Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring

Olivier Bouaziz[1]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

Abstract

In the context of right-censored and interval-censored data we develop asymptotic formulas to compute pseudo-observations for the survival function and the Restricted Mean Survival Time (RMST). Those formulas are based on the original estimators and do not involve computation of the jackknife estimators. For right-censored data, Von Mises expansions of the Kaplan-Meier estimator are used to derive the pseudo-observations. For interval-censored data, a general class of parametric models for the survival function is studied. An asymptotic representation of the pseudo-observations is derived involving the Hessian matrix and the score vector. Theoretical results that justify the use of pseudo-observations in regression are also derived. The formula is illustrated on the piecewise-constant-hazard model for the RMST. The proposed approximations are extremely accurate, even for small sample sizes, as illustrated on Monte-Carlo simulations and real data. We also study the gain in terms of computation time, as compared to the original jackknife method, which can be substantial for large dataset.

Keywords: Pseudo-observations; Restricted Mean Survival Time; Von Mises expansions; Jackknife; Interval-censoring.

1 Introduction

In order to study censored data in time to event analysis it is common to model the hazard rate. This allows to correctly take into account censoring in the estimation procedure and provides hazard ratio estimates in the framework of proportional hazard models. However, in some contexts other quantities, that have a more direct interpretation related to the studied problem, might be of interest. One example is the Restricted Mean Survival Time (RMST) which is defined as the average survival time up to a fixed point. In that case, it is common (see [1], [2], [3]) to first model the hazard rate using for instance a Cox model, to derive a survival estimator from the estimated hazard rate and to obtain an estimator of the RMST by integrating out this function. This procedure results in a cumbersome computation where it might be difficult to disentangle the effect of each covariate on the RMST. This is a serious drawback for medical applications and there is a need for more direct approaches. There are several other contexts that are concerned by the difficulty of direct modelling of the quantity of interest. This is typically the case for cumulative incidence functions in a competing risk setting or transition probabilities in a multi-state framework.

Pseudo-observations have been developed in the seminal work of [4] to answer this problem. Those pseudo-observations are constructed using the jackknife method from an estimator of the survival function. Theoretical results in [5] show that the pseudo-observations can then be used as response variables in a regression model for the quantity of interest, such as the conditional RMST, the cumulative incidence functions in a competing risk setting or the transition probabilities in a multi-state framework. This offers the possibility to directly model the quantity of interest and it is often performed by use of a generalised linear model.

Another more recent areas of development involving pseudo-observations concerns the study of machine learning methods for time to event analysis. In this context, the problematic is similar: one aims at deriving a complex model, based for instance on neural networks, for quantities of interest such as the survival function (see [6]), the cumulative incidence function (see [7], [8]) or the RMST (see for instance [9]). The use of pseudo-observations is then appealing since, once the pseudo-observations are obtained, it is possible to directly use any standard machine learning algorithm by considering those pseudo-observations as (non-censored) response variables.

Methods based on pseudo-observations are also attractive for interval-censored data. With those data, it is challenging to build a regression model based on semi-parametric methods, for quantities of interest such as the RMST. This is due to the lack of informations induced by interval-censoring. As a matter of fact, even in a nonparametric setting it may be problematic to perform estimation of the survival function. In this context, one usually relies on the Turnbull estimator or the convex minorant method which were introduced in [10] and [11], respectively. In [11] it has been proved that these estimators achieve the slow rate of convergence of order $n^{1/3}$ and their distribution is not Gaussian and cannot be explicitly computed. In a regression context, for the estimation of the hazard rate, the Cox model with nonparametric baseline was studied in [12] but again, the baseline survival function has the $n^{1/3}$ slow rate of convergence and the asymptotic distribution of this estimator could not be derived. As a result, it is common to rely on fully parametric models for modelling quantities such as the survival function or the hazard rate in a regression context. In [13] and [14] a Cox model was studied using parametric baselines such as Weibull or piecewise constant. The methods used to perform estimation are based on maximum likelihood theory where the parametric estimators are derived by maximising the likelihood of the observed data. This allows to recover the classical $\sqrt{n}$ rate of convergence of the parametric estimators. However, the derivation of the estimators is not explicit, even in the absence of covariates, and rely on a maximisation algorithm such as the Newton-Raphson procedure. In [15] a different approach was proposed based on the EM algorithm by considering the true event times as unobserved variables. This method has the advantage that direct estimators can be computed in the E-step of the algorithm when no covariates are present, which results in a stable and robust estimation procedure. All the aforementioned methods consider estimation of the survival function or of the hazard rate through proportional hazard assumptions, but they are not suited for direct modelling of the RMST, in a regression context. However, this can be achieved by using the pseudo-observations approach. In [16], an illness-death model was considered, and conditional transition probabilities or RMST were computed based on this approach. In order to compute the pseudo-observations, the cumulative transition intensities were estimated using either a penalised spline approach or assuming a Weibull distribution. Similarly, in [17] pseudo-observations were computed using a spline approach in order to estimate parameters related to the cumulative incidence function in a competing risk setting.

The key concept about pseudo-observations is that they are built based on the unconditional jackknife estimator of the quantity of interest. While applying the jackknife is straightforward in practice, a limitation of this method comes from the computation burden of calculating the initial estimator n times, where n is the sample size. There exists some R functions designed to improve the computation time, such as the `jackknife` function in the `prodlim` package, or the `pseudo_independent` function in the `eventglm` package, which rely on a C++ implementation but the gain is limited as the initial estimator still needs to be implemented n times. The computational burden is particularly important for interval-censored data where there is no direct calculation of the estimators, even in the absence of covariates. In this paper, we develop approximated formulas for pseudo-observations where the jackknife technique does not need to be implemented. In our formulas, the pseudo-observations can be directly computed based on the initial estimator. In the case of right-censored data, we provide formulas based on

Von Mises expansion of the Kaplan-Meier estimator. In the case of interval-censored data, we derive general formulas for parametric models that only involve the original estimator, the score function and the Hessian of the density. Those formulas are approximations of the original jackknife procedure in the sense that they are equal to the original pseudo-observations up to a remainder term that tends towards 0 as n tends to infinity.

However, they turn out to have a very high precision even for moderate sample sizes. Since they only involve the original estimator, the score vector and the Hessian matrix in a parametric context, they are extremely fast to compute thus resulting in a drastic reduction of time.

In the next section, we present a brief summary on the pseudo-values approach. In Section 3 we develop asymptotic formulas for computing pseudo-observations of the survival function and the RMST in the context of right-censored data. The case of interval censored data is studied in Section 4. We first discuss the context of nonparametric estimation of the survival function in Section 4.1. Then the asymptotic pseudo-observations formulas are developed for general parametric models in Section 4.2. In Section 4.3, theoretical validation of pseudo-observations for parametric models are provided: those results show that the conditional expectation of pseudo-observations approximate the conditional expectation of the response variable of interest. Simulations studies for modelling the conditional RMST in the context of right-censored or interval-censored data are conducted in Section 5 where precision and computation time of the approximate formulas are evaluated. Finally, two real data are analysed using the proposed methodology in Section 6.

2 Backgrounds on pseudo-regression estimation methods

Let $T_1^*, \ldots, T_n^*$ be n independent and identically distributed (i.i.d.) time to event variables of interest, let θ be a parameter of the form $\theta = \mathbb{E}[h(T_i^*)]$, where h is a known function. Then introduce $Z_1, \ldots, Z_n$ n i.i.d. covariates and define the conditional expectation $\theta_{(l)} = \mathbb{E}[h(T_l^*) \mid Z_l]$. We further assume there exists an invertible link function g such that $g(\theta_{(l)}) = Z_l^\top \beta$, where β is a vector of regression parameters of interest. Instead of observing the T_i^*'s one usually observes a sample $X_1, \ldots, X_n$ of i.i.d. variables, from which an estimator $\hat{\theta}$ is constructed. The l^{th} pseudo-observation is then given by:

$$\hat{\theta}_{(l)} = n\hat{\theta} - (n-1)\hat{\theta}^{(-l)}, \tag{1}$$

where $\hat{\theta}^{(-l)}$ is the jackknife estimator of $\hat{\theta}$, that is the estimator $\hat{\theta}$ computed on the sample where the l^{th} observation has been removed.

It has been suggested (see [4]) to estimate β based on the estimating equation

$$U(\beta) = \sum_{l=1}^{n} \left(\dot{\theta}_{(l)}\right)^\top V_l^{-1}(\hat{\theta}_{(l)} - \theta_{(l)}) = 0,$$

where $\dot{\theta}_{(l)}$ denotes the derivative with respect to β of $\theta_{(l)} = g^{-1}(Z_l^T\beta)$ and V_l is a weight matrix. As a result, the estimator $\hat{\beta}$ verifies the equation $U(\hat{\beta}) = 0$ and it has been suggested to use a sandwich estimator to estimate the variance of $\hat{\beta}$ (see for instance [4] for more details).

In the context of right-censored data, where the θ parameter is the survival function evaluated at some time point and $\hat{\theta}$ is its Kaplan-Meier estimator, it has been proved in [18] and [5] that the resulting estimating function has a mean asymptotically equal to zero. More specifically, one observes $T = \min(T^*, C)$, $\Delta = I(T^* \leq C)$, where C is a right-censoring variable and we define $X = (T, \Delta)$. Let $\theta = S(t) = \mathbb{P}(T^* > t)$, for $t \in [0, \tau]$. We further assume:

(i) $C \perp\!\!\!\perp (T^*, Z)$

(ii) $\mathbb{P}(T \geq \tau) > 0$.

We set $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$ be n i.i.d replications of (T, Δ). The authors have proved that:

$$\hat{\theta}_{(l)} = \theta + \dot{\psi}(X_l) + o_{\mathbb{P}}(1), \tag{2}$$

where $\dot{\psi}$ is a first order influence function that verifies $\mathbb{E}(\dot{\psi}(X_l) \mid Z_l) = \theta_{(l)} - \theta$. On the other hand, [5] have shown that the sandwich estimator used to estimate the variance of $\hat{\beta}$ is asymptotically biased. However, the authors have concluded that the difference between their corrected variance estimator and the usual sandwich estimator is of minor importance and as a consequence it is customary to use the sandwich estimator for pseudo-regression.

Once the pseudo-observations have been computed, implementation of the estimating equation along with the sandwich variance estimator can be easily performed from the `geese` function in the `geepack` R package.

In this article, we will present approximate formulas for computing pseudo-observations in the context of right-censoring in Section 3 and in the context of interval-censoring in Section 4. Instead of directly using Equation (1) to obtain the pseudo-observations, we will present approximated formulas that only involve the estimator $\hat{\theta}$ computed on the whole sample. In both Sections 3 and 4, we will first focus our attention on the problem of modelling $\theta_{(l)} = S(t \mid Z_l)$, the conditional survival function evaluated at time t given the covariate Z_l. The pseudo-observations can be computed using an estimator of $\theta = S(t)$ the unconditional survival function. A standard link function is $g(\cdot) = \log(-\log(\cdot))$ which gives rise to the Cox model. More complex functions can be chosen for g, such as neural-networks (see for instance [19], [6], [20]), which will provide performant prediction methods for the conditional survival function. Based on those results we will also consider the problem of modelling

$$\theta_{(l)} = \mathbb{E}(T^* \wedge \tau \mid Z_l) = \int_0^\tau S(t \mid Z_l) dt, \tag{3}$$

for some $\tau > 0$. This allows to estimate the RMST in a regression context by considering for instance the identity function for g or again more complex link functions such as neural-networks (see for instance [9]). In the context of right-censored data only, θ will be estimated based on the Kaplan-Meier estimator and in the context of interval-censored data, it will be estimated based on parametric models.

3 Approximate pseudo-observations for right-censored data

In this section, we use the same notations as in Section 2 for right-censored data. We denote by $\hat{S}$ the Kaplan-Meier estimator of S and we define for $l = 1, \ldots, n$, the l^{th} jackknife estimator $\hat{S}^{(-l)}$ of $\hat{S}$ as the estimator constructed when omitting the l^{th} observation $X_l = (T_l, \Delta_l)$. Introduce $\hat{H}(t) = \sum_i I(T_i \geq t)/n$ and the observed counting process $N_i(t) = I(T_i \leq t, \Delta_i = 1)$, where $I(\cdot)$ is the indicator function. Let

$$\hat{\Lambda}(t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\hat{H}(u)}$$

be the standard Nelson-Aalen estimator (see [21]) of the cumulative hazard function and define the martingale residuals

$$\hat{M}_l(t) = N_l(t) - \int_0^t I(T_l \geq u) d\hat{\Lambda}(u).$$

Proposition 1. *Under Assumptions (i) and (ii) in Section 2 the following results hold:*

$$\begin{aligned}\hat{S}_{(l)}(t) &= n\hat{S}(t) - (n-1)\hat{S}^{(-l)}(t) \\ &= \hat{S}(t) - \hat{S}(t)\int_0^t \frac{d\hat{M}_l(u)}{\hat{H}(u)} + o_{\mathbb{P}}(1),\end{aligned}$$

and

$$\int_0^\tau \hat{S}_{(l)}(t)dt = \int_0^\tau \hat{S}(t)dt - \int_0^\tau \int_u^\tau \hat{S}(t)dt\frac{d\hat{M}_l(u)}{\hat{H}(u)} + o_{\mathbb{P}}(1)\cdot$$

From those formulas it is clear that the pseudo-observations can be approximated from quantities computed on the original sample. In other words pseudo-observations can be computed without performing the jackknife procedure. This results in a drastic reduction of the computation time of those pseudo-observations as illustrated in the simulation section. We will also see that this approximation is very accurate even for moderate sample sizes. Besides, the main interest for using this formula is for large sample sizes, in particular in a machine learning context where computing pseudo-observations is the first step of the procedure before applying algorithms such as neural networks. In those contexts, the order of the sample size is often in millions. Even though Proposition 1 is a direct consequence of the results from [18] and [5], a separate proof is provided in the Appendix section.

It should be noted that a different approach for obtaining a fast approximation of pseudo-observations has been considered, based on the infinitesimal jackknife method (see [22]). It has been implemented in the `survival` package through the `pseudo` function. Two versions of the infinitesimal jackknife have been implemented, depending on the method used to estimate the survival function. The `pseudo` function takes as input a survival function that has either been estimated using the Kaplan-Meier estimator or using the Breslow estimator (the exponential of minus the Nelson-Aalen estimator). Both give very similar formulas as the one proposed in this paper. In particular, all three formulas are asymptotically equivalent. Details about the infinitesimal jackknife and its implementation in the `survival` package are given in the Supporting Information.

4 Approximate pseudo-observations for interval-censored data

In this section, we suppose that instead of directly observing T^* we observe a random interval $[L, R]$, $L \geq 0$ and $L \leq R$, which almost surely contains the event time: $\mathbb{P}(T^* \in [L, R]) = 1$. The right end interval is allowed to take the infinite value such that:

- if $0 < L < R < \infty$ the data are strictly interval-censored,
- if $0 = L < R < \infty$ the data are left-censored,
- if $0 < L < R = \infty$ the data are right-censored,
- if $0 < L = R < \infty$ the data are exactly observed.

Using the notations of Section 2 the data then consist of i.i.d. replications $X_i = (L_i, R_i)$, $i = 1, \ldots, n$. This situation is often called interval-censoring case 2 (see [23]) when exact observations are not allowed and mixed interval censoring (see [24]) or partly interval censoring (see [25]) otherwise. In order to derive consistent estimators of the survival function under interval censoring one will usually assume independent censoring in the following way (see for instance [26]): $\mathbb{P}(T^* \leq t \mid L = l, R = r) = \mathbb{P}(T^* \leq t \mid l \leq T^* \leq r)$. This supposes that the variables (L, R) do not convey additional information on the law of T^* apart from assuming T^* to be bracketed by L and R.

4.1 Comments on the nonparametric case

It seems appealing to use the same methodology for interval-censored data as in Section 3. A natural nonparametric estimator in the context of interval-censored data is the Turnbull estimator which can be seen as an EM estimator and is consequently rather slow to compute. The gain for avoiding computing n times the Turnbull estimator would therefore be highly significant.

However, in [23] and [27] it has been showed that this nonparametric maximum likelihood estimator converges at the $n^{1/3}$ or $(n\log(n))^{1/3}$ rates. Therefore it will not be possible to derive a relation of the following type:

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}(X_i) + o_{\mathbb{P}}(n^{-1/2}),$$

with ψ verifying $\mathbb{E}(\dot{\psi}(X_l) \mid Z_l) = \theta_{(l)} - \theta$ such as derived in [5] and [18]. Because if that would be the case, the convergence rate of $\hat{\theta}$ would be of the order $n^{1/2}$ due to the central limit theorem. However, this result is crucial to derive Equation (2) and assess the validity of the procedure.

An alternative could be to use results from [25] where it is further assumed that n_1/n tends to a positive constant as n tends to infinity, where n_1 is the number of exact observations. Under this assumption, the authors retrieved a $n^{1/2}$ rate of convergence for the nonparametric maximum likelihood estimator which converges toward a centred gaussian process. However, the covariance function of this process is not explicit and can only be determined as the solution of two integrals. This is caused by the construction of the nonparametric estimator that has no closed form but verifies a self-consistency equation. The asymptotic distribution of the nonparametric estimator was derived using results from infinite dimensional M-estimators in [28]. The same properties in M-estimators could be used here to derive approximated formulas for the nonparametric survival estimator. However, a careful examination of the proofs in [25] shows that such formulas would lead again to implicit expressions of the pseudo-observations in the same form as the asymptotic limit of the nonparametric survival estimator. Since it does not seem possible to approach those expressions in a straightforward manner we will not pursue this idea. We will focus instead in the next section in modelling the survival function using parametric models.

4.2 Parametric modelling of the survival function

We now assume that the common density function of $T_1^*, \ldots, T_n^*$ depends on $\alpha_0 \in \Theta \subset \mathbb{R}^d$, the true model parameter of dimension d. We will denote by $f^*(t;\alpha_0)$, $\lambda(t;\alpha_0)$, $\Lambda(t;\alpha_0)$ and $S(t;\alpha_0) = \exp(-\Lambda(t;\alpha_0))$ the true density, hazard, cumulative hazard and survival functions of T^*, respectively. Instead of directly observing the variables of interest, one usually observes a sample of i.i.d. variables $X_1, \ldots, X_n$ which are assumed to have a common density $f(t;\alpha_0)$. We will use the notations $\nabla \log f(t;\alpha_0)$ and $\nabla^2 \log f(t;\alpha_0)$ to represent the score vector and the Hessian matrix of this log-density where the derivatives are taken with respect to the model parameter α and are evaluated at $\alpha = \alpha_0$. The same notations will be used for $f^*(t;\alpha_0)$. It is important to emphasise the distinction between the notation f^*, which represents the density of the true data, and the notation f which represents the density of the observed data. As an illustration, in the general context of mixed interval-censored data, $X_i = (L_i, R_i)$ with $0 \leq L_i < R_i \leq \infty$ and we have (see [14] or [15])

$$f(X_i;\alpha) = (S(L_i;\alpha) - S(R_i;\alpha))I(L_i \neq R_i) + (\lambda(L_i;\alpha)S(L_i;\alpha))\, I(L_i = R_i),$$

with the slight abuse of notation $S(R_i;\alpha) = 0$ if $R_i = \infty$. In the following, we will consider maximum likelihood estimation for the parameter α_0 based on the observed variables $X_1, \ldots, X_n$.

The results derived in this section are not limited to the case of interval-censored data and can be applied to any parametric framework for incomplete data.

The maximum likelihood estimator $\hat{\alpha}$ of α_0 maximises with respect to α the log-likelihood $\sum_i \log f(X_i;\alpha)$ and, subject to regularity conditions, verifies the following equality:

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = \frac{1}{\sqrt{n}}\left(-\frac{1}{n}\sum_{i=1}^{n}\nabla^2 \log f(X_i;\tilde{\alpha})\right)^{-1}\sum_{i=1}^{n}\nabla \log f(X_i;\alpha_0),$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. Let $I = -\mathbb{E}\left(\nabla^2 \log f(X;\alpha_0)\right)$ be the Fisher information and consider the jackknife version $\hat{\alpha}^{(-l)}$ of the maximum likelihood estimator. It is then straightforward to write:

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = \frac{1}{\sqrt{n}}\, I^{-1}\sum_{i=1}^{n}\nabla \log f(X_i;\alpha_0) + \varepsilon_n,$$

$$\sqrt{n-1}(\hat{\alpha}^{(-l)}-\alpha_0) = \frac{1}{\sqrt{n-1}}\, I^{-1}\sum_{i\neq l}^{n}\nabla \log f(X_i;\alpha_0) + \varepsilon_n^{(-l)},$$

where

$$\varepsilon_n = \frac{1}{\sqrt{n}}\left(\left(-\frac{1}{n}\sum_{i=1}^{n}\nabla^2 \log f(X_i;\tilde{\alpha})\right)^{-1} - I^{-1}\right)\sum_{i=1}^{n}\nabla \log f(X_i;\alpha_0),$$

and $\varepsilon_n^{(-l)}$ is the jackknife version of ε_n. As a result, the l^{th} pseudo-observation of $\hat{\alpha}$ verifies the relation:

$$n\hat{\alpha} - (n-1)\hat{\alpha}^{(-l)} = \alpha_0 + I^{-1}\nabla \log f(X_l;\alpha_0) + \sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}. \tag{4}$$

In the Appendix section, it is proved that the term $\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}$ tends towards 0 in probability as n tends to infinity. This entails that asymptotically, the pseudo-observation of $\hat{\alpha}$ only depends on the true parameter, the Fisher information and the score vector, this latter quantity being only evaluated at the observation l. Since $\hat{\alpha}$ is a consistent estimator of α_0 and $\hat{I} = -\sum_{i=1}^{n}\nabla^2 \log f(X_i;\hat{\alpha})/n$ is a consistent estimator of I, a natural asymptotic approximation for the pseudo-observation of $\hat{\alpha}$ is simply:

$$\hat{\alpha} + \hat{I}^{-1}\nabla \log f(X_l;\hat{\alpha}).$$

While this result is interesting on its own, more work needs to be done in order to derive the pseudo-observations of $S(t;\hat{\alpha})$. The following proposition is derived based on this latter expression of the approximate pseudo-observation for $\hat{\alpha}$. The notation $\cdot^\top$ is used to denote the transpose of a vector or a matrix.

Proposition 2. *Under standard regularity conditions for maximum likelihood theory, the following relations hold:*

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\hat{\alpha}) - S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})^\top \hat{I}^{-1}\nabla \log f(X_l;\hat{\alpha}) + o_{\mathbb{P}}(1),$$

and for $\tau > 0$,

$$\begin{aligned} &n\int_0^\tau S(t;\hat{\alpha})dt - (n-1)\int_0^\tau S(t;\hat{\alpha}^{(-l)})dt \\ &\quad = \int_0^\tau S(t;\hat{\alpha})dt - \int_0^\tau S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})^\top dt\, \hat{I}^{-1}\nabla \log f(X_l;\hat{\alpha}) + o_{\mathbb{P}}(1). \end{aligned}$$

The same results also hold when replacing $\hat{\alpha}$, $\hat{I}$ by α_0, I in the right-hand side of the equations.

The proof of this proposition can be found in the Appendix section. As in Section 3, the main interest in this result lies in the fact that the approximated version of the pseudo-observation only depends on the parameter estimator $\hat{\alpha}$ and not on its jackknife version. This means that pseudo-observations in parametric models can be obtained without actually computing the n jackknife estimators. Only the estimator of α_0, along the Hessian matrix, the gradient of Λ and of the log-density are needed. This is particularly interesting in the context of interval-censored data since parametric estimators cannot be derived explicitly and numeric methods must be implemented. Two different strategies exist for those types of data: either a direct maximisation of the likelihood can be performed using the Newton-Raphson algorithm (see [13] and [14] for instance) or the complete likelihood (based on the unobserved true times) can be used through the EM algorithm in order to maximise the likelihood (see [15]). But in either case the method is iterative. Also, it should be noted that the Newton-Raphson algorithm requires to compute the score vector and Hessian matrix. Therefore the computational cost for implementing the approximated pseudo-observations is similar to the cost of simply computing the pointwise estimate from the Newton-Raphson algorithm.

Those approximated formulas are general and work for any parametric model. As an illustration, the piecewise-constant hazard (pch) model will be used in the simulation section. This model assumes that the hazard function verifies $\lambda(t;\alpha) = \sum_{k=1}^{K} \alpha_k I_k(t)$ where $I_k(t) = I(c_{k-1} < t \leq c_k)$, $c_0 = 0 < c_1 < \cdots < c_K = +\infty$ represent $K+1$ cuts and $I(\cdot)$ denotes the indicator function. We do not specify precisely the regularity conditions for maximum likelihood theory to hold. However, two important assumptions are first to assume the model identifiable and second to impose the Fisher information to be positive definite in a neighbourhood of the true parameter. For the pch model in the context of interval-censored data, two necessary conditions for those regularity assumptions to hold are:

$$\begin{aligned} \mathbb{P}(R < +\infty, [L,R] \cap (c_{k-1}, c_k] \neq \emptyset) > 0, \forall k = 1, \ldots, K, \\ \mathbb{P}(L > c_{k-1}) > 0, \forall k = 1, \ldots, K. \end{aligned} \tag{5}$$

The first assumption is quite natural: in order to estimate α_k, the probability that an interval intersects $[c_{k-1}, c_k]$ should be positive. The second assumption is necessary for the existence of a maximum of the likelihood function. It should be noted that those conditions are also valid when exact observations $L = R$ are allowed. Exact expressions of the score vector and Hessian matrix for the pch model along with the derivation of condition (5) are detailed in Section 9.4 of the Appendix. Details on the implementation of Proposition 2 for the pch model are given in Section 9.3 of the Appendix.

Precision and computational cost of the approximation for the RMST are evaluated and compared to the actual jackknife version of the pseudo-observations in the simulation section. In particular, it is seen that the approximation is much faster than the jackknife method and is very accurate even for small sample sizes.

4.3 Theoretical validation of pseudo-observations for parametric models

In this section we want to investigate if the approximated formula derived in Proposition 2 provides valid observations for performing pseudo-regression, similarly to the Kaplan-Meier estimator in the context of right-censored data (see Equation (2)). In other words, if we set

$$\varphi(X_l;\alpha_0) = \int_0^\tau S(t;\alpha_0)dt - \int_0^\tau S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top dt\, I^{-1}\nabla \log f(X_l;\alpha_0), \tag{6}$$

we want to investigate under which conditions we may have

$$\mathbb{E}(\varphi(X_l;\alpha_0) \mid Z_l) = \mathbb{E}(T_l^* \wedge \tau \mid Z_l). \tag{7}$$

It is easily seen that this equality will generally not hold by considering the simple scenario of exact observations. In that case, $X_i = T_i^*$ and $f = f^*$ is simply the density of the true variable T^*. If we further assume for instance that T^* follows a Weibull distribution, with shape parameter $a > 0$ and scale parameter $b > 0$ (the true parameters are noted a_0, b_0) such that:

$$f(X;\alpha) = f(T^*;\alpha) = \frac{a}{b}\left(\frac{T^*}{b}\right)^{a-1}\exp\left(-\left(\frac{T^*}{b}\right)^{a}\right),$$

with $\alpha = (a \;\; b)^\top$, then $\nabla \log f(X_l;\alpha)$ will depend on $\log(T_l^*)$, $(T_l^*)^{a-1}$ and $(T^*)^a$, when $a \neq 1$. As a result, $\mathbb{E}(\varphi(X_l;\alpha_0) \mid Z_l)$ will be a function of $\mathbb{E}(\log(T_l^*) \mid Z_l)$, $\mathbb{E}((T_l^*)^{a_0-1} \mid Z_l)$ and $\mathbb{E}((T_l^*)^{a_0} \mid Z_l)$ when $a_0 \neq 1$. When $a_0 = 1$ (the exponential model), then $\mathbb{E}(\varphi(X_l;\alpha_0) \mid Z_l)$ will be a function of $\mathbb{E}(T_l^* \mid Z_l)$, but it will still not verify Equation (7) unless $\tau = \infty$. Performing the same calculation for other distributions, we can similarly conclude that Equation (7) will not hold in general.

Nevertheless, even though Equality (7) is not verified in most cases, it is still possible to prove that the formula given by Proposition 2 provides a good approximation. The key is to assume that there exists a value α_z of the parameter α such that the conditional distribution of T^* given $Z = z$ follows a distribution with density $f^*(t;\alpha_z)$. In that case, $\mathbb{E}(T_l^* \wedge \tau \mid Z_l = z) = \int_0^\tau S(t;\alpha_z)dt$ and by expanding $S(t;\alpha_z)$ around $S(t;\alpha_0)$ from a Taylor development, we can prove that $\mathbb{E}(T_l^* \wedge \tau \mid Z_l = z)$ is equal to $\mathbb{E}(\varphi(X_l;\alpha_0) \mid Z_l = z)$ up to two remainder terms. Those remainder terms measure the distance between α_0 and α_z, and between the inverse of the Fisher information I^{-1} and the quantity $(-\mathbb{E}(\nabla^2 \log f(X;\alpha) \mid Z = z))^{-1}$ for any α that is on the real line between α_0 and α_z. Define $\Theta_z = \{\alpha \in \Theta : \|\alpha - \alpha_0\| \leq \|\alpha_0 - \alpha_z\|\}$ which represents the set of parameters that are on the real line between α_0 and α_z. In the next proposition, we denote for $k = 1, \ldots, K$, by α_k, $\alpha_{0,k}$, $\alpha_{z,k}$ the kth component of α, α_0 and α_z, respectively.

Proposition 3. *Assume there exists α_z such that the conditional distribution of T^* given $Z = z$ follows a distribution with density $f^*(t;\alpha_z)$. Assume also there exists $M_z < +\infty$ such that*

$$\forall k, k' \in \{1, \ldots, d\}, \forall \alpha \in \Theta_z, \frac{\partial^2}{\partial\alpha_k \partial\alpha_{k'}} \int_0^\tau S(t;\alpha)dt \leq M_z.$$

Then,

$$\mathbb{E}(T_l^* \wedge \tau \mid Z_l = z) = \mathbb{E}(\varphi(X_l;\alpha_0) \mid Z_l = z) + R_{1,z} + R_{2,z},$$

where

$$R_{1,z} \leq \frac{M_z}{2}\left(\sum_{k=1}^{d}(\alpha_{z,k} - \alpha_{0,k})\right)^2,$$
$$R_{2,z} \leq \max_{\alpha\in\Theta_z}\left|\int_0^\tau (\nabla S(t;\alpha_0))^\top dt \left(I_{\alpha,z}^{-1} - I^{-1}\right)\mathbb{E}(\nabla \log(f(X;\alpha_0)) \mid Z = z))\right|,$$

with

$$I_{\alpha,z} = -\mathbb{E}(\nabla^2 \log(f(X;\alpha)) \mid Z = z).$$

The proposition makes the strong assumption that the conditional distribution of T^* given $Z = z$ follows a distribution with density $f^*(t;\alpha_z)$. While this will not be true in general, it seems reasonable to assume that, if the chosen parametric distribution for T^* is rich enough, there will exist a value of the parameter α such that the parametric distribution is not too far from the conditional distribution of T^* given $Z = z$. This advocates for flexible parametric models such as the pch model or a spline approach such as proposed in [17]. However, this needs

to be imposed for all possible values of Z, which again seems reasonable if discrete covariates are considered and the number of those covariates is not too large.

Besides, it is difficult to evaluate, from the proposition, how large the remainder terms are. Previous experiments on the Weibull distribution in [16] using the jackknife approach suggest that the approximation is quite accurate in practice. When using the pch model, we can establish a different type of theoretical result. In the context or right-censored data, we show in the next proposition that if the number of cuts in the pch model tends to infinity, then we exactly retrieve Equality (7). We were not able to prove this result in the context of interval-censored data, we conjecture however that this result still holds in this case.

Proposition 4. *Under the context of right-censored data, if T^* follows the pch model with cuts $c_0 = 0 < c_1 < \cdots < c_K = +\infty$ and if we assume standard regularity conditions for maximum likelihood theory then the function $\varphi(X_l; \alpha_0)$ defined in Equation (6) converges, as K tends to infinity and $max_{|c_{k+1}-c_k|}$ tends to 0, towards a function $\varphi^\infty(X_l; \alpha_0)$ that verifies*

$$\mathbb{E}(\varphi^\infty(X_l; \alpha_0) \mid Z_l) = \mathbb{E}(T_l^* \wedge \tau \mid Z_l).$$

This result is interesting as it shows that it is theoretically possible for Equality (7) to hold true when using the pch model. Of course, in practice one has to choose a finite number of cuts. However, there are some strategies to choose the number of cuts from the data. In particular, in [15] the authors have developed a penalised method based on the adaptive-ridge to choose the number of cuts in an efficient way. In the simulation study (Section 5.2), we show that the approximation formula or the original jackknife method provide very similar and very performant results in pseudo-regression, when modelling the distribution of T^* with the pch model. The proofs of Propositions 3 and 4 are deferred to the Appendix section.

5 Simulation studies for the Restricted Mean Survival Time

We study two different simulation scenarios for the RMST: one with right-censored data and another one with interval-censored data. In the first scenario, the approximate pseudo observations are based on the Kaplan-Meier estimator (using Proposition 1) while in the second scenario they are based on the pch model (using Proposition 2). In both settings, the performance of the estimators derived from the approximated formulas and the ones obtained from the standard jackknife method is compared based on 500 replications. Implementation of the generalised estimation equation is performed through the `geese` function in the `geepack` R package. Computation times of the estimators were evaluated on 100 replications, from 10 different samples with 10 replications on each sample using the `microbenchmark` R package.

5.1 Right-censored data

The simulation setting is based on the one in [29]. We assume that

$$T_i^* = \tilde{\beta}_0^\top Z_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

with $\tilde{\beta}_0 = (5.5, 0.25, 0.25)^\top$, $Z_i = (1, Z_{i,1}, Z_{i,2})^\top$, $Z_{i,1}$ and $Z_{i,2}$ are Bernoulli variables with parameter 0.5 and $\varepsilon_i \sim \mathcal{U}[-\sigma, \sigma]$, with $\sigma = 3$. Under this model it can easily be seen that

$$\mathbb{E}(T_i^* \wedge \tau \mid Z_i) = \beta_{00} + \beta_{01} Z_{i,1}(1 - Z_{i,2}) + \beta_{10} Z_{i,2}(1 - Z_{i,1}) + \beta_{11} Z_{i,1} Z_{i,2}, \tag{8}$$

where $\beta_0 = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})^\top$ can be determined computationally using Monte-Carlo samples with size 10 million. We further set $\tau = 6$ which corresponds to the 54.2% quantile of T^* and to the value $\beta_0 = (4.98, 0.14, 0.14, 0.27)^\top$. Right-censored data were simulated from an

exponential distribution with parameter 0.07 yielding 33% of censoring on average. The results are presented in Table 1.

It is seen that the approximated formula gives similar results as compared to the standard jackknife method for $n = 100$. For larger sample sizes, the results are almost identical. We also compared the difference between the two estimators of β_0 by looking at the standard deviation for all four components and taking the maximum: the maximum value over all four components is equal to 7.06×10^{-3}, 5.90×10^{-4}, 2.16×10^{-4}, 6.88×10^{-6} for $n = 100$, $n = 500$, $n = 1,000$, $n = 10,000$ respectively. This shows that there is very little variations between the estimator computed from the jackknife and the one computed from the approximated formula. In terms of computation times, there is a clear advantage for the approximated formula which goes 14.3, 27.5, 25.1 and 18.7 times faster for $n = 100$, $n = 500$, $n = 1,000$ and $n = 10,000$ respectively. Clearly the computation time for the original jackknife method is not a linear function of the sample size and the gain for using the approximated method is considerable for large sample sizes. It should be noted that the computation time was evaluated for the pseudo-regression procedure, but it does not include the computation of the initial survival estimator, it only takes into account the computation of the pseudo-observations along with the implementation of the generalised estimating equations. Finally, the infinitesimal jackknife implemented in the `pseudo` function of the `survival` package was also briefly compared to our approach. It seems that the `pseudo` function provides very similar results but with a faster computational time. However, since there is no available information on the implementation of the `pseudo` function, it is not possible to clearly evaluate which of the two formulas (the one obtained with the infinitesimal jackknife and the one obtained from the Von Mises formula) has the smallest computational complexity.

	Jackknife				Approximated formula			
n	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
100	-0.006	0.273	0.075	0.211 s	-0.004	0.269	0.073	0.015 s
	-0.020	0.368	0.136		-0.022	0.361	0.131	
	-0.003	0.368	0.135		-0.006	0.361	0.130	
	-0.009	0.364	0.132		-0.013	0.357	0.128	
500	0.007	0.115	0.013	1.490 s	0.007	0.115	0.013	0.054 s
	-0.005	0.152	0.023		-0.005	0.151	0.023	
	-0.010	0.157	0.025		-0.010	0.156	0.024	
	-0.003	0.154	0.024		-0.004	0.153	0.024	
1,000	0.002	0.081	0.007	4.084 s	0.002	0.081	0.007	0.163 s
	-0.003	0.110	0.012		-0.003	0.110	0.012	
	-0.004	0.113	0.013		-0.004	0.112	0.013	
	-0.003	0.107	0.011		-0.003	0.107	0.011	
10,000	0.002	0.026	0.001	4.429 min	0.002	0.026	0.001	14.194 s
	-0.001	0.037	0.001		-0.001	0.037	0.001	
	0.001	0.036	0.001		0.001	0.036	0.001	
	-0.001	0.034	0.001		-0.001	0.034	0.001	

Table 1: Simulation results for the estimation of β in the RMST model (8) based on pseudo-regression with the Kaplan-Meier estimator on 33% of right-censored data. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates.

5.2 Interval-censored data

For interval censored data the survival function is estimated from the pch model, as detailed in Section 4.2. Using this model, estimation of the model parameter α_0 is performed using the

EM algorithm, as presented in [15]. An alternative method could be to directly maximise the observed likelihood but this would result in implementing the Newton-Raphson algorithm for each jackknife sample with inversion of a Hessian matrix of full rank which, in turn, would result in unstable results. In the EM algorithm, the M-step is explicit and as a result the computation of the jackknife methods is always stable. We refer the reader to [15] for more details on the two methods. The approximated method is implemented from the result in Proposition 2 and details on the computation of the score vector and Hessian matrix are detailed in Section 9.3 of the Appendix.

We assume Model (8) with the same values of σ and τ. Then, in order to simulate interval-censored data, a total of $K = 5$ visits were simulated such that $V_1 \sim \mathcal{U}[0, 6]$ and $V_k = V_{k-1} + U[0, 2]$, for $k = 2, \ldots, K$. The observations for which $T_i^* < V_1$ correspond to left-censored observations with $L_i = 0$ and $R_i = V_1$, the observations for which $T_i^* > V_K$ correspond to right-censored observations with $L_i = V_K$ and $R_i = \infty$, and the observations for which $V_{k-1} < T_i^* < V_k$ $(k = 2, \ldots, K)$ correspond to strictly interval-censored observations with $L_i = V_{k-1}$ and $R_i = V_k$. This resulted in 14.6% of left-censored data, 52.07% of interval-censored data and 33.33% of right-censored data. For interval-censored data, the average length of the intervals was approximately equal to 1.34. The pch model with cuts equal to $4, 5, 6, 7$ was used for the computation of the survival estimator. The pseudo-observations were generated based on the standard jackknife and on the approximated formulas and the results for the RMST model are presented in Table 2.

Again the results between the jackknife and the approximate formula are almost identical while there is a huge gain in terms of computational time for the approximated formula. The approximated formula is 107, 198 and 310 times faster than the jackknife method for $n = 200$, $n = 500$ and $n = 1,000$ respectively. It should be noted that the cuts must be carefully chosen in the pch model. In particular, the regularity conditions of Equation (5) must be satisfied. If there are only few values of L_i and R_i that intersect a cut $[c_{k-1}, c_k]$ or if the proportion of L_i's such that $L_i > c_{k-1}$ is too low then the pseudo-values can be incorrect (both for the jackknife method or using our approximated formula) which will in turn result in a poor performance of the parameters estimation. On the other hand, if the regularity conditions hold, the choice of the cuts will only have a minor impact on the performance of the estimator of β_0 and will lead to similar results. A supplementary simulation scenario for interval-censored data with τ equal to infinity is also presented in the Supporting Information.

6 Illustrative real data examples

6.1 The Cardiovascular Health Study (CHS)

In this data example, we mimic the analysis of the Cardiovascular Health Study (CHS) as it was performed in [9]. This study was initiated in 1987 to determine the risk factors for development and progression of cardiovascular disease (CVD) in older adults. The event of interest was time to CVD. In [9], the author considers a subsample of $5,380$ individuals of whom 65.2% had CVD during the study period and the others were right-censored. The aim of the study was to estimate the conditional RMST with 29 covariates and $\tau = 5$ years.

The methodology proposed in [9] uses pseudo-observations and implements a deep neural network directly on the pseudo-observations of the RMST, that is the g link function presented in Section 2 is a neural network. Moreover, a training dataset including 75% of the observations and a test set based on the remaining 25% of the data are built in order to evaluate the prediction performance of the method. This split of the data between training and test sets is repeated 10 times. At each repetition, the pseudo-observations must be entirely computed but only on the training datasets. This results in computing the pseudo-observations for the RMST for 10 samples of size $4,035$. We computed those pseudo observations from the jackknife method and

	Jackknife				Approximated formula			
n	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
200	-0.169	0.220	0.077	42.047s	-0.167	0.219	0.076	0.392s
	0.019	0.310	0.096		0.018	0.308	0.095	
	0.015	0.305	0.094		0.013	0.304	0.093	
	0.057	0.293	0.089		0.055	0.293	0.089	
500	-0.181	0.141	0.053	2.408 min	-0.181	0.140	0.052	0.731s
	0.036	0.190	0.037		0.036	0.190	0.037	
	0.035	0.191	0.038		0.035	0.191	0.038	
	0.077	0.182	0.039		0.076	0.182	0.039	
1,000	-0.184	0.101	0.044	6.675 min	-0.183	0.101	0.044	1.292s
	0.035	0.136	0.020		0.035	0.136	0.020	
	0.033	0.142	0.021		0.032	0.142	0.021	
	0.072	0.130	0.022		0.072	0.130	0.022	

Table 2: Simulation results for the estimation of β in the RMST model (8) based on pseudo-regression with 14.6% of left-censored data, 52.07% of interval-censored data and 33.33% of right-censored data. The piecewise constant hasard model with cuts equal to $4, 5, 6, 7$ was used for the estimation of the survival function in the computation of the pseudo-observations. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates.

the approximated formula. The former was computed in approximately 21.3 seconds while the latter took 1.9 seconds. Therefore, our approximated formula is more than 11 times faster than the original jackknife method. Since building a neural network is computationally expensive and needs to be implemented for all the training samples, this reduction in computation time is a major advantage for our approximated formula. Of note, the results of the analysis implemented with the approximated formula are identical to the original analysis (based on the jackknife method) and are therefore omitted.

6.2 The Signal Tandmobiel® data

In this section, our aim is to analyse the Signal Tandmobiel® data using three different models: the standard Cox model, a logistic model for the conditional survival function and the conditional RMST model presented in Equation (3). This dataset is part of the `bayesSurv` R package. Those data were collected from a longitudinal dental survey of $4,468$ school children born in 1989, who were annually examined by a dentist. The time scale is age in years. The dataset is composed of 0.68% of left-censored data, 61.69% of strictly interval-censored data and 37.63% of right-censored data. Our aim is to study the emergence of the tooth number 14 which is a permanent first premolar. The covariates used for the analysis are: gender (binary variable equal to 1 for boys, 0 for girls) and the number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84 of the dataset. This covariate is thus discrete taking values between 0 and 4. These data were previously studied by [30] using the Accelerated Failure Time model (AFT). In our analysis, the survival function is estimated from the pch model using the whole dataset and the pseudo-observations are then computed from the approximated formula in Proposition 2. There are 126 individuals with missing covariates and the generalised estimating equation used to implement our models is therefore applied to this reduced dataset composed of $4,342$ pupils.

In the pch model, the number of cuts and locations were chosen using the adaptive-ridge algorithm developed in [15]. This led to the selection of the four cuts 7.6, 8.4, 9 and 10. Since the maximum likelihood estimator has converged this entails that the regularity conditions of Equation (5) are satisfied. We can also easily check them empirically: in particular there are 3%

of strictly interval censored observations whose left intervals fell before 7.6, 40% of left intervals that fell after 10 and the percentage of strictly interval censored observations that intersect each other is high (values not shown). The corresponding estimated hazard and survival functions are displayed in Figure 1. We observe a low estimated hazard value (equal to $6{\cdot}10^{-4}$) from age 0 until age 7.6 due the low percentage of left intervals that fell before 7.6. This yields a very flat decay of the survival function on this time period, then the decay increases drastically for the four other time periods $[7.6, 8.4]$, $[8.4, 9]$, $[9, 10]$ and $[10, \infty)$. For illustration, we estimate from the survival function that approximately 83.39% of the teeth will emerge between age 7.6 and 12.

In order to implement a Cox model from the pseudo-observations, we need to set a grid of time points on which the baseline hazard rate is estimated (see [4]). We choose the time points $t_1 = 8, t_2 = 9, t_3 = 10, t_4 = 11, t_5 = 12$ and we use the link function $g(x) = \log(-\log(x))$ in the pseudo-regression which leads to the Cox model:

$$\log(-\log(S(t_m \mid Z_i)) = \log \Lambda_0(t_m) + Z_i^\top \beta, \quad m = 1, \ldots, 5,$$

where $\Lambda_0(t_m)$ represents the cumulative baseline hazard function at time t_m. The results are presented in Table 3. The second column provides the cumulative baseline hazard at different time points and the hazard ratios of the two covariates. The last column displays the Wald tests which are all extremely significant. We clearly see that the hazard for the emergence of the tooth is an increasing function of time. Also, under the proportional hazards assumption, boys have an increased risk as compared to girls with an hazard ratio equal to 1.47 and the hazard ratio for one supplementary decayed or missing deciduous first molar due to caries equals 1.13.

For the logistic regression model, we study the probability for the emergence of the tooth before a fixed time point. We conduct two separate analysis, one for the time point 9 and another one for the time point 12 which correspond to the 11% and 84% estimated quantiles of T^*, respectively, thus corresponding to early and late emergence of the tooth. There is no guarantees that the pseudo-observations are in the interval $[0, 1]$ and we therefore set the negative values to 0 and the values greater than 1 to 1. The model is the following:

$$\text{logit}(1 - S(t \mid Z_i)) = \gamma + Z_i^\top \beta,$$

where t is either equal to 9 or 12, γ is the intercept and β is a two dimensional vector representing the effect of the covariates and logit is the classical logistic function ($\text{logit}(x) = \log(x/(1-x))$). It should be noted that the model can be directly implemented with the `glm` or `geese` functions by using one minus the pseudo-observations as input. The results are presented in Table 4. We first observe that the effect of the two covariates are highly significant but they differ depending on the time endpoint: the odds ratio for the number of decayed or missing deciduous first molars decreases from $\exp(0.2808) \approx 1.32$ before age 9 to $\exp(0.1080) \approx 1.11$ before age 12 while the odds ratio for the gender effect increases from $\exp(0.2978) \approx 1.35$ before age 9 to $\exp(0.5284) \approx 1.70$ before age 12. This strongly suggests that the number of decayed or missing deciduous first molars is mostly responsible for the early teeth emergence while gender (with boys having a higher risk) is mostly responsible for the late tooth emergence. It is also interesting to compare the probabilities of the tooth emergence before age 9 between girls with no decayed or missing deciduous first molars (5.49%), boys with no decayed or missing deciduous first molars first molars (7.26%) and boys with 4 decayed or missing deciduous first molars (19.39%). We can similarly compare the probabilities of the tooth emergence before age 12 between girls with no decayed or missing deciduous first molars (70.63%), boys with no decayed or missing deciduous first molars (80.31%) and boys with 4 decayed or missing deciduous first molars (86.27%).

Finally, two RMST analysis were conducted with $\tau = 9$ and $\tau = 12$. The estimated regression parameters in the RMST model along with their Wald test are presented in Table 5. For $\tau = 9$ we observe a weak effect of the covariates with an intercept that is almost equal to τ,

highlighting that most emergences of the tooth will occur after 9 years of age. As a matter of fact, gender is not significant and the number of decayed or missing deciduous first molars is highly significant but with a weak effect. The number of decayed or missing deciduous first molars will accelerate the emergence of the tooth with 1 decayed molar (respectively 4 decayed molars) yielding a reduction of 0.0097 years (respectively 0.0390 years) for the emergence of the tooth. For $\tau = 12$ the effect of gender is now highly significant, meaning that gender only plays a role for late emergence of the tooth (a finding that was also observed with the logistic regression model). The emergence of the tooth for boys arrives on average 0.3336 years earlier than for girls. The number of decayed or missing deciduous first molars is also highly significant with 1 decayed molar (respectively 4 decayed molars) yielding a reduction of 0.1303 years (respectively 0.5211 years) for the emergence of the tooth.

We also tried to repeat the procedure using different cut values in the pch model and as already observed in the simulation study, this led to very similar results, for all three models. The results from [30] obtained using the AFT models were similar to our findings except that the authors did not provide statistical tests for the effects of the covariates and it was not possible from their method to detect that gender had mainly a role for late emergence of the tooth.

Finally, based on the approximated formulas developed in this paper, the whole procedure (computation of the pseudo-observations and implementation of the generalised estimating equations) took about 1.78 seconds for the RMST analysis (the computation times are similar for the other two models). The method was not implemented using the classical jackknife method but according to the simulation study it would have taken 29.8 minutes to obtain the pseudo-observations, since in the simulation study the time for the jackknife procedure was evaluated at 6.7 minutes for $n = 1,000$ (see Sections 5.2). Also, the results would have been identical, thus highlighting the relevance of the proposed approach in practical situations.

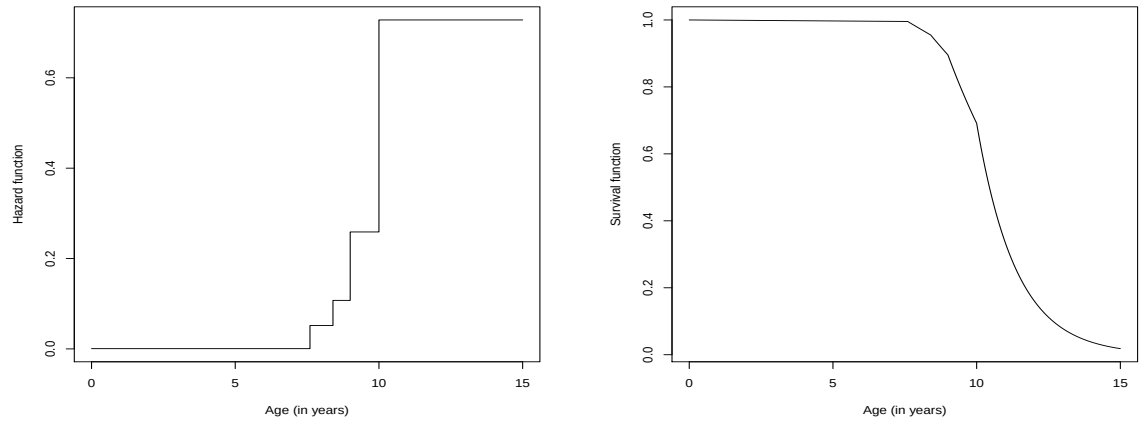

Figure 1: Distribution of time to emergence of the tooth number 14. On the left: estimated hazard function. On the right: estimated survival function. Those estimates were obtained from the pch model with cuts equal to $7.6, 8.4, 9$ and 10.

7 Conclusion

In this paper, we presented asymptotic formulas for computing pseudo-observations for time to event data. In the context of right-censored data, those formulas are based on the Kaplan-Meier estimator of the survival function. When dealing with interval-censoring our formulas were developed for a general class of parametric models. Pseudo-regression is an appealing

Covariates	effect	exp. effect	se	p-value
Intercept at time 8	-4.4510	0.0117	0.1175	$< 10^{-15}$
Intercept at time 9	-2.6999	0.0672	0.0727	$< 10^{-15}$
Intercept at time 10	-1.4461	0.2355	0.0463	$< 10^{-15}$
Intercept at time 11	-0.3206	0.7257	0.0358	$< 10^{-15}$
Intercept at time 12	0.2293	1.2577	0.0340	$< 10^{-10}$
Gender (1 = boy)	0.3885	1.4747	0.0395	$< 10^{-15}$
Nb of decayed molars	0.1249	1.1330	0.0130	$< 10^{-15}$

Table 3: Cox Model for the time to emergence of the tooth 14 with the covariates gender and number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84. The time has been discretised at $t = 8, 9, 10, 11, 12$ in the pseudo-regression approach. **se** represents the standard estimate of the regression parameter.

	time = 9			time = 12		
Covariates	**effect**	**se**	**p-value**	**effect**	**se**	**p-value**
Intercept	-2.8458	0.0761	$< 10^{-15}$	0.8777	0.0501	$< 10^{-15}$
Gender (1 = boy)	0.2978	0.0819	0.0003	0.5284	0.0599	$< 10^{-15}$
Nb of decayed molars	0.2808	0.0260	$< 10^{-15}$	0.1080	0.0194	2.7×10^{-8}

Table 4: Two logistic models for the probability that the emergence of the tooth 14 occurred before time 9 and 12 with respect to the covariates gender and number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84. **se** represents the standard estimate of the regression parameter.

	$\tau = 9$			$\tau = 12$		
Covariates	**effect**	**se**	**p-value**	**effect**	**se**	**p-value**
Intercept	8.9851	0.0047	$< 10^{-15}$	10.8755	0.0306	$< 10^{-15}$
Gender (1 = boy)	-0.0097	0.0066	0.1422	-0.3336	0.0361	$< 10^{-15}$
Nb of decayed molars	-0.0180	0.0024	8.1379×10^{-14}	-0.1303	0.0120	$< 10^{-15}$

Table 5: Restricted Mean Survival Time Model for the time to emergence of the tooth 14 with the covariates gender and number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84. Two values of τ are analysed in Equation (3). **se** represents the standard estimate of the regression parameter.

tool when the goal is to directly model a complex quantity of interest, such as the RMST, cumulative incidence functions in a competing risk setting or transition probabilities for multi-state models. Our formulas were precisely developed for the RMST but they could be easily extended for those other quantities of interest. While the pseudo-values approach is originally based on the jackknife procedure, our formulas only involve quantities computed on the initial sample. This results in a drastic reduction of the computational time, which is an interesting feature when dealing with large dataset or when the data are interval-censored, since in that case, the estimators are computationally intensive.

There has been an increasing interest of the pseudo-values approach in the machine learning community. After having computed the pseudo-observations, standard machine learning models can be applied to those new observations, by simply ignoring the censoring. In particular, this methodology has been applied for estimating the survival function in [19], [6], [20] or the RMST in [9] based on neural networks that were directly applied on the pseudo-observations. Therefore, our formulas are particularly interesting in those settings where the dataset can be extremely large and the algorithm usually relies on a cross-validation procedure. Using our approximated formulas results in a significant gain in terms of computation time as illustrated on the real

data analysis. Also, the approximations made by our formulas are extremely precise, even for moderate sample sizes, as shown in the simulation study. Surprisingly, we also saw that our formulas are more robust than the original jackknife method which sometimes fails due to some rare extreme values. For all these reasons, we advocate the use of our asymptotic formulas in practical situations.

As a reviewer pointed out, there exists an alternative to our fast formula in the case of right-censored data, in the `pseudo` function of the `survival` package. It is based on the infinitesimal jackknife and can be applied to the Kaplan-Meier estimator or to the Breslow estimator (defined as the exponential of minus the Nelson-Aalen estimator). A theoretical comparison of those approximations with the Von-Mises approach developed in this paper has been made in the Supporting Information. It is shown that all three approximations are very similar for small sample sizes and asymptotically equivalent. It would now be interesting to extend the infinitesimal jackknife to the parametric setting, in order to apply it to interval-censored data, as there already exists some approaches based on the infinitesimal jackknife to compute delta-beta residuals, which are very closely related to pseudo-observations.

Finally, new theoretical results for parametric pseudo-regression were also developed in this paper. Those results indicate that parametric pseudo-regression is only valid up to two extra terms but simulation studies suggest that they are reasonably small in practice and do not significantly impact the performance of the final estimator, as long as the number of covariates is not too large. They also suggest to use flexible parametric models such as the pch model. As an extension it would be interesting to study the theoretical validity of splines models for pseudo-regression such as the method developed in [17]. This is let to future research.

8 Software

The asymptotic formulas developed in this paper for the pseudo-values of the survival function and the RMST can be implemented using the GitHub package `FastPseudo` available at https://github.com/obouaziz/FastPseudo. The package can deal with both right-censored or interval-censored data. In the latter case, the formulas are implemented for the pch model.

Acknowledgments

We thank the reviewers for their very constructive criticisms and comments that have helped improve the paper. We also thank Per Kragh Andersen and Terry Therneau for their very interesting discussions about the connection between the infinitesimal jackknife and the Von Mises approximation for right-censored data.

9 Appendix

9.1 Proof of Proposition 1

Introduce $H_1(t) = \mathbb{P}(T \leq t, \Delta = 1)$, $H_0(t) = \mathbb{P}(T \leq t, \Delta = 0)$, $H(t) = \mathbb{P}(T \geq t)$ and their empirical counterparts, $\hat{H}_1(t) = \sum_i I(T_i \leq t, \Delta_i = 1)/n$, $\hat{H}_0(t) = \sum_i I(T_i \leq t, \Delta_i = 0)/n$, $\hat{H}(t) = \sum_i I(T_i \geq t)/n$. Let $\psi(A)(s,t] = \prod_{s<u\leq t}(1 + dA(u))$, such that $\psi(\Lambda)(0,t] = S(t)$, where $\Lambda(t)$ is the cumulative hazard function and $\psi(\hat{\Lambda})(0,t] = \hat{S}(t)$. We have the following Von-Mises expansion [see 31, 32]:

$$\hat{S}^{(-l)}(t) = \hat{S}(t) - \hat{S}(t)(\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)) + o_{\mathbb{P}}(\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)).$$

We now derive a Von-Mises expansion for $\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)$. The cumulative hazard function and its estimator can be defined as functions of H, H_1 and of $\hat{H}_1$, $\hat{H}$ respectively where $\Lambda(t) =$

$g(H_1, H) := \int_0^t dH_1(u)/H(u)$ and $\hat{\Lambda}(t) = g(\hat{H}_1, \hat{H})$. We have the following Von-Mises expansion:

$$\hat{\Lambda}^{(-l)}(t) = \hat{\Lambda}(t) + g'_{(\hat{H}_1,\hat{H})}(\hat{H}_1^{(-l)} - \hat{H}_1, \hat{H}^{(-l)} - \hat{H}) + o_{\mathbb{P}}(n^{-1}),$$

where g' is the Hadamard derivative of g, which is equal to [see 31, 32]:

$$g'_{(H_1,H)}(h_1, h) = \int_0^t \frac{dh_1}{H} - \int_0^t \frac{h_2 dH_1}{H^2}\cdot$$

The $o_{\mathbb{P}}(n^{-1})$ term above comes from the expressions:

$$\hat{H}_1^{(-l)}(t) - \hat{H}_1(t) = \frac{1}{n(n-1)} \sum_{i=1}^{n} I(T_i \leq t, \Delta_i = 1) - \frac{I(T_l \leq t, \Delta_l = 1)}{n-1}$$

and

$$\hat{H}^{(-l)}(t) - \hat{H}(t) = \frac{1}{n(n-1)} \sum_{i=1}^{n} I(T_i \geq t) - \frac{I(T_l \geq t)}{n-1},$$

which entail as a consequence that $\hat{H}_1^{(-l)}(t) - \hat{H}_1(t)$ and $\hat{H}^{(-l)}(t) - \hat{H}(t)$ are $O_{\mathbb{P}}(n^{-1})$. Moreover, using those expressions we have

$$\begin{aligned} g'_{(\hat{H}_1,\hat{H})}(\hat{H}_1^{(-l)} - \hat{H}_1, \hat{H}^{(-l)} - \hat{H}) &= \frac{1}{n-1}\int_0^t \frac{d\hat{H}_1(u)}{\hat{H}(u)} - \frac{1}{n-1}\frac{I(T_l \leq t, \Delta_l = 1)}{\hat{H}(T_l)} \\ &\quad - \frac{1}{n-1}\int_0^t \frac{d\hat{H}_1(u)}{\hat{H}(u)} + \frac{1}{n-1}\int_0^t \frac{I(T_l \geq u) d\hat{H}_1(u)}{(\hat{H}(u))^2}, \\ &= -\frac{1}{n-1}\frac{I(T_l \leq t, \Delta_l = 1)}{\hat{H}(T_l)} + \frac{1}{n-1}\int_0^t \frac{I(T_l \geq u) d\hat{H}_1(u)}{(\hat{H}(u))^2}\cdot \end{aligned}$$

Gathering all the different parts, we obtain

$$\begin{aligned} \hat{S}_{(l)}(t) &= \hat{S}(t) + \hat{S}(t)\left(\int_0^{T_l \wedge t} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} - \frac{I(T_l \leq t, \Delta_l = 1)}{\hat{H}(T_l)}\right) + o_{\mathbb{P}}(1) \\ &= \hat{S}(t) + \hat{S}(t)\left(\int_0^t \frac{I(u \leq T_l)}{\hat{H}(u)} d\hat{\Lambda}(u) - \int_0^t \frac{dN_l(u)}{\hat{H}(u)}\right) + o_{\mathbb{P}}(1) \\ &= \hat{S}(t) - \int_0^t \frac{d\hat{M}_l(u)}{\hat{H}(u)} + o_{\mathbb{P}}(1)\cdot \end{aligned}$$

The approximation for the RMST is then obtained by directly integrating the previous equation as it actually follows that the convergence holds in the Skorohod space $D[0, \tau]$ [see 31, 32] and therefore the convergence holds uniformly with respect to $t \in [0, \tau]$.

9.2 Proof of Proposition 2

Starting with Equation (4) we will first prove that $\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}$ tends to 0 in probability as n tends to infinity. Set

$$\tilde{I}_n = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2 \log f(X_i; \tilde{\alpha}), \quad \tilde{I}_n^{(-l)} = -\frac{1}{n-1}\sum_{i \neq l} \nabla^2 \log f(X_i; \tilde{\alpha}),$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. We have:

$$\begin{aligned}
&\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)} \\
&\quad = \left(\tilde{I}_n^{-1} - I^{-1}\right)\sum_{i=1}^{n}\nabla\log f(X_i;\alpha_0) - \left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\sum_{i\neq l}^{n}\nabla\log f(X_i;\alpha_0) \\
&\quad = \left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\nabla\log f(X_l;\alpha_0) + \left(\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1}\right)\sum_{i=1}^{n}\nabla\log f(X_i;\alpha_0).
\end{aligned}$$

Clearly for I positive definite, $\left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\nabla\log f(X_l;\alpha_0)$ tends to 0 in probability. Since $\sum_{i=1}^{n}\nabla\log f(X_i;\alpha_0)/n$ tends to $\mathbb{E}(\nabla\log f(X;\alpha_0)) = 0$ in probability, we just need to prove that $\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1} = O_{\mathbb{P}}(1/n)$ to conclude the proof. Write:

$$\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1} = \tilde{I}_n^{-1}(\tilde{I}_n^{(-l)} - \tilde{I}_n)\left(\tilde{I}_n^{(-l)}\right)^{-1}.$$

From the law of large numbers, $\tilde{I}_n^{-1}$ and $\left(\tilde{I}_n^{(-l)}\right)^{-1}$ tend towards I^{-1} in probability and

$$\tilde{I}_n^{(-l)} - \tilde{I}_n = -\frac{1}{n(n-1)}\sum_{i\neq l}\nabla^2\log f(X_i;\tilde{\alpha}) + \frac{1}{n}\nabla^2\log f(X_l;\tilde{\alpha}) = O_{\mathbb{P}}\left(1/n\right).$$

This proves that

$$n\hat{\alpha} - (n-1)\hat{\alpha}^{(-l)} = \alpha_0 + I^{-1}\nabla\log f(X_l;\alpha_0) + o_{\mathbb{P}}(1). \tag{9}$$

Using the consistency of $\Lambda(t;\hat{\alpha})$ towards $\Lambda(t;\alpha_0)$ from standard maximum likelihood theory, we now write a Taylor expansion for the cumulative hazard function around α_0:

$$\Lambda(t;\hat{\alpha}) = \Lambda(t;\alpha_0) + (\hat{\alpha}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0) + \frac{1}{2}(\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0), \tag{10}$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. We also write a Taylor expansion for the function $x\mapsto\exp(-x)$ around 0:

$$\begin{aligned}
\exp(-(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0))) = 1 - \big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) + \frac{1}{2}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 \\
- \frac{1}{6}e^{\xi_n}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^3,
\end{aligned}$$

with ξ_n tends to 0 in probability as n tends to infinity. This can be rewritten as:

$$S(t;\hat{\alpha}) = S(t;\alpha_0) + S(t;\alpha_0)\left(-\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) + \frac{1}{2}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2\right) + o_{\mathbb{P}}(1/n),$$

using the fact that $\sqrt{n}(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0))$ converges in distribution towards a centred gaussian variable with finite variance from standard results on maximum likelihood theory and the delta-method. As a result,

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\alpha_0) + A_{n,1} + A_{n,2} + o_{\mathbb{P}}(1), \tag{11}$$

where

$$\begin{aligned}
A_{n,1} &= -S(t;\alpha_0)\left(n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)\right), \\
A_{n,2} &= S(t;\alpha_0)\left(n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)^2\right)\frac{1}{2}.
\end{aligned}$$

We start with the $A_{n,2}$ term. From Equation (10) we have:

$$\begin{aligned} n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 &= n(\hat{\alpha}-\alpha_0)^\top \nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0) \\ &\quad + n(\hat{\alpha}-\alpha_0)^\top \nabla\Lambda(t;\alpha_0)(\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0) \\ &\quad + \frac{n}{4}\Big((\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0)\Big)^2. \end{aligned}$$

Using the consistency of $\hat{\alpha}-\alpha_0$ and the asymptotic normality of $\sqrt{n}(\hat{\alpha}-\alpha_0)$ from standard maximum likelihood theory, each of the last two terms in the above equation tends to 0 in probability as n tends to infinity. Therefore

$$\begin{aligned} &n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)^2 \\ &\quad = (\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(n(\hat{\alpha}-\alpha_0) - (n-1)(\hat{\alpha}^{(-l)}-\alpha_0)) \\ &\qquad + n(\hat{\alpha}-\hat{\alpha}^{(-l)})^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0) + o_{\mathbb{P}}(1) \\ &\quad = (\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(I^{-1}\nabla\log f(X_l;\alpha_0) + R_n) \\ &\qquad + (\alpha_0 - \hat{\alpha}^{(-l)} + I^{-1}\nabla\log f(X_l;\alpha_0) + R_n')^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0) + o_{\mathbb{P}}(1), \end{aligned}$$

where the last two lines were derived from Equation (9) and R_n, R_n' both tend to 0 in probability. The consistency of $\hat{\alpha}$ and $\hat{\alpha}^{(-l)}$ shows that $A_{n,2} = o_{\mathbb{P}}(1)$. We now study the term $A_{n,1}$. From Equation (10),

$$\begin{aligned} &n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big) \\ &\quad = (n(\hat{\alpha}-\alpha_0) - (n-1)(\hat{\alpha}^{(-l)}-\alpha_0))^\top\nabla\Lambda(t;\alpha_0) \\ &\qquad + \frac{1}{2}(\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(n(\hat{\alpha}-\alpha_0) - (n-1)(\hat{\alpha}^{(-l)}-\alpha_0)) \\ &\qquad + \frac{1}{2}n(\hat{\alpha}-\hat{\alpha}^{(-l)})^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0). \end{aligned}$$

Using similar arguments as before, the last two lines of this Equation tend to 0 in probability from Equation (9) and from the consistency of $\hat{\alpha}$ and $\hat{\alpha}^{(-l)}$. Finally, using again Equation (9)

$$A_{n,1} = -S(t;\alpha_0)\nabla\log f(X_l;\alpha_0)^\top I^{-1}\nabla\Lambda(t;\alpha_0) + o_{\mathbb{P}}(1).$$

This equality combined with Equation (11) give

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\alpha_0) - S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top I^{-1}\nabla\log f(X_l;\alpha_0) + o_{\mathbb{P}}(1).$$

The final result of Proposition 2 is obtained by simply replacing each quantity by its consistent estimator. Integrating the equation in Proposition 2 directly yields the approximation for the RMST. By careful examination of the remainder term, we directly see that its integral over $[0,\tau]$ is also $o_{\mathbb{P}}(1)$.

9.3 Log-likelihood, score vector and Hessian matrix in the piecewise constant hazard model

In this section, we study the parametric piecewise constant hazard model defined as follows: $\lambda(t;\alpha) = \sum_{k=1}^K \alpha_k I_k(t)$ where $I_k(t) = I(c_{k-1} < t \leq c_k)$, $c_0 = 0 < c_1 < \cdots < c_K = +\infty$. The cumulative hazard function is then equal to

$$\Lambda(t;\alpha) = \sum_{k=1}^K \alpha_k(c_k \wedge t - c_{k-1})I(c_{k-1} \leq t).$$

Under the mixed-case of interval-censored and exact data, we can directly write the log-likelihood as the sum between the log-likelihood of strictly interval-censored observations and the log-likelihood of exact observations. For the latter part see [33]. Recall that $X_i = (L_i, R_i)$ and $f(X_i; \alpha)$ denotes the density of the observations with parameter α evaluated at X_i. For strictly interval-censored data ($L_i \neq R_i$), the log-likelihood $\ell(\alpha)$ can be written as (see [14] or [15])

$$\ell(\alpha) = \sum_{i=1}^{n} \log f(X_i; \alpha) = \sum_{i=1}^{n} \Big\{ -(1-\Delta_i)\Lambda(L_i;\alpha) + \Delta_i \Big(\log\Big(1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big) - \Lambda(L_i;\alpha)\Big) \Big\},$$

where we used the notation $\Delta_i = I(R_i < +\infty)$ to denote uncensored observations. The k^{th} component of the score vector is equal to:

$$\frac{\partial \ell(\alpha)}{\partial \alpha_k} = \sum_{i=1}^{n} \frac{\partial \log f(X_i;\alpha)}{\partial \alpha_k} = \sum_{i=1}^{n} \Big\{ -(c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i) + \Delta_i \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)} \times \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big) \Big\}. \tag{12}$$

The $k \times k'$ component of the Hessian matrix is equal to:

$$\begin{aligned} \frac{\partial^2 \ell(\alpha)}{\partial \alpha_{k'} \partial \alpha_k} = -\sum_{i=1}^{n} \Delta_i \Bigg\{ & \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)} \\ & \times \big((c'_k \wedge R_i - c_{k'-1}) I(c_{k'-1} \leq R_i) - (c'_k \wedge L_i - c_{k'-1}) I(c_{k'-1} \leq L_i) \big) \\ & \times \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big) \\ & + \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{\Big(1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big)^2} \\ & \times \big((c'_k \wedge R_i - c_{k'-1}) I(c_{k'-1} \leq R_i) - (c'_k \wedge L_i - c_{k'-1}) I(c_{k'-1} \leq L_i) \big) \\ & \times \exp\Big(2\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big) \Bigg\}. \end{aligned} \tag{13}$$

The Fisher information is equal to the expectation of minus the Hessian matrix divided by n. Looking at its expression, we directly see that a necessary condition for the Fisher information to be positive definite is to assume that

$$\mathbb{P}(\Delta = 1, [L, R] \cap (c_{k-1}, c_k] \neq \emptyset) > 0, \forall k = 1, \ldots, K.$$

Another important condition for the model to be identifiable is to assume that $\mathbb{E}_{\alpha_0}[f(X;\alpha)]$ has a unique maximum with respect to α, equal to α_0, where the notation $\mathbb{E}_{\alpha_0}$ means the expectation is taken with respect to the true parameter α_0. However, it is clear from Equation (12) that $\mathbb{E}_{\alpha_0}[\partial f(X;\alpha)/\partial\alpha]$ cannot vanish if $\mathbb{P}(L > c_{k-1}) = 0$. Therefore, a second necessary condition for the model to be identifiable is to assume

$$\mathbb{P}(L > c_{k-1}) > 0, \forall k = 1, \ldots, K.$$

Those two conditions have opposite effects on the estimation method if they are violated. In case the first one is not valid for a given k then it will not be possible to compute the corresponding

estimator $\hat{\alpha}_k$ from the Newton-Raphson algorithm (since the Hessian will not be invertible) while using the EM algorithm (which does not involve the Score vector nor the Hessian matrix), the estimator $\hat{\alpha}_k$ will become smaller at each iteration step until eventually reaching the value 0. This situation can be numerically resolved in the latter case, by simply setting the iterated estimate $\hat{\alpha}_k$ to 0 when it reaches a value below a fixed threshold. However this situation is problematic for the computation of the pseudo-values. This can be easily seen by recalling that pseudo-values should average to the initial estimator. In Proposition 2 this simply follows from the fact that $\sum_{l=1}^{n} \nabla \log f(X_l; \hat{\alpha}) = 0$ from regularity conditions for maximum likelihood estimation. However, the k^{th} component of the score vector will never vanish if the first condition is not valid, leading to incorrect pseudo-values.

On the other hand, if the second condition is not valid for a given k, the algorithm will attempt to minimise the term $\exp(-\Lambda(R_i; \alpha))$ from Equation (12) and as a consequence the corresponding estimator $\hat{\alpha}_k$ will become larger at each iteration step of the EM algorithm, diverging to infinity.

Finally, note that if the log-likelihood only include exact observations $L = R$, the conditions then translate to $\mathbb{P}(c_{k-1} < L < c_k) > 0, \forall k = 1, \ldots, K$.

9.4 Implementation of the pseudo-observations for the survival function and the RMST in the pch model

In this section we provide the precise expression of the terms involved in Proposition 2 for the pch model. We have

$$S(t; \alpha) = \exp\Big(- \sum_{k=1}^{K} \alpha_k (t \wedge c_k - c_{k-1}) I(c_{k-1} \leq t) \Big)$$
$$\frac{\partial \Lambda(t; \alpha)}{\partial \alpha_k} = (c_k \wedge t - c_{k-1}) I(c_{k-1} \leq t),$$

while the expression of the gradient of the density $\nabla \log f(X_l; \alpha)$ is given by the term between brackets in Equation (12) and $\hat{I}$ is equal to minus the Hessian matrix (see Equation (13)) divided by n.

For the integrated version we need to precise how to compute the integral between 0 and τ of $S(t; \alpha)$ and the integral between 0 and τ of $S(t; \alpha)\nabla\Lambda(t; \alpha)$. We first notice that

$$S(t; \alpha) = \exp\Big(- \sum_{k=1}^{K} \alpha_k (c_k - c_{k-1}) I(c_k \leq t) \Big) \exp\Big(- \sum_{k=1}^{K} \alpha_k (t - c_{k-1}) I(c_{k-1} \leq t \leq c_k) \Big),$$

and

$$\begin{aligned}
\int_0^\tau S(t; \alpha) dt &= \sum_{l=1}^{K} \int_{c_{l-1}}^{c_l \wedge \tau} S(t; \alpha) dt \; I(\tau > c_{l-1}) \\
&= \sum_{l=1}^{K} \int_{c_{l-1}}^{c_l \wedge \tau} \exp\Big(- \sum_{k=1}^{K} \alpha_k (c_k - c_{k-1}) I(c_k \leq t) \Big) \exp\Big(- \alpha_l (t - c_{l-1}) \Big) dt \, I(\tau > c_{l-1}).
\end{aligned}$$

Set $A_1 = 0$ and for $l \geq 2$, define

$$A_l = - \sum_{k=1}^{l-1} \alpha_k (c_k - c_{k-1}) + c_{l-1} \alpha_l.$$

For the first term we now have:

$$\begin{aligned}\int_0^\tau S(t;\alpha)dt &= \sum_{l=1}^K \exp(A_l)\int_{c_{l-1}}^{c_l\wedge\tau}\exp(-\alpha_l t)dt\, I(\tau > c_{l-1})\\ &= \sum_{l=1}^K \exp(A_l)\alpha_l^{-1}\Big(\exp(-\alpha_l c_{l-1}) - \exp\big(-\alpha_l(c_l\wedge\tau)\big)\Big) I(\tau > c_{l-1}).\end{aligned}$$

For the second term we have:

$$\begin{aligned}\int_0^\tau S(t;\alpha)\frac{\partial\Lambda(t;\alpha)}{\partial\alpha_k}dt &= \sum_{l=1}^K \int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)(c_k\wedge t - c_{k-1})I(c_{k-1}\le t)dt\, I(\tau > c_{l-1})\\ &= \sum_{l=k+1}^K \int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)dt\,(c_k - c_{k-1})I(\tau > c_{l-1})\\ &\quad + \int_{c_{k-1}}^{c_k\wedge\tau} tS(t;\alpha)dt\, I(\tau > c_{k-1}) - c_{k-1}\int_{c_{k-1}}^{c_k\wedge\tau} S(t;\alpha)dt\, I(\tau > c_{k-1}).\end{aligned}$$

From the previous calculation on the first term, we easily see that on the one hand

$$\int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)dtI(\tau > c_{l-1}) = \exp(A_l)\alpha_l^{-1}\Big(\exp(-\alpha_l c_{l-1}) - \exp\big(-\alpha_l(c_l\wedge\tau)\big)\Big) I(\tau > c_{l-1}).$$

On the other hand, we have:

$$\begin{aligned}&\int_{c_{l-1}}^{c_l\wedge\tau} tS(t;\alpha)dtI(\tau > c_{l-1})\\ &= \exp(A_l)\int_{c_{l-1}}^{c_l\wedge\tau} t\exp(-t\alpha_l)dtI(\tau > c_{l-1})\\ &= \exp(A_l)\Big(\alpha_l^{-2}\Big(\exp\big(-c_{l-1}\alpha_l\big) - \exp\big(-(c_l\wedge\tau)\alpha_l\big)\Big)\\ &\quad + \alpha_l^{-1}\Big(c_{l-1}\exp\big(-c_{l-1}\alpha_l\big) - (c_l\wedge\tau)\exp\big(-(c_l\wedge\tau)\alpha_l\big)\Big)\Big) I(\tau > c_{l-1}),\end{aligned}$$

where the last equation was obtained using integration by parts. Gathering all elements allows to implement the second equation in Proposition 2.

9.5 Proof of Proposition 3

First, from a Taylor expansion of $\int_0^\tau S(t;\alpha_z)dt$ around $\int_0^\tau S(t;\alpha_0)dt$, we obtain:

$$\begin{aligned}\mathbb{E}(T_l^*\wedge\tau \mid Z_l = z) &= \int_0^\tau S(t;\alpha_z)dt\\ &= \int_0^\tau S(t;\alpha_0)dt + \int_0^\tau (\nabla S(t;\alpha_0))^\top dt(\alpha_z - \alpha_0)\\ &\quad + R_{1,z},\end{aligned}$$

with

$$R_{1,z} = \frac{1}{2}(\alpha_z - \alpha_0)^\top \int_0^\tau \nabla^2 S(t;\tilde{\alpha}_z)dt(\alpha_z - \alpha_0)$$

and $\tilde{\alpha}_z$ is on the real line between α_0 and α_z. Then, since α_z maximises with respect to α the expected log-likelihood $\mathbb{E}(\log f(X_l;\alpha) \mid Z_l = z))$ we have that $\mathbb{E}(\nabla \log f(X_l;\alpha_z) \mid Z_l = z)) = 0$. Then, from a Taylor expansion around α_0 we obtain:

$$0 = \mathbb{E}(\nabla \log f(X_l;\alpha_0) \mid Z_l = z)) + (\alpha_z - \alpha_0)^\top \mathbb{E}(\nabla^2 \log f(X_l;\tilde{\tilde{\alpha}}_z) \mid Z_l = z)),$$

where $\tilde{\tilde{\alpha}}_z$ is on the real line between α_0 and α_z. As a result, we have:

$$\begin{aligned}\alpha_z - \alpha_0 = I^{-1}\mathbb{E}(\nabla \log f(X_l;\alpha_0) \mid Z_l = z) \\ + \left\{(-\mathbb{E}(\nabla^2 \log f(X_l;\tilde{\tilde{\alpha}}_z) \mid Z_l = z)))^{-1} - I^{-1}\right\}\mathbb{E}(\nabla \log f(X_l;\alpha_0) \mid Z_l = z).\end{aligned}$$

Gathering all parts, we have proved that

$$\begin{aligned}\mathbb{E}(T_l^* \wedge \tau \mid Z_l = z) = \int_0^\tau S(t;\alpha_0)dt + \int_0^\tau (\nabla S(t;\alpha_0))^\top dt I^{-1}\mathbb{E}(\nabla \log f(X_l;\alpha_0) \mid Z_l = z) \\ + R_{1,z} + R_{2,z},\end{aligned}$$

with

$$R_{2,z} = \int_0^\tau (\nabla S(t;\alpha_0))^\top dt \left\{(-\mathbb{E}(\nabla^2 \log f(X_l;\tilde{\tilde{\alpha}}_z) \mid Z_l = z)))^{-1} - I^{-1}\right\}\mathbb{E}(\nabla \log f(X_l;\alpha_0) \mid Z_l = z).$$

Finally, writing $S(t;\alpha_0) = \exp(-\Lambda(t;\alpha_0))$, we directly obtain

$$\int_0^\tau \nabla S(t;\alpha_0)dt = -\int_0^\tau S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)dt,$$

which concludes the proof.

9.6 Proof of Proposition 4

Let $X_i = (T_i, \Delta_i)$ and assume the pch model for λ. In this model, the cumulative hazard function evaluated at T_i, is equal to

$$\Lambda(T_i;\alpha) = \sum_{k=1}^K \alpha_k(c_k \wedge T_i - c_{k-1})I(c_{k-1} \leq T_i).$$

However, since we want to prove the result when the mesh of the partition $0 = c_0 < c_1 < \cdots < c_K = +\infty$ tends to zero, we can write without loss of generality that there exists a $\delta > 0$, such that for any partition whose mesh is less than δ we have for all T_i, $i = 1, \ldots, n$,

$$\Lambda(T_i;\alpha) = \sum_{k=1}^K \alpha_k(c_k - c_{k-1})I(c_k \leq T_i).$$

Therefore, for a partition $0 = c_0 < c_1 < \cdots < c_K = +\infty$ such that $\max_k |c_k - c_{k-1}| < \delta$,

$$\begin{aligned}\log f(X_i;\alpha) &= \Delta_i \sum_{k=1}^K \log(\alpha_k) I_k(T_i) - \sum_{k=1}^K \alpha_k(c_k - c_{k-1})I(c_k \leq T_i), \\ \frac{\partial}{\partial\alpha_k}\log f(X_i;\alpha) &= \frac{\Delta_i}{\alpha_k} I_k(T_i) - (c_k - c_{k-1})I(c_k \leq T_i), \\ \frac{\partial^2}{\partial\alpha_k^2}\log f(X_i;\alpha) &= -\frac{\Delta_i}{\alpha_k^2} I_k(T_i), \\ \frac{\partial^2}{\partial\alpha_k'\partial\alpha_k}\log f(X_i;\alpha) &= 0, \text{ for } k \neq k'.\end{aligned}$$

We therefore have that the Fisher information I is a diagonal matrix whose kth element, $k = 1, \ldots, K$, is equal to $\mathbb{E}(\Delta_i I_k(T_i))/(\alpha_k^0)^2$, where α_k^0 is the kth component of α_0. Also under standard maximum likelihood regularity conditions, the true parameter α_0 verifies that

$$\mathbb{E}\left(\frac{\partial}{\partial \alpha_k} \log f(X_i; \alpha_k^0)\right) = 0,$$

which is equivalent to

$$\alpha_k^0 = \frac{\mathbb{E}(\Delta_i I_k(T_i))}{(c_k - c_{k-1})\mathbb{P}(T_i \geq c_k)}.$$

Now, we have that $\mathbb{E}(\Delta_l I_k(T_l))/(c_k - c_{k-1})$ tends to $H_1'(c_{k-1})$ as $(c_k - c_{k-1})$ tends to 0, where $H_1(t) = \mathbb{P}(T \leq t, \Delta = 1)$. Therefore, as the limit $(c_k - c_{k-1})$ goes to 0, α_k^0 tends to $\lambda(c_k)$, the true hazard rate evaluated at c_k. In other words, the parametric hazard $\lambda(t; \alpha_0)$ tends to the true hazard function $\lambda(t)$ as $(c_k - c_{k-1})$ goes to 0.

Then, $\nabla\Lambda(t; \alpha_0)$ is a vector whose kth component is equal to $(c_k - c_{k-1})I(c_k \leq t)$. As a consequence,

$$\begin{aligned}
&\nabla\Lambda(t; \alpha_0)^\top I^{-1} \nabla \log f(X_l; \alpha_0) \\
&= \sum_{k=1}^{K} I(c_k \leq t) \frac{(\alpha_k^0)^2 (c_k - c_{k-1})}{\mathbb{E}(\Delta_l I_k(T_l))} \left(\frac{\Delta_l}{\alpha_k^0} I_k(T_l) - (c_k - c_{k-1}) I(c_k \leq T_l)\right).
\end{aligned}$$

In the first term of this equation, there can only be one interval of the form $[c_{k-1}, c_k]$ that contains T_l and therefore the sum over k is not zero for only one value of k. As the limit of $(c_k - c_{k-1})$ goes to 0, T_l gets closer to c_k and c_{k-1} and α_k^0 tends to $\lambda(T_l)$. More precisely,

$$\lim_{c_k - c_{k-1} \to 0} \sum_{k=1}^{K} I(c_k \leq t) \frac{(\alpha_k^0)^2 (c_k - c_{k-1})}{\mathbb{E}(\Delta_l I_k(T_l))} \frac{\Delta_l}{\alpha_k^0} I_k(T_l) = \frac{\Delta_l I(T_l \leq t) \lambda(T_l)}{H_1'(T_l)}.$$

On the other hand, the second term of the equation is simply a Riemann integral. We have:

$$\lim_{c_k - c_{k-1} \to 0} \sum_{k=1}^{K} I(c_k \leq t) \frac{(\alpha_k^0)^2 (c_k - c_{k-1})^2}{\mathbb{E}(\Delta_l I_k(T_l))} I(c_k \leq T_l) = \int_0^{T_l \wedge t} \frac{(\lambda(u))^2}{H_1'(u)} du.$$

Noticing that $\lambda(t) = H_1'(t)/H(t)$, with $H(t) = \mathbb{P}(T \geq t)$, we therefore have proved

$$\lim_{c_k - c_{k-1} \to 0} \nabla\Lambda(t; \alpha_0)^\top I^{-1} \nabla \log f(X_l; \alpha_0) = \frac{\Delta_l I(T_l \leq t)}{H(T_l)} - \int_0^{T_l \wedge t} \frac{dH_1(u)}{(H(u))^2}.$$

We now want to compute the conditional expectation with respect to Z_l of this quantity. Let $H_1^Z(t) = \mathbb{P}(T \leq t, \Delta = 1 \mid Z)$, $H^Z(t) = \mathbb{P}(T \geq t \mid Z)$ and write

$$\begin{aligned}
\mathbb{E}\left(\frac{\Delta_l I(T_l \leq t)}{H(T_l)} - \int_0^{T_l \wedge t} \frac{dH_1(u)}{(H(u))^2} \mid Z_l\right) &= \int_0^t \frac{dH_1^{Z_l}(u)}{H(u)} - \int_0^t \frac{H^{Z_l}(u) dH_1(u)}{(H(u))^2} \\
&= \int_0^t \left(\frac{(H_1^{Z_l}(u))'}{H^{Z_l}(u)} - \frac{(H_1(u))'}{H(u)}\right) \frac{H^{Z_l}(u)}{H(u)} du \\
&= \int_0^t (\lambda^{Z_l}(u) - \lambda(u)) \frac{H^{Z_l}(u)}{H(u)} du,
\end{aligned}$$

where λ^{Z_l} represents the conditional hazard function. Setting $q_{Z_l}(u) = (\lambda^{Z_l}(u) - \lambda(u)) H^{Z_l}(u)/H(u)$ as in [5] we obtain from their proofs in Section A.2, Equation (11), that $-S^{Z_l}/S$ is a primitive of q_{Z_l}, where S^{Z_l} is the conditional survival function of T^*. As a consequence,

$$\int_0^t (\lambda^{Z_l}(u) - \lambda(u)) \frac{dH^{Z_l}(u)}{H(u)} = 1 - \frac{S^{Z_l}(t)}{S(t)}.$$

Gathering all the parts, we have proved that

$$\begin{aligned}\lim_{c_k - c_{k-1} \to 0} \mathbb{E}\left(\varphi(X_l; \alpha_0) \mid Z_l\right) &= \int_0^\tau S(t)dt - \int_0^\tau S(t)\left(1 - \frac{S^{Z_l}(t)}{S(t)}\right)dt \\ &= \int_0^\tau S^{Z_l}(t)dt = \mathbb{E}(T_l \wedge \tau \mid Z_l),\end{aligned}$$

which concludes the proof.

References

[1] David M Zucker. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709, 1998.

[2] Pei-Yun Chen and Anastasios A Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

[3] Min Zhang and Douglas E Schaubel. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*, 67(3):740–749, 2011.

[4] Per Kragh Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

[5] Martin Jacobsen and Torben Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, 2016.

[6] Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics*, 24(11):3308–3314, 2020.

[7] Michael C Sachs, Andrea Discacciati, Åsa Everhov, Ola Olén, and Erin E Gabriel. Ensemble prediction of time to event outcomes with competing risks: A case study of surgical complications in crohn's disease. *arXiv preprint arXiv:1902.02533*, 2019.

[8] Pablo Gonzalez Ginestet, Philippe Weitz, Mattias Rantalainen, and Erin E Gabriel. A deep cnn approach for predicting cumulative incidence based on pseudo-observations. 2021.

[9] Lili Zhao. Deep neural networks for predicting restricted mean survival times. *Bioinformatics*, 2021.

[10] Bruce W Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

[11] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science and Business Media, 1992.

[12] Jian Huang and Jon A Wellner. Efficient estimation for the proportional hazards model with" case 2" interval censoring. Technical report, 1995.

[13] JK Lindsey. A study of interval censoring in parametric regression models. *Lifetime data analysis*, 4(4):329–354, 1998.

[14] Jianguo Sun. *The statistical analysis of interval-censored failure time data*. Springer Science and Business Media, 2007.

[15] Olivier Bouaziz, Eva Lauridsen, and Grégory Nuel. Regression modelling of interval-censored data based on the adaptive-ridge procedure. *Journal of Applied Statistics*, pages 1–25, In press.

[16] Camille Sabathe, Per K Andersen, Catherine Helmer, Thomas A Gerds, Hélène Jacqmin-Gadda, and Pierre Joly. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical methods in medical research*, 29(3):752–764, 2020.

[17] Martin Nygård Johansen, Søren Lundbye-Christensen, and Erik Thorlund Parner. Regression models using parametric pseudo-observations. *Statistics in Medicine*, 39(22):2949–2961, 2020.

[18] Frederik Graw, Thomas A Gerds, and Martin Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, 2009.

[19] Lili Zhao and Dai Feng. Dnnsurv: Deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:1908.02337*, 2019.

[20] Dai Feng and Lili Zhao. Bdnnsurv: Bayesian deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:2101.03170*, 2021.

[21] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[22] Louis A Jaeckel. *The infinitesimal jackknife*. Bell Telephone Laboratories, 1972.

[23] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

[24] Qiqing Yu, Linxiong Li, and George YC Wong. On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, 27(1):35–44, 2000.

[25] Jian Huang. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, pages 501–519, 1999.

[26] Zhigang Zhang, Liuquan Sun, Xingqiu Zhao, and Jianguo Sun. Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33(1):61–70, 2005.

[27] Jon A Wellner. Interval censoring, case 2: alternative hypotheses. *Lecture Notes-Monograph Series*, pages 271–291, 1995.

[28] AW Van der Vaart. Efficiency. of infinite dimensional m-estimators. *Statistica Neerlandica*, 49(1):9–30, 1995.

[29] Xin Wang and Douglas E Schaubel. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime data analysis*, 24(1):176–199, 2018.

[30] Emmanuel Lesaffre, Arnošt Komárek, and Dominique Declerck. An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, 14(6):539–552, 2005.

[31] Richard D Gill. Lectures on survival analysis. In *Lectures on Probability Theory*, pages 115–241. Springer, 1994.

[32] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[33] O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer, 2008.