Regression Models for Interval Censored Data Using Parametric Pseudo-Observations

Martin Nygård Johansen (martin.johansen@rn.dk)
Aalborg University Hospital https://orcid.org/0000-0001-9790-0985
Søren Lundbye-Christensen
Aalborg University Hospital
Jacob Moesgaard Larsen
Aalborg University: Aalborg Universitet
Erik Thorlund Parner
Aarhus University: Aarhus Universitet

Technical advance

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

Posted Date: September 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-78804/v1

Version of Record: A version of this preprint was published on February 15th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01227-8.

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

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

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

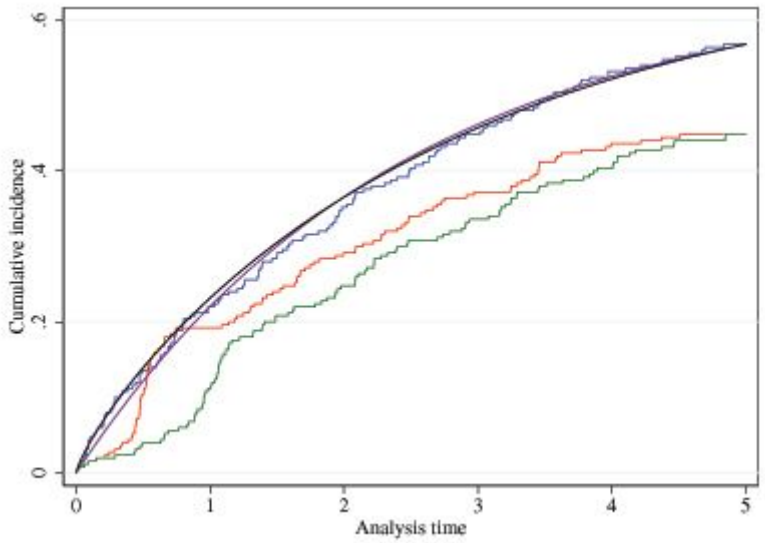

Figure 1

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 Sabathe et al. Black curve: Flexible parametric approach used in our proposed approach.

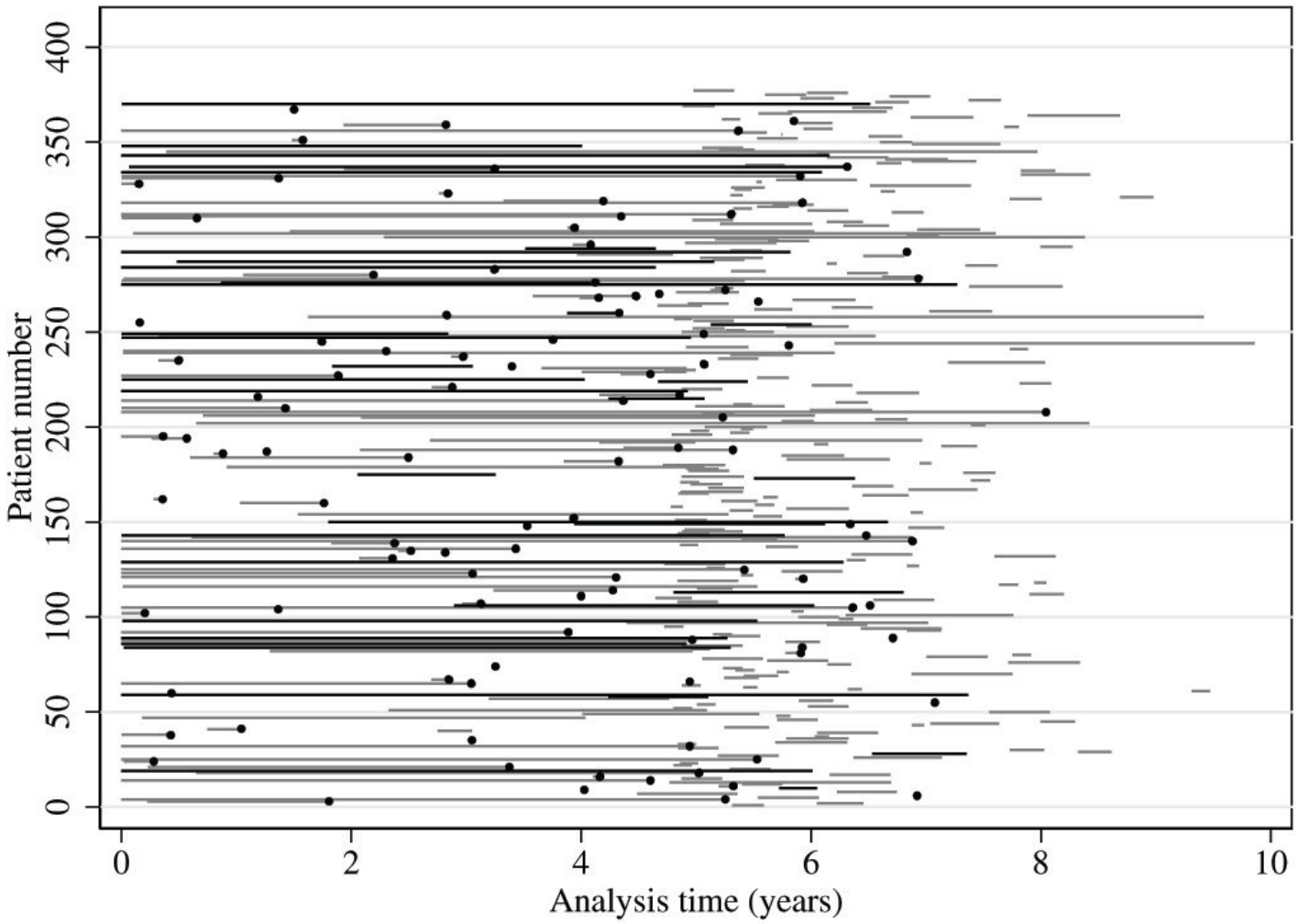

Figure 2

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.

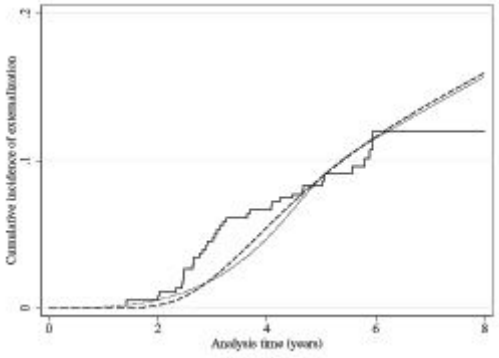

Figure 3

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.