Flexible parametric pseudo-observations for right and interval censored time-to-event data with competing risks

Johansen, Martin Nygård

DOI (link to publication from Publisher):
10.5278/vbn.phd.med.00135

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

FLEXIBLE PARAMETRIC PSEUDO-OBSERVATIONS FOR RIGHT AND INTERVAL CENSORED TIME-TO-EVENT DATA WITH COMPETING RISKS

BY
MARTIN NYGÅRD JOHANSEN

DISSERTATION SUBMITTED 2020

AALBORG UNIVERSITY
DENMARK

Flexible parametric pseudo-observations for right and interval censored time-to-event data with competing risks

Ph.D. Dissertation

Martin Nygård Johansen

Dissertation submitted November 2020

Dissertation submitted: November 18, 2020

PhD supervisor: Professor Sam Riahi, MD, PhD
Department of Clinical Medicine
Faculty of Medicine
Aalborg University

Assistant PhD supervisors: Associate Professor Søren Lundbye-Christensen, MSc, PhD
Department of Clinical Medicine
Faculty of Medicine
Aalborg University

Associate Professor Jacob Moesgaard Larsen, MD, PhD
Department of Clinical Medicine
Faculty of Medicine
Aalborg University

PhD committee: Professor Martin Bøgsted (chair)
Aalborg University

Professor Per Kragh Andersen
Copenhagen University

Professor Paul C Lambert
University of Leicester

PhD Series: Faculty of Medicine, Aalborg University

Department: Department of Clinical Medicine

ISSN (online): 2246-1302
ISBN (online): 978-87-7210-844-5

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2020

Abstract

It is very common in medical research to study the time to occurrence of a particular event such as death or a specific disease in a group of individuals. The data that are gathered in such studies is almost inevitably subject to what is called right censoring so that, for some individuals, the event is only known to not have occurred within a certain amount of time. An even more inhibiting type of censoring, interval censoring, emerges when the status of the event is only assessed at a number of distinct time points. In this case, the only available information when the event is first observed is that it has occurred at some point since the last assessment with negative status. Interval censored data is very often also complicated by competing risks of other events that may occur and thus preclude the event of interest. Most statistical methods to handle interval censoring typically impose assumptions on the distribution of the actual event times to enable the use of parametric models, which have mathematical properties that make them easy to employ to interval censored data. However, this limits the way in which we can measure the association between potential risk factors and the event of interest using regression models. An alternative way of handling censoring is to transform the incomplete data to a set of pseudo-observations, which can be analyzed with a versatile family of regression models, called generalized linear models, to measure associations with a range of easily interpretable measures. This pseudo-observation approach is based on non-parametric methods to define the transformation of the data, but by applying an alternative technique based on splines, which are very flexible parametric functions, we are able to formulate a set of pseudo-observations for interval censored data. These parametric pseudo-observations can then be used in generalized linear models to estimate associations with risk factors using measures that are relevant in the specific context of each study.

Resumé

I medicinsk forskning er det meget almindeligt at undersøge tiden til en bestemt hændelse, såsom død eller en specifik sygdom, forekommer i en gruppe individer. De data, som indsamles i sådanne studier, er næsten uundgåeligt underlagt såkaldt højrecensurering, således at hændelsen for nogle individer kun vides ikke at være forekommet indenfor en bestemt tid. En endnu mere begrænsende form for censurering, intervalcensurering, opstår når status for hændelsen kun opgøres på en række adskilte tidspunkter. I dette tilfælde er den eneste tilgængelige information, når hændelsen observeres første gang, at den er opstået på et tidspunkt siden seneste opgørelsestid med negativ status. Intervalcensurerede data er ofte yderligere kompliceret af konkurrerende hændelser, som kan forekomme og dermed udelukke den hændelse, man ønsker at undersøge. De fleste statistiske metoder til at håndtere intervalcensurering forudsætter typisk antagelser om fordelingen af de faktiske hændelsestider for at muliggøre brugen af parametriske modeller, som har nogle matematiske egenskaber, der gør dem lette at anvende på intervalcensurerede data. Dette begrænser dog mulighederne for ved hjælp af regressionsmodeller at måle sammenhængen mellem potentielle risikofaktorer og hændelsen, man ønsker at undersøge. En alternativ måde at håndtere censurering på er ved at transformere de ukomplette data til nogle såkaldte pseudo-observationer, som kan analyseres ved hjælp af en alsidig familie af regressionsmodeller, generaliserede lineære modeller, til at måle associationer med en række let fortolkelige mål. Denne tilgang med pseudo-observationer er baseret på ikke-parametriske metoder til at definere transformationen af data, men ved at anvende en alternativ teknik baseret på splines, som er yderst fleksible parametriske funktioner, kan vi formulere pseudo-observationer for intervalcensurerede data. Disse parametriske pseudo-observationer kan så anvendes i generaliserede lineære modeller til at estimere sammenhænge med risikofaktorer udtrykt som størrelser, der er relevante for den specifikke kontekst i det enkelte studie.

Contents

Preface

In January 2014, I had the pleasure of spending a few days in the northernmost part of Denmark along with my colleague Søren Lundbye-Christensen. The primary purpose of this most enjoyable stay was to work on a specific project that was methodologically challenging, but we also found time to work on secondary purposes of less academic but equally enjoyable content. Although we were not aware of it at the time, we were cutting the first sod to what would become the PhD project that forms the basis of this thesis.
The thesis that I present here is written as a summary of the entire process that has formed around our efforts to elucidate and understand the methodlogical challenges that arise from interval censored time-to-event data with competing risks. I have thus written the thesis primarily in singular first person as a reporting of the results of my work on this project, but since I have evidently been completely reliant on collaboration with a group of highly skilled people, I will frequently refer to work that has been carried out in collaboration in plural first person. On specific occasions, I will also be using plural first person to refer to you, the reader, and I, the author, collectively.
This PhD project has been an incredibly rewarding process for me both personally and professionally. It has only been possible to carry out the project due to a great amount of cooperation, help, and support from a number of people to whom I am immensely grateful. First and foremost, my close friend, colleauge, and assistant supervisor Søren Lundbye-Christensen who has been my irreplaceable partner for discussion of all things related to the project and life in general. I am forever thankful for your support and friendship. Thank you also to my two other very competent supervisors, Sam Riahi and Jacob Moesgaard Larsen, to whom I am indebted for their invaluable feedback on my work, and to Jens Cosedis Nielsen for willingly allowing me to work on data from his ongoing clinical trial. I have also profited from another inexhaustible source of feedback and counseling in Erik Thorlund Parner who has shown a formidable ability to help me keep a sense of perspective and save my mental health in particularly frustrating periods of the process. The project would never have been launched had it not been for the kind and whole-hearted support of then research director, Søren Hjortshøj. I

am also very grateful to Erik Berg Schmidt who is the main responsible for my close collaboration with the Department of Cardiology at Aalborg University Hospital and for an important decision in my life by persistently directing my attention towards Aalborg University Hospital and Forskningens Hus in the first place. Finally, I am thankful beyond words to my wife, Anne, for her support and to our two daughters, Thea and Tilde, who continuously bring joy into my life.

Martin Nygård Johansen
Aalborg University, November 18, 2020

Part I

Background

Background

1 Introduction

1.1 Outline

As the papers upon which this thesis is based describe different aspects of the developed methods in both statistical and epidemiological frameworks, the thesis itself will give a quite thorough introduction to the intended settings and an explanation of the core aspects of the methodology and the challenges it has been developed to overcome.
The thesis first presents the specific clinical setting that inspired the methodological work that follows. This clinical dataset was analyzed in Paper B and will also be considered in the thesis as a motivating example and an application of the methods.
The methodological part of the thesis first reviews some basic methods for time-to-event data, and then describes two core ideas that we employed in developing the proposed method, namely pseudo-observations and the flexible parametric modeling approach. Then, I will formulate an alternative to the non-parametric pseudo-observations for right censored time-to-event data with competing risks, which we proposed in Paper A. After that, I will show how we extended this method in Paper B to cover a situation with interval censoring on the event of interest and the methodological part will conclude with a section that reframes the methodological work in an epidemiologically applied setting. For this purpose, I will show the methods applied to a dataset with the illustratively advantageous feature that we can work with both a right censored and an interval censored version of the event time data. This dataset was also used in Paper C.

1.2 Motivating example

When a person is known to experience irregularities of the heart rhythm, known as arrhythmias, and is considered at risk for sudden cardiac death as a consequence, the treatment strategy is often the use of an implantable

cardioverter-defibrillator. This is a specialized device, which functions both as a pacemaker to help maintain a stable heart rhythm by giving electrical stimuli but is also capable of giving a high-voltage shock if needed. These devices consist of a generator box and two leads that are connected to the generator, while the other end is inserted into the heart through a large blood vessel. Fig. 1 illustrates an implanted device. The leads are complex con-

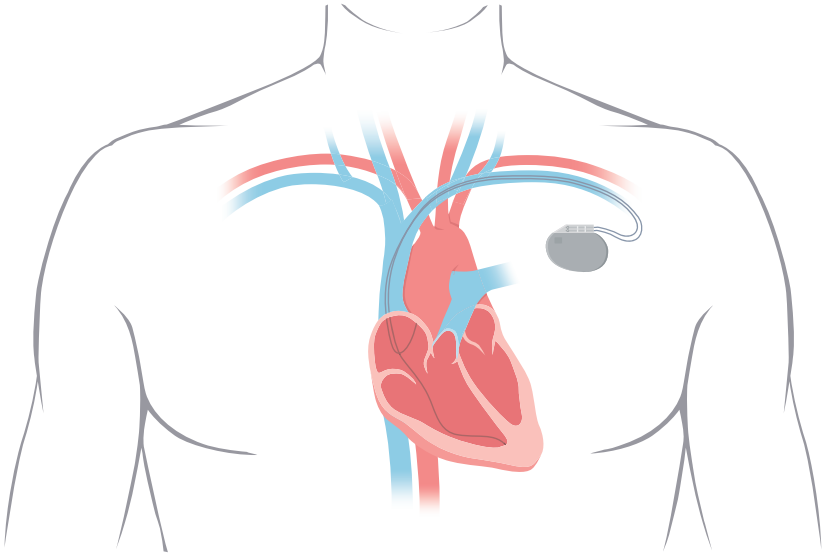


Fig. 1: Illustration of an implantable cardiac device.

structions consisting of shock coils that connect to multiple wires, which are protected by an insulation material. A particular type of leads has been under scrutiny due to a potential risk of a specific insulation failure in which the wires wear through the insulation material as a result of the stress that the ordinary movement of the patient's torso puts on them. These mechanical failures, called *externalizations*, are potentially very dangerous for the patient because the leads may become sensitive to false signals with the consequence that the device might give a shock in an inappropriate time.

In an effort to understand the mechanisms behind this critical phenomenon, researchers have been interested in performing regression analyses on the time to an externalization event. The event, however, can go unnoticed with no detectable symptoms and is usually not noticed until a routine examination in the form of an X-ray or fluoroscopy imaging is performed. This means that the available data has the form of a number of examination times and the externalization status (present or not present) at each examination. This type of data structure is known as *interval censored* data. Traditional methods for right censored time-to-event data, however, require knowledge of the actual

time of the occurrence of the event or a censoring time if the event did not occur within the observed period, whereas interval censored data requires quite different techniques.
One of the factors that have been suspected of being associated with the externalization risk is the amount of slack in the lead. The hypothesis behind this is that without a sufficient amount of lead slack the patient's muscle movements put more strain on the lead and thus increase the wear in the long run. But there have also been speculations that too much slack would entail a more immediate increase in risk due to sharper bends of the lead, which might cause excessive abrasion leading to externalization. Hence, modeling of the externalization rate should be performed with due care to reflect the potentially quite different risk patterns over time.
The dataset that I will analyze for illustrative purposes in this thesis has been described by Larsen *et al.* [20]

1.3 Time-to-event data and basic notation

The fundamental real-life circumstance that must be dealt with when considering time-to-event data is that *right censoring* of study subjects can prevent the observation of the actual time of the event of interest. In such cases, the information is limited to the knowledge that the event has not yet occurred at the time of censoring. Right censoring is very common in studies where the outcome is the time to some specified event, and it is generally caused by either loss-to-follow-up or an administrative censoring due to ending of the study or data availability. It is important to note that censoring is only related to the observation of the event and not the actual occurrence of the event. Most of the methods developed for right censored data thus assumes *independent censoring* meaning that the time to censoring and the time to the event of interest are independent processes.

Functions related to the distribution of time-to-event data

If we let the random variable T denote the time to event with density function f, we can characterize the extent of the risk at any particular time during follow-up, t, by the *hazard function*, which can be considered as a measure of the instantaneous risk and is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

Another commonly studied function related to the survival distribution for T is the *survival function*,

$$S(t) = \int_t^\infty f(u)du,$$

or the related *cumulative incidence function*, which is simply the distribution function $F(t) = 1 - S(t)$. The survival function describes the probability of a random individual from the relevant population to survive at least until the time point at which the function is evaluated. Similarly, the cumulative incidence function describes the probability of having experienced the event prior to a given time point.
The survival function and the hazard function have a simple relationship involving the density function, since

$$h(t) = \frac{f(t)}{S(t)} = \frac{d}{dt} \ln(S(t)),$$

where the second equality follows from the fact that $f(t) = \frac{d}{dt}F(t)$. By integration of this equation, we obtain what is known as the *cumulative hazard function*,

$$H(t) = -\ln(S(t)). \tag{1}$$

Data structure of right censored data

In a right censored setting, we observe data of the form (t_i, d_i), where t_i denotes the observed time to event or censoring for subject i in a study population of n individuals, and d_i is an indicator for observing the event of interest. This notation implies that we have defined a time scale with a well-defined origin on which the functions are defined. In practice, the choice of a time scale will depend on the specific characteristics of the given study setting, but for the remainder of this thesis, I will be using the simple time scale, usually referred to as time-on-study, where time zero is defined as the time that the individual enters the study.

1.4 Classical methods

Non-parametric estimators of survival and cumulative incidence

In contrast to a fully observed continuous outcome, for a time-to-event outcome it is not obvious how to give a simple descriptive summary of the observed data. We are unable to measure the mean of the time to event due to the censoring of some individuals. This is very often handled by using the non-parametric product-limit estimator of the survival function, usually referred to as the *Kaplan-Meier estimator*. [15] For a particular dataset there will

be a set of distinct observed event times, $t_1, t_2, \ldots$ and a corresponding number of events occurring at these time points, $d_1, d_2, \ldots$, and the Kaplan-Meier estimator, $\hat{S}(t)$, is then defined as

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where n_i is the number of subjects still under observation and not yet having experienced the event of interest at time t_i. This estimator is a piecewise constant function of time with jumps at the observed event times. For the Kaplan-Meier estimator to give unbiased estimates of the survival function, the independent censoring assumption is paramount because it ensures that the subjects who are still under observation and considered at risk of the event of interest are representative for the entire subset of individuals who can still experience the event whether we can observe it or not. From the Kaplan-Meier estimator, we can obviously obtain an estimator of the cumulative incidence function as $1 - \hat{S}(t)$.
The cumulative hazard function can be estimated by the Nelson-Aalen estimator [1,24],

$$\hat{H}(t) = \sum_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

A non-parametric estimator of the cumulative incidence function that can also incorporate competing risks, which we will consider in Section 1.6, is the Aalen-Johansen estimator [2], which can be obtained by combining the Kaplan-Meier and Nelson-Aalen estimators.

The Cox proportional hazards regression model

Performing regression modeling on right censored time-to-event data has been completely dominated by the commonly applied Cox proportional hazards model. [9] As the name suggests, this model relies on the assumption that hazard functions for different values of the independent variables are proportional. Inference in the model is based on a partial likelihood approach where the general shape of the hazard, determined by a baseline hazard function, is unspecified in the model and is estimated semi-parametrically when the model is fit to a given dataset using the Breslow estimator. [5] For a vector of explanatory variables, $\boldsymbol{z}$, the model can be formulated on the log hazard scale as

$$\ln(h(t; \boldsymbol{z})) = \ln(h_0(t)) + \boldsymbol{\beta}^\intercal \boldsymbol{z},$$

where $\boldsymbol{\beta}$ are the regression coefficients and h_0 is the baseline hazard function. In the Cox proportional hazards model, the association between the outcome and an explanatory variable, say the j'th element of $\boldsymbol{z}$, is measured by the ratio

of the hazards known as the *hazard ratio*, $\text{HR}_j = \exp(\beta_j)$, which is interpreted as the relative instantaneous risk for a one-unit difference in the explanatory variable. The Cox model does not assume the same strictly independent censoring as the non-parametric estimators, but only independence given the regression variables.

Parametric models

Fully parametric models, i.e. models that are completely specified by a finite number of parameters, have some mathematical advantages in that inference can be based on a full likelihood approach. For a parametric model determined by a set of parameters, $\boldsymbol{\theta}$, the likelihood function can be written as

$$L(\boldsymbol{\theta}; \boldsymbol{z}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{z}_i)^{\delta_i} S(t_i; \boldsymbol{z}_i)^{(1-\delta_i)}, \tag{2}$$

where $\boldsymbol{z}_i$ is the values of the explanatory variables for the i'th subject and δ_i is an event indicator.
For any assumed parametric distribution of the time to event, this likelihood function can be maximized to find estimates, $\hat{\boldsymbol{\beta}}$, of the regression coefficients. Some of the simpler parametric models include the Weibull model with hazard function $h(t; k, b) = bkt^{k-1}$, and a piecewise version of the Exponential distribution with constant hazard function. Due to the rather restrictive distributional assumptions in these models, their application in medical research has been quite limited compared to the Cox proportional hazards model.

1.5 The flexible parametric model

With the increasing computer performance, new methods relying on numerical maximization of complex likelihood functions with more parameters have become feasible alternatives to the traditional distribution-based parametric models. A number of different approaches using splines to model different functions related to the distribution of a time-to-event variable have been proposed. [6, 12, 18, 19, 31, 32] The most commonly applied of these approaches is probably the *flexible parametric model* suggested by Royston & Parmar. [32] This model can be formulated in either a proportional odds or a proportional hazards version but I will only be using the proportional hazards version where the model can be written as

$$\ln(H(t; \boldsymbol{z})) = s(\ln(t); \boldsymbol{\gamma}) + \boldsymbol{\beta}^{\intercal} \boldsymbol{z}.$$

In this formulation, $s(\ln(t); \boldsymbol{\gamma})$ is a restricted cubic spline with m internal knots evaluated in log time that is used to model the log cumulative baseline hazard function and has parameters $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{m+1})$. A restricted

cubic spline is determined by a set of basis functions, $v_1(\cdot), \ldots, v_m(\cdot)$, and the parameters as

$$s(x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_m(x).$$

More details on restricted cubic splines can be found in the Appendix of Paper A. Using a spline to model the log cumulative baseline hazard gives the advantage of being able to model almost any realistic shape of the underlying distribution function while allowing for full likelihood estimation owing to the parametric nature of the model.
The flexible parametric model provides a smooth estimator of the cumulative baseline hazard function based on the spline parameters using the relation

$$H_0(t) = \exp\left(s(\ln(t); \boldsymbol{\gamma})\right). \tag{3}$$

This has particular advantages in a competing risk setting with interval censored data, as we shall see.

1.6 Competing risks

So far, I have only been considering right censored data where all individuals are bound to eventually experience the event of interest. For studies of survival time, this is a reasonable approach but for most non-fatal outcomes in living creatures, the omnipresent risk of dying interferes with this simple order of things, whether we like it or not. For instance, individuals are certainly no longer at risk of developing a specific disease once they have died. Hence, in a wealth of different research questions the presence of one or more *competing risks* should be accounted for. This is not to be confused with censoring where, as I have already pointed out, the risk of the event of interest should be the same regardless of the censoring status.
The competing risk phenomenon is often considered as a simple multi-state model in which each individual is in an event-free state at the beginning of follow-up, but can then move to one of K different states during follow-up. This can be visualized as in Fig. 2, which also appears in Paper C. Each of the arrows in Figure 2 represents a potential *transition* from one state to another, and they can be modeled by individual *cause-specific hazard functions*, $h_1(\cdot), \ldots, h_K(\cdot)$. There are no arrows out of any other state than State 0 so for the purpose of the model we consider each of the states $1, \ldots, K$ to be *absorbing* states.
In a competing risk setting, there is not a simple relationship between the cause-specific hazard and the cumulative incidence as in (1) for right censored data. The cause-specific cumulative incidence for cause d, $F_d(\cdot)$, can,

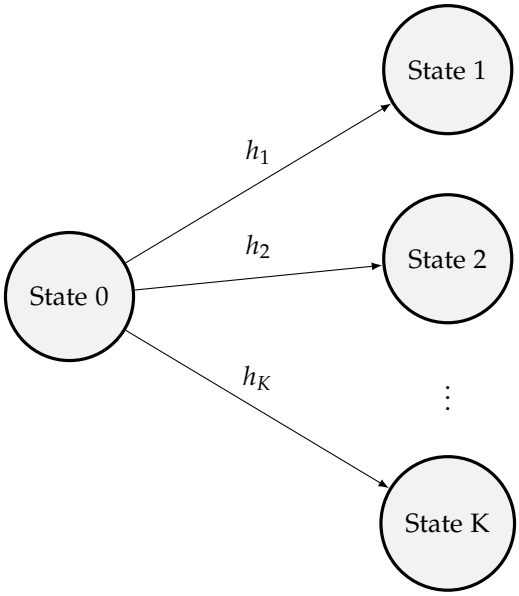


Fig. 2: A multi-state model with K competing risks.

however, be calculated from all of the K cause-specific hazard functions,

$$F_d(t) = \int_0^t h_d(u) \exp\left(- \int_0^u \sum_{k=1}^{K} h_k(v) dv \right) du = \int_0^t h_d(u) \cdot S(u) du, \quad (4)$$

where $S(\cdot)$ refers to the *overall survival*, which expresses the probability of not having experienced any of the K events at a given time.

1.7 Pseudo-observations

If we imagine for a while that we actually had a fully observed dataset of time-to-event data with no censoring, what kind of regression models would then be relevant for estimating associations? One obvious approach would be some kind of linear model on the event times, which would be relevant for assessing the difference in expected time to the event under different values of an exposure variable in a relevant scale. We might also be interested in modeling the probability of having experienced the event at a certain time point by perform regression on a binary outcome. Both these approaches would be feasible within the general theory of generalized linear models.
But confronted with the harsh reality that right censoring is an issue, which must be dealt with in practice perpetually, a transformation of a right censored dataset into some fully observed data with some appropriate math-

ematical properties would be a savior. A very clever transformation of a right censored dataset that provides just this has become increasingly popular in the recent decades. This transformation gives rise to a set of *pseudo-observations* and they are derived as follows. [3,4] Based on the full right censored data sample, we calculate an estimate of the function, f, that describes the scale at which we are ultimately interested in measuring associations by a parameter θ. This gives us a full-sample estimator, $\hat{\theta}$, of the expected value $\mathrm{E}[f(X)]$, where X is the vector containing the fully observed data. Similarly, we could obtain an estimator, $\hat{\theta}_{(-i)}$, based on the leave-one-out (or jackknife) sample where the i'th observation is left out. The transformation of the right censored dataset then consists of performing the transformation defined by calculating the pseudo-observations as

$$\theta_i = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}, \quad \text{for } i = 1, \ldots, n. \tag{5}$$

That is, the i'th pseudo-observation measures what the difference in the estimator of θ would be had the i'th observation in the dataset not been there.
The most common application of the pseudo-observation approach is to estimate the cumulative incidence function by one minus the Kaplan-Meier estimator or, in the presence of competing risks, by the Aalen-Johansen estimator. In this situation, it can be shown that the resulting non-parametric pseudo-observations have the property that, for large samples, $\mathrm{E}[\theta_i|\boldsymbol{z}_i] \approx \mathrm{E}[\mathrm{I}(T_i \leq t, D_i = d)|\boldsymbol{z}_i] = F_d(t|\boldsymbol{z}_i)$, where $F_d(t|\boldsymbol{z}_i)$ is the cause-specific cumulative incidence for cause d given the values of the explanatory variables for the i'th subject. [11] This means that we can use the pseudo-observations as the outcome variable in a generalized linear model to obtain estimates of associations with the explanatory variables.
More specifically, we can formulate regression models for a range of different association measures such as hazard ratios, risk differences, relative risks, and difference in restricted mean survival time. This can be achieved using the generalized linear model [4]

$$g(\mathrm{E}[f(T)]) = \beta_0 + \sum_j \beta_j \boldsymbol{z}_j,$$

for suitable choices of the function, $f(\cdot)$, and link function, $g(\cdot)$. In the above example where $f(T) = \mathrm{I}(T_i \leq t, D_i = d)$, we can use the identity link function to model the absolute difference in cause-specific cumulative incidence, or we can use the log function to model relative associations in terms of the cause-specific relative risks. We can even obtain an estimate of the hazard ratio in a proportional hazards model at a specific time point by using the link function $g(t) = \ln(-\ln(t))$. Another useful choice is to use the overall τ-restricted mean function, $f(T) = \min(T, \tau)$, by the integral of the Kaplan-

Meier estimator, $\int_0^\tau \hat{S}(u)du$, and apply a model with identity link function,

$$\mathrm{E}[\min(T,\tau)] = \beta_0 + \sum_j \beta_j z_j,$$

to obtain estimates of the amount of expected loss of lifetime due to a given explanatory variable.
Estimating the variance of the regression coefficients in these models has been the subject of a considerable amount of research in itself. For pseudo-observations based on the Aalen-Johansen estimator, it was originally proposed to use a robust sandwich estimator of the variance or applying a non-parametric bootstrap procedure, and the sandwich estimator has become the standard in practical applications. However, Jacobsen *et al.* [13] have shown that this is generally a conservative estimator and Overgaard *et al.* [26, 27] have suggested and thoroughly studied an alternative estimator. For the purpose of this thesis however, I will be using the slightly conservative sandwich estimator.
The modeling framework that I have presented here covers the case of pseudo-observations calculated at one time point only, but models can also be formulated for a finite number of time points. Such models should take into account the correlation between the pseudo-observations for each individual. Since the pseudo-observations are based on overall estimators of the cumulative incidence function, the assumption of independent censoring is crucial for the above results to hold. If this is not fulfilled, however, pseudo-observations can be calculated in strata of the relevant variables to obtain unbiased estimates. A thorough discussion on assumptions on possible remedies for violations is given by Mortensen *et al.* [23]

1.8 Interval censoring

The methods described above that rely on observing exact event times, at least for some individuals, are not directly transferable to interval censored data. As mentioned in Section 1.2, interval censoring implies that we only have information on the event status at a given set of *examination times* such that when we observe an event of interest in one individual, all we know is that it has occurred since the last time we observed that it had not yet occurred. So, if the event has not occurred at visit time L but has occurred at time R, we know that it occurred somewhere in the interval (L, R). If, on the other hand, the event had not occurred at the last examination time L, this situation corresponds to a right censoring at this time and we could cover this by setting $R = \infty$ such that the event is only known to occur at some point after L. All examinations prior to L or after R give no further relevant

information.
Interval censoring is not limited to the study of time-to-event outcomes but can also occur when measurements are for other reasons only known to lie within an interval. This could be the case when measurement of an outcome is only possible by assessing whether or not the value exceeds certain thresholds. Although the methods I will be discussing are applicable more generally, I will stick to the terminology from the time-to-event setting.
For any of the methods discussed in the following, a crucial assumption is that the examination times are independent of the event status. That is, the examination pattern must be determined either by a fixed schedule or governed by a random process, which is independent of the event time process. In practical applications, this is an important limitation.
One other important thing to note about interval censored data is that there is a greater amount of unknown information than just the exact time of the events, since an individual for which we do not observe the event at the last examination time will subsequently be censored with unknown event status. This means that the event might have occurred somewhere between the last examination time and the censoring time, or it might not. This additional uncertainty should be accounted for when we perform any kind of statistical inference.

Data structure

A simple dataset of interval censored time-to-event data can then be characterized by two variables where one contains the last known time when the event had not occurred (potentially at time zero) and the other contains the first known time when the event had occurred (potentially never; e.g. coded as a missing value).
In practice, it is quite common to have a dataset consisting of a mix of (right censored) exactly observed and interval censored event times. This can happen if, for instance, data are gathered from different sources, but it can also happen if the event of interest can either be detected immediately or go unnoticed until a specific examination is performed. These two different mechanisms both leading to a mix of right and interval censored data differ in the sense that in the first case individuals are predetermined to be observed by either continuous or pointwise follow-up whereas in the second case exactly observed event times are added to otherwise interval censored data. Right censored data can be considered a special case of interval censored data with $R = \infty$ for censored individuals and $L = R$ for exactly observed event times. [22]

Non-parametric estimation of survival

The extremely popular Kaplan-Meier estimator for right censored data does not have a direct counterpart for interval censored data. However, both Peto [28] and Turnbull [36] derived the same non-parametric estimator of the survival function. The procedure for finding this estimator is not very straightforward and applications of the estimator has been limited. The estimator is defined by a set of intervals between which the value is constant, but it is not explicitly defined within the intervals since any function with the appropriate increase within the intervals has the same likelihood.

Example, cont. In the ICD dataset introduced in Section 1.2, we can obtain a naive estimator of the survival function for the externalization event by erroneously right censoring at death and employ the Peto-Turnbull approach. In Fig. 3, I show the Peto-Turnbull estimator for the two exposure

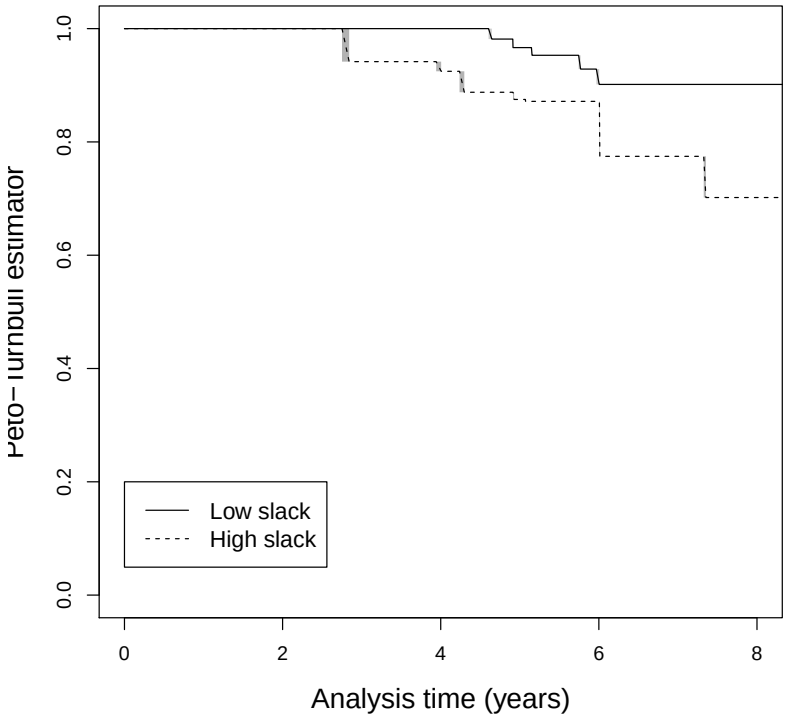


Fig. 3: The Peto-Turnbull estimator of the externalization survival in the ICD data.

groups, low and high slack, separately. We can see that, most of the intervals

(shaded grey) where the Peto-Turnbull is undefined are very short. Although there are 11 and 26 observed externalization events in the group with high, respectively low, slack, the curves jump at 5 and 6 intervals. The estimated externalization survival after 8 years is about 0.90 for the low slack group and 0.70 for the high slack group, indicating that the estimated cumulative incidence at 8 years is about 10 and 30%, respectively. However, by censoring at death, we are assuming that the patients who die before having experienced an externalization event still have the same risk after their death as those who are still observed at similar time points.

Parametric regression models

The likelihood contribution for different individuals depend on the trajectory of that individual. For an event that is observed to occur in the interval (L, R), the likelihood contribution is the difference in the survival function over the interval, i.e. $S(L) - S(R)$, while for right censored or exactly observed event times the contributions are as described in (2). [17] With these analytical expressions for the likelihood contributions, it is rather straightforward to apply standard maximum likelihood methods to interval censored data by using a distribution-based parametric model. The Weibull and piecewise Exponential models mentioned in Section 1.4 are then easily extended to an interval censored setting. As is the case for right censored data, the parametric models yield precise estimates of the relevant quantities if the model provides a good fit to the data and thus adequately describes the underlying risk patterns. In such situations, parametric models also offer the opportunity to produce smooth estimators of the functions describing the distribution and calculating out-of-sample predictions. However, if the model is misspecified in terms of the distributional assumptions the result will be apparently precise estimates that do not have the intended interpretation.

Example, cont. If we stick to the naive approach of right censoring at death, we can fit a parametric model to the ICD dataset by assuming that the externalization event times follow a Weibull distribution and that the hazard functions for the two slack groups are proportional. This model fits the data reasonably well, and the estimated hazard ratio is 3.1 (1.5 to 6.2). Based on this model, we can also estimate the externalization survival functions for low and high slack, and these are shown in Fig. 4. The parametric nature of the Weibull model facilitates a smooth estimate of the baseline hazard function such that, in contrast to the Peto-Turnbull curves in Fig. 3, the estimated survival curves are smooth.

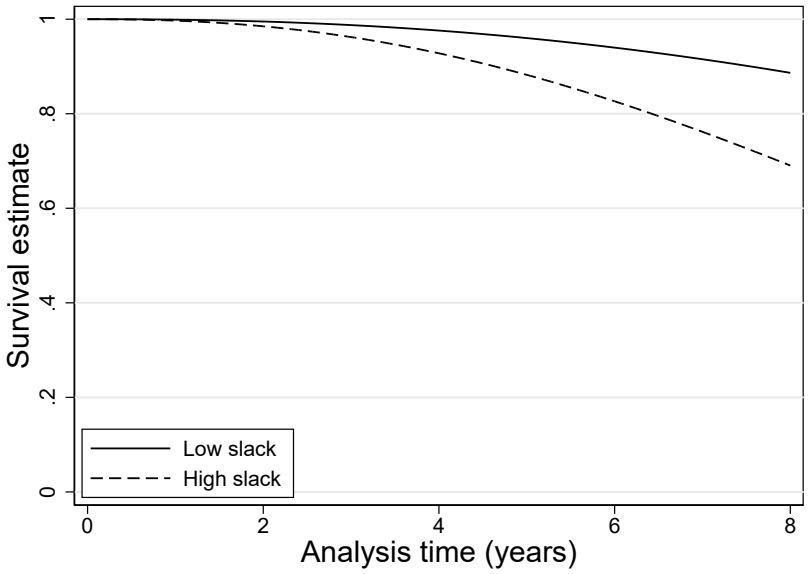


Fig. 4: The estimator of the externalization survival in the ICD data based on a Weibull model.

Midpoint and right endpoint methods

The relatively strict assumptions of the parametric models and the lack of generally accepted non- or semi-parametric regression modeling approaches for interval censored data has led to a common practice of imputing either the interval midpoint or the right endpoint of the interval for the event time and use standard methods for right censored data. Using either of these methods also requires a crucial, yet often neglected, decision about how to handle individuals for whom we do not observe the event of interest. One way to handle these would be to censor at the last examination, since we do not know whether the event of interest has occurred after that. However, this would entail that all individuals that die during follow-up would be considered as censored and we would thus not see any competing events in the imputed dataset. If we, on the other hand, censor individuals without an observed event of interest at the end of their observable follow-up, we would be assuming that they did not experience the event of interest after their last examination even though we do not know this. In this PhD project, I have chosen to use a combination of these two methods so that individuals without an observed event of interest are coded as dead at the observed time of death if any and are otherwise censored at their last examination time.
The validity of the midpoint and right endpoint imputation approaches has been discussed by several authors [10,22,25,30] and are generally considered

questionable, at least when the underlying hazard function is not flat. The right endpoint is the last point at which the event could have occurred, and this obviously leads to an underestimation of the cumulative incidence at any particular time point. The midpoint imputation approach is also known to lead to potentially biased results and underestimate the standard error of parameter estimates [16,21,25], and the censoring at the last examination time for individuals without an observed event of interest means ignoring some information that is actually present in the data.

Example, cont. If we use midpoints or right endpoints in the ICD data, we can apply standard right censored methods by estimating externalization survival with the Aalen-Johansen estimator. This enables us to account properly for the competing risk of death by estimating cause-specific cumulative incidence of externalization. Fig. 5 shows the estimated cumulative incidence of externalization based on midpoints (blue curves) and right endpoints (red curves). The curves show some interesting features. First, we can clearly

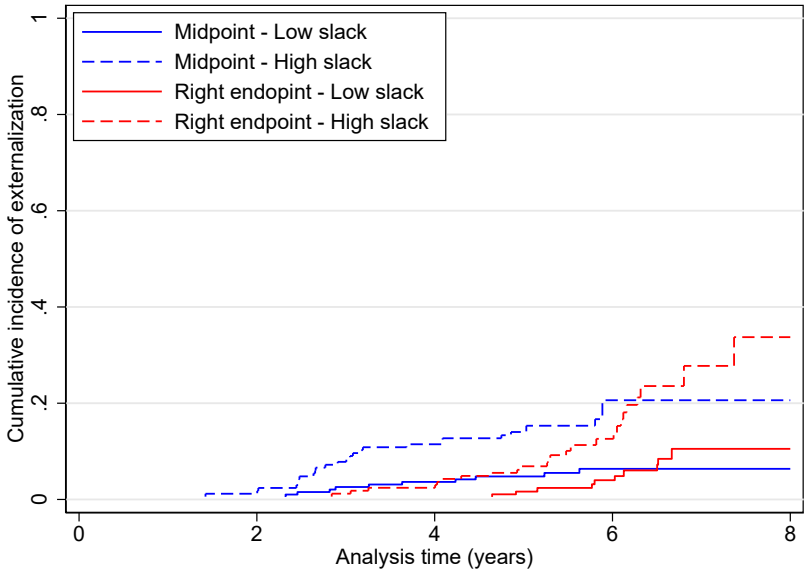


Fig. 5: Aalen-Johansen estimators of the cumulative externalization incidence in the ICD data based on midpoints and right endpoints.

see that when we use right endpoints, the events occur later than with the midpoints such that the estimated cumulative incidence remains low until much later in follow-up. Second, at later follow-up times, we notice that the estimate based on right endpoints increases to a higher level than that of the midpoints. This phenomenon occurs because more individuals have

already been censored at the event times, which causes the jumps in the Aalen-Johansen estimator to become larger.
If we fit a Cox proportional hazards model to the midpoints or the right endpoints, we can obtain estimates of the cause-specific hazard ratio comparing the externalization rate with high slack to that with low slack. This gives an estimated hazard ratio of 2.9 (1.4 to 5.9) for the midpoints and 3.5 (1.7 to 7.1) for the right endpoints, which reflects the differences we observed in the Aalen-Johansen curves quite well.

Competing risks and the illness-death model

Since the event of interest in an interval censored setting is almost never death (for which we generally observe an exact time), competing risks will inevitably be a phenomenon that should be taken into account. As I have already mentioned in Section 1.6, this is usually handled by considering a multi-state model with a number of different states that each subject can move to from the initial state. For simplicity, I will, henceforth, combine all event types other than the event of interest into a common competing event, thus simplifying the model in Fig. 2 to a situation with $K = 2$.
When the event of interest is interval censored, although it is still true that there are two possible transitions out of the initial state, the limited information for some individuals renders the competing risks multi-state model inadequate. If an individual is not seen at an examination time with the event of interest before the competing event occurs, we simply do not know to which state this individual moved from the initial state because the event of interest may or may not have occurred. Similar uncertainty exists if the individual is censored without a preceding positive examination for the event of interest. To fully exploit the information that we do have, we need to model the complete set of possible transitions so we can model the potential transitions as well as the observed transitions. This gives rise to expanding the multi-state model to an *irreversible illness-death model* [7] where it is also possible to experience the competing event after the event of interest, such that we are no longer just considering the time to whatever event occurs first. I have illustrated this in Fig. 6, which also appears in Paper C. In this model, we consider only State 2 to be an absorbing state, and the term *irreversible* refers to the fact that we do not allow transitions from State 1 back to State 0. In the terminology of an illness and death, this means that the disease we are considering is incurable. I will model the transitions in the irreversible illness-death model using *transition-specific hazard functions*, $h_{01}(\cdot)$, $h_{02}(\cdot)$, and $h_{12}(\cdot)$, under a semi-Markov assumption meaning that each transition-specific hazard function, $h_{kl}(\cdot)$, only depends on the time since entry into the current state, k. The transition from State 1 to State 2 may seem redundant for the purpose of modeling the event of interest, i.e. the transition from State 0

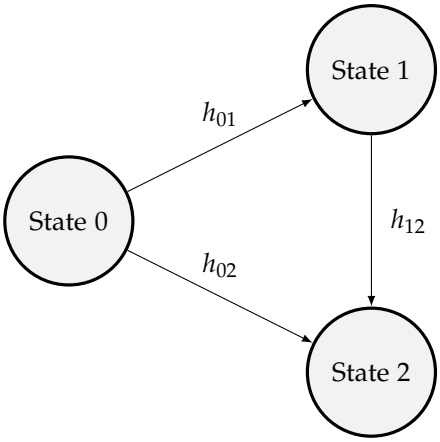


Fig. 6: The irreversible illness-death model.

to State 1, but we will see in Section 2.2 that it is imperative to model this transition as well.

Data structure with competing risks

Since we are no longer just considering the time to and type of the first event to occur, the data structure needs to be extended to accommodate the additional information for each subject. One possible way to do this is to use multiple data records to describe the different transitions for each individual, but in the following I will be using a different structure where the trajectory of each individual is described by a set of five variables in place of the two variables needed in the right censored competing risk setting, since there are now two relevant event types where one is potentially interval censored. The notation I will use follows that used in Paper B where the variables are defined as follows for the i'th individual.

List 1.

d_{1i} is the indicator for an observed event of interest (either exactly observed or interval censored)

l_{1i} is the last known negative time point (potentially at time zero)

t_{1i} is the observation time for the event of interest (either the exact time or the first positive examination time)

d_{2i} indicates a competing event (exactly observed)

t_{2i} is the observation time for the competing event

To distinguish between events of interest, which are exactly observed and those that are observed in an interval, I will use the convention that if the event of interest is observed exactly at time t_{1i} then l_{1i} is set to missing as it is then no longer relevant.

2 Methodology

In this section, the fundamentals of the proposed analysis approach are explained. The parametric approach for calculating pseudo-observations for right censored competing risks data is detailed in Section 2.1, and in Section 2.2 I explain how this method can be extended to cover the case of an interval censored event of interest. In Section 2.3 I aim to frame the proposed methods in a more epidemiological context and give some practical comments and caveats.

2.1 Parametric pseudo-observations for right censored data with competing risks

This section is based on Paper A and it introduces the basic methodology in which this PhD project has resulted. The general idea behind the method is to use the flexible parametric approach to estimate the marginal cumulative incidence function and define pseudo-observations based on this estimator. I consider a setting with competing risks to cover the more general case and work with the cause-specific versions of the relevant functions.

Definition

If we consider the non-parametric pseudo-observations defined in (5), they are most often based on the Aalen-Johansen estimator of the cause-specific cumulative incidence function. This function has two main limitations in relation to the variability of the point estimates. First, it is a piecewise constant function so, particularly when events are scarce, small differences in the observed event times may cause rather large differences in point estimates. Second, it is constructed in such a way that it uses only information up to, but not beyond, the time point at which it is evaluated. These circumstances might suggest that by fitting a spline to the entire follow-up data we will be able to make better use of the information and decrease the sensitivity to small fluctuations in the observed data.

To obtain such an estimator of the cumulative incidence function, I have followed the approach of Royston & Parmar [32] in their flexible parametric model to fit a restricted cubic spline to the log cumulative hazard function in log time. Fitting such splines for each transition in the competing risk model yields a set of estimated spline parameters, $\hat{\gamma}_1, \ldots, \hat{\gamma}_K$, which can be used to obtain an estimator for the cause-specific cumulative incidence function for the event of interest by applying the formula in (4). I will be using the flexible parametric model only marginally such that the estimator of the cumulative baseline hazard function in (3) will work as the overall estimator, $\hat{\theta}$, in (5).
By denoting the spline-based estimator of the baseline cause-specific cumulative hazard for the event of interest as $\hat{\theta}^p$ and the corresponding i'th leave-one-out estimator as $\hat{\theta}^p_{(-i)}$, we can then define a set of *parametric pseudo-observations* for the cause-specific cumulative incidence as

$$\theta_i^p = n\hat{\theta}^p - (n-1)\hat{\theta}^p_{(-i)}, \quad \text{for } i = 1, \ldots, n. \tag{6}$$

These parametric pseudo-observations can then be used in generalized linear models as described for the non-parametric pseudo-observations in Section 1.7. I suggest that the splines fitted in the leave-one-out samples should be based on the same knot points as for the full-sample fit. The reasoning behind this suggestion is that if we aim to find the contribution of each individual to the full-sample estimator of the cumulative incidence function for which the spline knots have been defined, we would not accurately reflect this if the spline knots were redefined for the leave-one-out estimator.
I have not theoretically justified that $\hat{\theta}^p$ is an approximately unbiased estimator of the cumulative incidence function, but in Paper A we have performed an extensive simulation study that examines the empirical properties of the parametric pseudo-observation approach in terms of bias and estimate variability and compared these to the properties of non-parametric pseudo-observations.

Simulation results

In our simulation study, we generated datasets in different scenarios to assess the influence of different elements in a practical application of the method. The evaluations are based on 5 000 repetitions in each specific setting. The general conclusion from the simulation study was that the parametric pseudo-observation approach gave unbiased estimates with appropriate coverage. In the simplest settings with no competing risks and constant event intensities, we observed that the standard deviation of the regression coefficient estimates based on the parametric pseudo-observations was about 7 or 8% lower than those based on non-parametric pseudo-observations. This relative efficiency decreased when we introduced a competing risk. By further investi-

gations, we found that the magnitude of the relative efficiency was mainly determined by two factors.

1. The size of the risk set at the analysis time point
2. The amount of information in the dataset that lie beyond the analysis time point

Since the size of the jumps in the non-parametric estimators is directly related to the number of individuals still at risk at each observed event time, the first point supports our initial hypothesis that a smooth estimator provides more stability in the estimation. The second can be explained by a fundamental difference in the way the estimators are constructed. When evaluated at a given time point, the non-parametric estimators do not take into account what happens after this time point, whereas the spline is always fitted to the entire follow-up data and is then evaluated at the given time point. This means that the non-parametric estimators basically use less of the information available in the dataset, which further adds to the instability of the estimates. The specific results from Paper A that give rise to these deductions are those from the scenarios in which we adjusted the amount of additional information after the analysis time point and adjusted the size of the risk set at the analysis time point.

Conclusion

The contents of this section constitute the crux of the methodological contribution in this thesis. The idea of replacing the non-parametric estimators typically used to calculate pseudo-observations for the spline-based parametric estimators was based on the hypothesis that this would produce an improvement in terms of more stability in the final coefficient estimates. We have developed the methods in a right censored setting with competing risks and performed simulations that suggest that the method performs at least as well as the traditional non-parametric pseudo-observation approach and in some situations provide a valuable gain in efficiency.
Settings where the observed gain would occur are rather common in applied medical research. Examples include register-based studies where there is often a surplus of follow-up readily available, which could be included in the dataset to add information after the analysis time point. If, on the other hand, the observed follow-up data is utilized to full extent in terms of follow-up time, the risk set might be diminishing at the analysis time point such that the smooth estimation will provide an improvement in stability of the non-parametric methods.
The price we pay for the improvement in stability is the added complexity of computation of the pseudo-observations. Fitting splines to each of the

leave-one-out samples is a rather demanding procedure that requires use of numerical optimization of likelihood functions, which entails quite intensive computer processing in comparison to the non-parametric counterpart. In the datasets that I have worked with in relation to this thesis where the number of observations has been between 250 and 400, the computations can typically be performed on an ordinary PC in less 5 minutes, though. Implementing the methods in a statistical software program with existing procedures for fitting restricted cubic splines requires some programming knowledge, but it should be feasible to achieve a working solution without using overly advanced programming. In Paper A, I have provided an example of an implementation in Stata software.

Application to the ICD data

In the ICD dataset, we can calculate both non-parametric as well as parametric pseudo-observations based on the midpoint imputation method previously described. Since we are considering death as a competing risk, the calculation of non-parametric pseudo-observations is based on the Aalen-Johansen estimator using the midpoint method. The full-sample estimators of the cumulative incidence based on the Aalen-Johansen estimator and spline-based estimators with 3 through 5 spline knots are show in Fig. 7. Adding

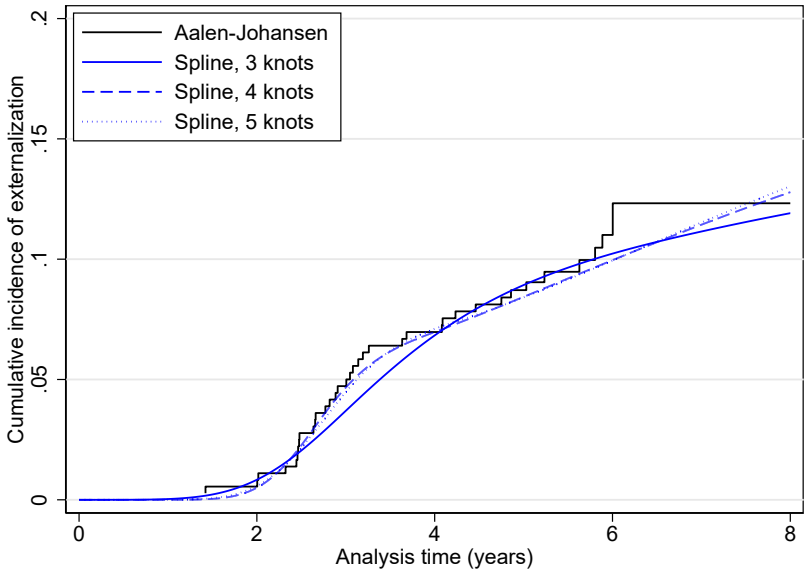


Fig. 7: The Aalen-Johansen and spline-based estimators of the cumulative externalization incidence in the ICD data.

more than 4 knots does not seem to increase the fit of the spline-based es-

timators and I will calculate parametric pseudo-observations using 4 knots in this illustration. I thus calculated non-parametric and parametric pseudo-observations at 5 years of follow-up and used generalized linear models with identity and log link functions to compare the cumulative incidence at 5 years between the two slack groups. The non-parametric approach gives an estimated risk difference of 9.2% (3.2% to 15.1%) and the parametric approach gives 9.0% (3.2% to 14.7%). The estimates of the relative risk are 3.2 (1.5 to 7.0) and 3.1 (1.5 to 6.5), respectively. This suggests that the parametric and non-parametric approaches in the setting give very similar point estimates and that the parametric approach might give a slight increase in efficiency.

2.2 Extending the parametric pseudo-observations to interval censored data

In Paper B, the parametric pseudo-observation method is extended to cover the case of interval censored data. As I have discussed in Section 1.8, methods to deal with a competing risks setting are not directly transferable to an interval censored setting and data should instead be considered in an illness-death model. Since datasets from such settings may in practice consist of a mixture of right and interval censored data for the event of interest, I will consider the general case where both types of data can occur.

Definition

The definition of parametric pseudo-observations is basically the same for interval censored data as in the right censored case with competing risks. The difference lies in how we estimate the cumulative incidence function that describes the transition into the state defined by the event of interest. In the right censored competing risks setting, this was obtained by estimating the cause-specific cumulative incidence function using the K cause-specific hazard functions, which can be estimated separately by considering the risk set and event times for each event type. In the more complex case of an illness-death model with an interval censored event of interest, the transition-specific hazard functions have to be estimated simultaneously in one likelihood maximization procedure because we must take the potential transitions of individuals into account.
The contribution of each individual to this likelihood function will depend on the patient's trajectory through the states in the illness-death model. There are six such trajectories that can occur. [35]

Trajectory 1 An exactly observed event of interest followed by right censoring

Trajectory 2 A negative examination followed by censoring

Trajectory 3 An interval censored event of interest followed by censoring

Trajectory 4 An exactly observed event of interest followed by death

Trajectory 5 A negative examination followed by death

Trajectory 6 An interval censored event of interest followed by death

Each trajectory gives rise to a particular contribution to the likelihood function, L_i for the i'th individual, and I will describe these using the notation defined in **List 1** on page 19. I have already introduced the transition-specific hazard functions, $h_{01}(\cdot)$, $h_{02}(\cdot)$, and $h_{12}(\cdot)$, and following that notational style, I will refer to the transition-specific cumulative hazard functions, $H_{01}(\cdot)$, $H_{02}(\cdot)$, and $H_{12}(\cdot)$, from which we can define the *event-free survival function* as

$$S(t) = \exp\left(-H_{01}(t) - H_{02}(t)\right).$$

We can now specify the likelihood function as the product of the contributions [17] defined as follows. Since Trajectories 4–6 are similar to Trajectories 1–3 with the exception that the individual dies instead of being censored, we can write the contributions in three expressions.
For an individual following Trajectory 1 or 4,

$$L_i = S(t_{1i})h_{01}(t_{1i})\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(t_{1i}))}h_{12}(t_{2i})^{d_{2i}},$$

for an individual following Trajectory 2 or 5,

$$L_i = S(t_{2i})h_{02}(t_{2i})^{d_{2i}} + \int_{l_{1i}}^{t_{2i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})^{d_{2i}}du,$$

and for an individual following Trajectory 3 or 6,

$$L_i = \int_{l_{1i}}^{t_{1i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})^{d_{2i}}du.$$

This is obviously a rather complex likelihood function and I have used a Newton-Raphson algorithm to maximize it over the parameter space of spline parameters, $\gamma_{01}, \gamma_{02}, \gamma_{12}$. In practice, it is worthwhile to obtain starting values for the maximization procedure by fitting a simpler model. We have presented a suggested method for this in Paper B.
When this likelihood function has been maximized, the resulting estimates can be used to find an estimate of the cumulative incidence function for the event of interest using the formula

$$F_{01}(t) = \int_0^t h_{01}(u)S(u)du. \tag{7}$$

Letting $\hat{\theta}^{IC}$ denote the full-sample estimate of the cumulative incidence function, and $\hat{\theta}^{IC}_{(-i)}$ the leave-one-out counterpart, we can then define the parametric pseudo-observations for an interval censored event of interest with competing risks on the same form as in the right censored case in (6).

$$\theta_i^{IC} = n\hat{\theta}^{IC} - (n-1)\hat{\theta}^{IC}_{(-i)}, \quad \text{for } i = 1, \ldots, n. \tag{8}$$

The pseudo-observations thus defined constitute the suggested solution to the original methodological challenge that this PhD project has sought to overcome; an interval censored event of interest that is expected to follow risk patterns that do not follow the same fundamental shape for different values of explanatory variables. The pseudo-observation framework provides an appealing versatility in terms of association measures and the spline-based modeling approach enables these pseudo-observations to be calculated for interval censored data without any assumptions about the distribution of the event times.
Since the calculation of these pseudo-observations involves maximization of a complex likelihood function in each leave-one-out sample, the method requires considerably more computational time than is needed in the case with right censored data only.

The penalized likelihood approach

Simultaneously with the development of our method, a quite similar method was proposed by Sabathé *et al.* [34], which builds upon a penalized likelihood approach to modeling interval censored data in an illness-death model. [14] In this approach, the transition-specific cumulative hazard functions are modeled directly by linear combinations of monotone I-splines [29] and the parameter estimates are obtained by maximizing a penalized likelihood function with a penalty parameter for each transition to control the smoothness of the estimated hazard functions. In addition to the knot positions and the penalty parameters, this model contains a high number of parameters equal to the order plus the number of internal knots and the authors therefore recommend against using it for right censored data.
The pseudo-observation approach by Sabathé *et al.* is very similar to our suggested approach and we have calculated pseudo-observations using both approaches in our simulation study in Paper B to compare their performance.

Simulation results

In a relatively simple setting with exponentially distributed event times and a mix of exactly observed and interval censored event times for the event of interest in an illness-death model, we have performed a simulation study that is presented in detail in Paper B. For the simulated datasets, we compared

the estimates of the cumulative incidence of the event of interest as well as the association with a binary exposure measured by the risk difference and relative risk. Due to the intensity of the calculations we had to limit the number of repetitions in this simulation study to 1 000.
The parametric pseudo-observations in (8) showed no notable biases and the coverage probabilities were quite close to the nominal value of 95%. Compared to the penalized likelihood approach, our approach gave slightly larger standard deviations of the regression coefficients.

Conclusion

By establishing a method for calculating parametric pseudo-observations for interval censored data with competing risks, I have presented a workable solution to the challenge of formulating regression models for such data. The method is solidly grounded in the sense that it has proven quite effective in the simpler case of right censored data where it offers a potential gain in terms of lower uncertainty of regression coefficient estimates.
There are two main limitations of the applicability of the method. First, it requires great amounts of computer power, which might be a deal breaker in some settings. However, with the ever-increasing computational capacity in general-purpose computers and the developments within high-performance computing, this limitation is likely to decrease over time. Second, the complexity of the likelihood function requires rather complex programming in order to implement the methods in practice. I have not developed a publicly available software solution as part of this PhD project because my main objective was chosen to be the development of the methodology. I have implemented the method in Stata software by making use of a number of packages and the built-in command, `ml`, for maximizing a user-specified likelihood function. The penalized likelihood approach has been implemented in an unpublished R package, `pseudoICD`, that is available on GitHub. [33]
Despite these limitations, I believe the method has the qualities to become a valuable addition to the vast toolbox of biostatistical methods. The fields of application of the method are not limited to studies of time-to-event data; interval censoring occurs as a consequence of practical limitations to measurements of an outcome in other situations as well.

Application to the ICD data

To conclude the analysis of the ICD example data, I have calculated parametric pseudo-observations based on the examination times in the dataset. The ICD dataset contains no exactly observed events of interest, so all of the 37 observed events of interest are interval censored. The estimated cumulative incidence of externalization based on the splines fit to the full-sample interval censored data (Fig. 8) is very similar to the spline-based curves obtained

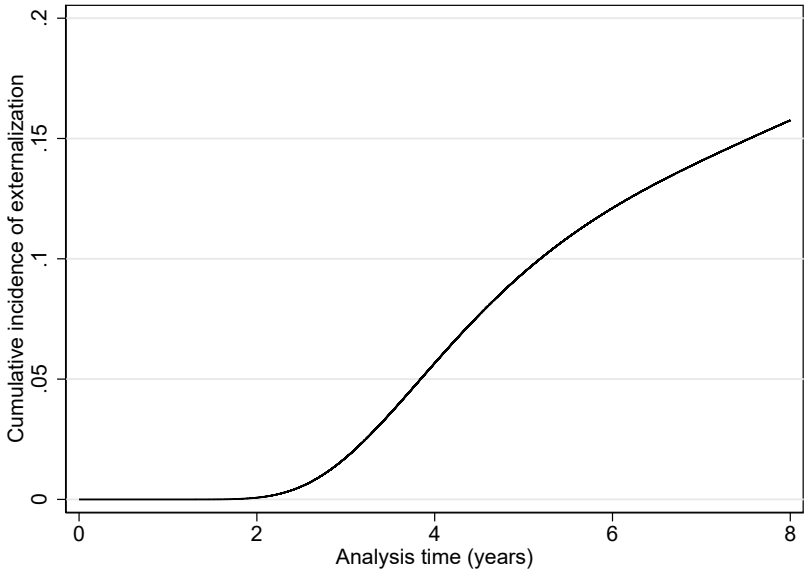


Fig. 8: The full-sample spline-based estimator of the cumulative incidence of externalization in the ICD data.

using the midpoints and a right censored competing risks approach (Fig. 7). The estimate of risk difference at 5 years between high slack and low slack using the parametric pseudo-observations for the interval censored data is 7.1% (0.4% to 13.8%) and the estimate of relative risk is 3.0 (1.1 to 8.1). This is a slightly higher estimate of the absolute difference in risk than we found using the midpoints, while the relative risk is slightly lower.

2.3 Epidemiological context

Interval censored data may occur in a wide range of different epidemiological settings. The data that I have been working with in this thesis is an example of a patient population that is followed by routine examinations to monitor, among other things, the occurrence of the particular event of interest. But any kind of systematic assessments over time can give rise to interval censored data.

An extreme form of interval censoring arises from population screening data where each individual might have a maximum of one examination time; this special case of interval censoring is sometimes referred to as *current status data*. These data are often only used to assess the prevalence of a certain condition in a population measured by the proportion of positive examinations, but by considering the data as interval censored and applying the paramet-

ric pseudo-observation approach it would be possible to estimate cumulative incidence and perform regression analyses and use different meaningful association measures.

Methodological considerations

As already argued in Section 1.8, imputation of either right endpoints or interval midpoints are not guaranteed to produce unbiased estimates of the cumulative incidence and that alone should constitute a serious caution against using these methods if it can be avoided. Using a spline-based method in an illness-death model provides a way to use the available information to the full extent by taking into account the potential events that have not been observed to occur.
The parametric pseudo-observation approach does not impose assumptions about specific shapes of the risk over time but it does require that this can be captured by the spline that is fit to the logarithm of the cumulative hazard function. Although this covers a very wide range of shapes, there are some functional forms that a restricted cubic spline is not technically able to model exactly. In Paper B we simulated event times from a log-logistic distribution, which the flexible parametric approach does not cover, and our simulations indicated that this does not constitute a problem in practice as we neither saw signs of bias nor problems with the width of the confidence intervals. In contrast, all our simulations have shown that the flexible parametric approach produces unbiased estimates with reasonable coverage. However, there are some important assumptions, which are essential for the parametric pseudo-observation approach to produce valid results.

Independent censoring The censoring mechanism(s) must be independent of the event process(es). This assumption can be relaxed by calculating the pseudo-observations within strata of the data. [4] This assumption is the same for non-parametric and parametric pseudo-observations for right censored data as well as for parametric pseudo-observations for interval censored data.

Conditionally independent examination process For interval censored data, the process that controls each examination time must be independent of the event processe(s) given the history since the previous examination. [7] This is a universal assumption for methods for handling interval censored data.

Semi-Markovian illness-death model The illness-death model we are assuming must satisfy the assumption that the transition-specific hazard functions do not depend on the history prior to entering the current state.

It is not possible to evaluate these assumptions on the basis of the collected dataset, and they should thus be justified by considering the particular circumstances and what is known or can reasonably be assumed about the different mechanisms.

Worked example: DANPACE II

In Paper C, we present an analysis of the ongoing trial, DANPACE II, in which patients implanted with a pacemaker are followed for up to two years after the implantation date. The pacemaker devices monitor the heart rhythm of patients and for this analysis we record the date of the first episode of atrial fibrillation after implantation. The atrial fibrillation event is defined as atrial fibrillation lasting for at least 6 minutes. The pacemaker records the date of an atrial fibrillation event and this provides data that can be analysed using methods for right censored data. Moreover, the patients are scheduled to have their pacemaker checked in the hospital routinely after 3, 12, and 24 months of follow-up. If we evaluate the event status only at these checks, we can consider the data as interval censored. Since we do not have the actual examination dates, we have added some random variation to the examination times from a normal distribution with a mean value of zero and a variance of 10 days such that 95% of the patients will have examination times within roughly one week of the scheduled date.
To illustrate the use of the parametric pseudo-observation approach, I will use these data to estimate the cumulative incidence of atrial fibrillation and evaluate the association between atrial fibrillation and the age of the patient by the time of pacemaker implantation.
The overall cumulative incidence of atrial fibrillation can be estimated using both the right censored and the interval censored spline-based methods and Fig. 9 shows both of these curves. It is clear from the exact data that a lot of the pacemaker patients experience an atrial fibrillation within the first couple of months after having their pacemaker implanted. The interval censored data does not contain the necessary information to capture this sharp increase. Consequently, any method will underestimate the increase in the very early follow-up.
First, I will take a look at the data in the dataset and then I will go through the steps needed to perform an analysis based on parametric pseudo-observations. Fig. 10 shows a graphical representation of the trajectories for four of the participants in the trial. They are chosen randomly to represent the four different trajectories that are possible to observe in the study. The time scale is chosen such that each patient enters at time zero. Patient A has an examination after 1 year of follow-up with no atrial fibrillation episode and is then censored after 1.3 years (Trajectory 2). Patient B has an atrial fibrillation episode after roughly 6 months, which results in a negative examination at about 3

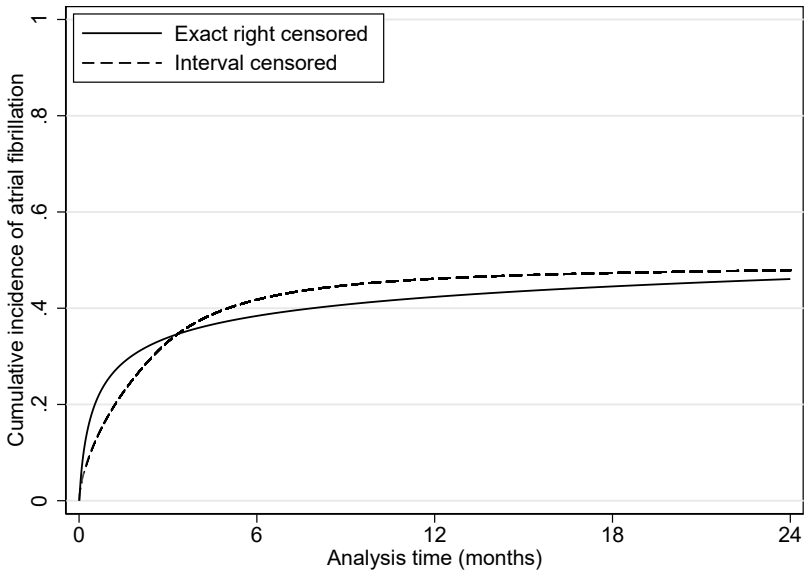


Fig. 9: The spline-based estimators of the cumulative incidence of atrial fibrillation in the DANPACE II data.

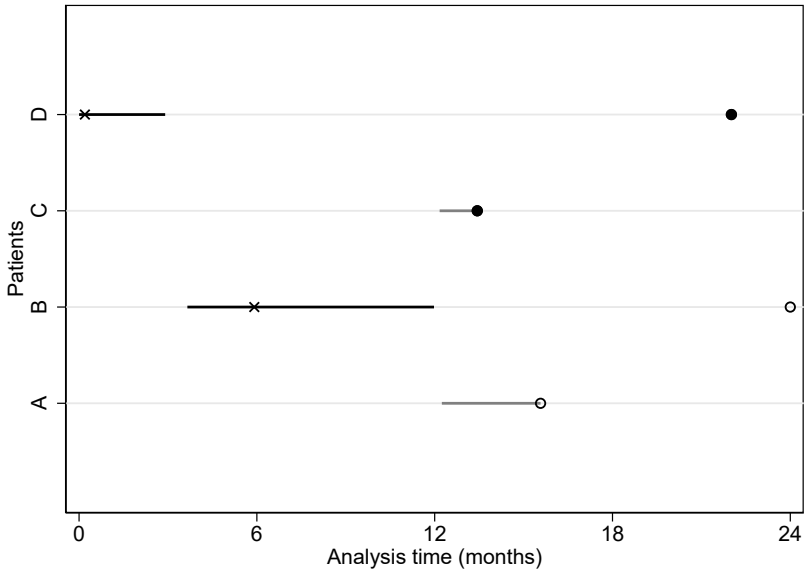


Fig. 10: The trajectories of four patients in the DANPACE II data. Crosses indicate the exact time of atrial fibrillation, black lines indicate the interval in which the event is observed, grey lines indicate the last observed interval for participants in which no atrial fibrillation episode is observed, and black dots indicate death times.

months followed by a positive examination at 1 year, giving rise to an interval censored observed event, and survives for the remainder of the follow-up (Trajectory 3). Patient C has a negative examination after 1 year and then dies shortly after (Trajectory 5). Patient D has an atrial fibrillation episode just after having the pacemaker implantation, which is seen at the 3-month examination, and then dies almost 2 years after the implantation (Trajectory 6). These four patients are coded in the dataset by the variables described in **List 1** on page 19 as shown in Table 1.

Table 1: The values of the variables describing the trajectories of one patient following each of the four trajectories in the DANPACE II data.

Patient	Trajectory	d_{1i}	l_{1i}	t_{1i}	d_{2i}	t_{2i}
A	2	0	1.02	1.30	0	1.30
B	3	1	0.30	1.00	0	2.00
C	5	0	1.01	1.12	1	1.12
D	6	1	0.00	0.24	1	1.83

Suggested workflow

I here present a simple step-by-step workflow for using the parametric pseudo-observation approach for interval censored data. I do not advocate use of any specific statistical software and thus only indicate how to implement each step in general terms.

1. The validity of the assumptions in Section 2.3 should be assessed. Stratification of the steps 3–6 may be necessary if the assumption of independent censoring is violated.

2. The data must be organized in the appropriate way. The exact definitions may depend on the specific software used as different software may impose different conventions, but the basic information to be coded generally follows the content in **List 1**, page 19.

3. Determine number and positions of spline knots to be used. Note that these do not need to be the same for the different transitions. In practice, determining the knot points will be done by repeating the next step for different values and evaluate the trade-off between fit and smoothness by plotting the fitted splines.

4. The illness-death model should then be fit to the full dataset by maximizing the likelihood function defined in Section 2.2. This implies fitting the splines for each of the transitions $0 \to 1$, $0 \to 2$, and $1 \to 2$, thus obtaining estimates, $\hat{\gamma}_{01}, \hat{\gamma}_{02}, \hat{\gamma}_{12}$, of the spline parameters. Use these estimates and (7) to find the full-sample estimator of the cumulative incidence for the event of interest, $\hat{\theta}^{IC}$.

5. Using the spline knots determined in step 3, step 4 should be repeated for each subsample, $i = 1, \ldots, n$, to find the leave-one-out estimators, $\hat{\theta}^{IC}_{(-1)}, \ldots, \hat{\theta}^{IC}_{(-n)}$. For efficiency, it is recommended to use the estimates from step 4 as starting values for the maximization process.

6. The parametric pseudo-observations, $\theta^{IC}_1, \ldots, \theta^{IC}_n$, can then be calculated using the definition in (8).

7. Use appropriate generalized linear models as discussed in Section 1.7 to estimate associations between the outcome of interest and the exposure.

This suggested workflow provides a simple overview of the construction of the parametric pseudo-observations for interval censored data and summarizes the overall conceptual framework that they are based upon.

Worked example, cont.

I conclude this section with an analysis of the DANPACE II data where I consider the association between the atrial fibrillation outcome and age as a continuous exposure variable.
The assumptions are relatively easy to justify in this setting. The only (right) censoring that occurs in the data is the administrative censoring induced by the data collection date by which time some of the participants did not reach 2

years of follow-up. The examination times are determined by a fixed schedule for all individuals with purely random deviations. The semi-Markov property of the assumed illness-death model is only relevant for the transition from having experienced an atrial fibrillation episode to death. The assumption then translates to an assumption that the risk of dying does not depend on the time from pacemaker implantation to the next episode of atrial fibrillation. Since the overall risk of dying in this trial is very low such that only 4 participants die after having experienced the atrial fibrillation event, we can safely assume that any violations of this assumption will not have severe consequences for the analysis.
Due to the homogeneity of the examination times in the study, I have used only 3 knots for the splines, and for the $0 \to 1$ transition they are placed at time 2.3, 3.0, and 23.9 months. I then found the full-sample and leave-one-out estimates of the cumulative incidence of atrial fibrillation evaluated at 6 months, 1 year, and 2 years, and calculated the parametric pseudo-observations on the basis of these values. Fig. 11 shows a histogram of the pseudo-observations calculated at 1 year of follow-up. Andersen & Perme studied the behavior of non-parametric pseudo-observations and saw that they generally take on values in a range that is somewhat larger than the interval from zero to one. [4] The parametric pseudo-observations in this dataset tend mostly to take values close to either zero or one with some larger deviations.
I analyzed the pseudo-observations separately in generalized linear regression models using both the identity link function to estimate the risk difference and the log link function to estimate relative risk while adjusting for the sex of the participants. I have scaled the estimates to correspond to a difference of 10 years and presented them in Table 2. The cumulative incidence is

Analysis time	CIP (95% CI)	RD (95% CI)	RR (95% CI)
6 months	41.4 (36.0 to 46.8)	6.4 (1.5 to 11.3)	1.17 (1.04 to 1.31)
1 year	45.7 (40.0 to 51.3)	6.9 (1.7 to 12.1)	1.16 (1.04 to 1.29)
2 years	47.4 (41.4 to 53.4)	7.1 (1.7 to 12.5)	1.15 (1.04 to 1.28)

CIP: Cumulative incidence proportion (as percentage)
RD: Risk difference (of percentages)
RR: Relative risk

Table 2: Results of regression analyses based on parametric pseudo-observations of the DANPACE II data.

already quite high after 6 months and it only increases moderately after that, which is also clear from Fig. 9. There is some increase in the absolute risk difference between differently aged participants, whereas the relative risk seems very stable over these three analysis time points.

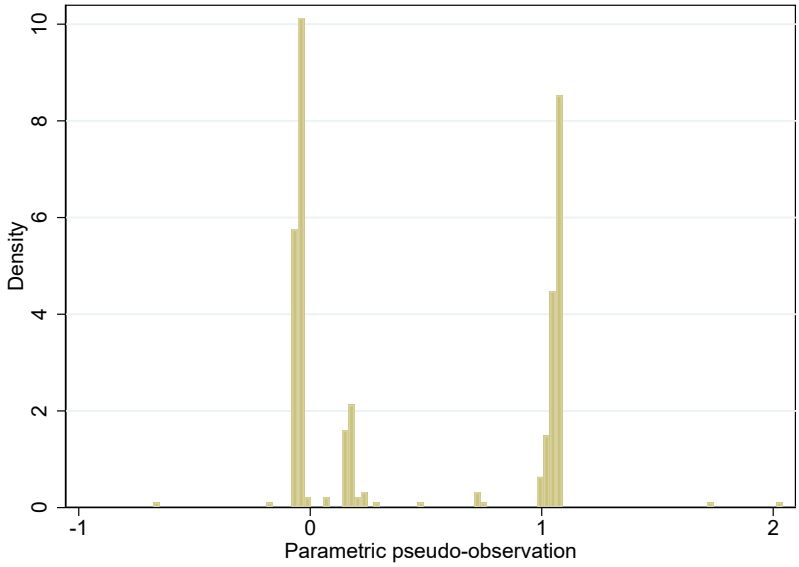


Fig. 11: Histogram of the pseudo-observations evaluated at 1 year.

3 Conclusion

With this project, I have sought to close a methodological gap by developing a workable solution to the practical problem of formulating regression models in a setting with an interval censored event of interest that is subject to competing risks without imposing strict distributional assumptions. In doing so, I have first established an alternative way of calculating pseudo-observations in a right censored competing risks setting. This proved to pose a valuable addition to the existing methods in itself, since the spline-based flexible parametric approach proved to provide benefits over the traditional non-parametric approach. I then continued to extend the methodology to settings with interval censoring of the event of interest and have shown that the method seems to work reasonably well for this more complicated situation. The complex calculations and maximization that this requires does make greater demands on the programming effort and the available computer processing power.

The application of the developed methods would certainly benefit from implementation in publicly accessible software packages. Such software should be developed with a great amount of attention on efficient programming. There is also a great deal of exception handling that should be implemented to make sure that the software creates reasonable output under different circumstances and provides sufficient details on potential problems.

Although the parametric pseudo-observation approach is developed with the intention to minimize the constraints from strict assumptions, there are still some potentially limiting assumptions that for the most part are irremediable. I have already mentioned how the independent censoring assumption can be relaxed by calculating pseudo-observations by stratification, and further ways to accommodate violations of assumptions would increase the applicability of the methods. Another potential way of handling dependence issues is by employing the flexible parametric approach with covariates when splines are fit in the illness-death model. I have not pursued this idea, but it might be considered as a generalization of the stratified pseudo-observation approach. Recently, Cook & Lawless [8] proposed a joint modeling approach to handling dependent loss-to-follow-up censoring with interval censored data that might be transferable to the pseudo-observation approaches. Joint modeling might even be utilized further to ease the assumption of independence of examination times. That would substantially expand the range of applications, since this would refine the method to cover analyses of disease occurrence in situations where examinations might be prompted by symptoms. Most register-based studies of disease incidence assume that the disease occurs at the date of diagnosis and that people who die without having the specific diagnosis do not have the disease.
In conclusion, the methodology that we have developed for both right and interval censored data with competing risks constitute a valid alternative and addition to the existing methods with some interesting possibilities for further development.

References

[1] O. Aalen, "Nonparametric inference for a family of counting processes," *The Annals of Statistics*, vol. 6, 1978.

[2] O. Aalen and S. Johansen, "An empirical transition matrix for non-homogeneous markov chains based on censored observations," *Scandinavian Journal of Statistics*, vol. 5, pp. 141–150, 1978.

[3] P. K. Andersen, J. P. Klein, and S. Rosthøj, "Generalised linear models for correlated pseudo-observations, with applications to multi-state models," *Biometrika*, vol. 90, no. 1, pp. 15–27, 2003.

[4] P. K. Andersen and M. P. Perme, "Pseudo-observations in survival analysis," *Statistical methods in medical research*, vol. 19, no. 1, pp. 71–99, 2010.

[5] N. E. Breslow, "Contribution to the discussion of paper by D.R. Cox," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 216–217, 1972.

[6] T. Cai and R. A. Betensky, "Hazard regression for interval-censored data with penalized spline," *Biometrics*, vol. 59, no. 3, pp. 570–579, sep 2003. [Online]. Available: https://doi.org/10.1111/1541-0420.00067

[7] R. J. Cook and J. F. Lawless, *Multistate models for the analysis of life history data*. Boca Raton, FL: CRC Press, 2018.

[8] ——, "Failure time studies with intermittent observation and losses to follow-up," *Scandinavian Journal of Statistics*, 2020.

[9] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[10] F. J. Dorey, R. J. A. Little, and N. Schenker, "Multiple imputation for threshold-crossing data with interval censoring," *Statistics in Medicine*, vol. 12, no. 17, pp. 1589–1603, 1993. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780121706

[11] F. Graw, T. A. Gerds, and M. Schumacher, "On pseudo-values for regression analysis in competing risks models," *Lifetime Data Analysis*, vol. 15, no. 2, pp. 241–255, 2009.

[12] J. E. Herndon II and F. E. Harrell Jr., "The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables," *Statistics in Medicine*, vol. 14, no. 19, pp. 2119–2129, 1995. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780141906

[13] M. Jacobsen and T. Martinussen, "A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations," *Scandinavian Journal of Statistics*, vol. 43, no. 3, pp. 845–862, 2016.

[14] P. Joly, D. Commenges, C. Helmer, and L. Letenneur, "A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia," *Biostatistics*, vol. 3, pp. 433–443, 2002.

[15] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[16] J. S. Kim, "Maximum likelihood estimation for the proportional hazards model with partly interval-censored data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, pp. 489–502, 5 2003. [Online]. Available: http://doi.wiley.com/10.1111/1467-9868.00398

[17] J. P. Klein and M. L. Moeschberger, *Survival Analysis*. Springer-Verlag, 2003. [Online]. Available: https://doi.org/10.1007/b97377

[18] C. Kooperberg and D. B. Clarkson, "Hazard regression with interval-censored data," *Biometrics*, vol. 53, no. 4, p. 1485, dec 1997. [Online]. Available: https://doi.org/10.2307/2533514

[19] C. Kooperberg and C. J. Stone, "Logspline density estimation for censored data," *Journal of Computational and Graphical Statistics*, vol. 1, no. 4, pp. 301–328, dec 1992. [Online]. Available: https://doi.org/10.1080/10618600.1992.10474588

[20] J. M. Larsen, S. Riahi, J. C. Nielsen, R. Videbaek, J. Haarbo, K. M. Due, D. A. M. J. Theuns, and J. B. Johansen, "Nationwide fluoroscopic screening of recalled riata defibrillator leads in Denmark," *Heart Rhythm*, vol. 10, no. 6, pp. 821–827, 2013.

[21] C. G. Law and R. Brookmeyer, "Effects of mid-point imputation on the analysis of doubly censored data," *Statistics in Medicine*, vol. 11, no. 12, pp. 1569–1578, 1992. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780111204

[22] J. C. Lindsey and L. M. Ryan, "Methods for interval-censored data," *Statistics in Medicine*, vol. 17, no. 2, pp. 219–238, jan 1998. [Online]. Available: https://doi.org/10.1002/(sici)1097-0258(19980130)17:2<219::aid-sim735>3.0.co;2-o

[23] L. M. Mortensen, C. P. Hansen, K. Overvad, S. Lundbye-Christensen, and E. T. Parner, "The pseudo-observation analysis of time-to-event data. example from the danish diet, cancer and health cohort illustrating assumptions, model validation and interpretation of results," *Epidemiologic Methods*, vol. 7, no. 1, oct 2018. [Online]. Available: https://doi.org/10.1515/em-2017-0015

[24] W. Nelson, "Theory and applications of hazard plotting for censored failure data," *Technometrics*, vol. 14, 1972.

[25] P. M. Odell, K. M. Anderson, and R. B. D'Agostino, "Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model," *Biometrics*, vol. 48, no. 3, pp. 951–959, 1992.

[26] M. Overgaard, E. T. Parner, and J. Pedersen, "Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations," *Ann. Statist.*, vol. 45, no. 5, pp. 1988–2015, 10 2017. [Online]. Available: https://doi.org/10.1214/16-AOS1516

[27] ——, "Estimating the variance in a pseudo-observation scheme with competing risks," *Scandinavian Journal of Statistics*, vol. 45, no. 4, pp. 923–940, may 2018.

[28] R. Peto, "Experimental survival curves for interval-censored data," *Applied Statistics*, vol. 22, no. 1, p. 86, 1973. [Online]. Available: https://doi.org/10.2307/2346307

[29] J. O. Ramsay, "Monotone regression splines in action," *Statistical Science*, vol. 3, pp. 425–441, 1988.

[30] G. Rücker and D. Messerer, "Remission duration: An example of interval-censored observations," *Statistics in Medicine*, vol. 7, no. 11, pp. 1139–1145, 1988. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780071106

[31] P. S. Rosenberg, "Hazard function estimation using b-splines," *Biometrics*, vol. 51, no. 3, p. 874, sep 1995. [Online]. Available: https://doi.org/10.2307/2532989

[32] P. Royston and M. K. B. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in Medicine*, vol. 21, no. 15, pp. 2175–2197, aug 2002. [Online]. Available: http://doi.wiley.com/10.1002/sim.1203

[33] C. Sabathé, "pseudoICD," https://github.com/camsabathe/pseudoICD, 2019, [Online; accessed 7 November 2019].

[34] C. Sabathé, P. K. Andersen, C. Helmer, T. A. Gerds, H. Jacqmin-Gadda, and P. Joly, "Regression analysis in an illness-death model with

interval-censored data: A pseudo-value approach," *Statistical Methods in Medical Research*, p. 096228021984227, Apr. 2019. [Online]. Available: https://doi.org/10.1177/0962280219842271

[35] C. Touraine, T. A. Gerds, and P. Joly, "SmoothHazard: An r package for fitting regression models to interval-censored observations of illness-death models," *Journal of Statistical Software*, vol. 79, no. 7, 2017. [Online]. Available: https://doi.org/10.18637/jss.v079.i07

[36] B. W. Turnbull, "The empirical distribution function with arbitrarily grouped, censored and truncated data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 38, no. 3, pp. 290–295, jul 1976. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1976.tb01597.x

Part II

Papers

Paper A

Regression models using parametric pseudo-observations

Martin N. Johansen[1], Søren Lundbye-Christensen[1], and Erik T. Parner[2]

1. Unit of Clinical Biostatistics, Aalborg University Hospital, Aalborg, Denmark; 2. Section for Biostatistics, Department of Public Health, Aarhus University, Aarhus, Denmark.

Published in *Statistics in Medicine*. 2020; 39: 2949– 2961.
https://doi.org/10.1002/sim.8586

Description
In this paper, we develop the concept of flexible parametric pseudo-observations for right censored data with competing risks. The validity of the method is assessed in a large simulation study in which we evaluate the impact of a number of different factors on the performance of the proposed method.

Received: 22 February 2019 | Revised: 1 March 2020 | Accepted: 5 May 2020

DOI: 10.1002/sim.8586

RESEARCH ARTICLE

Statistics in Medicine WILEY

Regression models using parametric pseudo-observations

Martin Nygård Johansen[1] | **Søren Lundbye-Christensen**[1] | **Erik Thorlund Parner**[2]

[1]Unit of Clinical Biostatistics, Aalborg University Hospital, Aalborg, Denmark

[2]Section for Biostatistics, Department of Public Health, Aarhus University, Aarhus, Denmark

Correspondence
Martin Nygård Johansen, Unit of Clinical Biostatistics, Aalborg University Hospital, Forskningens Hus, Sdr Skovvej 15, Aalborg 9000, Denmark.
Email: martin.johansen@rn.dk

Pseudo-observations based on the nonparametric Kaplan-Meier estimator of the survival function have been proposed as an alternative to the widely used Cox model for analyzing censored time-to-event data. Using a spline-based estimator of the survival has some potential benefits over the nonparametric approach in terms of less variability. We propose to define pseudo-observations based on a flexible parametric estimator and use these for analysis in regression models to estimate parameters related to the cumulative risk. We report the results of a simulation study that compares the empirical standard errors of estimates based on parametric and nonparametric pseudo-observations in various settings. Our simulations show that in some situations there is a substantial gain in terms of reduced variability using the proposed parametric pseudo-observations compared with the nonparametric pseudo-observations. The gain can be measured as a reduction of the empirical standard error by up to about one third; corresponding to an additional 125% larger sample size. We illustrate the use of the proposed method in a brief data example.

KEYWORDS
flexible parametric models, pseudo-observations, time-to-event

1 | INTRODUCTION

The Cox model proposed by Cox[1] is by far the most applied regression model for time-to-event data in the medical literature for comparing rates of events. The model is often formulated as a regression model of the hazard function, $h(t)$, on a set of regression variables, $\boldsymbol{z}$, as

$$\ln\left(h(t;\boldsymbol{z})\right) = \ln\left(h_0(t)\right) + \boldsymbol{\beta}^{\intercal}\boldsymbol{z}, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. The Cox model in (1) assumes proportionality of the hazard functions for different values of regression variables and independent censoring conditionally on the regression variables. The model is a semiparametric model in that the baseline hazard function, $h_0(t)$, is left unspecified, whereas the comparisons between rates are described by the hazard rate ratios $\exp(\boldsymbol{\beta})$. The hazard rate ratios are estimated using Cox partial likelihood and the integrated baseline hazard function, $H_0(t) = \int_0^t h_0(s)ds$, is estimated semiparametrically by the Breslow estimator[2] which is a step function with jumps at observed event times. An important application of the Cox model is in the presence of competing risks where the model can be applied to the cause-specific hazard functions.[3]

As an alternative to the semiparametric Cox model, Royston and Parmar have suggested to perform a fully parametric regression analysis by formulating a *flexible parametric* model based on modeling of the log cumulative hazard function

using splines. This model can be written on the log cumulative hazard scale as

$$\ln\left(H(t;\boldsymbol{z})\right) = s(\ln(t);\boldsymbol{\gamma}) + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{z}, \tag{2}$$

where $s(\cdot;\boldsymbol{\gamma})$ is a restricted cubic spline (for further details on restricted cubic splines, see Appendix A1). One advantage of this approach is that a parametric model yields more precise estimates of model parameters.[4] It also provides smooth estimates of the baseline survival and hazard functions which may be favorable, particularly for datasets with few events. Smooth estimates are also preferable if data protection authorities do not permit presenting data based on single individual data as a step function with jumps at observed single event times.

A different modeling approach is the use of *pseudo-observations* which was proposed with the objective of analyzing other effect measures than the hazard rate ratio. The idea is to create a transformation of the time-to-event data, the pseudo-observations, that are analyzed using estimating equations for a generalized linear model (GLM).[5] The original proposed method is based on a nonparametric estimator of a parameter of interest and, hence, we will refer to it as the *nonparametric pseudo-observation* approach. The pseudo-observation approach offers great versatility in terms of effect measures in addition to the hazard rate ratio, for example, also risk difference (RD), risk ratio (RR), difference in t_0-restricted mean life time,[6] and life-years lost due to a specific cause.[7] In a setting with competing risks, it is most common to work on the cause-specific cumulative incidence scale to obtain estimates of RRs. In the following, we will hence discuss pseudo-observations on the cumulative incidence scale only even though the choice of scale is not fixed in general applications.

In this article, we propose a modeling approach that uses a combination of the spline-based estimation of the hazard function and the pseudo-observation approach by calculating *parametric pseudo-observations* using an estimate of the cumulative incidence proportion (CIP) obtained from a flexible parametric model in order to estimate different effect measures. By combining the approaches of the flexible parametric model and pseudo-observations we aim to obtain pseudo-observations with greater precision than the nonparametric version. The suggested approach models effects on the cumulative incidence scale similarly to the nonparametric pseudo-observation approach while it takes advantage of the features of a parametric model similarly to the flexible parametric model. In Section 2, we will give a short introduction to the nonparametric pseudo-observations and present the proposed modeling approach. In Section 3, we present a simulation study that compares the potential bias and the efficiency of the proposed approach to that of the nonparametric pseudo-observations. We then show an application of the method in Section 4 and conclude the article with a discussion of the observed advantages and disadvantages of the different modeling approaches in Section 5.

2 | METHODOLOGICAL DETAILS

2.1 | Nonparametric pseudo-observations

The basic idea behind the pseudo-observation approach is to make a transformation of the time-to-event data that provides a dataset without censoring which can be used in place of the original censored observations.

Let (T_i, D_i) denote the survival time and the event type indicator for K event types for the i'th subject. We are interested in modeling the cause-specific cumulative incidence function for a given event type d, $F_d(t)$, and we will estimate this using the Aalen-Johansen estimator.[8] The pseudo-observations can then be thought of as the contribution for each observation to this estimator.

If we denote the nonparametric estimator based on all n observations as $\hat{\theta}^{np}$, and the leave-one-out estimator based on all observations except the i'th as $\hat{\theta}^{np}_{(-i)}$, the i'th pseudo-observation is defined as

$$\theta^{np}_i = n\hat{\theta}^{np} - (n-1)\hat{\theta}^{np}_{(-i)}. \tag{3}$$

To calculate the pseudo-observations we must choose one or more time points at which to evaluate the cumulative incidence function. For simplicity, we will refer to just one time point, t, in the following and we restrict the focus of the remainder of this article to one time point.

In the special case with no competing risks, the Aalen-Johansen estimator of the cumulative incidence function reduces to one minus the Kaplan-Meier estimator of the survival function.[9]

The pseudo-observations based on the Aalen-Johansen estimator have the property that, for large samples, $\mathrm{E}[\theta_i^{np}|\boldsymbol{z}_i] \approx \mathrm{E}[\mathrm{I}(T_i \leq t, D_i = d)|\boldsymbol{z}_i] = F_d(t|\boldsymbol{z}_i)$, where $F_d(t|\boldsymbol{z}_i)$ is the conditional probability that an event of type d has occurred in the i'th subject prior to time t given the covariates.[10] Hence, the pseudo-observations provide a set of data points that can be used in a generalized linear regression model to obtain estimates of effects on the cause-specific cumulative incidence

$$g\left(F_d(t|\boldsymbol{z}_i)\right) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i, \tag{4}$$

where g is a link function. When several time points are considered, the GLM should take the correlation between pseudo-observations at different time points for the same individual into account.

The GLM regression for the pseudo-observations provides an unbiased, though probably not optimal, estimate of the regression parameter $\boldsymbol{\beta}$ and the standard deviation of this can be estimated using a sandwich estimator. Jacobsen and Martinussen[11] and Overgaard et al[12] have recently proved that estimation of the standard error can be improved slightly. However, the gain from this improvement has been found to be negligible. Here, we will use the computationally much simpler, but slightly conservative sandwich estimator.

Unlike in the Cox and flexible parametric models, simple uses of the pseudo-observation approach requires that censoring is independent of covariates but, following Andersen and Perme,[13] this extra assumption may be avoided when censoring only depends on categorical covariates by stratifying the calculation of pseudo-observations on the relevant variables. Binder et al have suggested an alternative approach to handling covariate-dependent censoring which is based on inverse probability of censoring weighting.[14] Mortensen et al[15] have performed a comprehensive review of the validation of assumptions for nonparametric pseudo-observations and potential remedies for violations.

2.2 | Proposed approach: Parametric pseudo-observations

To obtain a potential improvement of the nonparametric pseudo-observations, we will use a very simple version of the flexible parametric model without any regressors to estimate the baseline log cumulative hazard function. This will then be transformed to an estimator of the marginal cumulative incidence function which can be used to calculate parametric pseudo-observations. These parametric pseudo-observations could then be used to form regression models using GLM regression as is done for nonparametric pseudo-observations.

The transformation of the estimator of the log cumulative hazard function to an estimator of the cumulative incidence function is based on the relation between the cause-specific cumulative incidence function and the cause-specific hazard functions, $h_1(\cdot), \ldots, h_K(\cdot)$,

$$F_d(t) = \int_0^t h_d(u) \exp\left(-\int_0^u \sum_{k=1}^{K} h_k(v)dv\right) du.$$

This expression cannot be solved analytically, and estimation of the cumulative incidence function must be based on numerical methods using estimates of each of the K cause-specific hazard functions.[16] However, in a setting with no competing risks, the expression simplifies to

$$F(t) = 1 - S(t) = 1 - \exp\left(-H(t)\right),$$

and the cumulative incidence function can be estimated directly using the fitted spline from the model in (2) by $1 - \exp\left(-\exp\left(s(\ln(t);\boldsymbol{\gamma})\right)\right)$. We will denote the parametric spline-based estimator of the cumulative incidence function thus obtained as $\hat{\theta}^p$ and its i'th leave-one-out counterpart as $\hat{\theta}^p_{(-i)}$. The splines based on leave-one-out samples are fitted using the same knot positions as the full-sample version.

Having obtained this, we can define a set of parametric pseudo-observations as

$$\theta_i^p = n\hat{\theta}^p - (n-1)\hat{\theta}^p_{(-i)}. \tag{5}$$

This implies that calculation of the pseudo-observations requires n spline estimations of the log cumulative hazard function. The parametric pseudo-observations can then be analyzed using a GLM regression as in (4) and the standard errors of the regression coefficients, $\boldsymbol{\beta}$, are obtained by a sandwich estimator as in Andersen and Perme.[13]

The assumption about marginal independence of censoring that applies to nonparametric pseudo-observations is also applicable for the proposed parametric pseudo-observations. For the parametric pseudo-observations, calculating stratified pseudo-observations corresponds to the estimation of log cumulative hazard by a different spline for each level of the stratification variable. However, the censoring process can also be modeled using the approach of Binder et al[14] for nonparametric pseudo-observations.

The proposed approach requires the analyst to determine the number and location of knots to use for fitting the splines. This obviously introduces an additional uncertainty that is not accounted for analytically. Care should be given to the choice of knots and it is recommended to investigate the potential impact of alternative choices.[17]

3 | SIMULATION STUDY

3.1 | Simulation strategy

We have evaluated the performance of the nonparametric and parametric pseudo-observation approaches in seven different scenarios. Each scenario is a variation of a general set-up in which one aspect is varied. The general set-up is chosen to mimic a typical clinical or register-based study comparing two exposure groups of individuals, denoted as exposed and nonexposed, with staggered entry. Individuals enter the study during an accrual period of 6 years which is followed by a follow-up until 13 years after the beginning of the accrual period. This implies that individuals have a potential follow-up for at least 7 and at most 13 years. Where nothing else is stated, we have used a fixed sample size of $n = 500$. The aim in each setting was to estimate the overall event probability (the CIP) at 10 years as well as the effect measures RD and RR comparing the two exposure groups also evaluated at 10 years. This means that there is some information in the observed data which lies after the analysis time point; in this case up to 3 years of follow-up.

In each scenario, we generated 5000 replications of datasets for each variation. We considered 50% of each sample exposed and 50% nonexposed according to a nonrandom allocation. We then simulated time-to-event data from exponential distributions such that the cumulative incidence of the event at $t = 10$ was $\frac{1}{3}$ in nonexposed individuals (ie, a rate parameter of $\frac{1}{10}\log(\frac{1}{3})$) and $\frac{2}{3}$ in exposed individuals (rate parameter of $\left(\frac{1}{10}\log\left(\frac{2}{3}\right)\right)$. This gives an overall CIP of 0.5, a RD of $\frac{1}{3}$, and a RR of 2 at $t = 10$. We imposed a uniform censoring on the interval from 7 to 13 to serve as the administrative censoring assuming a constant accrual rate. In addition, we simulated independent loss-to-follow-up censoring from an $\text{Exp}\left(\frac{1}{10}\log\left(\frac{1}{6}\right)\right)$ distribution. We then estimated CIP, RD, and RR at time $t = 10$ in each sample using both nonparametric and parametric pseudo-observations with three knots for the splines. After 10 years of follow-up, the expected size of the risk set is 20.8% of the original sample size.

The seven scenarios we consider are designed to assess the influence of (1) the sample size, (2) the number of knots for the fitted splines with more complex time-to-event distributions, (3) competing risks of different intensity, (4) model misspecification, (5) the amount of additional information after the analysis time point, (6) using a smooth estimator, and (7) covariate-dependent censoring.

All simulations and analyses were performed in Stata/MP version 15.1 and an example of how to calculate parametric pseudo-observations and estimate effect measures is shown in the Supporting Information. We used the Stata packages `survsim`,[18] `stpsurv`, `stpci`[19] and `stpm2`.[17]

Scenario (1) *Sample size* In this scenario, we simulated data based on the general set-up with varying sample sizes of $n = 50, 100, 250, 500$, and 1000.

Scenario (2) *Complex distributions* The time-to-event data were simulated from a two-component mixture Weibull distribution instead of exponential distributions to simulate more complex time-to-event distributions. In this scenario, we varied the number of internal knots from 1 to 9.

Scenario (3) *Competing risks* We introduced a competing risk by simulating a second event process which is associated to the exposure similarly to the event of interest but with rate parameters yielding hazard ratios relative to the event of interest equal to .5, .75, 1, 1.5, and 2.

Scenario (4) *Model misspecification* The time-to-event data were simulated from a log-logistic distribution in which the cumulative hazard function cannot be represented as a cubic spline and we varied the number of internal knots from 1 to 9.

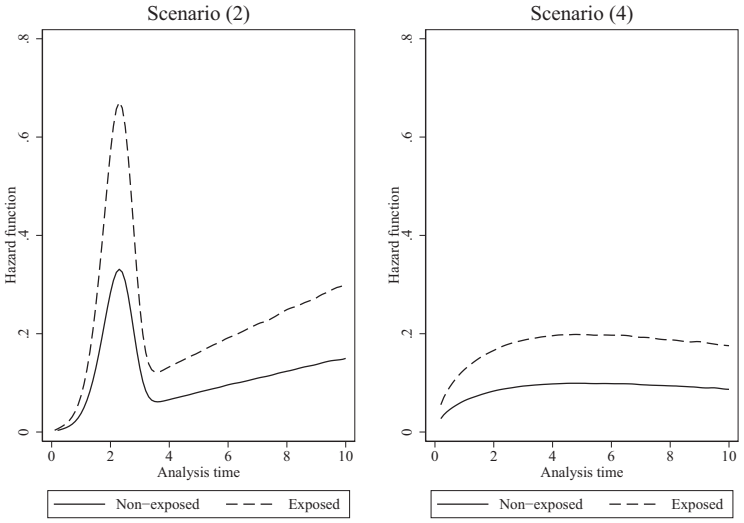


FIGURE 1 Hazard functions from the mixture Weibull and the log-logistic distribution used in Scenarios (2) and (4)

Scenario (5) *Additional information* In this scenario there was no accrual period such that by varying the administrative censoring time over the range $10, 11, \ldots, 20$, we adjusted the amount of additional information used when fitting the splines.

Scenario (6) *Smoothing* Without an accrual period and with administrative censoring at the analysis time point, we introduced a uniform censoring on the interval from $t = 0$ to $t = 10, 11, \ldots, 20$ to reduce the risk set to varying percentages of the original sample size at the analysis time point.

Scenario (7) *Covariate-dependent censoring* We introduced a covariate-dependent censoring by simulating the loss-to-follow-up censoring to have twice of the original rate in the exposed group while we retained the original rate in the nonexposed group. In this scenario, we performed both a naive analysis ignoring the dependent censoring and an analysis in which we calculated pseudo-observations with stratification on the exposure variable. For completeness, we also included an analysis with stratification but no dependent censoring.

The two-component mixture Weibull distribution in Scenario (2) is determined by two sets of scale and shape parameters, $(\lambda_1, \gamma_1), (\lambda_2, \gamma_2)$, and a mixture probability, p. The cumulative incidence function can then be expressed as

$$F(t) = 1 - (p \exp(-\lambda_1 t^{\gamma_1}) + (1 - p) \exp(-\lambda_2 t^{\gamma_2})) .$$

We used the parameters $(\lambda_1, \gamma_1) = (.01, 1.9)$, $(\lambda_2, \gamma_2) = (.015, 5)$, $p = .7$. The cumulative incidence function for the log-logistic distribution used in Scenario (4) can be expressed as

$$F(t) = \frac{t^{\beta}}{\alpha^{\beta} + t^{\beta}},$$

and we simulated data with $(\alpha, \beta) = (8, 1.5)$. The hazard functions from the mixture Weibull and the log-logistic distributions are shown in Figure 1.

We analyzed each dataset using both nonparametric and parametric pseudo-observations to obtain parameter estimates and 95% confidence intervals at time $t = 10$.

For each modeling approach, we report the empirical standard error, bias calculated as the difference between the median of the estimates and the corresponding theoretical value, the coverage of the confidence intervals, the relative efficiency calculated as the ratio of the empirical standard errors from the two approaches, and the corresponding ratio of required sample sizes.[20,21]

TABLE 1 Precision of the cumulative incidence estimates based on 5000 replications in Scenarios (1) to (3)

		Nonparametric POs			Parametric POs				
		SD	Bias	Cov. (%)	SD	Bias	Cov. (%)	Relative efficiency	Relative sample size
Scenario (1)	*n*								
	50	0.077	−0.002	95.2	0.072	0.000	95.3	1.07	1.16
	100	0.054	0.001	95.6	0.050	0.001	96.1	1.08	1.17
	250	0.034	0.000	95.9	0.032	0.000	95.8	1.07	1.14
	500	0.024	0.000	96.3	0.022	0.000	96.3	1.08	1.17
	1000	0.017	0.000	95.3	0.016	0.000	95.7	1.08	1.17
Scenario (2)	Knots								
	1	0.021	0.000	95.6	0.017	−0.005	95.3	1.19	1.42
	2	0.021	0.000	95.8	0.018	−0.005	95.5	1.15	1.31
	3	0.020	0.000	95.7	0.019	0.007	93.8	1.10	1.21
	4	0.021	0.000	95.4	0.019	0.002	94.8	1.12	1.24
	5	0.021	0.001	95.2	0.019	0.002	95.1	1.13	1.27
	6	0.021	0.000	95.3	0.019	0.001	95.3	1.13	1.27
	7	0.021	0.000	95.3	0.019	0.002	95.5	1.13	1.27
	8	0.021	0.000	95.5	0.019	0.001	95.4	1.11	1.23
	9	0.021	0.000	95.1	0.019	0.001	95.7	1.12	1.26
Scenario (3)	HR								
	0.5	0.024	−0.001	95.2	0.023	0.000	95.4	1.05	1.10
	0.75	0.023	0.000	95.8	0.022	0.000	95.3	1.04	1.09
	1	0.023	0.000	95.2	0.023	0.001	94.8	1.03	1.06
	1.5	0.022	0.000	94.4	0.022	0.000	94.5	1.02	1.05
	2	0.021	−0.001	95.0	0.021	0.000	95.2	1.02	1.05

Abbreviations: SD: Empirical standard error, defined as standard deviation of parameter estimates.
Bias: Defined as absolute deviation of median parameter estimates from true parameter values.
Cov.: Coverage, defined as percentage of estimated 95% confidence intervals containing the true parameter values.
Relative efficiency: Defined as standard error of nonparametric divided by standard error of parametric estimates.
Relative sample size: Defined as relative efficiency squared.
HR: Hazard ratio between the competing event and the event of interest.

3.2 | Nonvalid estimates

We recorded how many times the estimation procedures failed to produce valid estimates for each of the three parameters. We defined as nonvalid estimates cases where the GLM estimation procedure did not converge or the estimate was unrealistically small or large (CIP estimates outside the range $(0; 1)$, RD estimates outside the range $(-1; 1)$, and RR estimates outside the range $(\frac{1}{10}; 10)$). The remaining estimates were considered valid.

3.3 | Simulation results

In the general set-up, the simulated datasets contained approximately 15% and 30% observed events in the nonexposed and exposed groups, respectively. Since the cumulative incidence function is the foundation of the effect measures RD and RR, we show the simulation results for the CIP estimates only in tables and figures. The corresponding results for RD and RR are quite similar and can be found in the Supporting Information. We show empirical standard error, bias, coverage probability, relative efficiency, and relative sample size of CIP estimates for Scenarios (1) to (6) in Tables 1 and 2.

TABLE 2 Precision of the cumulative incidence estimates based on 5000 replications in Scenarios (4) to (6)

		Nonparametric POs			Parametric POs				
		SD	Bias	Cov. (%)	SD	Bias	Cov. (%)	Relative efficiency	Relative sample size
Scenario (4)	Knots								
	1	0.023	0.001	95.0	0.021	0.002	94.9	1.09	1.18
	2	0.023	0.000	95.2	0.021	−0.001	95.6	1.10	1.21
	3	0.023	0.001	95.6	0.021	0.000	95.6	1.09	1.20
	4	0.023	−0.001	95.3	0.021	−0.002	95.5	1.09	1.18
	5	0.023	0.000	95.7	0.021	−0.001	96.0	1.10	1.20
	6	0.023	0.000	95.8	0.021	−0.001	96.0	1.09	1.20
	7	0.023	0.000	95.3	0.021	−0.001	95.8	1.08	1.18
	8	0.023	0.000	95.2	0.021	0.000	95.6	1.09	1.18
	9	0.023	0.000	95.4	0.021	−0.001	95.7	1.07	1.15
Scenario (5)	Years								
	0	0.022	0.000	96.1	0.022	0.000	96.1	1.00	1.00
	1	0.022	0.000	96.2	0.021	−0.001	96.2	1.05	1.10
	2	0.022	0.000	96.4	0.020	−0.002	96.9	1.11	1.22
	3	0.022	0.000	96.2	0.020	−0.002	96.2	1.11	1.24
	4	0.023	0.000	95.7	0.020	−0.002	96.1	1.13	1.29
	5	0.022	0.001	95.8	0.019	−0.003	96.2	1.15	1.33
	6	0.022	−0.001	95.8	0.020	−0.004	95.4	1.14	1.29
	7	0.022	−0.001	95.9	0.019	−0.004	96.0	1.16	1.35
	8	0.022	0.000	96.1	0.019	−0.003	96.6	1.18	1.39
	9	0.022	0.000	96.3	0.019	−0.004	96.1	1.17	1.37
	10	0.022	0.000	96.1	0.019	−0.004	96.0	1.16	1.35
Scenario (6)	Average risk set (%)								
	0	0.056	−0.009	92.2	0.037	0.004	95.2	1.51	2.27
	4	0.040	−0.001	95.4	0.034	0.003	95.6	1.17	1.37
	7	0.035	0.000	95.9	0.032	0.002	96.0	1.10	1.22
	10	0.033	0.000	95.2	0.031	0.002	95.0	1.06	1.13
	12	0.032	0.000	95.4	0.030	0.001	95.9	1.06	1.13
	14	0.030	−0.001	95.8	0.029	0.001	95.8	1.04	1.07
	16	0.029	0.001	95.7	0.028	0.002	95.9	1.03	1.07
	17	0.028	0.000	96.1	0.028	0.002	96.2	1.02	1.05
	19	0.029	−0.001	95.2	0.028	0.000	95.2	1.03	1.05
	20	0.028	0.000	95.7	0.027	0.001	95.9	1.02	1.05
	21	0.027	0.000	95.8	0.027	0.001	95.6	1.02	1.04

Abbreviations: SD: Empirical standard error, defined as standard deviation of parameter estimates.
Bias: Defined as absolute deviation of median parameter estimates from true parameter values.
Cov.: Coverage, defined as percentage of estimated 95% confidence intervals containing the true parameter values.
Relative efficiency: Defined as standard error of nonparametric divided by standard error of parametric estimates.
Relative sample size: Defined as relative efficiency squared.
HR: Hazard ratio between the competing event and the event of interest.

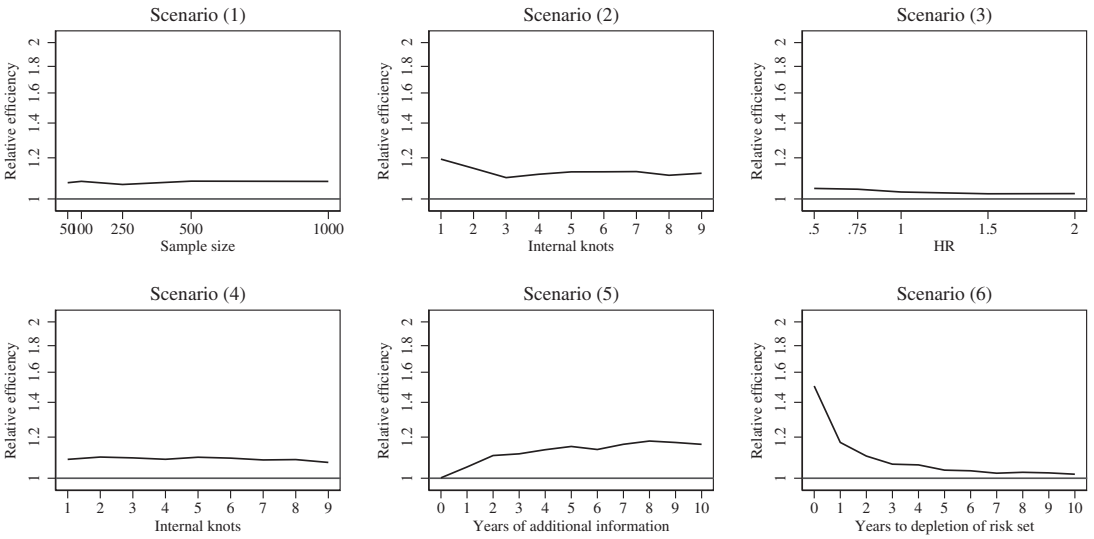


FIGURE 2 Empirical relative efficiency between parametric and nonparametric cumulative incidence estimates in each of the simulation Scenarios (1) to (6)

TABLE 3 Empirical bias and relative efficiency of the cumulative incidence estimates based on 5000 replications in Scenario (7)

		Nonparametric POs			Parametric POs				
Dep. cens.	Strat.	SD	Bias	Cov. (%)	SD	Bias	Cov. (%)	Relative efficiency	Relative sample size
No	No	0.024	0.000	95.5	0.022	0.000	95.6	1.09	1.18
No	Yes	0.024	−0.001	96.3	0.022	0.000	96.4	1.09	1.18
Yes	No	0.024	−0.008	95.2	0.023	−0.007	95.2	1.06	1.13
Yes	Yes	0.024	0.000	96.4	0.022	0.001	96.8	1.09	1.19

Abbreviations: Dep. cens.: Covariate-dependent censoring.
Strat.: Pseudo-observations calculated using stratification.
SD: Empirical standard error, defined as standard deviation of parameter estimates.
Bias: Defined as absolute deviation of median parameter estimates from true parameter values.
Cov.: Coverage, defined as percentage of estimated 95% confidence intervals containing the true parameter values.
Relative efficiency: Defined as standard error of nonparametric divided by standard error of parametric estimates.
Relative sample size: Defined as relative efficiency squared.

The relative efficiency of the CIP estimates for Scenarios (1) to (6) are visualized in Figure 2. The results for Scenario (7) are shown in Table 3.

3.3.1 | Empirical standard error

For both the nonparametric and parametric pseudo-observations, the empirical standard error decreases regressively with increasing sample size. For the different amounts of additional information in Scenario (5), the nonparametric pseudo-observation empirical standard error is completely stable since no additional information is used, but for the parametric pseudo-observations the empirical standard error decreases slightly at first with increasing amount of additional information and then seems to level off. In Scenario (6) with varying sizes of the risk set, the

empirical standard errors of both the nonparametric and the parametric pseudo-observations are increased with small risk sets, though the nonparametric pseudo-observations show this tendency to a greater extent than the parametric pseudo-observations. In the other scenarios, the empirical standard errors of the two methods show quite similar behaviors.

3.3.2 | Bias

In general, the biases are very low. The only situations that produced noteworthy biases in our simulations were the complex time-to-event distributions in Scenario (2) with an inappropriate number of spline knots, the settings with a very small risk set in Scenario (6) and the settings with covariate-dependent censoring in Scenario (7) when it was not accounted for in the analyses. In Scenario (2), using too few knots, in this case 3 or less, seemed to induce some bias in the parametric pseudo-observation approach. In Scenario (7), ignoring a covariate-dependent censoring lead to bias of similar magnitude in either approach.

3.3.3 | Coverage

Coverage probabilities are quite similar for the two modeling approaches and keep within 94% to 96% in almost all cases. Our simulations showed no systematic deviations in any of the simulation scenarios, the only exceptions being that both approaches seem to produce a slight overcoverage in Scenario (5).

3.3.4 | Relative efficiency

Across the different sample sizes in Scenario (1), the relative efficiency is quite stable at 1.07 to 1.08. With the more complex time-to-event distributions in Scenario (2), we generally observe quite stable larger relative efficiency with slightly higher values for low numbers of spline knots. The relative efficiency is mitigated when a competing event is introduced in Scenario (3) depending on the rate of the competing event process. The model misspecification in Scenario (4) does not seem to influence the relative efficiency regardless of the chosen number of spline knots. Scenario (5) shows that the relative efficiency increases with increasing amount of additional information used from almost exactly 1 with no additional information to a level of about 1.15 which is reached after adding 5 years of additional information. In Scenario (6), we see that the size of the risk set exerts the greatest influence on the relative efficiency when the data is analyzed at the time when the risk set is depleted. In this setting the relative efficiency reaches a level of approximately 1.5 which then wears off with larger risk sets.

3.3.5 | Nonvalid estimates

We only observed estimates that were not considered valid in some of the most extreme of our simulation settings. Particularly for a sample size of only $n = 50$, the nonparametric approach resulted in 15 cases of nonvalid RR estimates of the 5000 replications, whereas the parametric approach resulted in 10 nonvalid RR estimates. When the risk set is completely depleted at the analysis time point, the nonparametric pseudo-observation approach failed to produce valid estimates of both CIP, RD, and RR in a few replications (14, 19, and 15, respectively) while the parametric approach always produced valid estimates.

4 | DATA EXAMPLE

We illustrate the use of both the nonparametric and the parametric pseudo-observations by estimating the effect of the patients' age among the 855 ICU patients in a publicly available simulated dataset.[22] In this dataset, there is information on length of ICU stay and death for each patient. We consider time from ICU admission to both a composite endpoint of either discharge alive or death as well as discharge alive where death is considered as a competing event. We measure the

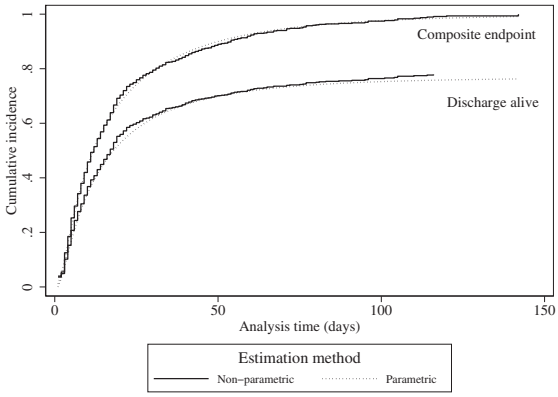


FIGURE 3 Estimated cumulative incidence of discharge before death and composite endpoint of death and discharge in the sample from the pneumonia dataset

TABLE 4 Parameter estimates and relative efficiency in the pneumonia dataset calculated at analysis time point $t = 120$ days using nonparametric and parametric pseudo-observations with three internal knots

Analysis approach	RR	SE	95% CI	Relative efficiency	Relative sample size
Composite endpoint					
Nonparametric	0.998	0.001	(0.997 to 1.000)	–	–
Parametric	0.997	0.001	(0.996 to 0.998)	1.50	2.24
Discharge alive					
Nonparametric	0.991	0.002	(0.987 to 0.995)	–	–
Parametric	0.989	0.002	(0.986 to 0.993)	1.11	1.23

Abbreviations: CI, confidence interval; RR, relative risk estimate; SE, robust standard error.

effect of age linearly on the log relative risk scale. The study population is comprised of ICU patients who are followed from admission to either death or discharge with the longest observed follow-up being 142 days. We evaluate the effect of age at the time point 120 days. After 120 days, 176 patients had died before being discharged, 641 had been discharged alive, 35 had been censored and the remaining three patients were still at risk in the study.

Figure 3 shows the cumulative incidence functions for both outcomes estimated nonparametrically and by a flexible parametric estimator with three internal knots. We then estimated the RR using both nonparametric and parametric pseudo-observations with three internal knots. We chose three internal knots to allow the hazard functions to show some variation over follow-up time but still limit the risk of overfitting. To calculate nonparametric pseudo-observations, we used the Stata packages `stpsurv` for the composite endpoint and `stpci`[19] for discharge alive with competing risk. For the parametric pseudo-observations, we used the `stpm2` package[17] to fit the full-sample and leave-one-out spline-based estimators of the cumulative incidence functions. The results of the analyses are shown in Table 4. The RR comparing a 1-year difference in age for the composite endpoint estimated by the nonparametric approach is 0.998 with a robust standard error of 0.001 and the estimate from the parametric approach with three internal knots is 0.997 with a relative efficiency of 1.50. As sensitivity analyses, we examined the influence of varying the number of internal knots from 1 to 9. The RR estimates for different numbers of internal knots are quite similar and the relative efficiency range from 1.63 achieved using only one internal spline knot to 1.36 with nine internal knots. The estimates for discharge alive show similar tendencies also giving larger relative efficiency with fewer internal spline knots.

In this example, we observe relative efficiencies of the RR estimates without competing events that are comparable to the results from the setting in Scenario (6) of our simulation study where the risk set is depleted at the analysis time point while the relative efficiency in the competing risk setting is mitigated which is also in accordance with our simulation results.

5 | DISCUSSION

We suggest parametric pseudo-observations to improve precision as compared with nonparametric pseudo-observations. In our simulations, we observed a reduced variability of the parameter estimates using our proposed parametric pseudo-observations compared with that of the traditional nonparametric pseudo-observations. The reduction in estimate variability was evident for the estimation of absolute values of the cumulative incidence as well as the effect measures RD and RR. The reduction in estimate variability depends on both the amount of additional follow-up time after the analysis time point and the size of the risk set at the analysis time point. When pseudo-observations are calculated at the end of the potential follow-up, there is a substantial relative efficiency, and it also increases with increasing amount of additional follow-up beyond the analysis time point. The large relative efficiency at analysis time points where the risk set is small is primarily explained by an increased variability of the nonparametric pseudo-observations. The observed reduction in variability in our simulation settings translates to a reduction in required sample size of up to 127% for a specific precision and this could be a significant gain in interventional studies if gathering data is costfull or time-consuming. However, the gain is reduced when competing risks are introduced.

In our simulation study, we have identified two mechanisms that contribute to the gain in efficiency of the parametric pseudo-observations; use of additional information beyond the analysis time point and instability of the nonparametric estimators when the risk set is very small. The gain obtained from using additional information beyond the analysis time point is caused by the fact that the spline in the parametric approach can be fitted using information from events during the entire observed follow-up whereas the nonparametric pseudo-observations based on the Aalen-Johansen estimator do not take events occurring after the analysis time point into account. When the risk set is very small, the size of the jumps in the nonparametric estimators at the observed event times depends heavily on the size of the risk set at that specific time which gives rise to a greater uncertainty in the estimated cumulative incidence.

In the presence of competing risks, the gain in efficiency decreases with increasing intensity of the competing event. This might be explained by the lower number of observed events of interest and the larger number of parameters to be estimated than in the absence of competing risks.

We based most of our simulations on exponential distributions in which the log cumulative hazard function is linear but supplemented this with two settings with more complex hazard functions. A simulation study by Rutherford et al[23] has shown that the flexible parametric approach is well-suited for modeling more complicated shapes provided that an adequate number of knots are chosen for the splines. Furthermore, we observed that while small nonsystematic biases were introduced when analyzing a complex setting with few internal knots the relative efficiency was generally higher than what we observed in the simple setting using exponential distributions at the same sample size of 500. We suspect that this is because the spline-based approach is best suited to capture the shape more complex hazard functions as compared with a more trivial setting.

We chose to study small and moderate sample sizes only as we suspected the potential decrease in variability would be most pronounced for smaller sample sizes. However, our results show quite similar magnitudes of relative efficiency for our main analysis time point over our range of samples sizes from 50 to 1000.

Another variable factor that can influence the stability of the pseudo-observations is the proportion of observed events vs censored individuals. In our simulations, we did not vary the intensities of either events or censoring but only studied a rather common event which was observed in 15% and 30% of the exposed and nonexposed individuals, respectively. Very rare events or situations with more dominating censoring mechanisms will give rise to greater instability of both the nonparametric and parametric pseudo-observations but we have no reason to suspect that the relative efficiency should be influenced by this.

A crucial decision when using a spline-based method is the choice of both the number and positioning of knots. Royston and Lambert[16] recommend using between one and five internal knots with positions which are based on centiles of the observed event times. In a simulation study that was designed specifically to investigate the impact of knot selection, Rutherford et al[23] similarly concluded that even for complex hazard functions choosing more than two internal knots adds very little to the accuracy with which a flexible parametric model fits the true hazard function. In our simulations, both the bias and relative efficiency were quite stable when using three or more knots.

In any specific setting, the choice of knots should of course be made in consideration of the given sample size and the consequence of alternative choices should always be assessed by sensitivity analyses. Particularly, it would be advisable to use a larger number of knots when the shape of the hazard function is expected to be complex. On the other hand, specifying too many knots increases the risk of overfitting.

Our proposed method is not implemented directly in any software packages but, using the Stata package `stpm2` or corresponding software, the implementation is not very difficult. We have provided an example of Stata syntax in the Supporting Information that shows how to calculate parametric pseudo-observations and estimate CIP, RD, and RR at a specific time point.

The need to fit a spline to the log cumulative hazard function for each leave-one-out sample leads to a computationally rather intensive procedure. However, once the pseudo-observations have been calculated for a given dataset, the parameter estimation in different GLM regression models can be performed without need for repeating the time-consuming task of calculating the parametric pseudo-observations. In our experience, calculating parametric pseudo-observations in a dataset of reasonable size with a competing risk is usually feasible within a matter of minutes. In the data example provided, the calculations took about 4 minutes on an ordinary laptop of current standards.

One of the most appealing advantages of the pseudo-observation method is the ability to estimate different effect measures. Depending on the context, the most relevant effect measure might be the RD or the RR. Furthermore, pseudo-observations can be formulated to model the restricted mean survival defined as $\mu_t = \mathrm{E}(\min(T, t))$ by defining the i'th pseudo-observation as the integral of the contribution to the estimated survival function, $\int_0^t \hat{S}(u)du$. On this scale, we can perform GLM regression using the identity link function to estimate difference in restricted mean survival which is a clinically very intuitive and meaningful effect measure.[24] We have not investigated the efficiency of the parametric pseudo-observations defined on this scale but since they are calculated as an average of the pseudo-observations on the cumulative incidence or survival scale we expect our results to be transferable to this setting as well.

Similarly to the observation that nonparametric pseudo-observations are unstable when the risk set is diminishing near the end of follow-up, Mortensen et al pointed out that traditional pseudo-observations perform poorly if the first events occur when the risk set is small due to delayed entry.[15] These instabilities are not shared by the parametric pseudo-observations since the estimation of the log cumulative hazard function is performed considering the entire follow-up data simultaneously.

ACKNOWLEDGEMENTS

The authors are thankful to the two reviewers of the original submission who provided valuable comments resulting in substantial improvements to the article.

DATA AVAILABILITY STATEMENT

The simulated data that support the findings of the simulation study are available on request from the corresponding author. The pneumonia data that support the findings of the example are available at https://www.stata-press.com/data/r16/st.html.

ORCID

Martin Nygård Johansen https://orcid.org/0000-0001-9790-0985
Søren Lundbye-Christensen https://orcid.org/0000-0002-9420-2783
Erik Thorlund Parner https://orcid.org/0000-0003-3661-1922

REFERENCES

1. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B (Methodol)*. 1972;34(2):187-220.
2. Breslow NE. Contribution to the discussion of paper by D.R. Cox. *J Royal Stat Soc Ser B (Methodol)*. 1972;34(2):216-217.
3. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978;34(4):541-554.
4. Hjort NL. On inference in parametric survival data models. *Int Stat Rev/ Revue Internationale de Statistique.* 1992;60(3):355. https://doi.org/10.2307/1403683.
5. Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*. 2003;90(1):15-27. https://doi.org/10.1093/biomet/90.1.15.
6. Andersen PK, Syriopoulou E, Parner ET. Causal inference in survival analysis using pseudo-observations. *Stat Med*. 2017;36(17):2669-2681. https://doi.org/10.1002/sim.7297.
7. Andersen PK. Decomposition of number of life years lost according to causes of death. *Stat Med*. 2013;32(30):5278-5285. https://doi.org/10.1002/sim.5903.
8. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov Chains based on censored observations. *Scandinavian J Stat*. 1978;5:141-150. https://doi.org/10.2307/4615704.

9. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481. https://doi.org/10.2307/2281868.
10. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal*. 2009;15(2):241-255. https://doi.org/10.1007/s10985-008-9107-z.
11. Jacobsen M, Martinussen T. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian J Stat*. 2016;43(3):845-862. https://doi.org/10.1111/sjos.12212.
12. Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *Ann Stat*. 2017;45(5):1988-2015. https://doi.org/10.1214/16-AOS1516.
13. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res*. 2010;19(1):71-99. https://doi.org/10.1177/0962280209105020.
14. Binder N, Gerds TA, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Anal*. 2014;20(2):303-315. https://doi.org/10.1007/s10985-013-9247-7.
15. Mortensen LM, Hansen CP, Overvad K, Lundbye-Christensen S, Parner ET. The pseudo-observation analysis of time-to-event data. example from the Danish diet, cancer and health cohort illustrating assumptions, model validation and interpretation of results. *Epidemiol Methods*. 2018;7(1). https://doi.org/10.1515/em-2017-0015.
16. Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press; 2011.
17. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stat J*. 2009;9(2):265-290.
18. Crowther MJ, Lambert PC. Simulating complex survival data. *Stata J*. 2012;12(4):674-687.
19. Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: an update. *Stata J*. 2015;15(3):809-821.
20. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279-4292. https://doi.org/10.1002/sim.2673.
21. Morris TP, White IR, Crowther MJ. *Using Simulation Studies to Evaluate Statistical Methods*. Statistics in Medicine38. 11. 2019;2074-2102. https://doi.org/10.1002/sim.8086.
22. StataCorp. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press; 2017.
23. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Simulat*. 2015;85(4):777-793. https://doi.org/10.1080/00949655.2013.845890.
24. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(1):152. https://doi.org/10.1186/1471-2288-13-152.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nygård Johansen M, Lundbye-Christensen S, Thorlund Parner E. Regression models using parametric pseudo-observations. *Statistics in Medicine*. 2020;39:2949–2961. https://doi.org/10.1002/sim.8586

APPENDIX A. RESTRICTED CUBIC SPLINES

A restricted cubic spline with internal knots $k_1 < \cdots < k_m$ and boundary knots $k_{\min}$ and $k_{\max}$ is defined as

$$s(x;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_m(x),$$

where the v_j's are basis functions defined as

$$v_j(x) = \left(x - k_j\right)_+^3 - \delta_j \left(x - k_{\min}\right)_+^3 - (1 - \delta_j)\left(x - k_{\max}\right)_+^3, \quad j = 1, \ldots, m,$$

with

$$\delta_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}, \quad j = 1, \ldots, m.$$

Supporting Information

			Non-parametric POs			Parametric POs				
Parameter			**SD**	**Bias**	**Cov. (%)**	**SD**	**Bias**	**Cov. (%)**	**Relative efficiency**	**Relative sample size**
Scenario (1)		***n***								
	CIP	50	0,077	-0,002	95,2	0,072	0,000	95,3	1,07	1,16
		100	0,054	0,001	95,6	0,050	0,001	96,1	1,08	1,17
		250	0,034	0,000	95,9	0,032	0,000	95,8	1,07	1,14
		500	0,024	0,000	96,3	0,022	0,000	96,3	1,08	1,17
		1000	0,017	0,000	95,3	0,016	0,000	95,7	1,08	1,17
	RD	50	0,153	-0,003	94,1	0,140	-0,002	94,2	1,09	1,19
		100	0,109	0,001	94,6	0,100	0,000	94,7	1,08	1,18
		250	0,069	0,001	93,9	0,065	0,002	94,0	1,07	1,15
		500	0,048	0,000	94,9	0,045	0,000	95,3	1,08	1,16
		1000	0,034	0,000	95,2	0,032	0,000	95,1	1,07	1,15
	ln(RR)	50	0,391	-0,005	96,4	0,364	0,000	96,4	1,08	1,16
		100	0,268	0,000	95,1	0,248	-0,002	95,3	1,08	1,17
		250	0,166	0,006	94,7	0,157	0,005	94,5	1,06	1,12
		500	0,114	-0,001	95,0	0,107	-0,001	95,1	1,07	1,14
		1000	0,080	0,000	95,3	0,076	0,001	94,9	1,06	1,13
Scenario (2)		**Knots**								
	CIP	1	0,021	0,000	95,6	0,017	-0,005	95,3	1,19	1,42
		2	0,021	0,000	95,8	0,018	-0,005	95,5	1,15	1,31
		3	0,020	0,000	95,7	0,019	0,007	93,8	1,10	1,21
		4	0,021	0,000	95,4	0,019	0,002	94,8	1,12	1,24
		5	0,021	0,001	95,2	0,019	0,002	95,1	1,13	1,27
		6	0,021	0,000	95,3	0,019	0,001	95,3	1,13	1,27
		7	0,021	0,000	95,3	0,019	0,002	95,5	1,13	1,27
		8	0,021	0,000	95,5	0,019	0,001	95,4	1,11	1,23
		9	0,021	0,000	95,1	0,019	0,001	95,7	1,12	1,26
	RD	1	0,042	0,000	94,2	0,035	-0,006	93,9	1,19	1,41
		2	0,041	-0,001	95,0	0,036	-0,005	94,2	1,14	1,30
		3	0,041	0,000	95,0	0,037	0,002	95,1	1,10	1,22
		4	0,042	0,001	94,9	0,038	0,000	94,9	1,11	1,23
		5	0,041	0,000	95,1	0,036	-0,001	95,0	1,14	1,29
		6	0,041	-0,001	94,8	0,037	-0,001	94,8	1,14	1,29
		7	0,042	-0,001	94,5	0,037	-0,002	95,2	1,14	1,29
		8	0,042	-0,001	95,2	0,037	-0,001	94,9	1,13	1,28
		9	0,041	0,001	95,3	0,037	0,000	95,3	1,12	1,25
	ln(RR)	1	0,057	0,000	94,4	0,049	-0,006	94,0	1,18	1,38
		2	0,056	-0,002	95,1	0,050	-0,005	94,6	1,13	1,28
		3	0,056	0,001	95,4	0,051	0,001	95,2	1,11	1,22
		4	0,058	0,001	94,8	0,052	-0,001	94,9	1,11	1,23
		5	0,057	0,000	94,9	0,050	-0,003	95,3	1,13	1,28
		6	0,057	-0,001	94,6	0,050	-0,002	95,0	1,13	1,28
		7	0,058	-0,001	94,5	0,051	-0,004	95,3	1,13	1,29
		8	0,057	0,000	95,3	0,051	-0,002	95,1	1,13	1,27
		9	0,056	0,001	95,6	0,051	-0,001	95,4	1,11	1,23
Scenario (3)		**HR**								
	CIP	0,5	0,024	-0,001	95,2	0,023	0,000	95,4	1,05	1,10

		0,75	0,023	0,000	95,8	0,022	0,000	95,3	1,04	1,09
		1	0,023	0,000	95,2	0,023	0,001	94,8	1,03	1,06
		1,5	0,022	0,000	94,4	0,022	0,000	94,5	1,02	1,05
		2	0,021	-0,001	95,0	0,021	0,000	95,2	1,02	1,05
	RD	0,5	0,048	0,000	94,8	0,046	0,001	94,8	1,05	1,11
		0,75	0,047	0,001	94,6	0,045	0,001	94,6	1,04	1,09
		1	0,046	0,002	94,5	0,044	0,003	94,5	1,03	1,07
		1,5	0,045	-0,001	94,2	0,044	0,001	94,1	1,03	1,06
		2	0,042	0,000	94,8	0,042	0,002	95,0	1,02	1,04
	ln(RR)	0,5	0,127	0,001	95,0	0,120	0,002	95,1	1,06	1,11
		0,75	0,131	0,001	95,1	0,125	0,004	94,9	1,05	1,10
		1	0,138	0,003	95,0	0,132	0,006	95,0	1,04	1,08
		1,5	0,152	-0,001	94,3	0,146	0,004	94,6	1,04	1,08
		2	0,159	0,003	95,4	0,154	0,007	95,3	1,03	1,06
Scenario (4)		**Knots**								
	CIP	1	0,023	0,001	95,0	0,021	0,002	94,9	1,09	1,18
		2	0,023	0,000	95,2	0,021	-0,001	95,6	1,10	1,21
		3	0,023	0,001	95,6	0,021	0,000	95,6	1,09	1,20
		4	0,023	-0,001	95,3	0,021	-0,002	95,5	1,09	1,18
		5	0,023	0,000	95,7	0,021	-0,001	96,0	1,10	1,20
		6	0,023	0,000	95,8	0,021	-0,001	96,0	1,09	1,20
		7	0,023	0,000	95,3	0,021	-0,001	95,8	1,08	1,18
		8	0,023	0,000	95,2	0,021	0,000	95,6	1,09	1,18
		9	0,023	0,000	95,4	0,021	-0,001	95,7	1,07	1,15
	RD	1	0,046	0,001	95,0	0,043	0,003	94,7	1,08	1,17
		2	0,045	0,002	95,0	0,042	0,001	95,1	1,09	1,18
		3	0,046	0,000	94,7	0,042	0,001	95,0	1,09	1,18
		4	0,046	0,000	95,0	0,042	-0,001	95,2	1,09	1,19
		5	0,046	0,000	95,1	0,042	0,000	95,2	1,09	1,19
		6	0,046	-0,001	95,1	0,042	0,000	95,2	1,10	1,20
		7	0,045	0,000	95,4	0,042	0,001	95,4	1,08	1,17
		8	0,046	-0,001	95,3	0,042	-0,001	94,8	1,09	1,18
		9	0,046	0,000	95,0	0,043	0,001	94,9	1,08	1,16
	ln(RR)	1	0,072	0,001	94,9	0,066	0,003	95,1	1,08	1,17
		2	0,070	0,002	95,1	0,065	0,003	95,5	1,08	1,17
		3	0,071	0,000	95,1	0,065	0,001	94,9	1,08	1,16
		4	0,071	0,000	95,0	0,066	0,000	95,5	1,08	1,18
		5	0,071	-0,001	95,3	0,065	0,000	95,5	1,09	1,18
		6	0,071	-0,001	95,3	0,065	0,001	95,2	1,09	1,19
		7	0,070	0,001	95,6	0,065	0,001	95,1	1,08	1,16
		8	0,071	-0,003	95,3	0,066	-0,001	95,0	1,08	1,18
		9	0,071	-0,001	95,0	0,066	0,001	95,0	1,07	1,15
Scenario (5)		**Years**								
	CIP	0	0,022	0,000	96,1	0,022	0,000	96,1	1,00	1,00
		1	0,022	0,000	96,2	0,021	-0,001	96,2	1,05	1,10
		2	0,022	0,000	96,4	0,020	-0,002	96,9	1,11	1,22
		3	0,022	0,000	96,2	0,020	-0,002	96,2	1,11	1,24
		4	0,023	0,000	95,7	0,020	-0,002	96,1	1,13	1,29
		5	0,022	0,001	95,8	0,019	-0,003	96,2	1,15	1,33
		6	0,022	-0,001	95,8	0,020	-0,004	95,4	1,14	1,29

		7	0,022	-0,001	95,9	0,019	-0,004	96,0	1,16	1,35
		8	0,022	0,000	96,1	0,019	-0,003	96,6	1,18	1,39
		9	0,022	0,000	96,3	0,019	-0,004	96,1	1,17	1,37
		10	0,022	0,000	96,1	0,019	-0,004	96,0	1,16	1,35
	RD	0	0,045	0,000	94,8	0,045	0,001	94,7	1,00	1,00
		1	0,044	0,002	95,0	0,042	0,002	95,2	1,06	1,11
		2	0,044	-0,001	95,2	0,041	-0,001	95,5	1,09	1,19
		3	0,044	0,000	95,2	0,039	-0,002	95,1	1,13	1,27
		4	0,045	-0,001	95,2	0,039	-0,004	95,5	1,16	1,34
		5	0,045	0,001	94,6	0,038	-0,001	94,9	1,17	1,37
		6	0,044	0,000	95,0	0,037	-0,004	95,1	1,18	1,40
		7	0,044	0,000	95,3	0,037	-0,004	96,3	1,20	1,44
		8	0,044	0,001	95,1	0,037	-0,003	95,9	1,20	1,44
		9	0,044	0,000	95,1	0,037	-0,005	95,1	1,20	1,45
		10	0,045	0,000	94,6	0,037	-0,004	95,1	1,20	1,44
	ln(RR)	0	0,107	0,000	94,8	0,107	0,001	94,7	1,00	1,00
		1	0,105	0,005	95,4	0,100	0,004	95,5	1,05	1,10
		2	0,105	-0,002	95,6	0,098	-0,001	95,5	1,08	1,16
		3	0,105	0,002	95,6	0,095	-0,001	95,5	1,11	1,23
		4	0,108	-0,003	95,0	0,094	-0,007	95,8	1,14	1,30
		5	0,107	0,002	95,0	0,094	0,000	95,1	1,14	1,31
		6	0,106	-0,001	95,2	0,091	-0,002	95,3	1,16	1,34
		7	0,107	0,001	95,3	0,090	-0,004	95,6	1,18	1,39
		8	0,105	0,002	95,5	0,090	-0,002	95,7	1,17	1,37
		9	0,106	0,001	95,4	0,090	-0,006	95,7	1,19	1,41
		10	0,108	0,001	95,0	0,092	-0,003	94,9	1,17	1,38
Scenario (6)		**Years**								
	CIP	0	0,056	-0,009	92,2	0,037	0,004	95,2	1,51	2,27
		1	0,040	-0,001	95,4	0,034	0,003	95,6	1,17	1,37
		2	0,035	0,000	95,9	0,032	0,002	96,0	1,10	1,22
		3	0,033	0,000	95,2	0,031	0,002	95,0	1,06	1,13
		4	0,032	0,000	95,4	0,030	0,001	95,9	1,06	1,13
		5	0,030	-0,001	95,8	0,029	0,001	95,8	1,04	1,07
		6	0,029	0,001	95,7	0,028	0,002	95,9	1,03	1,07
		7	0,028	0,000	96,1	0,028	0,002	96,2	1,02	1,05
		8	0,029	-0,001	95,2	0,028	0,000	95,2	1,03	1,05
		9	0,028	0,000	95,7	0,027	0,001	95,9	1,02	1,05
		10	0,027	0,000	95,8	0,027	0,001	95,6	1,02	1,04
	RD	0	0,111	-0,002	95,8	0,072	0,002	95,2	1,54	2,36
		1	0,080	-0,003	95,6	0,066	0,001	95,8	1,21	1,46
		2	0,070	-0,001	95,5	0,063	0,001	95,9	1,11	1,23
		3	0,065	0,000	95,4	0,061	0,003	95,4	1,07	1,15
		4	0,061	0,001	95,4	0,059	0,001	95,5	1,05	1,10
		5	0,059	0,002	95,5	0,056	0,003	95,9	1,04	1,08
		6	0,059	0,001	95,1	0,057	0,002	94,9	1,04	1,07
		7	0,057	0,000	95,0	0,055	0,002	95,3	1,03	1,06
		8	0,055	0,001	95,3	0,054	0,003	95,3	1,02	1,04
		9	0,055	0,002	95,4	0,054	0,002	95,1	1,02	1,03
		10	0,054	0,001	95,1	0,053	0,002	94,9	1,02	1,04
	ln(RR)	0	0,248	0,017	95,2	0,169	0,002	95,1	1,47	2,16

1	0,187	-0,002	95,6	0,156	0,001	95,7	1,20	1,43
2	0,166	0,000	95,7	0,150	0,001	95,5	1,11	1,23
3	0,155	0,001	95,3	0,145	0,004	95,3	1,07	1,15
4	0,145	0,003	95,7	0,138	0,001	95,7	1,05	1,10
5	0,140	0,006	95,8	0,134	0,005	95,4	1,04	1,08
6	0,140	0,001	95,2	0,135	0,002	95,2	1,04	1,07
7	0,135	0,003	95,5	0,131	0,002	95,5	1,03	1,06
8	0,132	0,005	95,5	0,129	0,004	95,4	1,02	1,04
9	0,131	0,006	95,1	0,128	0,005	95,3	1,02	1,04
10	0,128	0,002	95,3	0,126	0,004	95,2	1,02	1,04

			Non-parametric POs			Parametric POs				
Parameter	Dependent censoring	Strat	SD	Bias	Cov. (%)	SD	Bias	Cov. (%)	Relative efficiency	Relative sample size
CIP	0	0	0,024	0,000	95,5	0,022	0,000	95,6	1,09	1,18
		1	0,024	-0,001	96,3	0,022	0,000	96,4	1,09	1,18
	1	0	0,024	-0,008	95,2	0,023	-0,007	95,2	1,06	1,13
		1	0,024	0,000	96,4	0,022	0,001	96,8	1,09	1,19
RD	0	0	0,048	0,000	95,1	0,044	0,000	95,0	1,07	1,15
		1	0,048	0,001	94,6	0,045	0,001	94,8	1,08	1,17
	1	0	0,049	-0,002	95,3	0,045	-0,002	94,9	1,07	1,15
		1	0,049	0,000	94,9	0,045	0,001	94,6	1,09	1,18
ln(RR)	0	0	0,114	0,001	95,2	0,107	0,001	95,2	1,07	1,14
		1	0,115	0,001	95,1	0,107	0,002	95,1	1,07	1,15
	1	0	0,118	0,007	95,5	0,110	0,009	95,3	1,07	1,15
		1	0,115	0,000	95,0	0,107	0,001	95,0	1,08	1,16

Paper B

Regression models for interval censored data using parametric pseudo-observations

Martin N. Johansen[1], Søren Lundbye-Christensen[1,2], Jacob M. Larsen[4,2], and Erik T. Parner[3]

1. Unit of Clinical Biostatistics, Aalborg University Hospital, Aalborg, Denmark; 2. Department of Clinical Medicine, Aalborg University, Aalborg, Denmark; 3. Section for Biostatistics, Department of Public Health, Aarhus University, Aarhus, Denmark; 4. Department of Cardiology, Aalborg University Hospital, Aalborg, Denmark.

Submitted to *BMC Medical Research Methodology* (2020, under review).

Description

In this paper, we extend the flexible parametric pseudo-observation method to a setting with interval censored data. To accomodate the presence of competing risks, we employ an irreversible illness-death model for the event of interest. We evaluate the empirical properties of the method in a simulation study and apply the method to a dataset of patients with an implantable cardioverter-defibrillator who are followed by routine examinations considering the occurrence of a specific lead failure as the event of interest.

RESEARCH

Regression models for interval censored data using parametric pseudo-observations

Martin Nygård Johansen[1*], Søren Lundbye-Christensen[1,2], Jacob Moesgaard Larsen[4,2] and Erik Thorlund Parner[3]

*Correspondence:
martin.johansen@rn.dk
[1]Unit of Clinical Biostatistics, Aalborg University Hospital, Sdr Skovvej 15, 9000 Aalborg, DK
Full list of author information is available at the end of the article

Abstract

Background: Time-to-event data that is subject to interval censoring is common in the practice of medical research and versatile statistical methods for estimating associations in such settings have been limited. For right censored data, non-parametric pseudo-observations have been proposed as a basis for regression modeling with the possibility to use different association measures. In this article, we propose a method for calculating pseudo-observations for interval censored data.

Methods: We develop an extension of a recently developed set of parametric pseudo-observations based on a spline-based flexible parametric estimator. The inherent competing risk issue with an interval censored event of interest necessitates the use of an illness-death model, and we formulate our method within this framework. To evaluate the empirical properties of the proposed method, we perform a simulation study and calculate pseudo-observations based on our method as well as alternative approaches. We also present an analysis of a real dataset on patients with implantable cardioverter-defibrillators who are monitored for the occurrence of a particular type of device failures by routine follow-up examinations. In this dataset, we have information on exact event times as well as the interval censored data, so we can compare analyses of pseudo-observations based on the interval censored data to those obtained using the non-parametric pseudo-observations for right censored data.

Results: Our simulations show that the proposed method for calculating pseudo-observations provides unbiased estimates of the cumulative incidence function as well as associations with exposure variables with appropriate coverage probabilities. The analysis of the real dataset also suggests that our method provides estimates which are in agreement with estimates obtained from the right censored data.

Conclusions: The proposed method for calculating pseudo-observations based on the flexible parametric approach provides a versatile solution to the specific challenges that arise with interval censored data. This solution allows regression modeling using a range of different association measures.

Keywords: pseudo-observations; interval censoring; flexible parametric model

1 Background

In medical research, the outcome is often an event such as death, occurrence of a disease, or a treatment-related event during a follow-up period. Some individuals will be event-free throughout follow-up, but the event may occur after the end of follow-up. This kind of incomplete follow-up is called *right censoring* and methods

for dealing with this form of censoring are used very frequently in the medical literature. Right censored data thus consist of a mixture of exactly observed event times and censoring times. In other situations, the exact event times are never observed and the event status is only evaluated at certain time points, *examination times*, and the data are then said to be *interval censored*. This phenomenon occurs frequently when for example a specific group of individuals is monitored by routine examinations for a medical condition. In such cases, event times are known only to lie within a time interval from the last examination without the event to the first examination after the event has occurred. In practice, data can also consist of a mixture of right and interval censored data, e.g. when data are gathered from different sources. A standard assumption when analyzing interval censored data is that the examination times are independent of the event risk. In that case one can in the analysis ignore the distribution of the examination times, and treat the examination times as fixed. We will also assume that the examination times are independent of the event risk.

Interval censoring has posed a challenge to the medical research community that has proven hard to overcome. Regression models for interval censored data has traditionally mostly been concerned with basic parametric regression models where inference can be performed by standard maximum likelihood methods and in which the estimators converge at a rate of $\sqrt{n}$. Parametric models are easily fitted using most common statistical software but each distributional family imposes rather strict assumptions on the shape of the hazard function and it is our impression that their use in applications has diminished in recent years; most likely due to reluctance to impose such assumptions, although covariate adjustment is straightforward in parametric models. A parametric approach that can accomodate different distributional characteristics is the piece-wise exponential proportional hazards model or equivalently a Poisson log-linear model where the hazard is assumed constant in some set of intervals of the follow-up time[1]. When events are plentiful the follow-up intervals can be made small enough to give a reasonable fit to practically any shape of the hazard function but when the data is more sparse with few events or the hazard has a more complex shape during follow-up the piece-wise exponential model has obvious limitations[2].

As an example of an interval censored dataset, we consider a group of patients with an implantable cardioverter-defibrillator (ICD), which is a kind of pacemaker that can protect against slow heart rhythm but also fast arrhythmias, which otherwise can result in hemodynamic compromise with loss of consciousness and cardiac arrest. The fast arrhythmias can be treated by fast pacing or delivery of a high voltage shock that restores the heart rhythm to normal. The ICD is placed in the subcutaneous tissue on the front of the chest below the left collarbone and is connected to the inside of the heart through a large blood vessel. The ICD lead gives the ICD the ability to continuously monitor the heart rhythm and if needed deliver the high voltage shock inside the heart. The ICD lead is the most sensitive part of an ICD system and is the part with the highest risk of failure either due to insulation failures or conductor fractures. The particular lead investigated is prone to a rather unique type of insulation failure because of a design flaw where the inner conductors over time work their way through the outer insulation. Such outer insulation

failures, called *externalizations*, may be electrically silent at normal ICD follow-up and require dedicated fluoroscopic/X-ray imaging to be detected. The ICD is at risk of failing from such externalization events throughout follow-up, but patients can also have their ICD leads removed (extracted) for other reasons during follow-up, which obviously precludes an externalization event. We consider externalization as the event of interest and we are interested in estimating the association between the amount of slack in the lead body inside the heart and the time to externalization, since more lead slack puts the continuously moving lead body under more physical stress. In this setting, we have a combined competing risk of death or extraction of the ICD leads. To assess the association between lead slack and externalization, we are interested in comparing the cumulative risk of externalization at one or more time points.

In this application, interest lies in assessing the effect of the exposure on the cumulative risk of developing the outcome in the presence of the competing risks but existing methods are not well-equipped for this type of situation. However, in the right censored competing risk setting, *pseudo-observations* have been proposed[3] as a modeling approach which enables effect estimation on a number of different scales other than the hazard scale such as the cumulative incidence scale. This method is based on a transformation of the potentially censored time-to-event data into a set of complete data on which regression can be performed using generalized linear models to estimate the relevant effect parameters. When the aim is to model some function of the cumulative incidence, the transformation is based on the non-parametric Aalen-Johansen estimator of the cumulative incidence function.

A non-parametric estimator of the survival function based on interval censored data has been proposed by both Peto and Turnbull[4, 5]. The resulting Peto-Turnbull estimator is a piece-wise constant curve with relatively few jumps. A natural way to apply the pseudo-observation approach to interval censored data therefore seems to be to perform a transformation of the data based on the Peto-Turnbull estimator similarly to the pseudo-observation approach based on the Aalen-Johansen estimator. This approach has been investigated by Kim and Kim[6] in a competing risk setting. However, the asymptotic properties of the resulting pseudo-observations are unclear since the theory for pseudo-observations has been developed only for estimators with parametric $\sqrt{n}$ convergence rate[7], whereas the Peto-Turnbull estimator has slower $n^{1/3}$ convergence rate[8].

Royston and Parmar[9] have proposed a *flexible parametric model* which is applicable to both right censored and interval censored data. This is a regression modeling framework where the log cumulative hazard function is estimated using a restricted cubic spline in log time. In the most simple form with no covariates this approach provides a way to model the cumulative incidence function and when covariates are included the model can be formulated as either a proportional hazards or a proportional odds model.

As in our example above, the event of interest in interval censored data is often a non-fatal event, so methods for handling interval censoring should accomodate death as a competing risk. For the remainder of this article, we consider only competing events for which the event time is exactly observed and refer to competing events as death for ease of terminology. In a competing risk setting with a right

censored event of interest, we can model the cause-specific hazard functions separately by considering only the time to whatever event occurs first. But when the event of interest is interval censored, we are only observing the event if there is an examination after the event has occurred but before the individual is censored or dies. Hence, there might be some events of interest which are unobserved in the data. Because of this circumstance, the inference needs to take into account that the event of interest might or might not have occurred in the interval between the last examination time without the event of interest and time of death or censoring. To accomodate this, the data could be considered in an illness-death model[10] where the risk of death is also modeled after an event of interest has occurred.

Recently, an elegant approach to calculating pseudo-observations for interval censored data was proposed by Sabathé *et al.*[11] specifically for an illness-death model. This approach is based on modeling the three transition intensities using M-splines and applying a penalized likelihood approach where more roughly shaped intensity functions are penalized using the second derivatives of the three M-splines squared. This requires a high number of coefficients for each of the three splines depending on the order and the number of knots of the spline as well as three penalty parameters to be chosen by the analyst. Due to this high number of parameters, the authors do not recommend using their method in place of the traditional non-parametric pseudo-observation approach for right censored data.

For right censored competing risk data, we have recently shown that in some situations calculating *parametric pseudo-observations* based on a marginal flexible parametric estimator of the cumulative incidence function can provide less variability in the effect estimates than that of traditional non-parametric pseudo-observations[12]. In this article, we propose an extension of this approach that applies to the interval censored setting and is targeted directly at estimating associations between an exposure and the event of interest. In Section 2.1, we describe the proposed method in more detail and in Section 2.2 we describe a simulation study that compares our proposed method to the existing methods. We present the results of these simulations in Section 3.1 and present an analysis of the example data in Section 3.2. We conclude the article with a discussion and conclusion in Sections 4 and 5.

2 Methods

2.1 Proposed method

We now give details on how the parametric pseudo-observation approach can be extended to cover interval censored settings with competing risks using an illness-death model.

An illness-death model involves an event of interest and the competing event death which gives three different states; 0 where neither event has occurred, 1 where only the event of interest has occurred, and 2 which is death with or without having experienced the event of interest. In the following, we will assume that all individuals are initially in state 0 at time $t = 0$ and we let h_{kl} denote the hazard function describing transition from one state, k, to another, l and similarly we let H_{kl} denote the cumulative hazard function. To estimate the cumulative incidence function of the event of interest, $F_{01}(\cdot)$, we will use the estimates of the transition-specific hazard

functions and the relationship between these and the transition-specific cumulative incidence function,

$$F_{01}(t) = \int_0^t h_{01}(u)S(u)du, \tag{1}$$

where $S(\cdot)$ is the event-free survival function defined as

$$S(t) = \exp\Big(-H_{01}(t) - H_{02}(t)\Big).$$

We estimate the transition-specific hazard functions by modeling the transition-specific log cumulative hazard functions using restricted cubic splines in $x = \ln(t)$. According to Royston and Parmar[9], a natural cubic spline with m internal knots, $\xi_1, \ldots, \xi_m$, and external knots ξ_{min}, ξ_{max} can be expressed as

$$s(x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_m(x),$$

where $v_j(x) = (x - \xi_j)_+^3 - \lambda_j(x - \xi_{min})_+^3 - (1 - \lambda_j)(x - \xi_{max})_+^3$.
Hence, we are assuming the model

$$\begin{aligned}\ln(H_{kl}(t)) &= s_{kl}(x; \boldsymbol{\gamma}_{kl}) \\ &= \gamma_{kl,0} + \gamma_{kl,1} x + \gamma_{kl,2} v_{kl,1}(x) + \cdots + \gamma_{kl,m+1} v_{kl,m}(x),\end{aligned}$$

for going from state k to state l. For simplicity, we assume that the number of knots is m for all three splines. The model, hence, contains $m + 2$ spline coefficients, $\boldsymbol{\gamma}_{kl} = \gamma_{kl,0}, \ldots, \gamma_{kl,m+1}$, for each transistion and corresponding spline knots $\xi_{kl,min}, \xi_{kl,1}, \ldots, \xi_{kl,m}, \xi_{kl,max}$. Based on the spline coefficients, $\boldsymbol{\gamma}_{01}$, $\boldsymbol{\gamma}_{02}$, and $\boldsymbol{\gamma}_{12}$, we can express the transition-specific hazard function as

$$\begin{aligned}h_{kl}(t) &= \frac{ds_{kl}(x; \boldsymbol{\gamma}_{kl})}{dt} \cdot \exp(s_{kl}(x; \boldsymbol{\gamma}_{kl})) \\ &= \frac{1}{t} \cdot \frac{ds_{kl}(x; \boldsymbol{\gamma}_{kl})}{dx} \cdot \exp(s_{kl}(x; \boldsymbol{\gamma}_{kl})).\end{aligned}$$

The derivative of $s_{kl}(x; \boldsymbol{\gamma}_{kl})$ is

$$\begin{aligned}\frac{ds_{kl}(x; \boldsymbol{\gamma}_{kl})}{dx} = \gamma_{kl,1} + \sum_{j=2}^{m} \Bigg\{ \gamma_{kl,j} \cdot \Big(& 3(x - \xi_{kl,j})_+^2 \\ & - 3\lambda_{kl,j}(x - \xi_{kl,min})_+^2 - 3(x - \xi_{kl,max})_+^2 \Big) \Bigg\}.\end{aligned}$$

We consider a setting where the time to the event of interest can either be observed exactly (right censored) or interval censored but the time of death is always observed exactly (right censored). Estimation of the spline coefficients is performed using maximum likelihood methods and the contributions to the likelihood function, $L(\boldsymbol{\gamma}_{01}, \boldsymbol{\gamma}_{12}, \boldsymbol{\gamma}_{02})$, take different forms according to the event trajectory of each individual. These trajectories are determined by the occurrence and timing of the event of interest and death as described by Touraine *et al.*[13]

2.1.1 Maximum likelihood estimation

The observed trajectory of an individual can be described by the observed event status and observation time for both the event of interest, (d_1, t_1), and death, (d_2, t_2), as well as a time of the last examination time without the event of interest if any such has occurred, l_1. This last negative examination time might be at time $l_1 = 0$ if no negative examinations have occurred. For individuals with an interval censored event of interest, the event of interest is then known to occur in the interval (l_1, t_1). For individuals with an event of interest for which the time is observed exactly, l_1 is not defined and for individuals with right censored data but no event of interest, we let l_1 denote the time point at which follow-up ends for that individual. We now describe the contributions to the likelihood function for each trajectory. For the i'th individual, we use the following notation.

- d_{1i} indicates an observed event of interest (either exactly observed or interval censored)
- l_{1i} is the last known negative time point (potentially at time zero)
- t_{1i} is the observation time for the event of interest (either the exact time or the first positive examination time)
- d_{2i} indicates a competing event (exactly observed)
- t_{2i} is the observation time for the competing event

For short, we will denote each individual's contribution to the likelihood function as L_i.

Trajectory 1

If an individual has an exactly observed event of interest at time t_{1i} and is then right censored at time t_{2i}, the corresponding contribution to the likelihood function is

$$L_i = S(t_{1i})h_{01}(t_{1i})\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(t_{1i}))}.$$

Trajectory 2

If an individual has a negative examination at time l_{1i} and is then right censored at time t_{2i}, the contribution is

$$L_i = S(t_{2i}) + \int_{l_{1i}}^{t_{2i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}du.$$

This likelihood contribution also applies to individuals with right censoring of the event of interest, since this corresponds to the special case where $l_{1i} = t_{2i}$ and the integral is thus zero.

Trajectory 3

If an individual has an interval censored event of interest occurring between time l_{1i} and t_{1i} and is then censored at time t_{2i}, the contribution is

$$L_i = \int_{l_{1i}}^{t_{1i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}du.$$

Trajectory 4

If an individual has an exactly observed event of interest at time t_{1i} and then dies

at time t_{2i}, the contribution is

$$L_i = S(t_{1i})h_{01}(t_{1i})\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(t_{1i}))}h_{12}(t_{2i}).$$

Trajectory 5
If an individual has a negative examination at time l_{1i} and then dies at time t_{2i}, the contribution is

$$L_i = S(t_{2i})h_{02}(t_{2i}) + \int_{l_{1i}}^{t_{2i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})du.$$

Again, this applies to individuals with right censoring of the event of interest.
Trajectory 6
If an individual has an interval censored event of interest occurring between time l_{1i} and t_{1i} and then dies at time t_{2i}, the contribution is

$$L_i = \int_{l_{1i}}^{t_{1i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})du.$$

If we furthermore use the indicator, d_{2i}, for the competing event (exactly observed), we can write all likelihood contributions as one of the following three expressions.
Trajectories 1 and 4
For an individual with the event of interest observed at time t_{1i} exactly, followed by death or censoring at time t_{2i}, the contribution is

$$L_i = S(t_{1i})h_{01}(t_{1i})\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(t_{1i}))}h_{12}(t_{2i})^{d_{2i}}.$$

Trajectories 2 and 5
For an individual with an examination without the event of interest or right censoring of the event of interest at time l_{1i} followed by death or censoring at time t_{2i}, the contribution is

$$L_i = S(t_{2i})h_{02}(t_{2i})^{d_{2i}} + \int_{l_{1i}}^{t_{2i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})^{d_{2i}}du.$$

Trajectories 3 and 6
For an individual with an interval censored event of interest occurring between time l_{1i} and t_{1i} followed by a death or censoring at time t_{2i}, the contribution is

$$L_i = \int_{l_{1i}}^{t_{1i}} S(u)h_{01}(u)\frac{\exp(-H_{12}(t_{2i}))}{\exp(-H_{12}(u))}h_{12}(t_{2i})^{d_{2i}}du.$$

The likelihood function obtained by multiplying the relevant contributions for each individual can be maximized numerically by using e.g. the Newton-Raphson algorithm.

2.1.2 Initial values

For likelihood maximization in practice, it is worth considering how to provide initial values for the parameter vector $(\boldsymbol{\gamma}_{01}, \boldsymbol{\gamma}_{02}, \boldsymbol{\gamma}_{12})$ in order to achieve convergence in as few iterations as possible. We propose the following approach using midpoints for interval censored events of interest.

Modeling the transition from state 0 to 1 can be done by fitting a flexible parametric model with the spline knots chosen for this transition and using the midpoints between l_{1i} and t_{1i} for interval censored events of interest. From this fitted model we can calculate a predicted survival function to estimate 1 minus the cumulative incidence of the event of interest. For each individual that has not had an observed event of interst, we can then estimate the probability that they had an unobserved event of interest in the interval between their last negative examination time, l_{1i}, and their end of follow-up time, t_{2i}, as the difference in predicted survival between these two time points. We can then randomly assign these individuals as having had or not having had an unobserved event of interst based on their individual probabilities and then temporarily consider some of them as if they had an event of interest at the midpoint of the interval from l_{1i} to t_{2i}. This allows us to more accurately estimate the remaining two transitions.

The transitions from state 0 to 2 and from 1 to 2 can now be modeled, again using flexible parametric models with the relevant knots, using the updated event and status variables and imposing delayed entry at the time of the event of interest for the transition from state 1 to 2.

2.1.3 Parametric pseudo-observations for interval censored data

Once we have obtained estimates, $\hat{\boldsymbol{\gamma}}_{01}$, $\hat{\boldsymbol{\gamma}}_{02}$, and $\hat{\boldsymbol{\gamma}}_{12}$, of the parameters in the likelihood function described above, we can define a set of parametric pseudo-observations for interval censored data, $\theta_1^{IC}, \ldots, \theta_n^{IC}$, as

$$\theta_i^{IC} = n\hat{\theta}^{IC} - (n-1)\hat{\theta}_{(-i)}^{IC}, \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where $\hat{\theta}^{IC}$ denotes the estimate of the cumulative incidence function and $\hat{\theta}_{(-i)}^{IC}$ is the corresponding leave-one-out estimate based on all observations except the i'th with the same spline knots as for the full-sample estimate.

The pseudo-observations thus defined can be analyzed using generalized linear models with a sandwich estimator of the variance in the same way as both non-parametric and parametric pseudo-observations for right censored data[3, 12].

2.2 Simulation studies

2.2.1 Data generation

We simulated datasets imposing a non-random binary exposure, x, such that half of the individuals are exposed and the other half is non-exposed and an administrative censoring at time $t = 5$.

For the event of interest, we simulated realizations of a random variable $T_{01} \sim \text{Exp}(\lambda_{01}(x))$, where the intensities are $\lambda_{01}(0) = 0.3$ and $\lambda_{01}(1) = 0.2$. Similarly, we simulated death from a random variable $T_{02} \sim \text{Exp}(\lambda_{02})$ with intensity $\lambda_{02} = 0.1$. Based on these variables we define event indicators δ_{01} and δ_{02} according to which

event occurs first if $\min(T_{01}, T_{02}) < 5$. Hence, all individuals enter the study at time $t = 0$ in state 0.

For individuals who experience the event of interest, we simulate the transition from state 1 to state 2 as another random variable $T_{12} \sim \text{Exp}(\lambda_{12})$ with $\lambda_{12} = 0.4$. The time-to-event for this transition is then $T_{01} + T_{12}$ with censoring at $t = 5$ and the event indicator is δ_{12}.

To mimic a practical setting with a mixture of right and interval censored data, we consider the event of interest for some individuals to be interval censored and for the others to be right censored. This allocation follows a Bernoulli distribution with probability parameter p_{ic} for being interval censored. For individuals with interval censoring of the event of interest, we simulate examination times with a mean interval length of Δ and a random error following a normal distribution with mean zero and variance σ^2. We continue adding examinations until either the event of interest has occurred or the induvidual has died or has been censored following an iterative formula for examination times,

$$e_{i+1} = e_i + \delta_i,$$

where $\delta_i \sim N(\Delta, \sigma^2)$. This gives rise to the variable l_{1i} which is the last known time with a negative status for the event of interest and the variable t_{1i} which is the first known positive status. For individuals with an exactly observed event of interest, we let $l_{1i} = t_{1i}$ be the event time, and for right censored individuals in which we do not observe an event of interest will have $l_i = t_{1i} = t_{2i}$ which is the time of death or censoring.

For the simulations, we performed 1 000 repetitions of datasets of sample size $n = 250$, where $p_{ic} = 80\%$ of the events of interest are interval censored, and the mean time between examinations is $\Delta = 1$ with $\sigma^2 = 0.2$.

2.2.2 Data analysis

In each dataset, we calculated five sets of pseudo-observations for the event of interest based on five different approaches.

$\theta_1^E, \ldots, \theta_n^E$	Potentially unobservable exact right censored event times for all individuals. These will serve as a way to measure the empirically highest achievable precision.
$\theta_1^M, \ldots, \theta_n^M$	Midpoints of the examination intervals for interval censored events, exact right censored event times otherwise.
$\theta_1^R, \ldots, \theta_n^R$	Right endpoint of the examination intervals for interval censored events, exact right censored event times otherwise.
$\theta_1^{IC}, \ldots, \theta_n^{IC}$	Proposed method for taking interval censoring into account.
$\theta_1^S, \ldots, \theta_n^S$	Method for taking interval censoring into account proposed by Sabathé *et al.*

For each set of pseudo-observations we fitted the same generalized linear models to estimate the risk, risk difference, and relative risk of experiencing the event of interest before time $t = 3$. If the estimation of spline coefficients for either the full sample or one or more leave-one-out subsamples did not converge or if the generalized linear regression model gave unreasonable estimates (cumulative incidence not in $(0, 1)$, risk difference not in $(-1, 1)$, relative risk not in $(10^{-1}, 10)$), we considered the results to be unvalid and ignore them in the following. Based on the obtained estimates, we then calculated the median bias, the empirical standard error (empSE) and the confidence interval coverage probability[14]. We also calculated a relative empSE with the empSE of the θ_i^Es as the reference value to assess the amount of additional variation that is added by accounting for the interval censored nature of the data.

We generated data and performed all pseudo-observation calculations except the θ_i^Ss as well as regression modeling using Stata/MP version 16.1. To calculate the θ_i^Ss we used R version 3.6.3 and the packages `SmoothHazard` and `pseudoICD`.

3 Results

3.1 Simulation studies

To illustrate the five different estimation approaches, we have shown the full-sample estimators on which each of the compared approaches are based for a randomly chosen simulated dataset in Figure 1. It is clear that the Aalen-Johansen estimator based on either the midpoints (red curve) or the right endpoints (green curve) underestimate the cumulative incidence as estimated by the Aalen-Johansen estimator on the exact event times (blue curve). Both the penalized likelihood estimator (purple curve) and the flexible parametric estimator (black curve) follow the estimator based on the exact event times reasonably well. The results of the simulation study are shown in Table 1. In the 1 000 datasets, there were on average 146 events of interst but only 120 that we observe when considering the data as interval censored. We focus mainly on the estimates of absolute cumulative incidence of the event of interest. For the estimation of cumulative incidence, 18 of the 1 000 datasets resulted in an unvalid estimate for the interval censored method, 4 did so when we used the right endpoints and none did for the other methods. For both the risk difference and the relative risk, this happened in 13 and 4 of the subsamples for the interval censored and right endpoint methods respectively.

Using the exactly observed data, the parametric pseudo-observations perform very well and we obtain unbiased estimation of the true value of the cumulative incidence function at time $t = 3$, which is 0.460, with an empirical standard error of 0.028 and coverage probability close to the nominal value of 95%. Using the midpoints with right censored methods, we observe a substantial negative bias due to the unobserved events. This bias is exacerbated when we use the right endpoints due to the systematic over-estimation of the observation time. These biases cause both of the methods to yield useless coverage probabilities. Analysing the interval censored data using our proposed parametric pseudo-observations, we still get an unbiased estimator but the empirical error is roughly 50% higher due to the added uncertainty inherent in the interval censored data. The coverage of this method is also reasonably close to 95%. In terms of bias and coverage, the method proposed

by Sabathé *et al.* performs quite similarly to our proposed method while the empirical standard error of the cumulative incidence estimates is somewhat lower for the Sabathé *et al.* method. This might be explained by the additional three penalization parameters which control the smoothness of the fitted M-splines but must be provided explicitly or determined from the data using an approximate likelihood technique[13].
Estimating associations with the exposure gives small biases for both the risk difference and relative risk using either our proposed method and that of Sabathé *et al.* and the coverage probabilities are in good agreement with the nominal value.

3.2 Application to ICD data

Our ICD dataset holds data on 377 patients who are followed from the time of ICD implantation and for a maximum of about 10 years. During follow-up we have information on our event of interest, externalization status, at each fluoroscopic examination time and on the date of death or lead extraction if this occurred. The dataset, hence, consists only of interval censored data for the event of interest and right censored data for death or lead extraction. We show the trajectory for each patient in Figure 2 where lines indicate an observation interval colored black for intervals ending at a positive examination and grey if we do not observe externalization and black dots indicate death or lead extraction times. We observed 37 externalization events and 106 cases of death or lead extraction during follow-up.
We first estimated the cumulative incidence function for the externalization event based on a competing risk model using the non-parametric Aalen-Johansen estimator[15] applied to the midpoints of the intervals. This is illustrated by the solid step function in Figure 3. The dashed and dotted curves in the figure show the estimator based on the flexible parametric approach by fitting splines with 3 and 4 knots, respectively, to the interval censored data in an illness-death model. The three estimators seem to capture roughly the same shape of the cumulative incidence function although the Aalen-Johansen estimator based on midpoints shows a tendency to place the bulk of the events around 2–3 years due to a high number of patients having their first examination since implantation after roughly 5 years.
We then calculated parametric pseudo-observations for externalization events based on splines with 3 knots evaluated at 5 years after ICD implantation and estimated the cumulative incidence at this time point as well as the risk difference and relative risk comparing patients with high lead slack to those with low lead slack. The results of the regression analyses show an estimated cumulative incidence at 5 years of 0.07 with a 95% confidence interval (CI) of (0.04 to 0.10). The risk is quite different for the two exposure groups with an estimated risk difference of 0.07 (95% CI: (0.01 to 0.14)) and the estimated relative risk is 2.94 (95% CI: (1.11 to 7.75)).

4 Discussion

With the methods proposed in this article, we have provided a way to calculate pseudo-observations and hence perform regression modeling in data consisting of both right and interval censored data on an event of interest which is subject to competing risks. We have shown by simulations that this method avoids the bias that occurs when using methods for right censored data on either the midpoints or

the right endpoints of interval censored data. Our proposed methods also provides confidence intervals that have coverage probabilities close to the nomimal value. Our method is a further development of an approach for right censored competing risks data[12] and compared to the recently proposed method by Sabathé *et al.*[11] it requires relatively few parameters and does not require any analyst choices apart from determining the spline knots.

There are a number of considerations and assumptions for the parametric pseudo-observations for right censored data that also apply to the interval censored version. This concerns the assumption of independent censoring as well as the choice of number and positions of knots for the splines. For the interval censored data, we have imposed the additional assumption that the examination times are independent of the risk of the event of interest.

A practical limitation of our method is that it is a very computationally intensive task to estimate the spline coefficients in each leave-one-out subsample of the dataset. Fortunately, this need only be done once for each study. This is also the reason for our limited number of repetitions in our simulation study.

Although we allow that the event of interest is either right or interval censored or a mix of both, we have only considered the case where the time of the competing event is exactly observed. If this is not the case and the competing event is also interval censored, the situation is far more complicated. This is unlikely to be the case when death is the only competing event but it could be relevant if other events can preclude the event of interest. Our proposed methods do not cover this situation and are not easily extended to do so.

A special case of interval censored data to which our methods do apply is known as *current status* data in which we only have one examination for each individual. One example of such data is information from a systematic population screening for a specific condition. For a non-congenital condition, a positive screening would provide information that the condition has occurred at some point prior to the screening but nothing more yielding long intervals that reflect the uncertainty of the exact occurrence time of the condition.

5 Conclusion

In this article, we have shown how the previously proposed parametric pseudo-observations for right censored data can be extended to cover setting with both right and interval censored data. Since interval censored data are almost inevitably subject to the competing risk of death, we have formulated the methods in an illness-death model that accommodates this circumstance. We have demonstrated through simulations that the proposed method performs well with no noteworthy bias and satisfactory coverage probabilities for estimating the cumulative incidence as well absolute and relative associations with an exposure.

6 Abbreviations

ICD: Implantable cardioverter-defibrillator
CI: Confidence interval
empSE: Empirical standard error

Declarations

Ethics approval and consent to participate

Since the simulated datasets did not involve any human data, ethics approval was not applicable. The ICD study was approved by the local science ethics committee of the North Denmark Region (N-20110038) and the Danish Data Protection Agency (2008-58-0028). By Danish law, no informed consent is required for a register-based study of anonymized data.

Consent for publication

Not applicable.

Availability of data and materials

The simulated datasets used and analysed during the current study are available from the corresponding author on reasonable request.

The ICD data that support the findings of this study are available from the Danish Pacemaker and ICD Register but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Danish Pacemaker and ICD Register.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by Aalborg University Hospital and supported by a grant from the Danish Pacemaker and ICD Register. Neither of the funding sources had any role in the current research project.

Author's contributions

MNJ, SLC and ETP developed the methodology. JML provided the ICD data. MNJ performed simulations and analyzed both simulated data and the ICD data. All authors have provided critical comments to drafts of the manuscript and approved the final version.

Acknowledgements

Not applicable.

Author details

[1]Unit of Clinical Biostatistics, Aalborg University Hospital, Sdr Skovvej 15, 9000 Aalborg, DK. [2]Department of Clinical Medicine, Aalborg University, DK. [3]Section for Biostatistics, Department of Public Health, Aarhus University, DK. [4]Department of Cardiology, Aalborg University Hospital, DK.

References

1. Lindsey, J.C., Ryan, L.M.: Methods for interval-censored data. Statistics in Medicine **17**(2), 219–238 (1998). doi:10.1002/(sici)1097-0258(19980130)17:2¡219::aid-sim735¿3.0.co;2-o
2. Singh, R.S., Totawattage, D.P.: The statistical analysis of interval-censored failure time data with applications. Open Journal of Statistics **03**(02), 155–166 (2013). doi:10.4236/ojs.2013.32017
3. Andersen, P.K.: Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika **90**(1), 15–27 (2003). doi:10.1093/biomet/90.1.15
4. Peto, R.: Experimental survival curves for interval-censored data. Applied Statistics **22**(1), 86 (1973). doi:10.2307/2346307
5. Turnbull, B.W.: The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological) **38**(3), 290–295 (1976). doi:10.1111/j.2517-6161.1976.tb01597.x
6. Kim, S., Kim, Y.-J.: Regression analysis of interval censored competing risk data using a pseudo-value approach. Communications for Statistical Applications and Methods **23**(6), 555–562 (2016). doi:10.5351/CSAM.2016.23.6.555
7. Overgaard, M., Parner, E.T., Pedersen, J.: Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. The Annals of Statistics **45**(5), 1988–2015 (2017). doi:10.1214/16-aos1516
8. Groeneboom, P., Wellner, J.A.: Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhäuser, Basel, Switzerland (1992)
9. Royston, P., Parmar, M.K.B.: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine **21**(15), 2175–2197 (2002). doi:10.1002/sim.1203
10. Cook, R.J., Lawless, J.F.: Multistate Models for the Analysis of Life History Data. CRC Press, Boca Raton, FL (2018)
11. Sabathé, C., Andersen, P.K., Helmer, C., Gerds, T.A., Jacqmin-Gadda, H., Joly, P.: Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. Statistical Methods in Medical Research, 096228021984227 (2019). doi:10.1177/0962280219842271
12. Johansen, M.N., Lundbye-Christensen, S., Parner, E.T.: Regression models using parametric pseudo-observations. Statistics in Medicine **n/a**(n/a) (2020). doi:10.1002/sim.8586
13. Touraine, C., Gerds, T.A., Joly, P.: SmoothHazard: An r package for fitting regression models to interval-censored observations of illness-death models. Journal of Statistical Software **79**(7) (2017). doi:10.18637/jss.v079.i07

14. Burton, A., Altman, D.G., Royston, P., Holder, R.L.: The design of simulation studies in medical statistics. Statistics in Medicine **25**(24), 4279–4292 (2006). doi:10.1002/sim.2673
15. Aalen, O., Johansen, S.: An empirical transition matrix for non-homogeneous markov chains based on censored observations. Scandinavian Journal of Statistics **5**, 141–150 (1978)

Figures

Figure 1 Full-sample estimators of the cumulative incidence function in one of the simulated datasets. Blue curve: Aalen-Johansen estimator on exact event times. Red curve: Aalen-Johansen estimator on interval midpoints. Green curve: Aalen-Johansen estimator on right endpoints. Purple curve: Penalized likelihood estimator used in the approach by Sabathé *et al.* Black curve: Flexible parametric approach used in our proposed approach.

Figure 2 Visualization of the interval censored real example dataset. A black line indicates an interval with an observed externalization, a grey line indicates an interval with no observed externalization, black dots indicate deaths or lead extractions.

Figure 3 Estimated cumulative incidence of externalization. Solid curve: Aalen-Johansen estimator in a competing risk model. Dashed curve: Flexible parametric estimator with 3 knots based on an illness-death model fitted on the full sample. Dotted curve: Flexible parametric estimator with 4 knots based on an illness-death model fitted on the full sample.

Table 1 Results of the simulations in the general set-up based on estimation of cumulative incidence, risk difference and the logarithm of relative risk.

Method	Bias	empSE	Relative empSE	Coverage (95% CI)
Cumulative incidence (true value: 0.460)				
Exact	0.000	0.028	1 (ref.)	95.4 (93.9 to 96.5)
Midpoint	−0.067	0.033	1.16	35.6 (32.7 to 38.6)
Right endpoint	−0.105	0.029	1.02	4.3 (3.2 to 5.8)
IC	−0.001	0.043	1.51	94.2 (92.5 to 95.5)
Sabathé *et al.*	0.001	0.034	1.21	95.7 (94.2 to 96.8)
Risk difference (true value: -0.128)				
Exact	0.000	0.057	1 (ref.)	95.0 (93.5 to 96.2)
Midpoint	0.020	0.057	1.00	95.5 (94.0 to 96.6)
Right endpoint	0.025	0.056	0.99	94.0 (92.3 to 95.3)
IC	−0.002	0.076	1.33	95.3 (93.8 to 96.5)
Sabathé *et al.*	−0.001	0.070	1.24	95.2 (93.7 to 96.4)
Logarithm of relative risk (true value: -0.281)				
Exact	−0.001	0.128	1 (ref.)	95.5 (94.0 to 96.6)
Midpoint	0.002	0.149	1.16	95.7 (94.2 to 96.8)
Right endpoint	−0.012	0.165	1.29	96.0 (94.6 to 97.0)
IC	−0.006	0.168	1.31	94.6 (93.0 to 95.9)
Sabathé *et al.*	−0.004	0.158	1.23	95.3 (93.8 to 96.5)

Figures

Fig. B.1: Figure 1

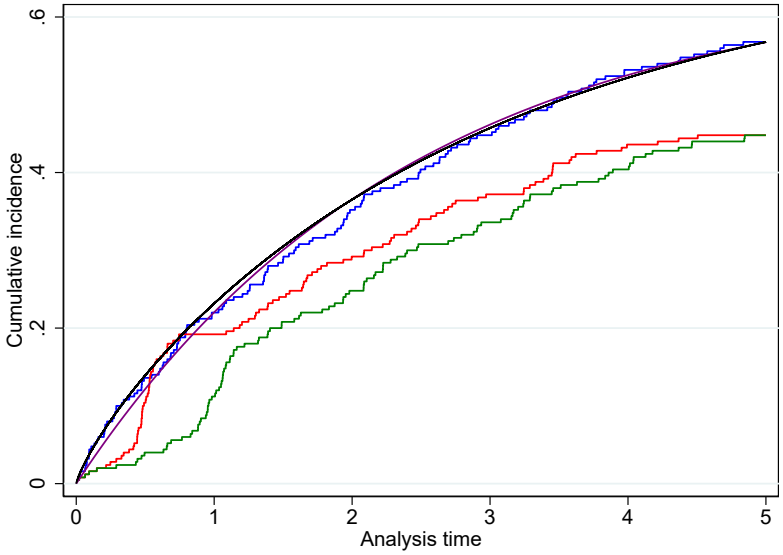


Fig. B.2: Figure 2

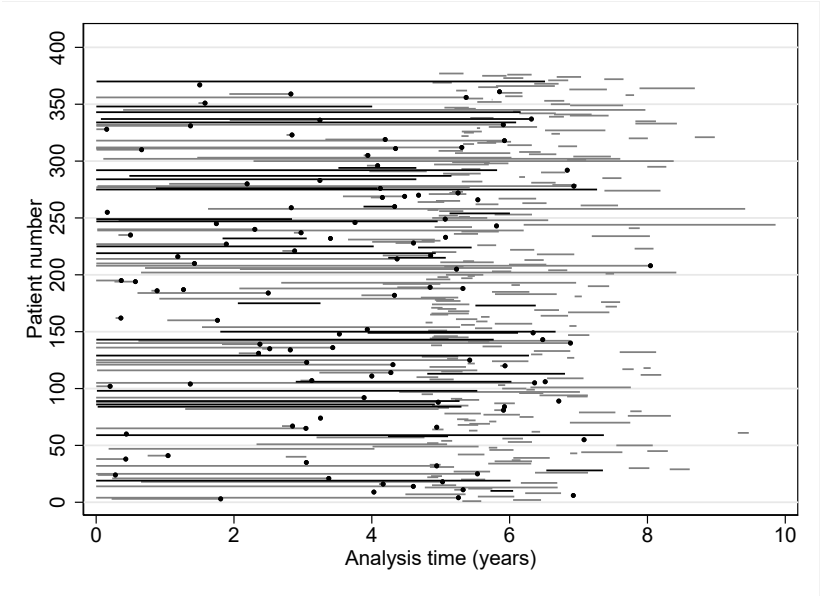
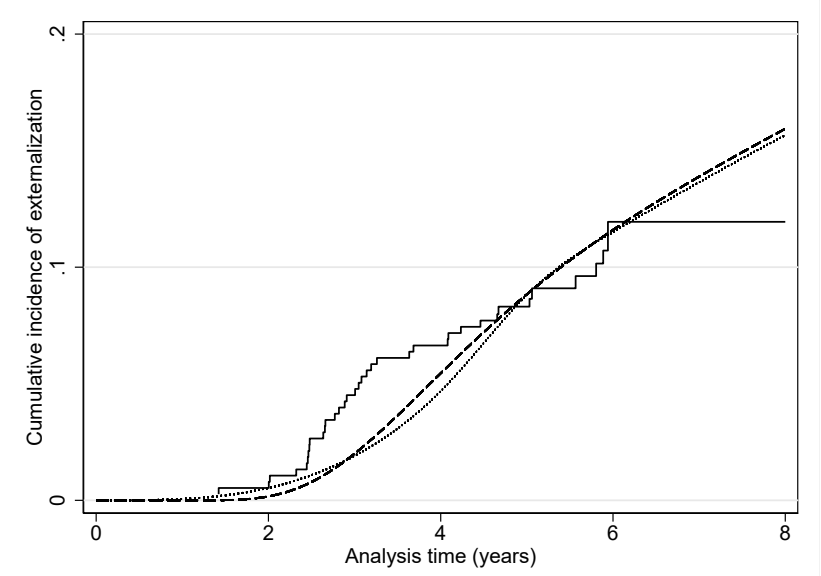


Fig. B.3: Figure 3



Paper C

Pseudo-observations for competing risks settings with interval censored time-to-event data

Martin N. Johansen[1], Søren Lundbye-Christensen[1], Jens C. Nielsen[2], Jacob M. Larsen[3], Erik T. Parner[4], and Sam Riahi[3]

1. Unit of Clinical Biostatistics, Aalborg University Hospital, Aalborg, Denmark; 2. Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; 3. Department of Clinical Medicine, Aalborg University, Aalborg, Denmark. 4. Section for Biostatistics, Department of Public Health, Aarhus University, Aarhus, Denmark.

Submitted to *Epidemiology* (2020, under review).

Description

This article focuses on the implications of having an interval censored time-to-event outcome of interest in terms of competing risks and the particular attention this circumstances require. The method of using flexible parametric pseudo-observations for both right and interval censored data is explained. The article aims to describe the methodological challenges and solutions as seen with the eyes of an applied medical researcher or epidemiologist.

Abstract

Recent developments in methods for handling right censored time-to-event data have introduced the pseudo-observation approach, which provides a methodological framework for formulating regression models. The pseudo-observation method provides the flexibility to measure associations on different scales such as absolute or relative risk differences while accounting for competing risks. The method enables researchers to avoid the typical assumption of proportional cause-specific hazards and present results of regression analyses in terms of association measures with a more simple interpretation. Calculation of pseudo-observations has traditionally been based on the non-parametric Aalen-Johansen estimator of the cumulative incidence function for right censored data. When the event status of the outcome of interest is only known at a finite number of examination times, the data is said to be interval censored, and the non-parametric pseudo-observation method is not easily extended to cover this situation. Due to a general lack of commonly applied methods for handling interval censored data, methods for right censored data have been applied to either the midpoints or the right endpoints of the observed intervals. In this paper, we argue that the inevitable presence of competing risks in interval censored data necessitates the use of an illness-death model and show how pseudo-observations for interval censored data can be calculated in this setting using a recently proposed spline-based parametric method. We illustrate the application of the methods by analyzing a dataset of 345 patients with pacemakers who are monitored for episodes of atrial fibrillation at routine check-ups.

I. Introduction

Inference for a non-fatal time-to-event outcome is a common methodological task in both observational studies and controlled trials. When the subjects under study are living creatures, this type of data is almost always subject to competing risk from death when the subjects are followed for a longer period. This has led to a widespread use of methods that can accommodate competing risks such as the cause-specific proportional hazards model[6] and the Fine & Gray proportional subdistribution hazards model.[5] An alternative to the proportional hazards models that has gained popularity in recent years is the use of *pseudo-observation* methods.[1] These methods are often applied to estimate associations on the relative risk scale but can also be used to estimate absolute risk differences, hazard rate ratios, cause-specific life-years lost and other association measures.

A special situation arises when the time of the event of interest is not observed directly but the event status is only assessed at a number of *examination times*. Such data are said to be *interval censored* and require special attention and special inferential methods. The most common methodological solution to this problem has been to impose distributional assumptions for the data and apply parametric regression models. However, a model that can handle interval censored data but remains free of distributional assumptions has been proposed by Royston and Parmar.[14] In a competing risk setting, this model relies on natural cubic splines to estimate the cause-specific hazard functions.

An example of data with competing risk is the study of patients with pacemakers who are monitored for episodes of atrial fibrillation. Atrial fibrillation is the most common arrhythmia in elderly patients and is associated with poor quality of life, stroke, heart failure and increased mortality. These patients are followed by routine check-ups to ensure the functionality of the pacemaker. At these check-ups it is also possible to detect whether the patient has experienced any atrial fibrillation episodes since the last check-up. This gives rise to interval censored information about the time to first atrial fibrillation episode. However, the device also records the actual time of atrial fibrillation episodes in continuous time which means that data can also be analyzed using methods for right censored time-to-event data for comparison. The ongoing Danish multi-centre

study DANPACE-II provides data from a cohort of patients with pacemakers which we will use for illustration in this paper.

In what follows, we discuss the methodological challenges arising from interval censoring in a competing risk setting and show some potential solutions using data from our data example. In Section II, we consider the right censored case, and describe the pseudo-observation approach in Subsection i. We introduce competing risks in Subsection ii and describe a parametric version of pseudo-observations in Subsection iii. In Section III we introduce interval censoring and we discuss the complications that arise with competing risks in Subsection i and describe a special version of pseudo-observations that accommodates an interval censored setting with competing risks in Subsection ii. We analyze data from the DANPACE-II study in both a right censored and an interval censored version in Section IV, and conclude the article with a summary and discussion in Section V.

II. Right censored time-to-event data

When interest involves a particular event that can occur during a follow-up period, it is very common to observe *right censored* data where we know the time of the event for some individuals but only know that the event has not occurred before end-of-follow-up (censoring) for other individuals. Methods for dealing with right censored time-to-event data are very widely used in medical research with the far most commonly applied regression model being the Cox proportional hazards model[4]. If we let T denote the random variable representing the time to the event, the *hazard function* expresses the instantaneous risk of experiencing the event given that it has not already occurred and can be defined as

$$h(t) = \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t},$$

for small values of Δt. The Cox model is based on an assumption of proportionality between hazard functions for different covariate values as well as the assumption that the individuals being censored at a given time point should be representative for the population at risk at that time with the same covariate values, an assumption known as *conditional independent censoring*. Under these assumptions, the hazard rate ratio for different values of the exposure of interest is estimated to assess the association between the exposure and the outcome.

i. Non-parametric pseudo-observations

Regression modeling without assuming proportionality of the hazard functions can be performed using pseudo-observations[1] which can be considered as a transformation of the censored data into a dataset of estimated uncensored event times. They can be thought of as the contribution of the data of each individual to the estimate of the cumulative incidence function,

$$F(t) = 1 - S(t).$$

In a setting without competing risks, the survival function, $S(\cdot)$, can be estimated by the non-parametric Kaplan-Meier estimator[9], say $\hat{S}(\cdot)$, and an estimate of the cumulative incidence function can be calculated as $\hat{F}(\cdot) = 1 - \hat{S}(\cdot)$, say.
The i'th non-parametric pseudo-observation calculated at time t is then defined as

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

where $\hat{F}^{-i}(\cdot)$ is one minus the Kaplan-Meier estimator calculated without the i'th individual, called the *leave-one-out* or *jackknife* estimator. These pseudo-observations have mathematical properties that enable us to use a versatile family of regression models, *generalized linear models* (GLMs), to estimate associations in terms of different association measures such as relative risk or risk difference. Since the pseudo-observations are calculated at one (or multiple) prespecified point(s) during follow-up, there are no assumptions concerning the shape or proportionality of the hazard functions. However, there is an important assumption requiring unconditional independent censoring because the pseudo-observations are calculated from a pooled version of the Kaplan-Meier estimator. This assumption can be relaxed in several ways to allow covariate-dependent censoring.[1, 2] A thorough description of pseudo-observations and their properties has been given in the article by Andersen & Perme[1] and a discussion of assumptions and interpretation of regression parameter estimates is given by Mortensen et al.[11]

ii. Death as a competing risk

If the event of interest is non-fatal, we may also observe that some patients die during follow-up and are therefore no longer at risk of experiencing the event of interest. This should not simply be considered as censoring since they cannot have the event of interest later on.[13] More generally, the competing risk setting can be formulated as a multi-state model in which subjects enter the study in an initial state being free of any of the relevant (competing) events and are then at risk of a number of different events that each preclude the others or otherwise makes further follow-up irrelevant. We will call the initial state 0, and consider K competing events leading to states $1, \ldots, K$. Figure 1 illustrates such a multi-state model. The arrows in this figure represent the potential *transitions* from one state to another which can be modeled as *cause-specific hazard functions*.

Letting D and T denote the random variables describing the type of the first event to occur for an individual and the time to this first event, respectively, the cause-specific hazard function for the transition into state k at time t is defined as

$$h_k(t) = \frac{P(t \leq T \leq t + \Delta t, D = k | T \geq t)}{\Delta t},$$

for $k = 1, \ldots, K$ and small values of Δt. As we shall see in (2), we can use these to calculate the cause-specific cumulative incidence functions, $F_1(\cdot), \ldots, F_K(\cdot)$, and they also have the convenient property that they can be used to calculate the overall survival function, i.e. the probability of not having experienced any of the K events prior to time t, as

$$S(t) = \exp\left(- \sum_{k=1}^{K} \int_0^t h_k(u) du \right).$$

In the special case where we are only considering an event of interest and a single competing event (death), there are only three states; 0 (alive and event-free), 1 (event of interest has occurred), and 2 (dead before event of interest).

Non-parametric pseudo-observations can then be calculated using the Aalen-Johansen estimator of the cause-specific cumulative incidence functions[16] and inference on cause-specific scales can be obtained using GLM regression methods as in the case without any competing risks.

iii. Parametric pseudo-observations

In a recent article,[7] we have proposed a parametric version of pseudo-observations for competing risk data which is based on the principles of the *flexible parametric model* advocated by Royston

& Parmar.[14] The non-parametric estimators of the cumulative incidence are step functions with larger and less precise steps when the risk set is small near the end of follow-up period and the non-parametric pseudo-observations will carry over this imprecision to the regression modeling. The underlying cumulative incidence function is presumably a smooth function and by estimating it by means of a smooth spline, we will be able to provide a more stable basis for pseudo-observations. More technically, the parametric pseudo-observations method uses a restricted cubic spline to estimate the log cumulative hazard function and then calculate pseudo-observations based on this parametric estimator. As an extension of the pseudo-observation approach for right censored data to a competing risk setting, we can base the inference on the cause-specific hazard functions, $h_1(\cdot), \ldots, h_K(\cdot)$. The cause-specific cumulative incidence function for event type k can then be obtained from the overall survival function and the cause-specific hazard function as

$$F_k(t) = \int_0^t S(u) \cdot h_k(u) du. \tag{2}$$

This provides an estimator that is flexible and fully parametric and allows us to define parametric pseudo-observations as in (1). We will denote the i'th parametric pseudo-observation derived using this method as $\hat{\theta}_i^p(t)$. Similarly to the non-parametric pseudo-observations, we can perform regression modeling using GLMs to estimate associations on the desired scale. We have shown by simulations that when pseudo-observations are calculated at time points where the risk set is small or when there is a large amount of events after the analysis time point, these parametric pseudo-observations reduce the variability of parameter estimates compared to that of the non-parametric pseudo-observations.[7]

III. Interval censored data

Considering again the case without competing risks, the special data structure that arises when the event of interest is interval censored requires methods that are specifically equipped for this situation. The most common method of estimating parameters is by use of *maximum likelihood estimation* where a likelihood function is used to find the estimates that give the maximum probability for observing the data actually observed. For an individual with an interval censored event in the semi-closed interval $(l, r]$, the contribution to the likelihood function is simply the difference in the survival function over the interval, $S(l) - S(r)$.[10] In parametric models, we have a closed form for the survival function which means that standard maximum likelihood methods can be applied to obtain parameter estimates in a regression model. However, most parametric models impose strict distributional assumptions and are confined to a given scale for measuring assocations. The parametric nature of the flexible parametric model[14] facilitates the handling of interval censored data without specific distributional assumptions, but it does require proportionality of either the hazard or the odds function.

In practical applications, it is common to use either the right endpoint, r, or the midpoint, $(l + r)/2$, of the interval as the time of event and apply methods for right censored data although these methods have known shortcomings. Firstly, using the right endpoint as the time of the event will inevitably cause an underestimation of the cumulative incidence since the right endpoint is the upper limit of the time to event. The midpoint method is also known to produce results that might be biased.[12] Both methods will also over-estimate the information in the data, producing too narrow confidence intervals and too low p-values, resulting in incorrect interpretation of the data. Using either the right endpoint or the midpoint, we will need to censor at the last examination time if no event of interest is observed. This means that we will not be making full use of the information in the data.

i. Competing risk problems with interval censored data

When an exactly observed (right censored) competing risk occurs for an interval censored event of interest, the methods for handling competing risks in right censored settings do not generalize directly for two reasons. First, the contribution to the likelihood function for an interval censored event of interest is no longer just the difference in the survival function at each interval endpoint but should instead be calculated as an integral over the interval in which the event is known to occur, $(l, r]$. Second, the inference we described in Section II for competing risk settings is based on the transition out of the initial state, 0, which can be characterized by the time of the first event, T, and the event type indicator, D. In the interval censored setting, this characterization is not possible because we do not know whether the event of interest has occurred if there is no later examination time. This is illustrated in Figure 2. If an individual enters the study at time 0, subsequently has a negative examination at time l, and then dies at time t, the event of interest may have occurred in the interval $(l, t]$ and the transition into state 2 can be from either state 0 or 1. This uncertainty has to be taken into account when we estimate the cumulative incidence of the event of interest. This can be accommodated in the likelihood function by computing the probability that the event of interest has occurred between l and t. A similar approach can accommodate the situation where the individual is censored at time t after having a negative examination at time l to take into account that the event may have occurred between l and t. A multi-state model that represents the situation with an event of interest and a competing risk and takes the transition from the event of interest to death into account is often called an irreversible *illness-death* model. Figure 3 illustrates this multi-state model. In an illness-death model all possible transitions between the three states are modeled separately and can be expressed as the *transition-specific hazard functions* $h_{01}(\cdot)$, $h_{02}(\cdot)$, and $h_{12}(\cdot)$ and regression modeling can be performed on all or some of these transitions.

ii. Pseudo-observations for interval censored data with competing risks

The parametric pseudo-observation approach can be extended to cover the interval censored competing risk setting by using an illness-death model in which each transition-specific hazard function is modeled in the same way as the cause-specific hazard functions for right censored data. Calculation of the pseudo-observations is still based on the full-sample as well as the jackknife estimates of the cumulative incidence function as in (1). If we restrict our interest to the transition from state 0 to state 1, we still need to model the other two transitions using cubic splines to calculate the pseudo-observations. This means that we must fit separate splines (without covariates) to each of the three log cumulative hazard functions. But once these are fitted to the data, we can define parametric pseudo-observations for the interval censored event of interest in the same way as in the right censored setting as

$$\theta_i^{ic}(t) = n \cdot \hat{F}_{01}(t) - (n-1) \cdot \hat{F}_{01}^{-i}(t), \tag{3}$$

where $\hat{F}_{01}^{-i}(t)$ is the jackknife estimator of the cumulative incidence of the event of interest. Once we have obtained the pseudo-observations, estimation of the association between an exposure variable and the event of interest can be obtained from a GLM regression model.

The assumption of unconditional independent censoring that applies to non-parametric pseudo-observations also applies to these parametric pseudo-observations for interval censored data. Since we are using splines to model each of the transitions, we must define a set of knots for each spline. The number of knots determines the smoothness of a spline and these should be chosen carefully. The precise positions of the knots is usually of less importance and standard methods based on

fractiles of the distribution of event times will generally suffice. The jackknife estimators should be based on the same spline knot points as the full-sample estimators.

IV. Analysis of the DANPACE-II data

We now consider the DANPACE-II dataset which consists of data from 345 individuals who had a pacemaker implantation between May 2014 and December 2018 and are followed until the data extraction was performed at 20 March 2019 to illustrate the use of the pseudo-observation methods. The study is an ongoing trial registered at clinicaltrials.gov as Study NCT02034526 performed as a follow-up to the original DANPACE study.[3] During follow-up, 156 of the patients experienced the outcome of interest which is defined as an atrial fibrillation episode lasting for at least 6 minutes. Furthermore, 10 individuals died during follow-up and among these 6 had not experienced the outcome of interest. For each individual with the outcome of interest, we have the exact date of the outcome, and all individuals are scheduled to have their pacemaker checked at 3, 12, and 24 months after implantation. Since we do not have information on the actual examination dates, we have simulated individual examination dates by adding a random variation to the scheduled examinations following a normal distribution with mean 0 and a variance of 10 days. This enables us to perform analyses on this dataset based on either the right censored data of exact outcome times or the interval censored data arising from only assessing the event status at the examination times. Since the patients in our dataset are no longer monitored after 2 years, we consider the individuals to be censored after 2 years of follow-up.

For the purpose of this analysis, we only have additional information on the sex and age at implantation of the study individuals. We will be estimating the overall cumulative incidence proportion at both 9 and 18 months after implantation as well as the association with sex and age as a binary and a continuous exposure, respectively. Figure 4 shows the individual patient trajectories in terms of both exact event times (black crosses) and examination intervals (black line for an observed event and grey line for no observed event) as well as death times (black dots). We can see that there is a bulk of events of interest occurring early in the follow-up period. Figure 5 shows the cumulative incidence function for atrial fibrillation episodes estimated by the non-parametric Aalen-Johansen estimator for a competing risk model based on the right censored event times (solid black curve), the interval midpoints (solid blue curve), the right endpoints (solid red curve) and the corresponding full-sample estimator in (3), $\hat{F}_{01}$, with one internal knot based on the interval censored data (dashed black curve). The estimated curve based on the exact event times increases most in the very early follow-up as a result of the high number of early events which none of the other methods are able to capture due to the limited information in the interval censored data. The curves based on interval midpoints or right endpoints both increase steeply at the imputed event times in remain constant in between reflecting the homogeneity of the examination patterns in this dataset. The spline-based estimated curve, on the other hand, estimates the cumulative incidence in reasonable accordance with the non-parametric estimator based on exact event times most of the time, though it also fails to capture the initial steep increase.

We then calculated four sets of pseudo-observations based on 1) the right censored exact event times, 2) right censoring using midpoints, 3) right censoring using right endpoints, and 4) interval censored event times. For the three right censored approaches, we calculated pseudo-observations in a competing risk model using the Aalen-Johansen estimator, whereas the interval censored approach is based on the flexible parametric pseudo-observations defined in (3). We used these pseudo-observations to fit generalized linear models in order to estimate the relevant parameters. The estimates of cumulative incidence proportion (CIP) at 9 and 18 months, the risk difference of experiencing an atrial fibrillation episode for males as compared to females (RD_{sex}) and the risk

difference associated with age using 10 years as the unit of comparison (RD_{age}) are shown in Table 1. The estimates confirm that the right endpoints can underestimate the CIP notably and that both interval midpoints and right endpoints somewhat underestimate the CIP at long-term follow-up. The associations with sex and age are rather consistently estimated by the four methods despite the differences in the overall CIP estimates indicating that in this dataset the biases more or less cancel out in the comparisons between patient groups.

V. Discussion

Traditionally, the available analytical tools to perform regression modeling on interval censored data have been parametric models based on distributional assumptions. However, they have only seen limited use in the medical literature whereas simplified approaches such as applying methods for right censored data to either the midpoints or the right endpoints have been more commonly applied. Since the outcome of interest in an interval censored setting is almost always subject to the competing risk of death, we have argued that the theoretically most justified approach to such modeling tasks is through an illness-death model. In this article, we have also shown one possible way to implement such a model through an extension of the flexible parametric pseudo-observation approach which was originally developed for a right censored competing risk setting.

Other possible ways to implement an illness-death model for interval censored data include the Royston & Parmar flexible parametric regression model,[14] a penalized likelihood approach based on M-splines[8] as well as a pseudo-observation approach based on this penalized likelihood method.[15] The cubic spline based method has the advantage that it facilitates the use of a wide range of association measures such as risk differences, relative risks or cause-specific life-years lost while it remains free of assumptions about the underlying hazard functions.

Since the calculation of pseudo-observations relies on jackknife estimators, the parametric pseudo-observations require quite intensive computing and can be lengthy or even infeasible for very large datasets. When we calculated pseudo-observations using the interval censored data in the DANPACE-II dataset using a single CPU core on a standard desktop PC, the calculations took about 1 h 50 m. Once the pseudo-observations have been calculated, however, different regression models can be fitted using standard regression procedures very fast on any standard computer.

The results of estimating the absolute cumulative incidence in our example dataset show the importance of account for the interval censored nature of the data. Analyzing the data in an illness-death model using an approach that correctly handles the interval censoring provides a way to make use of all relevant information in the data and ensure reasonable estimates of the relevant parameters. We acknowledge that the correct handling of an interval censored event of interest in the presence of competing risks might seem a little overwhelming for some applied researchers, and we encourage the collaboration between researchers with different professional competencies, e.g. medical doctors, epidemiologists, statisticians, and statistical programmers. The Stata syntax that we have used in this paper is available upon request.
As we have seen, there are different benefits and drawbacks related to using either of the methodological approaches to analyzing a dataset with an interval censored event of interest and competing risks. We have summarized these in Table 2. The steps required to calculate parametric pseudo-observations evaluated at time t for interval censored data can be summarized as follows.

1. *Organize the data such that the relevant time variables and event indicators are defined for each subject.*
2. *Fit the illness-death model to the full sample and save the estimated value of the CIP at time t,* $\hat{F}_{01}(t)$.
3. *For each subject, i, fit the illness-death model to the jackknife sample leaving out the i'th subject and save the estimated value of the CIP at time t,* $\hat{F}_{01}^{-i}(t)$.
4. *Generate the pseudo-observations defined in (3).*
5. *Use GLM regression on the pseudo-observations with the relevant link function and covariates.*

In our implementation, we have used the `ml` command in Stata/MP version 16.1 to maximize the likelihood function of the illness-death model.
The main drawbacks of using the parametric pseudo-observations for interval censored data are that it requires rather advanced programming and intensive computing power which might even call for the use of more powerful computers than general-purpose desktops. Although the methods yield no directly conflicting results in our example, we strongly recommend using appropriate handling of interval censored data whenever possible.

VI. Tables

Analysis time	Method	Overall CIP (95% CI)	RD_{sex} (95% CI)	RD_{age} (95% CI)
9 months	Exact	41.4 (36.1 to 46.6)	-10.5 (-21.0 to 0.1)	6.6 (2.0 to 11.1)
	Midpoint	43.8 (38.4 to 49.2)	-14.8 (-25.7 to -4.0)	5.7 (0.9 to 10.5)
	Right endpoint	38.0 (32.7 to 43.3)	-9.3 (-20.0 to 1.4)	6.2 (1.6 to 10.8)
	Interval censored	44.3 (38.8 to 49.9)	-12.9 (-24.2 to -1.6)	6.7 (1.6 to 11.9)
18 months	Exact	47.0 (41.5 to 52.4)	-15.2 (-26.1 to -4.3)	6.4 (1.5 to 11.2)
	Midpoint	46.6 (40.6 to 52.7)	-15.6 (-27.8 to -3.4)	6.5 (1.4 to 11.5)
	Right endpoint	50.2 (43.8 to 56.6)	-16.0 (-28.7 to -3.3)	7.7 (1.8 to 13.7)
	Interval censored	46.9 (41.1 to 52.6)	-13.1 (-24.8 to -1.3)	7.0 (1.7 to 12.3)

CIP: Cumulative incidence proportion (in percentage)
RD: Risk difference (of percentages)

Table 1: *Results of GLM regression analyses based on pseudo-observations.*

Method	Benefits	Drawbacks
Right censored interval midpoints	Fast computation Easy implementation	Downward bias in CIP estimation
Right censored right endpoints	Fast computation Even easier implementation	Further downward bias in CIP estimation
Interval censored data	Unbiased estimation of CIP	Slow computation Complicated implementation

Table 2: *Benefits and drawbacks of using different methods for handling interval censored datasets.*

VII. Figures

Figure 1: *A general competing risks multi-state model.*

Figure 2: *The two potential trajectories for an individual who enters the study at time 0, has a negative examination at time l and then dies at time t.*

Figure 3: *The irreversible illness-death model.*

Figure 4: *Individual trajectories for the 345 patients with pacemakers. Black lines indicate an interval censored event of interest, grey lines indicate an interval without the event of interest, black crosses indicate the exact event times for the event of interest, and black dots indicate death times.*

Figure 5: *Estimated cumulative incidence functions based on the non-parametric Aalen-Johansen estimator using exact event times (solid black), interval midpoints (solid blue), right endpoints (solid red) and spline-based estimated curved based on interval censored data (dashed black). Vertical reference lines indicate the analysis time points.*

References

[1] Andersen P, Perme M. Pseudo-observations in survival analysis. *Statistical methods in medical research*. 2010;19:71–99.

[2] Binder N, Gerds TA, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*. 2014;20(2):303–315.

[3] Nielsen JC, Thomsen PEB, Højberg S, Møller M, Vesterlund T, Dalsgaard D, Mortensen LS, Nielsen T, Asklund M, Friis EV, Christensen PD, Simonsen EH, Eriksen UH, Jensen GVH, Svendsen JH, Toff WD, Healey JS, Andersen HR, on behalf of the DANPACE Investigators. A comparison of single-lead atrial pacing with dual-chamber pacing in sick sinus syndrome *European Heart Journal*. 2011;32:686–696.

[4] Cox D. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*. 1972;34:187–220.

[5] Fine J, Gray R. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*. 1999;94:496–509.

[6] Holt J. Competing Risk Analyses with Special Reference to Matched Pair Experiments. *Biometrika*. 1978;65:159–165.

[7] Johansen M, Lundbye-Christensen S, Parner E. Regression models using parametric pseudo-observations. *Statistics in Medicine*. 2020;39:2949–2961.

[8] Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*. 2002;3(3):433–443.

[9] Kaplan E, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. 1958;53:457–481.

[10] Klein J, Moeschberger M. Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer; 2003.

[11] Mortensen L, Hansen C, Overvad K, Lundbye-Christensen S, Parner E. The Pseudo-Observation Analysis of Time-To-Event Data. Example from the Danish Diet, Cancer and Health Cohort Illustrating Assumptions, Model Validation and Interpretation of Results. *Epidemiologic Methods*. 2018;20170015.

[12] Odell PM, Anderson KM, D'Agostino RB. Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull-Based Accelerated Failure Time Model. *Biometrics*. 1992;48(3):951–959.

[13] Putter H, Fiocco M, Gekus R. Tutorial in biostatistics: Competing risk and multi-state models. *Statistics in Medicine*. 2007;26:2389–2430.

[14] Royston P, Parmar M. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*. 2002;21:2175–2197.

[15] Sabathé C, Andersen P, Helmer C, Gerds T, Jacqmin-Gadda H, Joly P. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical Methods in Medical Research*. 2020;29(3):752–764.

[16] Aalen O, Johansen S. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*. 1978;5:14–150.

Figures

Fig. C.1: Figure 1

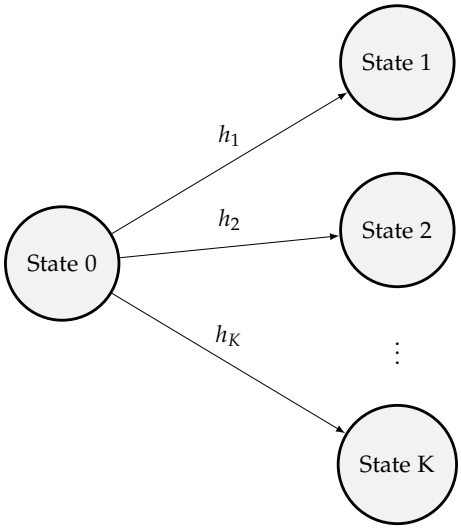


Fig. C.2: Figure 2

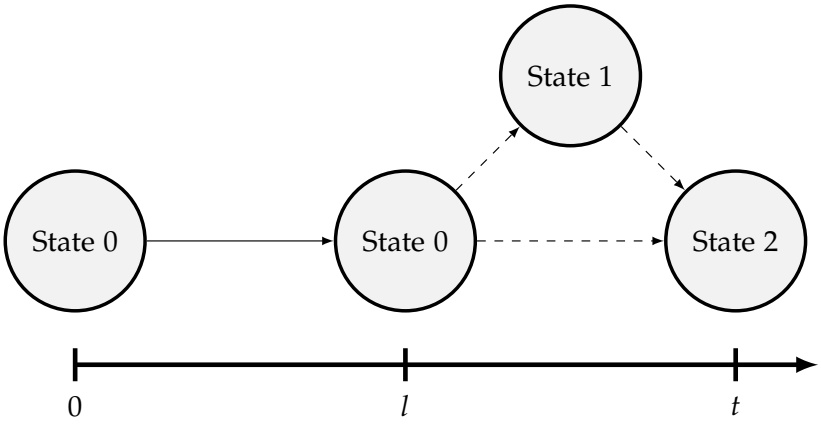


Fig. C.3: Figure 3

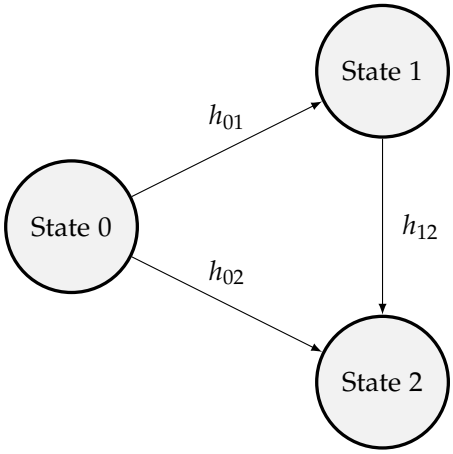


Fig. C.4: Figure 4

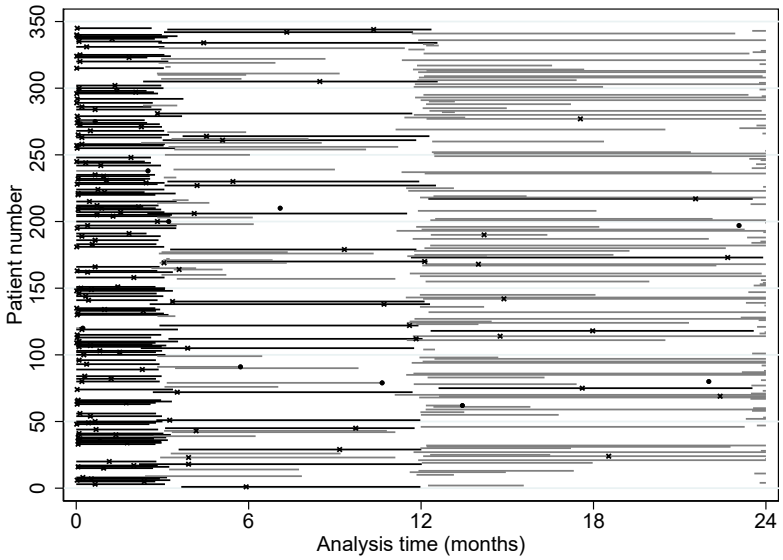
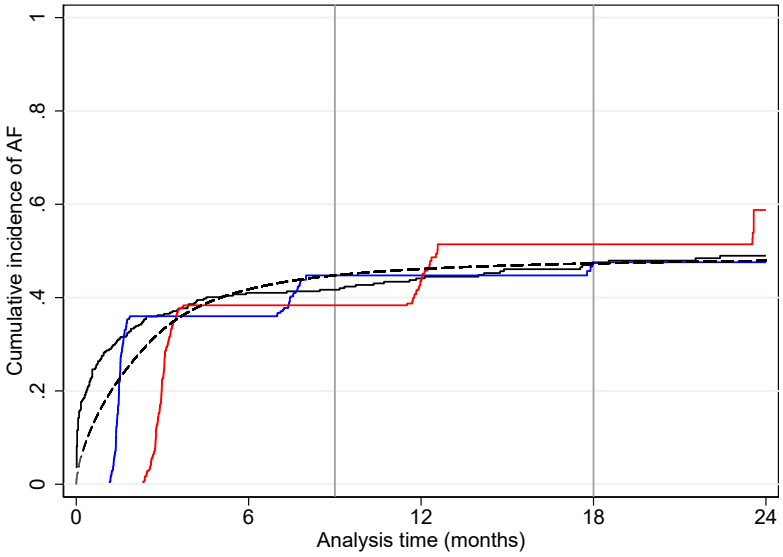


Fig. C.5: Figure 5



ISSN (online): 2246-1302
ISBN (online): 978-87-7210-844-5

AALBORG UNIVERSITY PRESS