0.253
				200	0.153	0.181	0.193	0.195	0.175	0.181	0.193	0.171	0.181
0.5	0.2	0	1	100	0.089	0.264	0.254	0.195	0.252	0.260	0.188	0.247	0.258
				200	0.093	0.173	0.179	0.194	0.166	0.183	0.190	0.166	0.182
		0.2		100	0.077	0.286	0.263	0.192	0.271	0.268	0.180	0.261	0.265
				200	0.074	0.207	0.187	0.186	0.194	0.190	0.179	0.190	0.189

Table 3: Results from simulation study based on 1000 simulations for pseudo-observation model types. The average of the $\widehat{\beta}$ parameter estimates is denoted $E(\widehat{\beta})$. Moreover, $E(\text{se}(\widehat{\beta}))$ denotes the average of the estimated standard errors of the $\widehat{\beta}$ estimates. The standard deviation of the $\widehat{\beta}$ estimates is denoted $SD(\widehat{\beta})$. Censoring times follow a Cox model per equation (11).

				Pseudo 1: $\widehat{G}(t)$ **KM**			**Pseudo 2:** $\widehat{G}_Z(t)$ **Strat. Cox**			**Pseudo 3:** $\widehat{G}(t \mid Z)$ **Cox**		
β	γ	ρ	n	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$	$E(\widehat{\beta})$	$SD(\widehat{\beta})$	$E(\text{se}(\widehat{\beta}))$
0.2	0	1	100	−0.016	0.228	0.242	0.192	0.203	0.261	0.143	0.213	0.256
			200	−0.010	0.161	0.172	0.196	0.146	0.185	0.156	0.150	0.182
	0.2		100	−0.032	0.241	0.246	0.203	0.214	0.266	0.141	0.225	0.259
			200	−0.022	0.183	0.176	0.215	0.164	0.189	0.161	0.172	0.185

Table 4: Results from simulation study based on 1000 simulations for pseudo-observation model types. The average of the $\widehat{\beta}$ parameter estimates is denoted $E(\widehat{\beta})$. Moreover, $E(\text{se}(\widehat{\beta}))$ denotes the average of the estimated standard errors of the $\widehat{\beta}$ estimates. The standard deviation of the $\widehat{\beta}$ estimates is denoted $SD(\widehat{\beta})$. Censoring times follow piece-wise constant hazard ratios, see equation (12).

Table 3 displays the results from the simulation studies based on the pseudo-observation models where the true censoring distribution follows a Cox model. Pseudo 1 is only unbiased for $\delta = 0$ but becomes biased as $\delta > 0$. Pseudo 2 and 3 are unbiased for all $\delta \geq 0$. The results are very similar to those in Table 1 where the Ghosh-Lin models were considered. Table 4 displays the results for the pseudo-observation models based on a true piece-wise constant censoring hazard ratios per equation (12). Pseudo 1 is biased and Pseudo 3 also exhibits some bias. Pseudo 2 is unbiased as expected. Again, the results for the GL and pseudo-observation type models are very similar (Tables 2 and 4). In general, the models based on pseudo-observations have larger variance estimates compared to Ghosh-Lin type models. This is not surprising as the true model follows the Ghosh-Lin structure for the true mean and the pseudo-observation model imposes less assumptions. The variance for the pseudo-observation models can be reduced by computing pseudo-observations at several time points, e.g., at times $(5, 7, 10)$, but this is more computationally expensive (see Furberg et al. (2023)). ***Intention: Additional simulations for*** $n = 500$ ***in Tables 3 and 4 are intended for the final manuscript.***

5 Application

The methods discussed in this paper will be illustrated using two examples.

Application 1: Hospital readmission among colorectal cancer patients The first application considers hospital readmissions among patients diagnosed with colorectal cancer (González et al. (2005)). A total of 403 patients that were diagnosed with colorectal cancer between January 1996 and December 1998 were actively followed up until 2002 in Hospital Universitary in L'Hospitalet (Barcelona, Spain). The purpose was to investigate sex-based differences in hospital admission. This was an epidemiological study without randomisation but will be used for illustration of the methods in this paper. Of the 403 patients, 112 died during follow-up and 294 were censored. Information on whether patients received chemotherapy ('Yes', 'No') was available. The data set can be downloaded from the R-package `frailtypack` (Rondeau, Mazroui, et al. (2012), Rondeau and Gonzalez (2005), Król et al. (2017), and Rondeau, Gonzalez, et al. (2019)).

For this application, we focus on modelling the marginal expected number of hospital admissions per chemotherapy group. Figure 3 displays non-parametric estimates of the expected number of recurrent events ($\mu(t)$), survival function ($S(t)$) and censoring distribution ($G(t)$) per chemotherapy treatment group. The chemotherapy group has less expected hospital admissions, higher probability of surviving as well as a higher probability of being censored.

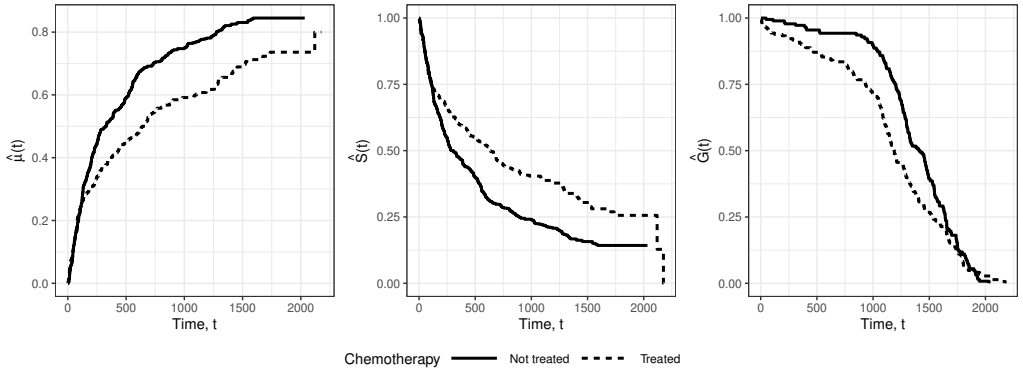


Figure 3: Non-parametric estimates based on the readmission data. Left figure: Marginal mean estimate of the expected number of hospital admissions per chemotherapy treatment computed using equation (2). Middle figure: Estimated survival function, $S(t) = P(D^* > t)$, per treatment computed using Kaplan-Meier. Right figure: Estimated censoring distribution, $G(t) = P(C > t)$, per treatment computed using Kaplan-Meier.

Three different Ghosh-Lin models are fitted to the data set to model the expected number of hospital admissions. Thus, the expected number of hospital admissions given treatment is assumed to follow the model,

$$E(N^*(t) \mid Z) = \mu(t \mid Z) = \mu_0(t)\exp(\beta Z),$$

where Z denotes treatment ($Z = 0$ corresponds to no chemotherapy and $Z = 1$ corresponds to chemotherapy). Three types of censoring weights are used in the estimation, corresponding to those discussed in the simulation studies, where the censoring distribution is estimated using a: 1. Kaplan-Meier estimate, 2. Stratified Cox model or 3. Cox model. Figure 4 displays the estimated censoring distributions for each situation. Situation 1 corresponds to assuming that $N^*(\cdot)$ and C are independent whereas situation 2 and 3 corresponds to a conditional independence between $N^*(\cdot)$ and C given Z. Table 5 displays the estimates derived from each model. Overall, the estimates look very similar. Models 2 and 3 produce almost identical results. Models 2 and 3 correspond to exploring if there is any covariate dependent censoring given treatment, and thus a different censoring pattern. The results indicate that this does not seem to be the case. Figure 5 displays the estimated marginal means per treatment group imposing the Ghosh-Lin models using the different weights. The estimates are almost indistinguishable. Assume that the Ghosh-Lin model type 1 (GL 1) was pre-specified as the primary analysis and either GL 2 or GL 3 as a sensitivity analysis exploring the independent censoring assumption. Then, this example corresponds to a situation where the results from the primary analysis (GL 1) seem robust compared to the results from the sensitivity analyses (GL 2 or GL 3).

The regression model based on pseudo-observations could also be fitted to the readmission data. Using a log-link, this model assumes that,

$$E(N^*(t) \mid Z) = \mu(t \mid Z) = \mu_0(t)\exp(\beta Z).$$

Model	$\widehat{\beta}$	se($\widehat{\beta}$)	exp($\widehat{\beta}$)	95% CI for exp(β)
GL 1 ($\widehat{G}(t)$ KM)	-0.520	0.167	0.595	[0.428; 0.826]
GL 2 ($\widehat{G}_Z(t)$ Strat. Cox)	-0.507	0.168	0.602	[0.433; 0.837]
GL 3 ($\widehat{G}(t \mid Z)$ Cox)	-0.507	0.168	0.602	[0.434; 0.836]

Table 5: Parameter estimates from Ghosh-Lin type models based on various censoring weights for the readmission data. CI: Confidence interval.

The pseudo-observations are computed at one time points, $t = (1000)$, and three time points, $t = (500, 1000, 1500)$ (see Figure 3). The censoring weights have been calculated in three ways using the fast leave-one-out estimator suggested in equation (7). Results are displayed in Table 6. The pseudo-observation models computed at $t = (1000)$ and $t = (500, 1000, 1500)$ produce similar results but the standard errors are smaller for the model based on three time points as expected. Overall, the results based on the pseudo-observation models are very similar to those based on the Ghosh-Lin models in Table 5.

Time points	Model	$\widehat{\beta}$	se($\widehat{\beta}$)	exp($\widehat{\beta}$)	95% CI for exp(β)
$t = (1000)$	Pseudo 1 ($\widehat{G}(t)$ KM)	-0.541	0.183	0.582	[0.407; 0.833]
	Pseudo 2 ($\widehat{G}_Z(t)$ Strat. Cox)	-0.539	0.183	0.583	[0.407; 0.835]
	Pseudo 3 ($\widehat{G}(t \mid Z)$ Cox)	-0.536	0.183	0.585	[0.409; 0.838]
$t = (500, 1000, 1500)$	Pseudo 1 ($\widehat{G}(t)$ KM)	-0.535	0.169	0.585	[0.421; 0.815]
	Pseudo 2 ($\widehat{G}_Z(t)$ Strat. Cox)	-0.530	0.169	0.589	[0.423; 0.820]
	Pseudo 3 ($\widehat{G}(t \mid Z)$ Cox)	-0.523	0.169	0.593	[0.426; 0.825]

Table 6: Parameter estimates from pseudo-observation type models based on various censoring weights for the readmission data. CI: Confidence interval.

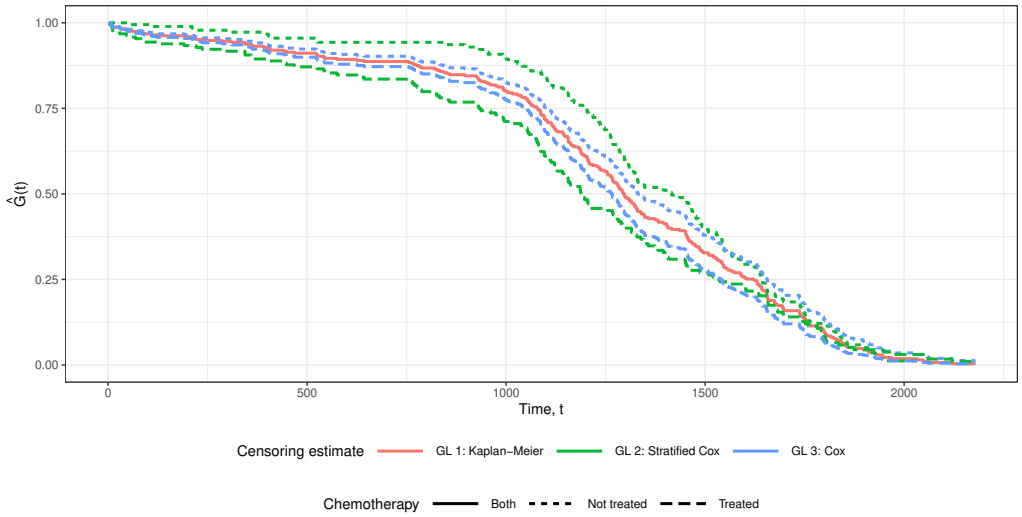


Figure 4: Probability of remaining uncensored, $G(t) = P(C > t)$, estimated using a Kaplan-Meier estimator, a Cox model given treatment group, a stratified Cox model per treatment, corresponding to the censoring distributions used in the inverse censoring weights for the three Ghosh-Lin and pseudo-observation models.

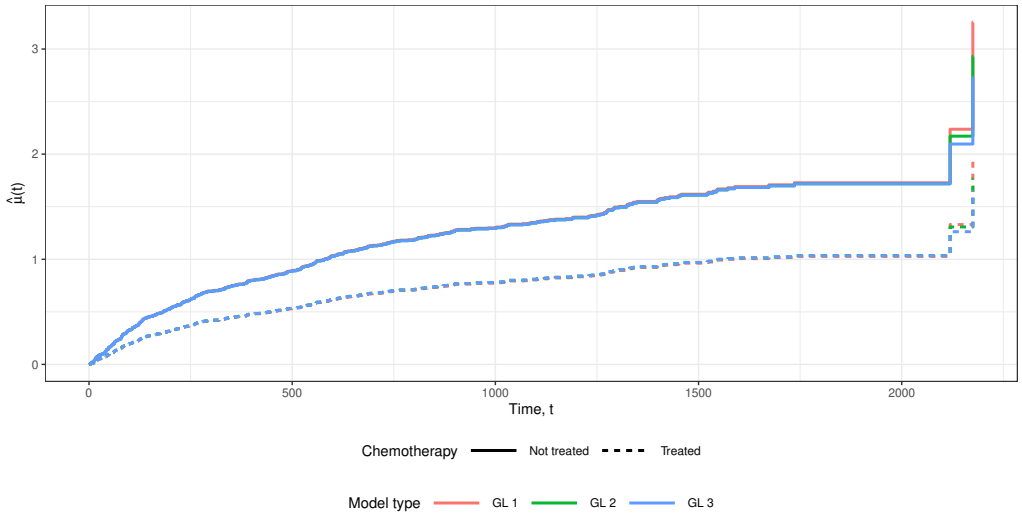


Figure 5: Estimated $\widehat{\mu}(t \mid Z) = \widehat{\mu}_0(t)\exp(\widehat{\beta}Z)$ for $Z = \{0,1\}$ for the three Ghosh-Lin models computed using different inverse censoring weights.

Application 2: Recurrent cardiovascular events for patients with diabetes ***Intention: Space for application to data from Novo Nordisk with covariate dependent censoring. Potentially, the LEADER data could be used with focus on recurrent cardiovascular events occurring while on-treatment focusing on differential treatment adherence for US versus non-US patients (Marso et al. (2016)). This needs to be explored in more detail.***

6 Discussion

The presented methods and models are intended as useful analyses for recurrent events that can explore departures from completely independent censoring. The suggested approaches accommodate covariate dependent censoring induced by baseline covariates. Emphasis has been placed on marginal models for recurrent events focusing on the expected number of recurrent events with or without the influence of terminal events. This is motivated by the intention to extract a simple treatment effect from recurrent event data from randomised controlled trials as discussed by Cook and Lawless (2007). Other treatment effects than the effect of being randomised may be explored which would entail the causality framework to adequately describe the target parameter(s) of interest. Su et al. (2022) discuss causal inference for recurrent event data without terminal events based on g-formula estimators, estimators based on inverse probability of treatment weights, doubly robust estimators as well as pseudo-observations. In order to describe an on-treatment estimand, censoring may be caused by treatment discontinuation leading to a potentially dependent censoring pattern (EMA (2020a)). Description of such an estimand may require both considerations from causal inference as well as dependent censoring as discussed in this paper.

The suggested marginal models for recurrent events qualify as candidates for pre-specified sensitivity analyses of recurrent events which explores the censoring assumption. This would be relevant for the Novo Nordisk clinical trial ZEUS (ClinicalTrials.gov (2023)). ZEUS is a randomised controlled trial exploring the effect of ziltivekimab versus placebo in a population with cardiovascular disease, chronic kidney disease and inflammation. A secondary confirmatory endpoint in this trial is a recurrent event endpoint, namely the number of heart failure hospitalisations or urgent heart failure visits. In line with the ICH E9 (R1) addendum on estimands, this implies that robustness of the secondary confirmatory analysis should be explored through a sensitivity analysis which targets the same estimand (EMA (2020a)).

Covariate dependent censoring has been introduced in several simulation studies either through a Cox model or with piece-wise constants hazard ratios of censoring accommodated by a stratified Cox model. With dependent censoring, standard models which assume completely uninformative censoring are biased. If the censoring distribution given baseline covariates are correctly specified, this bias disappears when the inverse censoring weights are updated correspondingly. The proportional means or rates regression model suggested by Ghosh and Lin (2002) was explored while focusing on different censoring models. Moreover, a regression model based on pseudo-observations obtained by a weighted Nelson-Aalen estimator was explored. This is an extension of the pseudo-observation regression model for recurrent

events discussed in Andersen, Angst, et al. (2019) and Furberg et al. (2023). Theoretical properties and large sample behaviour of the proposed pseudo-observation method should be explored in future work.

The models have been applied to two real data examples: hospital readmissions among patients with colorectal cancer and recurrent cardiovascular outcomes for patients with diabetes. Only the latter was a randomised trial comparing two treatments. The two examples illustrate two potential outcomes of the sensitivity analysis: either the results indicate robustness or not. For the first application to the hospital readmissions for patients with cancer, the results were very similar. Such a situation would indicate robustness. ***Intention: Whereas, for the second application, cardiovascular outcomes for patients with diabetes, the results differed more. This would be an example of the opposite, i.e., a situation which indicates some level of covariate dependent censoring being present to the extent that the results are impacted.***

Zhao et al. (2014) suggested a tipping-point style sensitivity analysis of time-to-event outcomes which imputes event times for individuals that drop-out prior to study closure. Imputations may be done differently in each treatment arm allowing for dependent censoring and thus 'penalizing' the active treatment arm relative to the placebo arm. The tipping point is the penalty required in order for the conclusion of the analysis to change (i.e., assuming that the analysis was met). A tipping-point type of analysis for recurrent events with terminal events is a topic for further research.

Extensions of the suggested pseudo-observation method to bivariate or higher dimensional measures that accommodates covariate dependent censoring are a topic for further research. An example could be focusing on the marginal mean and the survival function, $\theta = (\mu(t), S(t))$, or the marginal mean and cumulative incidences for causes 1 and 2, $\theta = (\mu(t), F_1(t), F_2(t))$, as suggested in Furberg et al. (2023).

References

Aalen, O (1978). "Nonparametric inference for a family of counting processes". In: *The Annals of Statistics*, pp. 701–726.

Akacha, M et al. (2018). *Request for CHMP Qualification Opinion: Clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses*. URL: `https://www.ema.europa.eu/en/documents/other/qualification-opinion-treatment-effect-measures-when-using-recurrent-event-endpoints-applicants_en.pdf` (visited on 03/28/2023).

Andersen, PK, J Angst, and H Ravn (2019). "Modeling marginal features in studies of recurrent events in the presence of a terminal event". In: *Lifetime Data Analysis* 25, pp. 681–695.

Andersen, PK, JP Klein, and S Rosthøj (2003). "Generalised linear models for correlated pseudo-observations, with applications to multi-state models". In: *Biometrika* 90, pp. 15–27.

Andersen, PK and MP Perme (2010). "Pseudo-observations in survival analysis". In: *Statistical Methods in Medical Research* 19, pp. 71–99.

Binder, N, TA Gerds, and PK Andersen (2014). "Pseudo-observations for competing risks with covariate dependent censoring". In: *Lifetime Data Analysis* 20.2, pp. 303–315.

Breslow, N (1974). "Covariance analysis of censored survival data". In: *Biometrics*, pp. 89–99.

Bühler, A, RJ Cook, and JF Lawless (2022). "Multistate models as a framework for estimand specification in clinical trials of complex processes". In: *Statistics in Medicine*.

ClinicalTrials.gov (2023). *ZEUS - A Research Study to Look at How Ziltivekimab Works Compared to Placebo in People With Cardiovascular Disease, Chronic Kidney Disease and Inflammation (ZEUS)*. URL: `https://clinicaltrials.gov/ct2/show/NCT05021835` (visited on 03/06/2023).

Cook, RJ and JF Lawless (1997). "Marginal Analysis of Recurrent Events and A Terminating Event". In: *Statistics in Medicine* 16, pp. 911–924.

— (2018). *Multistate Models for the Analysis of Life History Data*. 1st ed. CRC Press.

— (2007). *The Statistical Analysis of Recurrent Events*. 1st ed. Springer, New York.

Cook, RJ, JF Lawless, et al. (2009). "Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone". In: *Journal of the American Statistical Association* 104.485, pp. 60–75.

Cox, DR (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, pp. 187–220.

European Medicines Agency (EMA) (2020a). *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (Step 5)*. URL: `https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf` (visited on 03/15/2023).

— (2020b). *Qualification opinion of clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses*. URL: `https://www.ema.europa.eu/en/documents/other/qualification-opinion-clinically-interpretable-treatment-effect-measures-based-recurrent-event_en.pdf` (visited on 12/14/2020).

Food and Drug Administration (FDA) (2019). *Treatment for Heart Failure: Endpoints for Drug Development. Guidance for Industry. Draft Guidance.* URL: `https://www.fda.gov/media/128372/download` (visited on 01/31/2023).

Furberg, JK et al. (2023). "Bivariate pseudo-observations for recurrent event analysis with terminal events". In: *Lifetime Data Analysis* 29, pp. 256–287.

Ghosh, D and DY Lin (2002). "Marginal regression models for recurrent and terminal events". In: *Statistica Sinica* 12, pp. 663–688.
— (2000). "Nonparametric analysis of Recurrent Events and Death". In: *Biometrics* 56, pp. 554–562.
— (2003). "Semiparametric analysis of recurrent events data in the presence of dependent censoring". In: *Biometrics* 59.4, pp. 877–885.
González, JR et al. (2005). "Sex differences in hospital readmission among colorectal cancer patients". In: *Journal of Epidemiology & Community Health* 59.6, pp. 506–511.
Kaplan, EL and P Meier (1958). "Nonparametric estimation from incomplete observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481.
Król, A et al. (2017). "Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event". In: *Journal of Statistical Software* 81.3, pp. 1–52.
Lawless, JF and C Nadeau (1995). "Some simple robust methods for the analysis of recurrent events". In: *Technometrics* 37.2, pp. 158–168.
Lin, DY et al. (2000). "Semiparametric regression for the mean and rate functions of recurrent events". In: *Journal of the Royal Statistical Society* 62.4, pp. 711–730.
Marso, SP et al. (2016). "Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes". In: *New England Journal of Medicine* 375.4, pp. 311–322.
Nelson, W (1969). "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1.1, pp. 27–52.
— (1972). "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14.4, pp. 945–966.
Overgaard, M, ET Parner, and J Pedersen (2019). "Pseudo-observations under covariate-dependent censoring". In: *Journal of Statistical Planning and Inference* 202, pp. 112–122.
Rogers, JK et al. (2014). "Analysing recurrent hospitalisations in heart failure: a review of statistical methodology, with application to CHARM-Preserved". In: *European Journal of Heart Failure* 16.1, pp. 33–40.
Rondeau, V and JR Gonzalez (2005). "Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation". In: *Computer Methods and Programs in Biomedicine* 80.2, pp. 154–164.
Rondeau, V, JR Gonzalez, et al. (2019). *frailtypack: General Frailty Models: Shared, Joint and Nested Frailty Models with Prediction; Evaluation of Failure-Time Surrogate Endpoints*. R package version 3.0.3. URL: `https://CRAN.R-project.org/package=frailtypack`.
Rondeau, V, Y Mazroui, and JR Gonzalez (2012). "frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation". In: *Journal of Statistical Software* 47.4, pp. 1–28.
Su, C-L, RW Platt, and J-F Plante (2022). "Causal inference for recurrent event data using pseudo-observations". In: *Biostatistics* 23.1, pp. 189–206.
Zhao, Y et al. (2014). "A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring". In: *Journal of Biopharmaceutical Statistics* 24.2, pp. 229–253.

Acknowledgements

This research was conducted as a part of the PhD education of JK Furberg from which she received funding from Novo Nordisk A/S. JK Furberg and H Ravn are employed at Novo Nordisk A/S.

Software availability

`R` code for the models are available upon request to the corresponding author.