CAUSAL REASONING WITH SURVIVAL DATA

PÅL CHRISTIE RYALEN

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-484-9

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

1. Acknowledgements

I want to thank my main supervisor Kjetil Røysland for all the guidance and support I have received from him. I would also like to thank my co-supervisor Mats J. Stensrud for all the helpful feedback and reinforcement he has provided. I am particularly delighted with the cooperation we three have had the last three years.

I would also like to thank all my co-authors. In addition to Kjetil and Mats, I am thankful for the work and insight Sophie D. Fosså provided for **Paper 4**.

I want to thank the Research Council of Norway for funding this project through the research grant "NFR239956/F20 - Analyzing clinical health registries: Improved software and mathematics of identifiability."

I would also like to thank my colleagues at the Department of Biostatistics for creating such an enjoyable and exciting working environment.

Lastly, I would like to thank all the people who have encouraged me during this period of my life.

2. List of papers

Paper 1: Ryalen P.C., Stensrud M.J., Røysland K. "Transforming cumulative hazard estimates", *Biometrika*, vol. 105, no. 4, pp. 905-916, 2018

Paper 2: Stensrud M.J., Røysland K., Ryalen P.C. "On null hypothesis in survival analysis", re-submitted in *Biometrics*

Paper 3: Ryalen P.C., Stensrud M.J., Røysland K. "The additive hazard estimator is consistent for continuous time marginal structural models", accepted for publication in *Lifetime data analysis*

Paper 4: Ryalen P.C., Stensrud M.J., Fosså S., Røysland K. "Causal inference in continuous time: An application on prostate cancer therapy", *Biostatistics*, 2018

Contents

3. Preface

In this thesis, I summarize and discuss four research articles. The text is organized as follows: Section 4 provides a brief introduction to the main problems we have studied and an overview of the four articles. In Section 5 and 6, a short description of some key topics in survival analysis and causal inference is provided before the research articles are described in detail in Section 7. Section 8 contains some of the main contributions of each article, comments to put the findings in a broader context, and possible directions for future work. In Section 9, one can find worked examples that show how to use the software I have developed. The four papers are shown consecutively in Section 10.

4. Introduction and overview

Statisticians often face the problem of making inference on parameters that describe a population of interest. However, the parameters that are studied do not always have a straightforward and transparent interpretation, even if the statistical model fits the data well. This is unfortunate because statistical analyses are often used to aid decision making. Notions of causal effects usually guide decisions. When defining parameters that describe causal relationships, it is useful to specify them as functions of hypothetical manipulations of some exposure. However, such causal parameters are often hard to identify. Randomized controlled trials (RCTs) have been the gold standard for testing causal relationships. In real life, RCTs are often unavailable, and one has to analyze observational data, where spurious associations may be present. The field of causal inference has offered a framework for causal reasoning when observational data is at hand. Causal inference techniques are often used to analyze observational studies. The causal inference methods are, however, by no means restricted to the analysis of observational data; the tools from causal inference should be more widely applied throughout statistics, especially when issues of interpretability are important, also in RCTs.

Problems with interpreting the hazard function. Hazards and hazard ratios are commonly reported measures of association in survival analysis [1, 2]. The use of hazard ratios, usually estimated by Cox's proportional hazard (PH) model, has been appealing. Statistical analyzes are often summarized by numerical estimates of hazard ratios along with confidence intervals, and hazard ratios have traditionally been thought to have easily understood interpretations [3, 4]. It has been pointed out, however, that hazards can be difficult to interpret, even in the absence of unmeasured confounding and model misspecification [5, 6, 7]. A problem arises, since the hazard function at time t by definition is conditioned on survival up to t. This conditioning may induce selection bias, as it unblocks non-causal pathways from the exposure to future survival through any unobserved heterogeneity. The interpretation will therefore be complicated in many situations [5, 6, 8]. Hazard ratios are often the only statistical summary that is reported in epidemiological studies [5], so the fact that they can be difficult to interpret may have unfortunate consequences. Statistical hypotheses in survival analysis are often stated on the hazard scale, e.g. the rank tests. In some cases, it is difficult to understand such null hypotheses thoroughly.

Marginal structural models. Marginal structural models (MSMs) provide a tool for performing causal analysis of longitudinal data [9]. These models are particularly useful for problems with time-dependent confounding, i.e. when a process L is present that is affected by the exposure a patient has already received, while also influencing future exposure and the outcome of interest. The MSMs can be fitted by

weighting observational data according to the inverse probability of treatment weights (IPTWs). When the IPTWs are known, one can use weighting to obtain estimates with causal interpretations [9]. The MSMs were initially developed as discrete time models [9]. More recently continuous-time MSMs were introduced, building on theory from stochastic analysis such as Girsanov's theorem, as a continuous-time analog of the discrete models [10]. Similarly to the discrete models, causal estimates can be obtained by weighting observational data according to a weight process [10]. The continuous-time MSMs fit well with survival analysis, as they are defined in continuous time.

Objective. In this thesis, I point out that hazards are sometimes difficult to interpret and discuss alternative methods for analyzing survival data. A general method for estimation and inference for other survival analysis parameters is presented, with a possibility for covariate adjustment. Furthermore, I describe how the continuous-time MSMs has been further developed both concerning methodology and software. An example that demonstrates the continuous-time MSMs' practical feasibility is presented, through an application on a substantive prostate cancer treatment problem using registry data.

Paper 1. We develop a general method that uses hazard models as a stepping stone for modeling a class of other parameters in survival analysis. Examples of such parameters include the survival function, cumulative incidence functions, and the restricted mean survival function. We utilize the fact that these parameters belong to a class of functions that solve systems of ordinary differential equations (ODEs) driven by the integrated, or cumulative, hazard. We suggest a plugin estimator that solves naturally associated stochastic differential equation (SDE) systems driven by cumulative hazard estimates. We show that this SDE estimator is consistent, and write down an estimator for the pointwise covariance. Some asymptotic results have previously been found by focusing on one parameter at a time using the functional delta-method. We demonstrate the results for our class of parameters using stability theory for differential equations [11]. Our focus in on cumulative hazard estimators that are counting process integrals, as is the case for the additive hazard estimator. It is then possible to adjust for covariates using additive hazard regression.

Paper 2. We point out that there can be issues with the interpretation of the rank tests while arguing that hypothesis testing should be performed on parameters that have clear interpretations. Using results from **Paper 1**, we derive a general nonparametric test statistic for hypothesis testing pointwise in time. We use simulations to compare the power functions of our test statistics with conventional nonparametric

methods tailor-made for each parameter, and show that the performance is comparable. We also show using simulations that the rank tests are often unsuitable for testing our hypotheses. Hence, one should not use rank tests for testing e.g. five-year survival. Finally, we use our method to compare two adjuvant chemotherapy regimes among patients with stage III colon cancer. We thereby demonstrate how to use the methodology in practice.

Paper 3. We further develop the continuous-time MSMs. An estimator for the continuous-time weights is proposed, based on additive hazard regression. We show that causal cumulative hazards can be estimated consistently by weighting additive hazard regressions with our weight estimator. Furthermore, we show that the class of ODE parameters studied in **Paper 1** can be estimated consistently, by defining an SDE estimator driven by the weighted cumulative hazard estimates. The ODE parameters can thus be given causal interpretations if a sufficient set of (time-dependent) confounders are measured, and an additive hazard model correctly specifies their impact on the exposure hazard. We provide a simulated example showing how the methodology works in practice.

Paper 4. We investigate the problem of comparing the treatment regimens Radical Prostatectomy (RP) and Radiation (Rad), from the time of treatment assignment to death/treatment failure. We utilize the results in **Paper 1** and **Paper 3** on Norwegian prostate cancer registry data and use weighting to estimate causal cumulative incidences. While the conclusion of a naive analysis could be that the Rad treated individuals have a significantly better prognosis than the RP treated individuals, we find that the treatment regimens perform similarly when accounting for the competing risk of death. This is a clinically relevant finding, as it suggests that a large group of diagnosed patients should worry less about prognosis, and more about side effects when choosing treatment type.

Software. I have developed `R` software for all of the above papers, freely available for anyone to use at the GitHub repository `github.com/palryalen`. The package `transform.hazards` has been developed together with **Paper 1** and **Paper 2**. The main function takes cumulative hazard estimates, initial values, as well as the integrand function F with Jacobian matrices $\nabla F_1, \nabla F_2, \cdots$ (see Section 7.1 for details) as input, and gives parameter and covariance estimates as output. The methods for estimating the continuous-time weights, used in **Paper 3** and **Paper 4**, are implemented in the `ahw` package. The weight estimator assumes that the time to exposure follows an additive hazard model, estimated using the `aalen` function from the `timereg` package. Worked examples of both packages can be found in Section 9, and in the `transform.hazards` package vignette. Detailed examples are also available in the

four papers, particularly in the Supplementary material of **Paper 2**, as well as in the main text of **Paper 3**, and **Paper 4**.

5. Survival analysis

Survival analysis is the subfield of statistics that deals with modeling the time to events of interest, often death. Statistical methods in survival analysis are created to handle censoring, which arises when some of the subjects are unobserved in parts of the study period. Using survival analysis techniques, one can thus include subjects that later moves out of the study population, which is a real problem in many follow-up studies. Survival analysis has thus become an essential tool for most health and social scientists; the seminal papers by Kaplan and Meier [12] and Cox [1] remain two of the most cited statistics papers in all of science.

5.1. **Concerns with the interpretation of hazards.**

The hazard function is a fundamental quantity in survival analysis. It may be understood as the rate of deaths as a function of time conditional on survival up to that time, and a standard textbook definition [2] is

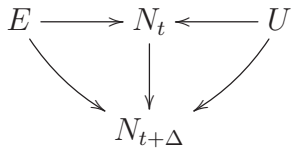
$$\lim_{\Delta \longrightarrow 0+} \frac{P(t \leq T < t + \Delta | t \leq T)}{\Delta},$$

where T is the event time of interest.

The most frequently used method for modeling the hazard is Cox's semi-parametric proportional hazards model [1]. The PH assumption is that covariates act multiplicatively on the hazard, or equivalently that the ratio between covariate-adjusted hazards is constant as a function of time. Some researchers have pointed out that this assumption often is unrealistic, in particular when adjusting for many covariates [13]. Still, the PH assumption is convenient as the statistical analysis can then be succinctly summarized by numerical estimates of hazard ratios with confidence intervals.

Hazards have traditionally been thought to have appealing interpretations [3,4]. In particular, the hazard ratio is often loosely interpreted as a relative risk [3]. However, these interpretations tend to neglect the fact that the hazard at each time point t is defined conditional on survival (i.e. $t \leq T$), and is therefore based on a specific subpopulation. It is not clear to what extent the hazard describes the population as a whole.

A problem arises if an unobserved heterogeneity, often called frailty, affects the survival times in the population under study; the frail subjects tend to die early, while the less frail subjects tend to live longer. The subpopulation still alive will thus consist of a diminishing proportion of frail individuals as time progresses, and will therefore not be representative of the population as a whole. By drawing a causal graph, one can see that the frailty opens non-causal pathways from the exposure to the outcome. Such a situation is depicted below, where U is a frailty variable:



Conditioning on survival N_t at time t opens the non-causal pathway $E \to N_t \leftarrow U \leftarrow N_{t+\Delta}$ between the exposure E and future survival $N_{t+\Delta}$ at time $t+\Delta$, i.e. a collider in conditioned upon in the graph. Hazards will then be subject to this "collider bias," and they can therefore in general not be understood as causal effect measures in the presence of a frailty U. Since frailty is present in many follow-up studies, also when the treatment is randomized, a collider bias may often be an issue. Several researchers have discussed this problem in the causal inference literature [5, 6, 8].

The problem induced by conditioning on survival is distinct from concerns regarding confounding, measured or otherwise. Instead, it stems from an intrinsic feature of the hazard function itself. Using the causal graphs, one can formally and clearly describe why frailty complicates the interpretation of hazards, and get clues on how to avoid the collider bias.

5.2. **Targeting other survival parameters.** A possible workaround for the collider bias mentioned in Section 5.1 is to study other survival analysis parameters. Many of the widely used measures of association in survival and event history analysis, such as the survival function, the cumulative incidence function, and the restricted mean survival function, do not condition on survival and are therefore easier to interpret than hazards. These parameters may also be more clinically relevant than hazards in some situations (see e.g. **Paper 2**; the colon cancer example in Section 6 and the discussion in Section 8).

However, hazard functions are often used to obtain other parameters. The survival function, for instance, is the exponential of minus the cumulative hazard. One can fit a Cox model for the hazard, possibly conditional on covariates, and use the estimated cumulative hazard for calculating survival curves. However, this type of survival curve modeling is restrictive, as the covariate-adjusted survival curves are not allowed to cross, which is unrealistic in many situations. Some have advocated models for the restricted mean survival time [14], and the attributable fraction function [15]. In particular, modeling procedures based on Cox regression is proposed. Such modeling is convenient for researchers whose exposure to regression in survival analysis is limited to the Cox model. Still, when targeting parameters different from the hazard (ratio), the restrictions imposed by the PH assumption seem undesirable.

We will use hazard modeling as a stepping stone for obtaining other survival parameters. A suitable candidate for this is Aalen's nonparametric additive hazard model [16]. As the name suggests, the hazard may vary in response to (time-varying) covariates Z without any restriction other than that the covariates act additively, i.e.

that the hazard is on the form

$$\alpha_t^0 + \sum_{j=1}^{p} Z_{t-}^j \alpha_t^j = Z_{t-}^\intercal \alpha_t, \tag{1}$$

for $Z = (1, Z^1, \cdots, Z^p)^\intercal$, and hazard coefficients $\alpha = (\alpha^0, \alpha^1, \cdots, \alpha^p)^\intercal$. The additivity assumption is less restrictive than the PH assumption, and the additive model is more able to capture detailed information in the data. Using martingale techniques, Aalen developed a consistent least squares estimator for the cumulative coefficients $\int_0^t \alpha_s^j ds$, $j = 0, \cdots, p$ [17], enabling estimation of cumulative hazards.

One obvious drawback with the additive hazard model is that the cumulative hazard estimates can have negative increments, indicating "negative risk" in a time interval. If the model fits the data well, this is not likely to be a problem; negative increments may just be an indication of model misspecification, at least when the sample size is large. Some researchers have deemed the cumulative coefficients hard to interpret [3]. I agree that these coefficients can be hard to understand profoundly, but this is may not be an issue if additive hazard regression is used as an intermediary step for obtaining other parameters.

Conditioning on survival does not seem to be problematic if the exposure and frailty acts additively on the hazard; if the exposure and frailty are marginally independent, they will remain independent when conditioning on survival [6]. Still, the additive model coefficient that corresponds to the exposure can often not be interpreted causally; see e.g. [18]. Nevertheless, this coefficient is only one of many quantities that can be studied, and other parameters may provide more clinically relevant information for a given situation.

Some authors have considered modeling on other scales directly, without using the hazard. In such cases, one needs additional techniques for including censored subjects. One example is the weighted approach discussed by Sun and Zhang, who focused on the mean residual life function [19]. They used inverse probability weighting for model estimation.

6. Causal inference

A traditional view in statistics is that statistical methods can deal with associations and not causal relationships. However, this attitude contradicts the actual practice of statistics in many settings, and the way statistics is perceived from the outside. In various fields, such as the social and health sciences, statistical analyses are performed not only to predict but also to obtain mechanistic understandings, i.e. to describe causal relationships. Moreover, statistical analyses are increasingly used to guide decision making. Decisions based on spurious associations could lead to choosing treatments without the desired effect, or imposing regulations that make no difference.

The causal claim "the treatment has an effect" can be assessed by performing randomized experiments such as RCTs, where treatment assignment is given at random in the population under study. However, RCTs are often expensive, unethical, or impractical to conduct. Meanwhile, there is a considerable amount of existing non-randomized, or observational, data available. Health registries, for instance, are becoming increasingly more extensive. There is therefore a clear interest in developing methods for making causal inferences when observational data is at hand.

Epidemiologists and clinicians have accumulated knowledge about different forms of confounding and selection bias that can arise in observational studies, and developed methods for identifying and estimating treatment effects. This knowledge has been progressively clarified and formalized by various key contributors (see e.g. Pearl [20], Robins et al. [9], or Rubin [21]), and modeling assumptions and issues of identifiability can now be stated and determined with mathematical rigor. A notable reference that provides a rigorous framework for causal modeling is that of Pearl [20]. There, subject matter knowledge on the observational study of interest is encoded in directed acyclic graphs (DAGs).

6.1. **DAGs.**

The graphs of Pearl consist of a collection of random variables with directional edges, that together form Bayesian networks. The edges indicate dependencies between the variables; each variable X_i in the graph is conditionally independent of its non-descendants (the variables that cannot be found by following directed edges from X_i) given its parents $\text{pa}(X_i)$ (the variables that have an edge pointing into X_i). The joint density of the variables then allows for a recursive factorization on the form

$$P(x_1, \cdots, x_k) = \prod_{j=1}^{k} P(x_j|\text{pa}(x_j)), \tag{2}$$

a product of the distribution of each node conditional on its parents, sometimes called the local characteristics. If a model follows a Bayesian network, it is thus possible to characterize the full model by the local characteristics.

6.1.1. *Causal validity and identifying causal effects.*

A Bayesian network is causally valid with respect to a hypothetical intervention if the density of each node on which

we do not intervene, conditional in its parents, retain its functional form in the hypothetical intervened scenario [20, Definition 1.3.1]. In other words, if the intervention gave rise to a joint distribution $\tilde{P}$, then

$$\tilde{P}(x_j|\text{pa}(x_j)) = P(x_j|\text{pa}(x_j)) \tag{3}$$

for the non-intervened x_j. Under causal validity one can use the do-calculus [20] to determine if causal effects are identifiable, so that a causal effect estimate can be obtained from observational data [22, 23].

Causal diagrams have allowed for visualization of causal dependencies, deriving tests for these dependencies, identifying causal effects, and obtaining estimating formulas for the effects [20]. More recently, causal diagrams have been used for causal assessment of parameters directly, in the case of the hazard function [6].

6.2. **Causal longitudinal models and time-dependent confounding.** There is often a need to account for the longitudinal aspect. Individuals may be censored due to loss to follow-up, and there is frequently an involved relationship between the exposure, the outcome, and the covariates: a doctor prescribes medication depending on a patients health status. The prescribed medication will typically influence later covariate levels, medical prescriptions, and treatment of the patient. Time-dependent confounding is said to be present if there are time-varying covariates that are simultaneously confounding, while also mediating the effect of the exposure regime on the outcome [9]. Such confounding is often present in longitudinal observational data. Standard regression techniques cannot provide causally interpretable effect estimates under time-dependent confounding, even if all confounders are measured [9]. Thus, other methods for assessing causal effects are needed.

6.3. **Marginal structural models in discrete time.** The marginal structural models were introduced by Robins et al. [9], originally motivated by the problem of time-dependent confounding. There are other methods for handling such confounding, such as the parametric G-formula [24] or structural nested models [25], but the work in this thesis focuses on the MSMs.

The MSMs can be used to account for time-dependent confounding, but their usage is not limited to such scenarios; they can also be used for analyzing longitudinal studies where time-dependent confounding is negligible (in fact, this approach is taken in **Paper 4**), or in the simple case of adjusting for confounding given a point exposure.

The MSM concept involves modeling distributional quantities, often the mean, as a function of a counterfactual outcome variable. Consider an observational longitudinal study with variables L_k, A_k, U_k - observed confounders, exposure, and unobserved variables - ordered discretely by the times $\{t_k\}_{k=0}^K$ with $0 = t_0 < t_1 < \cdots < t_K$, and an outcome Y of interest. We let $\bar{\cdot}$ denote the history of a variable, so that

$\bar{A}_k = (A_0, \cdots, A_k)$. An MSM for the cumulative treatment effect could be

$$E[Y^{\bar{a}}] = \beta_0 + \beta_1 \sum_{k=0}^{K} a_k, \tag{4}$$

where $Y^{\bar{a}}$ is the counterfactual outcome under exposure history $\bar{a}$, i.e. the outcome variable if, possibly contrary to the fact, treatment regime $\bar{a} = (a_1, \cdots, a_K)$ were imposed. As the outcome variable is counterfactual, it is defined for all exposure histories simultaneously, and usually not observed on the individual level in real life. From the observational data, one can fit the similar looking associational model

$$E[Y|\bar{A} = \bar{a}] = \gamma_0 + \gamma_1 \sum_{k=0}^{K} a_k, \tag{5}$$

using standard statistical techniques. As noted in Section 6.2, the coefficients of model (5) cannot be interpreted causally if (time-varying) confounding is present.

Robins et al. showed how to obtain causal estimates if all confounders are measured [9]. The strategy involves applying the now famous IPTWs. Use of these weights can heuristically be thought of as creating a pseudo-population in which the L_k's no longer affect future values of the treatment. The stabilized IPTW for individual i at time t takes the form

$$w_t^i = \prod_{k:t_k \leq t} \frac{P(A_k = a_k^i | \bar{A}_{k-1} = \bar{a}_{k-1}^i)}{P(A_k = a_k^i | \bar{A}_{k-1} = \bar{a}_{k-1}^i, \bar{L}_{k-1} = \bar{l}_{k-1}^i)}. \tag{6}$$

It is common to model the numerator and denominator of (6) using (pooled) logistic regression, and to estimate the weight by inserting the predicted probabilities [9, 26]. Fitting the associational model (5) to the weighted (pseudo-)population gives regression coefficients that consistently estimate the MSM coefficients (β_0, β_1) in (4); i.e. the estimated coefficients can be given causal interpretations [9].

6.3.1. *Assumptions.* For this procedure to work, some key assumptions must be satisfied. We have already mentioned the assumption of no unmeasured confounding given the L_k's. Furthermore, we generally need that both the outcome model (5) and the models used for obtaining the probabilities in (6) are correctly specified. Lastly, we rely on the positivity assumption, which is that $0 < P(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1})$ for all $a_k, \bar{a}_{k-1}, \bar{l}_{k-1}$, and for all k.

6.3.2. *Marginal structural models and survival analysis.* Pearls calculus of interventions for Bayesian networks has been successful for causal reasoning of random variables. The longitudinal aspect that deals with continuous-time processes have received somewhat less attention. The marginal structural models as we have briefly outlined here were originally developed in discrete time. However, if events occur in continuous time, as is the case in survival and event history analysis, a discrete-time

approach may not be entirely satisfactory. Sometimes the discrete models give an inadequate approximation of the underlying dynamics. In particular, when handling event histories and continuous-time mechanisms, there could be feedback between processes that is not adequately described by a DAG. From a continuous-time process point of view, the Bayesian networks often give a too coarse-grained description of the underlying system [27]. If the treatment initiation process is continuous in time, but a discrete-time weighting procedure like (6) is utilized, the resulting weights and weighted estimators will be biased (we show this using simulations in **Paper 3**; see Thesis Figure 6). Since survival analysis is defined in continuous time, it is also conceptually natural to investigate continuous-time weighting strategies - "discretizing" time by binning observations into discrete intervals for weight calculation seems to be contrived. For an MSM approach to survival analysis, a continuous rather than a discrete time weighting procedure would be appealing.

6.4. **Marginal structural models in continuous time.** In [10], Røysland introduced continuous-time MSMs using martingale theory and continuous-time counting processes. The idea is to formulate the desirable randomized trials as "randomized" probability measures. Under absolute continuity, we can relate the randomized measure to the observational measure, i.e. the model we have observations from, through Radon-Nikodym derivatives and Girsanov's theorem. We can then obtain likelihood ratios between the "observational" and "randomized" measures using modeling procedures that are surprisingly straight forward.

6.4.1. *Density factorization.* We consider a collection of baseline variables and counting processes whose joint distribution is given by the observational measure P. The graph induced by removing the counting processes, so that only baseline variables and edges between them remain, forms a Bayesian Network. From Section 6.1 we thus know that the joint density at baseline has a recursive factorization, as a product of the local characteristics of the baseline variables, i.e. given by (2). Moreover, from [28, A.1 a)] we know that the joint density restricted to all the events that could occur before t will uniquely be determined by density at baseline, and the intensities and jump times of the counting processes up to time t. This factorization allows us to find an expression of the weights.

6.4.2. *Hypothetical interventions in continuous time.* Some of the baseline variables and counting processes are subject to a hypothetical intervention. Such an intervention would give rise to a perturbed distribution $\tilde{P}$, e.g. a distribution corresponding to a randomized trial, that we would like to emulate. The assumption that enables us to model $\tilde{P}$ is, heuristically, that the short-term mechanisms we do not intervene upon are left invariant from P to $\tilde{P}$, while the ones we intervene upon changes according to the intervention.

6.4.3. *Causal validity.* Similar to the definition of Pearl our model is causally valid if the local characteristics of the nodes we do not intervene upon are the same under P and $\tilde{P}$. From Section 6.1.1 we recall that the local characteristic of a baseline variable is the distribution of that variable conditional on its parents. The local characteristic of a counting process N is its intensity function λ, which may heuristically be defined by

$$\lambda_t dt = P(N(t+dt) - N(t) = 1 | \text{"past"}),$$

where "past" refers to the events that are observed up to time t [2].

Thus, we say that a baseline intervention is causally valid if, for each baseline variable x_j that is not intervened upon, the conditional distribution of x_j given its parents $\text{pa}(x_j)$ coincide under P and $\tilde{P}$, as in (3). Moreover, we require that this baseline intervention does not affect the counting processes, in the sense that the functional form of each counting process intensity is invariant with respect to the intervention, i.e. that the local characteristic of each counting process coincide under P and $\tilde{P}$.

If an intervention instead is aimed at a counting process, it will by causally valid if the local characteristics of the remaining variables and processes coincide under P and $\tilde{P}$. This means that

- The functional forms of the intensities of the counting processes we do not intervene upon are invariant, i.e. are the same under P and $\tilde{P}$.
- The conditional distribution of each baseline variable, given its parent nodes, is the same under P and $\tilde{P}$.

6.4.4. *Identifying causal parameters.* The likelihood ratio between P and $\tilde{P}$ has a simple form; if the intervention is targeted at subject i's counting process N^i, changing the intensity λ^i to $\tilde{\lambda}^i$, the contribution to subject i's likelihood ratio at time t, R^i_t, is

$$\begin{aligned} R^i_t &= 1 + \int_0^t R^i_{s-} dK^i_s \\ K^i_t &= \int_0^t (\theta^i_{s-} - 1) dN^i_s + \int_0^t \lambda^i_s ds - \int_0^t \tilde{\lambda}^i_s ds, \end{aligned} \tag{7}$$

where $\theta^i = \tilde{\lambda}^i / \lambda^i$ [10]. If the intervention is targeted at a baseline variable X_i, the contribution to the likelihood ratio is the standard propensity weight R^i_0;

$$R^i_0 = \frac{d\tilde{P}(x_i | \text{pa}(x_i))}{dP(x_i | \text{pa}(x_i))}. \tag{8}$$

By simultaneously intervening on several components, one obtains individual likelihood ratios that are products of propensity weights and terms like (7). For these expressions to be proper likelihood ratios, they should at least be uniformly integrable martingales.

The likelihood ratio (7) is used for estimating the continuous MSMs analogously to the way IPTW (6) is used for estimating Robins' discrete MSMs. Hence, we sometimes refer to (7) as the continuous or continuous-time weights, and IPTW as the discrete or discrete-time weights.

6.4.5. *Assumptions.* Three key assumptions are required for the continuous-time MSM approach to work. The first two are related to hypothetical intervention, which are,

- The intervention is causally valid.
- The intervened measure $\tilde{P}$ is absolutely continuous to the observational measure P, i.e. if Q is an event such that $P(Q) = 0$, then $\tilde{P}(Q) = 0$ also.

We also need the following assumption:

- The likelihood ratio is identifiable.

Graphical identifiability criteria for the true likelihood ratio, analogous to the back-door criterion [20], is discussed in detail in a forthcoming paper [29].

For the outcome estimand to have the desired interpretation, the marginal structural model must also be correctly specified. If these criteria are met, one can perform causal survival analysis by re-weighting standard estimators.

7. Description of the papers

7.1. **Paper 1.** We build on a well-known relationship between the cumulative hazard function and the survival function to derive a general estimating equation for a range of parameters in survival and event history analysis. Examples of such parameters include the survival function, cumulative incidence functions, the restricted mean survival function, (cumulative) prevalence functions, as well as several other parameters that we present in the article and the Supplementary material. Other examples can be found in **Paper 2**.

7.1.1. *ODE parameters.* The approach is motivated by the fact that these parameters solve systems of ordinary differential equations. Our article focuses on the formulation

$$X_t = X_0 + \int_0^t F(X_s) dA_s, \tag{9}$$

where X is a vector containing the parameter(s) of interest, X_0 is a vector of initial values, and $F = (F_1, F_2, \cdots)$ is a matrix-valued function. In our examples, the integrator A is a vector of cumulative hazard coefficients. We also wanted to include parameters that are Lebesgue integrals, so we included the case when $dA_t^i = dt$ for some i.

7.1.2. *Plugin estimation.* We can estimate the parameter X by utilizing the ODE structure: by merely replacing A by an estimate $\hat{A}$, and X_0 by a consistent estimator $\hat{X}_0$, we obtain a stochastic differential equation (SDE) plugin estimator

$$\hat{X}_t = \hat{X}_0 + \int_0^t F(\hat{X}_{s-}) d\hat{A}_s. \tag{10}$$

7.1.3. *The P-UT property.* We will study integrators that are predictably uniformly tight (P-UT). A general definition of P-UT can be found in [11], but we do not need the full generality of this definition. Instead we provide a sufficient condition that is suitable for our purposes in Lemma 1: if $\{Z_t^{(n)}\}_n$ is a sequence of semimartingales on $[0, \mathcal{T}]$, $\{\rho^{(n)}\}_n$ are predictable processes such that $M_t^{(n)} := Z_t^{(n)} - \int_0^t \rho_s^{(n)} ds$ define square integrable local martingales, and

$$\lim_{J\to\infty} \sup_n P\Big(\sup_{s\leq\mathcal{T}} |\rho_s^{(n)}|_1 \geq J \Big) = 0$$
$$\lim_{J\to\infty} \sup_n P\Big(\operatorname{Tr}\langle M^{(n)} \rangle_{\mathcal{T}} \geq J \Big) = 0,$$

then $\{Z_t^{(n)}\}_n$ is P-UT. Here, Tr is the trace function, and $|\cdot|_p$ is the p norm.

7.1.4. *Estimating the cumulative hazard.* We focus on estimators $\hat{A}$ that are counting process integrals. Furthermore, we require these estimators to be consistent and P-UT.

The additive hazard estimator is such an estimator; in Proposition 1 we show that it is P-UT under a momentum condition on the covariates. When A is a cumulative hazard, we thus propose inserting cumulative hazard estimates coming from the additive model. These estimates are piecewise constant, and we denote the ordered jump times by $\{\tau_k\}_k$.

If $dA_t^i = dt$ for some i, we may refine the "grid" of event times by adding more τ_k's to improve precision. The proposed estimator of A_t^i is in that case $\hat{A}_t^i = \max_k\{\tau_k : \tau_k \leq t\}$. A short argument shows that $\hat{A}^i$ is consistent and P-UT.

7.1.5. *Asymptotic results.* In Theorem 1, we show that the SDE plugin estimator (10) is consistent, assuming the estimator $\hat{A}$ is consistent and P-UT. The results build upon stability theory for SDEs and also require Lipschitz and linear growth bound conditions to be satisfied. An essential reference for this result is [11].

We identify Z, the limiting process (with respect to the Skorohod metric) of the root n residuals $Z^n = \sqrt{n}(\hat{X} - X)$. It solves the SDE

$$Z_t = Z_0 + \sum_{j=1}^{k} \int_0^t \nabla F_j(X_{s-}) Z_{s-} dA_s^j + \int_0^t F(X_{s-}) dW_s,$$

where W is a mean zero Gaussian martingale with independent increments. We also find an expression for V, the pointwise covariance of Z:

$$V_t = V_0 + \sum_{j=1}^{k} \int_0^t \Big(V_s \nabla F_j(X_s)^\intercal + F_j(X_s) V_s \Big) dA_s^j + \int_0^t F(X_s) d[W]_s F(X_s)^\intercal. \tag{11}$$

This gives a way to also estimate the covariance; by plugging in the cumulative hazard and parameter estimates we obtain the estimator

$$\begin{aligned} \hat{V}_t = \hat{V}_0 &+ \sum_{j=1}^{k} \int_0^t \Big(\hat{V}_{s-} \nabla F_j(\hat{X}_{s-})^\intercal + \nabla F_j(\hat{X}_{s-}) \hat{V}_{s-} \Big) d\hat{A}_s^j \\ &+ n \int_0^t F(\hat{X}_{s-}) d[B]_s F(\hat{X}_{s-})^\intercal, \end{aligned} \tag{12}$$

where $[B]_t$ is a matrix defined by

$$\Big([B]_t\Big)_{i,j} = \begin{cases} 0, \text{ if } dA_t^i = dt \text{ or } dA_t^j = dt, \\ \sum_{s \leq t} \Delta \hat{A}_s^i \Delta \hat{A}_s^j, \text{ otherwise.} \end{cases}$$

We can estimate the pointwise covariance of the plugin estimator $\hat{X}$ by dividing (12) by n.

We show that the plugin covariance estimator $\hat{V}$ is consistent in Theorem 2. The result rests on the assumption that the root n residuals of the cumulative hazard estimator, $W^n = \sqrt{n}(\hat{A} - A)$, is P-UT and converges weakly with respect to the Skorohod metric to a mean zero Gaussian martingale with independent increments [30]. We also need that the quadratic variation of W^n is P-UT. By Proposition 1, both these properties hold when $\hat{A}$ are additive hazard estimates under two conditions on the covariates: if the hazard for subject i is given by $Q_-^{i,\intercal}\alpha$, Q is the matrix where row i is $Q^{i,\intercal}$, and Y is the diagonal matrix with i'th diagonal element equal to Y^i (the at-risk indicator for subject i), then these properties are satisfied for the additive hazard estimator of $A = \int_0^\cdot \alpha_s ds$ if

(1) $E[\sup_{t\leq\mathcal{T}} |Q_t^i|_3^3] < \infty$ for each i,
(2)
$$\lim_{J\to\infty} \inf_n P\left(\sup_{t\leq\mathcal{T}} \operatorname{Tr}\left(\left(\frac{Q_{t-}^\intercal Y_t Q_{t-}}{n} \right)^{-1} \right) \geq J \right) = 0.$$

7.1.6. *Implementation.* The parameter estimator (10) can be represented as a difference equation, so that the value at time t for $\tau_k \leq t < \tau_{k+1}$ depends on the increment $\Delta\hat{A}_{\tau_k}$ and $\hat{X}_{\tau_{k-1}}$;

$$\hat{X}_t = \hat{X}_{\tau_{k-1}} + F(\hat{X}_{\tau_{k-1}})\Delta\hat{A}_{\tau_k}. \tag{13}$$

Similarly, the plugin variance estimator (12) can be written as a difference equation that reads

$$\begin{aligned} \hat{V}_t = \hat{V}_{\tau_{k-1}} &+ \sum_{j=1}^{k} \left(\hat{V}_{\tau_{k-1}} \nabla F_j(\hat{X}_{\tau_{k-1}})^\intercal + \nabla F_j(\hat{X}_{\tau_{k-1}}) \hat{V}_{\tau_{k-1}} \right) \Delta\hat{A}^j_{\tau_k} \\ &+ nF(\hat{X}_{\tau_{k-1}})\Delta[B]_{\tau_k} F(\hat{X}_{\tau_{k-1}})^\intercal. \end{aligned} \tag{14}$$

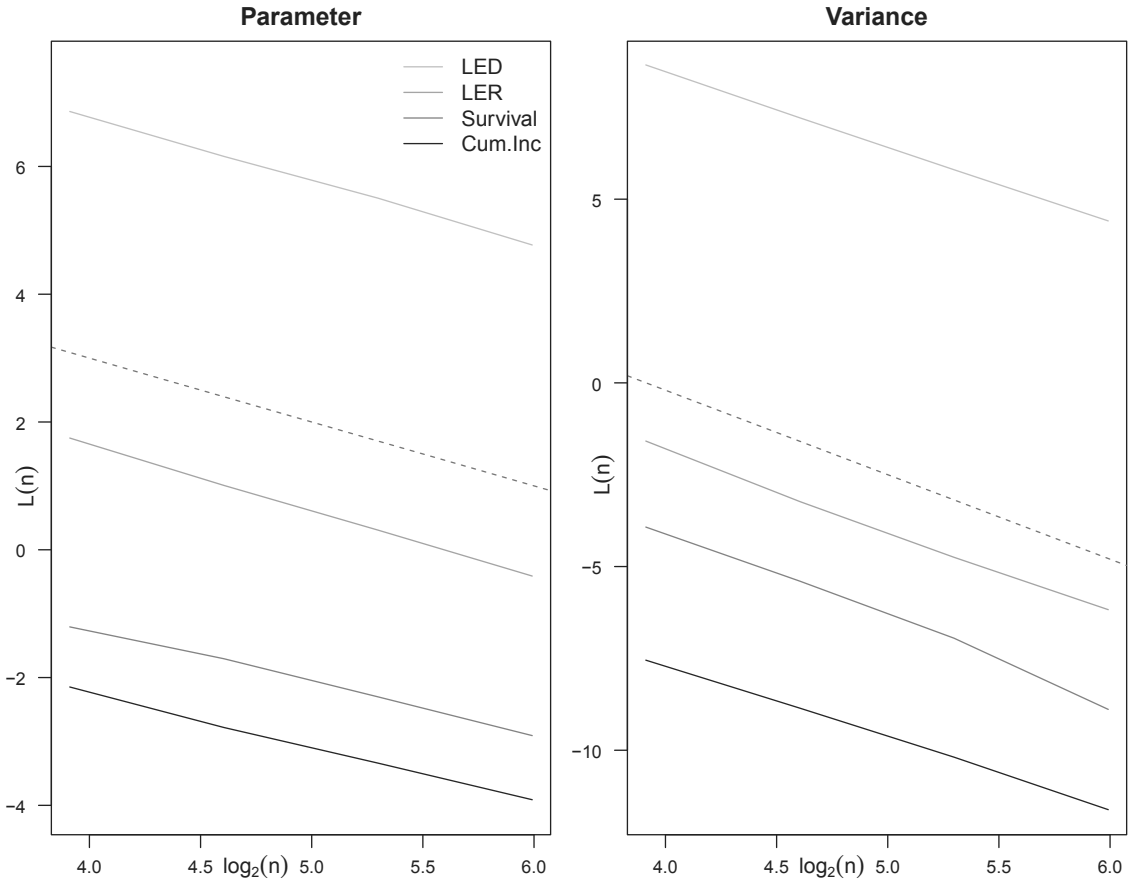
Thus, the parameter and covariance estimators at each event time τ_k depends on their values at the previous event time τ_{k-1}, and the increments $\Delta\hat{A}_{\tau_k}$. These equations can therefore be solved recursively on a computer: if we have $\hat{X}_0, \hat{A}$, F, and the ∇F_j's, we can perform the estimation using e.g. a `for` loop.

7.1.7. *Performance.* We check the performance of the estimators using simulations, by plotting the convergence order and coverage for several parameters. Convergence order is assessed using the L^2 criterion

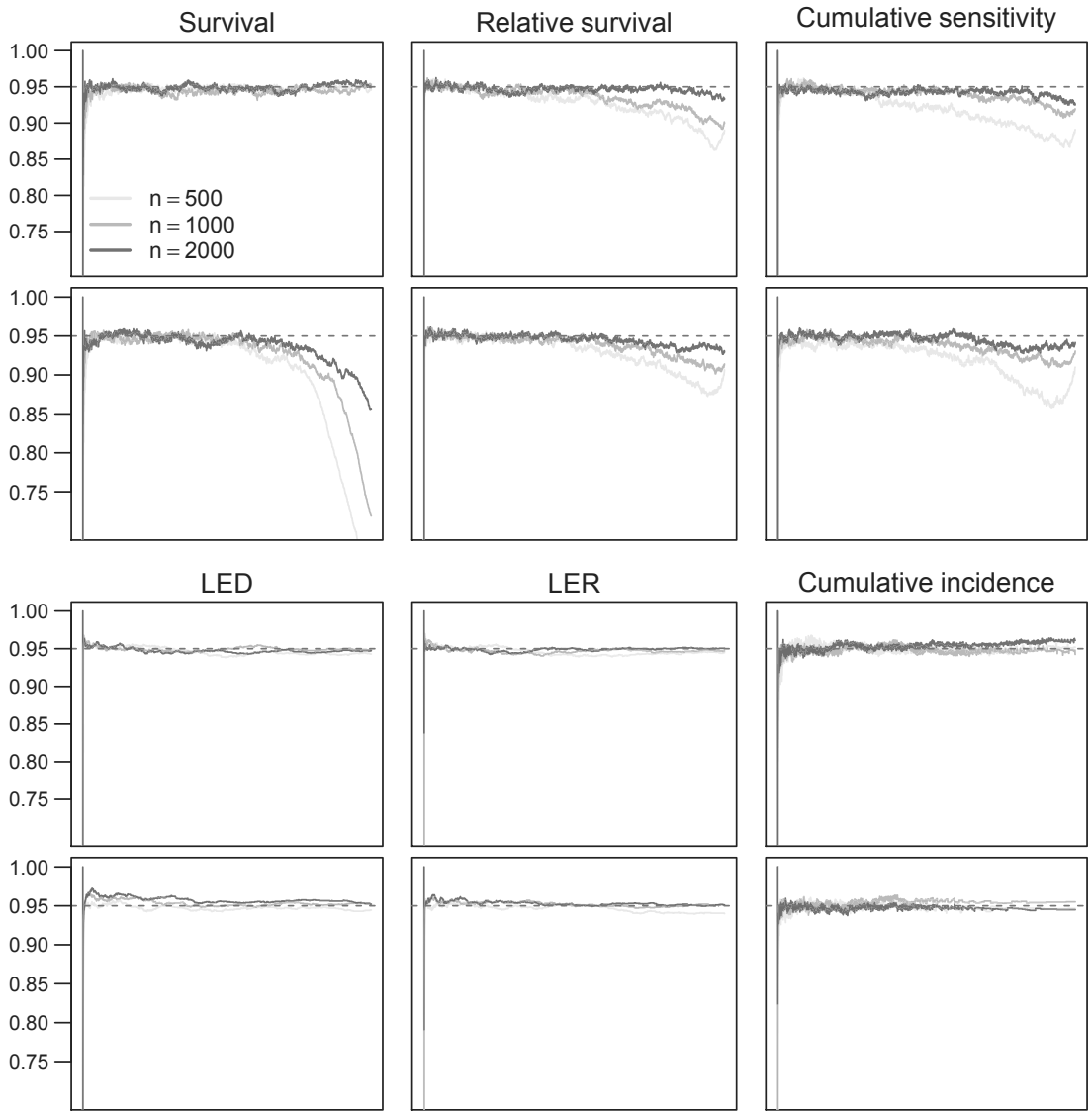
$$L(n) = \frac{1}{K} \sum_{j=1}^{K} \int_0^{\mathcal{T}} |X_s - \hat{X}_s^{n,j}|^2 ds,$$

where $\hat{X}^{n,j}$ is the realization of the plugin estimator for the j'th simulation with sample size n. Convergence order for the plugin estimators is indicated in Thesis

Figure 1. We show coverage for different sample sizes in Thesis Figure 2. The rejection rate appears to be well calibrated to the 95% confidence level.



THESIS FIGURE 1. Convergence of selected plugin estimators; parameter to the left and variances to the right. The dashed lines indicate convergence order 1 in the left panel and 2.3 in the right panel. LED and LER are abbreviations for life expectancy difference and life expectancy ratio, respectively; see the Supplementary material of **Paper 1** for details.



THESIS FIGURE 2. Estimated mean coverage for selected parameters simulated with constant (upper panel) and linearly crossing (lower panel) hazards. The crossing hazards were chosen such that they crossed in the middle of the x-axes. The dotted line indicates the confidence level, and n indicates the sample size. LED and LER are abbreviations for life expectancy difference and life expectancy ratio, respectively; see the Supplementary material of **Paper 1** for details.

7.2. **Paper 2.** There has recently been raised some concerns regarding the use of hazards and hazard ratios in survival analysis, as discussed in Section 5.1. Since hazards can be difficult to interpret, it is natural to re-evaluate the use of hypothesis tests in survival and event history analysis that compare hazards. The null hypothesis H_0^R for the rank test that compares two groups take the form

$$H_0^R : \alpha_t^1 = \alpha_t^2, \text{ for } t \in [0, \mathcal{T}],$$

where α^i is a hazard in group i and $[0, \mathcal{T}]$ is the study period. This null hypothesis is easy to interpret in specific situations, such as when survival functions are of primary interest. In other cases, it may be difficult to have a deep understanding of such null hypotheses, e.g. if competing risks are present.

We argue that null hypotheses should be stated on parameters that have clear interpretations, and use the results from **Paper 1** to obtain tests for a range of survival analysis parameters pointwise in time.

7.2.1. *Rank test malpractice.* We point out that the rank tests are sometimes incorrectly used, e.g. for testing null hypotheses that are different from the actual rank hypothesis H_0^R. One such example is when equality between two survival curves at a specific time point of interest, e.g. survival after five years of follow-up. Using the rank tests on such hypotheses will often lead to misleading p-values. Moreover, the rank tests are sometimes used when the research question is not adequately specified, e.g. when the researchers only vaguely indicate that they are interested in testing a difference between two groups.

7.2.2. *Alternative null hypotheses in survival analysis.* We focus on groupwise comparison of survival and event history analysis parameters at a pre-specified time point t_0, i.e. we study the null hypothesis

$$H_0 : X_{t_0}^1 = X_{t_0}^2,$$

where X^i is the true parameter in group i. If X solves an ODE like (9), we can use the plugin estimators developed in **Paper 1** to obtain the test statistic

$$(\hat{X}_{t_0}^1 - \hat{X}_{t_0}^2)\hat{V}_{t_0}^{-1}(\hat{X}_{t_0}^1 - \hat{X}_{t_0}^2)^\intercal, \tag{15}$$

where $\hat{V}_{t_0}$ is the plugin variance estimator of $\hat{X}_{t_0}^1 - \hat{X}_{t_0}^2$. Asymptotically, this follows a χ^2 distribution, which can be seen from an application of the continuous mapping theorem. We write this test statistic out for several parameters in the paper.

7.2.3. *Performance.* We estimate power functions for several parameters, and compare with the power functions of other nonparametric tests tailor-made for each parameter, as found in the literature. This is done for three hazard scenarios; constant, linearly crossing, and equal initially before deviating. We estimate power by optimizing the hazards such that $X_{t_0}^1 - X_{t_0}^2 = \kappa$ for several values of κ, and do this for each considered parameter and hazard scenario. The estimated power functions are

plotted in Thesis Figure 3, and the figure shows that our tests and the standard non-parametric tests have similar power. More extensive simulations can be found in the Supplementary material.

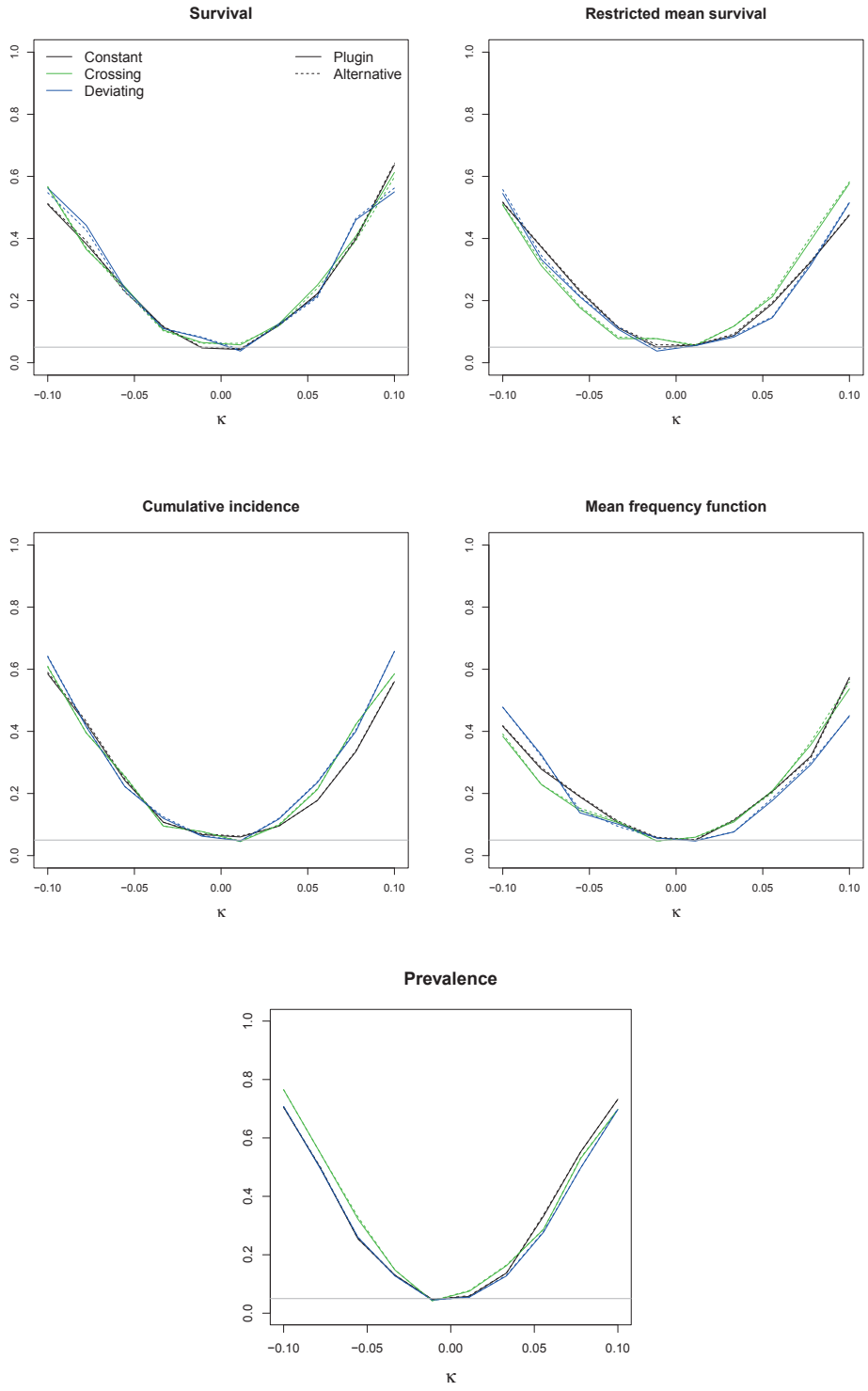
7.2.4. *Comparisons with the rank tests.* We compare the power of our tests with the rank tests in Thesis Table 1 when our null hypothesis is false, with $X^1_{t_0} - X^2_{t_0} = -0.05$ in each case. The power of the rank tests is sensitive to the shape of the underlying hazards, while the power of our tests varies less across the scenarios. This is not surprising, as the rank test hypothesis is different from our hypothesis; our null hypothesis is slightly violated in each case, but the rank null hypothesis may be arbitrary violated. We see less discrepancy between our tests and the rank tests when the hazards are proportional (or constant). The table indicates that our tests and the rank tests are fundamentally different, particularly when the hazards are not proportional.

In Thesis Figure 4 we compare the rejection rate of our tests with the rank tests when the hazards in the two groups are linearly crossing. The hazards are optimized such that our null hypothesis is true for each parameter under the restriction that the hazards are crossing. The crossing hazards are visualized in the uppermost row of panels. We estimate and plot the rejection rate as a function of the ratio of the slopes in the lower three rows of panels, so that the hazards are very different to the left of each panel and become identical to the right in each panel. Our tests provide rejection rates close to the 5% significance level for all combinations of crossing hazards and sample sizes, as desired. The rank test often falsely rejects the null hypothesis when the slopes are different but approaches the rejection rate of our test when the slopes approach each other, i.e. when the curves approach proportionality. Again, the rank tests perform poorly, as they are based on a null hypothesis that is different from the one we consider.

7.2.5. *Colon cancer example.* We demonstrate how to use our testing procedure on real data. The analysis can be found as a worked example in the Supplementary material.

We compare two adjuvant chemotherapy regimes (Lev and Lev+5FU) for patients with stage III colon cancer, using the `R` data set `colon` that is freely available for anyone. The data set included recordings of cancer recurrence and death for each subject. We plot survival curves and cumulative incidence curves of cancer recurrence in Figure 5.

The comparison is made at one and five years of follow-up using the test statistic (15) on several parameters. The results are shown in Thesis Table 2. At one year of follow-up, we are not able to find significant differences in survival or restricted mean survival between the two treatments. However, the cumulative incidence of recurrence and the number of subjects alive with recurrent disease are lower in the Lev+5FU



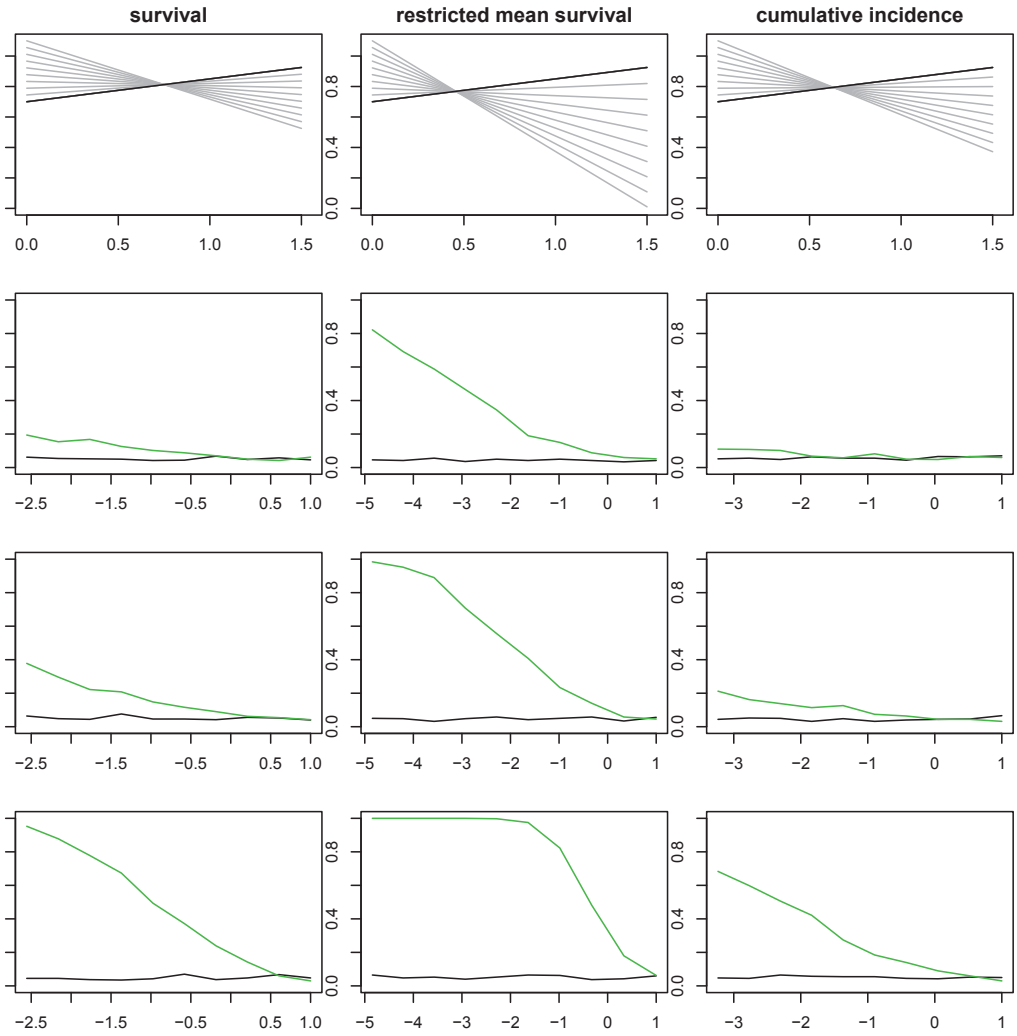
THESIS FIGURE 3. Estimated power functions for constant (black), crossing (green), and deviating (blue) hazards, based on 500 subjects with 10% random censoring. The dashed lines show test statistics derived from existing methods in the literature that are tailor-made for the particular scenario. The gray lines show the confidence level.

group. Moreover, the mean time spent alive and without recurrence is longer in the Lev+5FU group at one year of follow-up.

Survival and restricted mean survival are significantly better in the Lev+5FU group at five years of follow-up. Furthermore, the cumulative incidence of cancer recurrence, the prevalence of recurrence, and the restricted mean recurrence for survival are better in the Lev+5FU group. Restricted mean recurrence-free survival is also significantly longer after five years.

THESIS TABLE 1. Power comparisons of our tests and the log-rank test (our/log-rank), when testing equality between two groups of 1500 individuals each. In the cumulative incidence row, we also added power based on the Gray test [31] (our/log-rank/Gray). The cause-specific hazards for the competing event are held constant across the scenarios, while the cause-specific hazards for the event of interest are respectively constant, linearly crossing, and equal before deviating under the restriction that $X^1_{t_0} - X^2_{t_0} = -0.05$.

Parameter \ Hazard	Constant	Crossing	Deviating
Survival	0.81/0.88	0.79/0.96	0.83/0.70
Restricted mean survival	0.77/0.87	0.78/0.21	0.8/1
Cumulative incidence	0.85/0.94/0.88	0.86/0.80/0.70	0.86/0.83/0.76



THESIS FIGURE 4. In the upper row, we display hazard functions in scenarios where the hazard in group 1 is fixed (black line), and the hazards in group 2 vary (grey lines). The hazards are optimized such that our null hypothesis is true, i.e. $X^1_{t_0} = X^2_{t_0}$ at $t_0 = 1.5$ for each combination of black/gray hazards. In the lower rows we show the estimated rejection rate as a function of the ratio of the hazard slopes (slope of gray/slope of black). This is done for sample sizes sample sizes 500 (row 2), 1000 (row 3), and 5000 (row 4). The green curve shows the rejection rate of the log-rank test, and the black curve shows the rejection rate of our tests. If the sample size is large, the rank tests can falsely reject our null hypothesis even when the hazards are crossing. The cumulative incidence panels: we only show the cause-specific hazards for the event of interest (which we compare using the log-rank test). The cause-specific hazard for the competing event is equal to 0.4 in both groups.

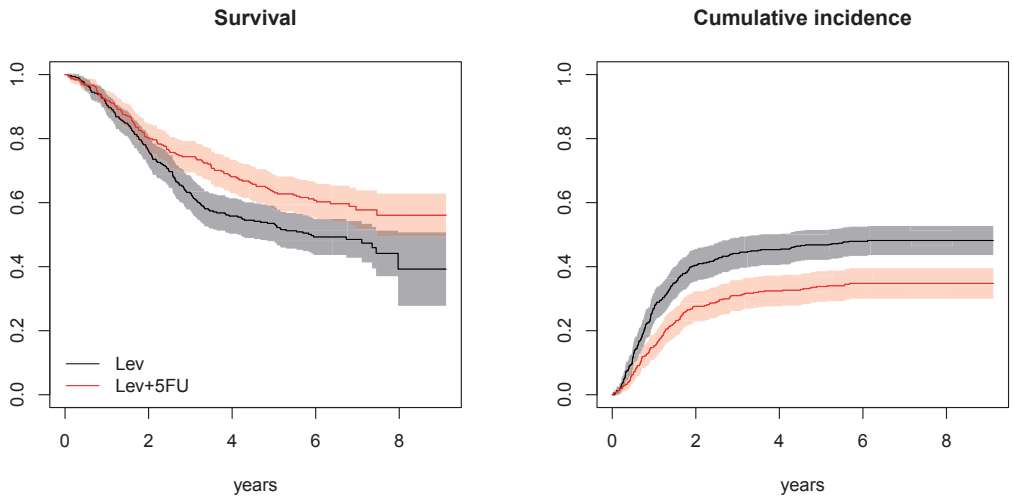
THESIS TABLE 2. Comparison of the colon cancer treatments Lev and Lev+5FU at one and five years of follow-up. 95% confidence intervals are shown in parenthesis, and the p-value of the test statistic (15) is provided in the rightmost column. The restricted mean recurrence-free survival (RMRFS) is $\int_0^{\cdot} S_s^{D \wedge R} ds$, i.e. derived from the "survival" function $S^{D \wedge R}$ for the composite endpoint of cancer recurrence and death.

At one year of follow-up

	Lev	Lev+5FU	p-value
Survival	0.91 (0.87,0.94)	0.92 (0.89,0.95)	0.62
Restricted mean survival	0.96 (0.95,0.98)	0.97 (0.95,0.98)	0.86
Cumulative incidence	0.27 (0.22,0.32)	0.15 (0.11,0.19)	0.00
Prevalence	0.19 (0.14,0.23)	0.09 (0.05,0.12)	0.00
RMRFS	0.85 (0.82,0.88)	0.90 (0.87,0.92)	0.01

At five years of follow-up

	Lev	Lev+5FU	p-value
Survival	0.54 (0.48,0.59)	0.63 (0.58,0.69)	0.01
Restricted mean survival	3.62 (3.44,3.81)	3.97 (3.79,4.15)	0.01
Cumulative incidence	0.47 (0.42,0.51)	0.34 (0.29,0.39)	0.00
Prevalence	0.07 (0.05,0.10)	0.03 (0.02,0.05)	0.02
RMRFS	2.29 (2.07,2.51)	2.95 (2.73,3.18)	0.00



THESIS FIGURE 5. Survival curves (left) and the cumulative incidence of recurrence (right) along with 95% pointwise confidence intervals (shaded) from the colon cancer trial.

7.3. **Paper 3.** We introduce an estimator for the continuous-time weights (7) based on the additive hazard estimator, and show that it is consistent. Furthermore, we show that the additive hazard estimator weighted with the estimated weights is consistent, i.e. we take into account that the weights are estimated. We also provide an argument showing that our weighted additive hazard estimator is P-UT. The results from **Paper 1** can then be used to target the ODE parameters that we would have seen if we could carry out the hypothetical intervention of interest. We include an example section describing how the `ahw` software works and provide a continuous-time MSM analysis of a simulated data set.

7.3.1. *A continuous-time weight estimator.* We utilize the SDE formulation (7). Assuming the time to treatment hazards are additive, we have under the multiplicative intensity model [2], that

$$\lambda_t^i = Y_t^i Z_{t-}^{i\intercal} h_t \text{ and } \tilde{\lambda}_t^i = Y_t^i \tilde{Z}_{t-}^{i\intercal} \tilde{h}_t,$$

where $Y_t^i = 1$ if subject i is at risk for treatment at time t, and 0 otherwise, Z^i and $\tilde{Z}^i$ are vectors of covariates for subject i and h and $\tilde{h}$ are vectors of additive hazard coefficients as in (1). By inserting these intensities we get that the system (7) reads

$$\begin{aligned} R_t^i &= 1 + \int_0^t R_{s-}^i dK_s^i \\ K_t^i &= \int_0^t \big(\theta_{s-}^i - 1\big) dN_s^i + \int_0^t Y_s^i Z_{s-}^{i\intercal} dH_s - \int_0^t Y_s^i \tilde{Z}_{s-}^{i\intercal} d\tilde{H}_s, \end{aligned}$$

where $H_t = \int_0^t h_s ds$ and $\tilde{H}_t = \int_0^t \tilde{h}_s ds$. Our estimation strategy is inspired by **Paper 1**; we let $\hat{H}$ and $\hat{\tilde{H}}$ be additive hazard estimates of H and $\tilde{H}$, and insert them into the above equation to obtain the SDE estimator

$$\begin{aligned} \hat{R}_t^i &= 1 + \int_0^t \hat{R}_{s-}^i d\hat{K}_s^i \\ \hat{K}_t^i &= \int_0^t (\hat{\theta}_{s-}^i - 1) dN_s^i + \int_0^t Y_s^i Z_{s-}^{i\intercal} d\hat{H}_s - \int_0^t Y_s^i \tilde{Z}_{s-}^{i\intercal} d\hat{\tilde{H}}_s, \end{aligned} \tag{16}$$

where $\hat{\theta}^i$ is an estimator of θ^i, given by

$$\hat{\theta}_t^i = \begin{cases} \theta_0^i, & 0 \le t < \frac{1}{\kappa} \\ \frac{\int_{t-\frac{1}{\kappa}}^t Y_s^i \tilde{Z}_{s-}^{i\intercal} d\hat{\tilde{H}}_s}{\int_{t-\frac{1}{\kappa}}^t Y_s^i Z_{s-}^{i\intercal} d\hat{H}_s}, & \frac{1}{\kappa} \le t \le \mathcal{T}, \end{cases} \tag{17}$$

for the bandwidth parameter κ.

7.3.2. *Weighted additive hazard regression.* We consider an additive hazard model for the outcome, where subject i has a vector of covariates Q^i, and hazard $Q^{i\intercal}_- \beta$. Let $N^{(n)} = (N^{1,D}, \cdots, N^{n,D})^\intercal$ be the vector of counting processes where $N^{i,D}$ counts when the outcome occurs for subject i, and $Q^{(n)}$ the matrix where row i is $Q^{i\intercal}$. Finally, let $Y_s^{(n),D}$ denote the $n \times n$-dimensional diagonal matrix where the i'th diagonal element is $Y_s^{i,D} \cdot \hat{R}^i_{s-}$; subject i's at risk-indicator for the outcome at time s multiplied by his weight estimate just before s. The weighted additive hazard regression is given by

$$\hat{B}_t = \int_0^t (Q^{(n)\intercal}_{s-} Y_s^{(n),D} Q^{(n)}_{s-})^{-1} Q^{(n)\intercal}_{s-} Y_s^{(n),D} dN_s^{(n)}, \tag{18}$$

which is an estimator of $B_t = \int_0^t \beta_s ds$.

7.3.3. *Consistency.* In Theorem 2, we show that our weight estimator (16) is consistent under standard assumptions for consistency of the additive hazard estimator [30]. We also assume the θ^i's are uniformly bounded, and right-continuous at $t = 0$, and that the bandwidth $\kappa = \kappa_n$ grows as a function of n such that $\lim_{n\longrightarrow\infty} \kappa_n = \infty$ and $\sup_n \kappa_n/\sqrt{n} < \infty$ are satisfied.

We show that (18) is consistent and P-UT in Theorem 1. The proof relies on the assumption that the true weights are uniformly bounded, and that the weight estimator converges in probability to the true weights at each time t. The latter is ensured by Theorem 2 for our estimator (16).

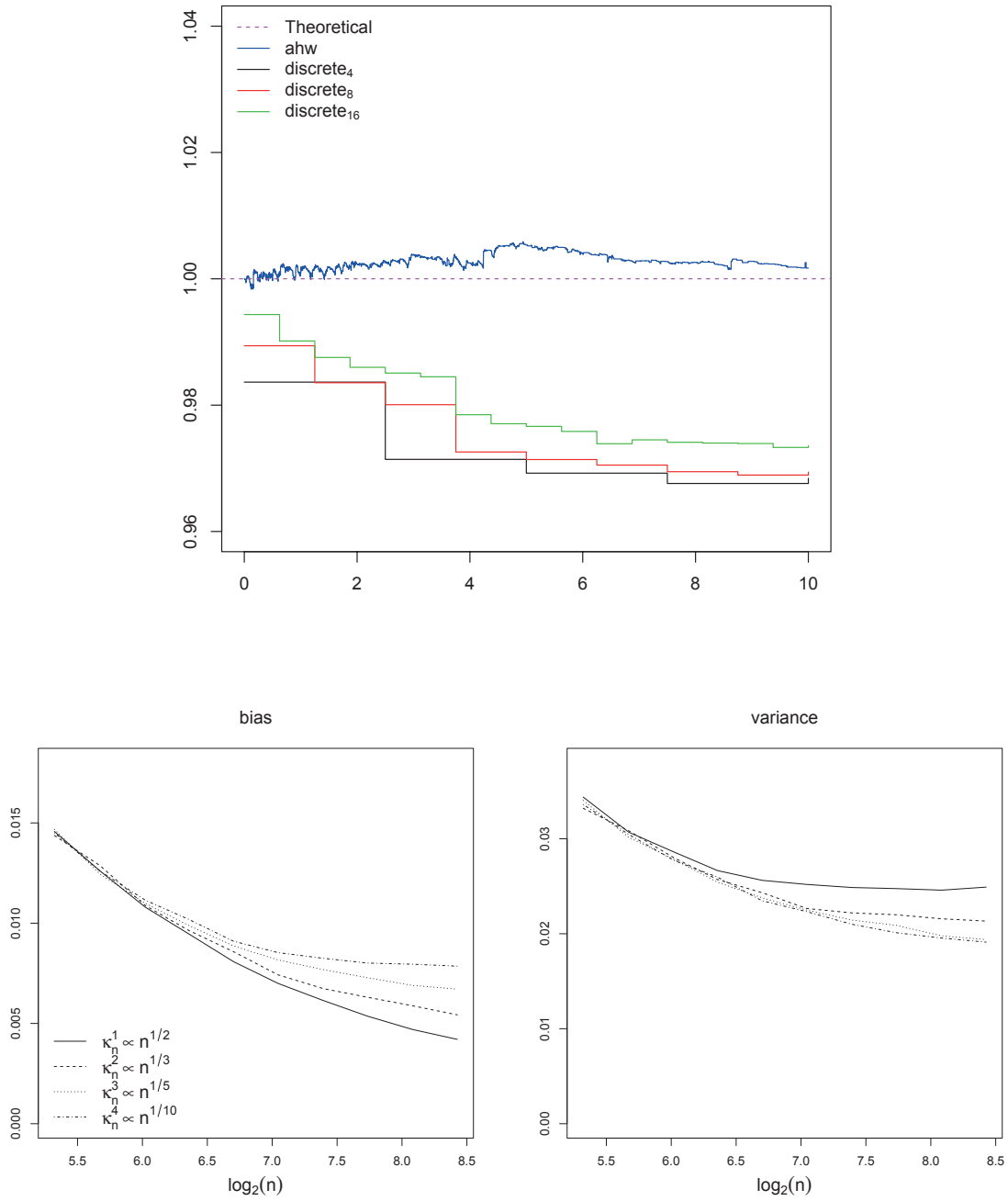
By combining Theorem 1 and 2, we conclude that (18) is consistent when our estimator (16) is used for obtaining the weights. In this way, we propose a two-step estimation procedure for obtaining causal cumulative hazards; we model the continuous-time weights using additive hazard regression and use the estimates to re-weight the additive hazard estimator for the outcome.

7.3.4. *Targeting MSM parameters.* Since the estimator (18) is consistent and P-UT, we may utilize plugin estimation to estimate causal parameters that solve ODEs on the form (9). These plugin estimators are consistent by Theorem 1 in **Paper 1**.

7.3.5. *Example.* We describe how our methods can be utilized for marginal structural modeling in practice through an example in Section 4 of the paper. We consider a data generating mechanism with time-dependent confounding, and we estimate the outcome hazard using the continuous-time weight estimator (18). We go into some detail on how the weight calculation software in the `ahw` package works, and how to perform weighted additive hazard regression in `R`. In Section 4.4 we specify a marginal structural relative survival model as a solution to an ODE, and plot the weighted cumulative hazard estimates and estimated MSM coefficients.

7.3.6. *Performance.* Usually, MSMs are estimated by use of the discrete weights (6). If treatment occurs in continuous time, the discrete weights may be a biased approximation of the true likelihood ratio (7). We compare our weight estimator (16) with the discrete weights obtained from a pooled logistic regression. We find that a discrete-time approach gives bias when the true exposure process is time-continuous, but that the bias diminishes as we refine the time discretization. The bias is plotted in the upper panel of Thesis Figure 6, and is based on the data example considered in Section 7.3.5. The curves are obtained by repeated simulations of the data, weight calculation for each simulation, before finally averaging over the simulations.

We investigate the convergence of our weight estimator (16) for four different bandwidth strategies κ, using simulations. The bandwidths are identical for the smallest sample n_0, i.e. $\kappa^1_{n_0} = \kappa^2_{n_0} = \kappa^3_{n_0} = \kappa^4_{n_0}$. Otherwise they satisfy $\kappa^1_n \propto n^{1/2}, \kappa^2_n \propto n^{1/3}, \kappa^3_n \propto n^{1/5}$, and $\kappa^4_n \propto n^{1/10}$. The bandwith choice is a bias-variance tradeoff, as can be seen in the two lower panels of Thesis Figure 6.



THESIS FIGURE 6. Upper panel: the true weights have expected value 1. Included are our additive hazard weights, as well as IPTW with 4, 8, and 16 time intervals. Our weight estimator (16) is less biased than IPTW in this scenario. Lower panels: bias and variance of our weight estimator as a function of the sample size n, for four bandwidth refinement strategies.

7.4. **Paper 4.** There is a debate about the relative effectiveness of the prostate cancer treatments regimens radical prostatectomy (RP) and radiation therapy (Rad) among nonmetastatic cancer patients [32, 33]. An RCT on patients with localized prostate cancer suggested that radiation therapy and radical prostatectomy give similar rates of disease progression after ten years of follow-up [34]. However, this study has been criticized for being too underpowered to detect clinically significant differences. On the other hand, observational studies have suggested that RP is favorable [32].

We employed a continuous-time MSM on Norwegian registry data to investigate the issue. The goal was to model a randomized scenario under which a comparison of the two treatment regimens would be fair.

7.4.1. *Description of the data.* The cohort consisted of Norwegian males ($n = 1296$) diagnosed with nonmetastatic prostate cancer in 2004/2005, all of whom received treatment. Diagnostic factors such as age, PSA, Gleason score, and T category were measured at baseline, i.e. the time of diagnosis, while the chosen treatment regimens and the dates of treatment/treatment failure/death and cause of death were registered until end of follow-up (June 2015). We extracted comorbidity variables from the Norwegian Prescription Registry, based on medication that was bought around the time of diagnosis. We also used data on the education level at the time of diagnosis, provided by Statistics Norway, to account for socioeconomic differences.

7.4.2. *Defining the Failure endpoint.* Ideally, we would like to compare the rates of death due to prostate cancer. However, this endpoint would not provide enough power to detect differences. To obtain higher power, we defined a surrogate failure endpoint, denoted "Failure," for events that indicated the given treatment was unsuccessful in some way. The Failure group therefore included individuals who received radiation treatment a long time after initial treatment, as such treatment is no longer thought to be adjuvant. We thus viewed the following two events as Failure:

- For RP patients: initiation of radiation therapy later than six months after surgery.
- For Rad patients: new radiation therapy later than eight weeks after initial treatment.

Furthermore, Rad patients received adjuvant hormone treatment three years after initial treatment. A longer period of hormone therapy indicates further signs of disease and was considered a Failure. Also, some patients stopped hormone treatment before three years, possibly due to side effects. If hormone treatment was re-initiated after a gap of six months, we considered the initial curative treatment to be unsuccessful, and therefore a Failure. Finally, a subject had Failure at his time of death if the cause of death was attributed to prostate cancer.

The cohort consisted of older men who were likely to die of reasons unrelated to prostate cancer during the eleven years of follow-up. We therefore accounted for the

competing risk of death from other causes, by studying the cumulative incidence of Failure. We referred to the competing endpoint as "Other death."

7.4.3. *Spurious effects and confounding.* The primary goal of the analysis was to make a fair comparison of the rates of Failure starting from the time of treatment assignment. When making this comparison, we faced two forms of systematic bias: first, patients received treatment based on their prognostic factors. These factors also influenced their long term prognosis - i.e. the prognostic factors at baseline were confounders. Second, the rate of treatment initiation depended on prognostic factors, the chosen treatment regimen, as well as socioeconomic factors. We attempted to mimic a scenario in which these two aspects - the treatment assignment and the rate of treatment initiation - were randomized in the population. We accounted for these aspects by use of weighting; propensity weights for balancing covariate distributions at baseline, and continuous-time weights to balance the systematic differences in time to treatment.

7.4.4. *Estimands of interest, and interpretation.* We proposed a marginal structural model for the cumulative incidence of Failure as a function of the two treatment regimens. This was done by considering an intervention g_l that put treatment mode to l while imposing the marginal treatment initiation rate in the population. The distribution of events under this intervention was denoted by $\tilde{P}^{g_l}$. We denoted the time from diagnosis to Failure by T_f, and the time to Other death by T_{od}, and expressed a marginal structural model G^l as a function of the treatment regime l:

$$\begin{aligned} G_t^l &= \tilde{P}^{g_{\mathrm{Rad}}}(t \geq T_f, T_f < T_{od}) I(l = \mathrm{Rad}) + \tilde{P}^{g_{\mathrm{RP}}}(t \geq T_f, T_f < T_{od}) I(l = \mathrm{RP}) \\ &= \tilde{C}_t^{\mathrm{Rad}} I(l = \mathrm{Rad}) + \tilde{C}_t^{\mathrm{RP}} I(l = \mathrm{RP}). \end{aligned} \tag{19}$$

The estimands of interest were the cumulative incidences of Failure we would have seen if treatment at baseline were randomized, and if we had ensured that the time to treatment initiation were a random draw from the marginal treatment time distribution of the cohort. These are the functions $\tilde{C}^{\mathrm{RP}}$ and $\tilde{C}^{\mathrm{RP}}$.

7.4.5. *Estimating the weights.* We calculated baseline weights using logistic regression, by fitting one marginal model, and one covariate-dependent model for the probability of receiving RP treatment. The covariate-dependent model for individual i was

$$\begin{aligned} \mathrm{logit}(p_i) = p_0 &+ p_{\mathrm{PSA}_{(8,15]}} I(\mathrm{PSA}_i \in (8, 15]) + p_{\mathrm{PSA}_{(15,22]}} I(\mathrm{PSA}_i \in (15, 22]) \\ &+ p_{\mathrm{age}>65} I(\mathrm{age}_i > 65) + p_{\mathrm{CAD}} I(\mathrm{CAD}_i = 1) + p_{\mathrm{HYP}} I(\mathrm{HYP}_i = 1) \\ &+ p_{\text{earlier cancer}} I(\text{earlier cancer}_i = 1) + p_{\mathrm{gleason}>6} I(\mathrm{gleason}_i > 6) \\ &+ p_{\text{T cat}} I(\text{T cat}_i = 1) + p_{< \text{ high school}} I(\mathrm{edu}_i < \text{ high school}) \\ &+ p_{>4 \text{ years college}} I(\mathrm{edu}_i > 4 \text{ years college}) + p_{\text{risk group}} I(\text{risk group}_i > 1), \end{aligned}$$

where CAD is Coronary artery disease, HYP is Hypertension, and risk group is a three-valued risk variable that is a combination of PSA, Gleason score, and T category. It takes the values 1 (low risk), 2 (intermediate risk) and 3 (high risk).

The estimated marginal probability of receiving RP treatment at baseline was denoted by $\hat{q}_i$. The propensity weight (8) for subject i was then estimated by inserting the predicted probabilities, which gave

$$\hat{R}_0^i = \frac{\hat{q}_i}{\hat{p}_i} I(\text{treat}_i = \text{RP}) + \frac{1-\hat{q}_i}{1-\hat{p}_i} I(\text{treat}_i = \text{Rad}),$$

where treat_i was the treatment subject i received.

The continuous-time treatment weights were estimated using the estimator (16) suggested in **Paper 3**. We thereby assumed that the observed time to treatment initiation in each treatment arm followed the additive hazard model

$$\begin{aligned}\alpha_t^i = \alpha_t^0 &+ \alpha_t^{\text{age}>65} I(\text{age}_i > 65) + \alpha_t^{\text{PSA}>5} I(\text{PSA}_i > 5) + \alpha_t^{\text{gleason}>6} I(\text{gleason}_i > 6) \\ &+ \alpha_t^{<\text{ high school}} I(\text{edu}_i < \text{ high school}) + \alpha_t^{>4 \text{ years college}} I(\text{edu}_i > 4 \text{ years college}) \\ &+ \alpha_t^{\text{risk group}} I(\text{risk group}_i > 1).\end{aligned}$$

This model was fitted separately for each treatment arm, i.e. for the patients who received Rad and the patients who received RP. We fitted a marginal model for the time to treatment in the pooled sample, i.e. a model for the time to treatment in the sample as a whole, regardless of covariates or treatment regimen. This model can be summarized by a marginal hazard $\tilde{\alpha}$. By inserting the fitted regressions as `aalen` objects into the function `makeContWeights` in the `ahw` package, we obtained continuous-time weight estimates $\hat{R}^i$ for each subject i. The final weights were estimated by $\hat{W}^i = \hat{R}_0^i \cdot \hat{R}^i$. We observed a minor instability, and we truncated the weights so that no weights were larger than seven in the final analysis. A weight trajectory plot is shown in Thesis Figure 7.

7.4.6. *Weighted analysis.* We calculated the weighted cause-specific cumulative hazards for Failure and Other death in each treatment arm. The estimands of interest were the cumulative incidences from (19), which are solutions of

$$\begin{pmatrix} \tilde{C}_t^l \\ \tilde{S}_t^l \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^t \begin{pmatrix} \tilde{S}_s^l & 0 \\ -\tilde{S}_s^l & -\tilde{S}_s^l \end{pmatrix} d \begin{pmatrix} \tilde{A}_s^{l,\text{Failure}} \\ \tilde{A}_s^{l,\text{Other death}} \end{pmatrix},$$

for group $l \in \{\text{Rad}, \text{RP}\}$. $\tilde{A}^{l,\text{Failure}}$ and $\tilde{A}^{l,\text{Other death}}$ are the hypothetical cumulative hazards (i.e. cumulative hazards under $\tilde{P}^{g_l}$) for the transition from state l to the Failure and Other death endpoints, while $\tilde{S}^l$ is the hypothetical survival function (i.e. the survival function under $\tilde{P}^{g_l}$). We estimated the hypothetical cumulative hazards by weighting the Nelson-Aalen estimator for each of the endpoints, i.e. a special case of the weighted additive hazard estimator (18). We furthermore estimated

the hypothetical cumulative incidences by utilizing plugin estimation as outlined in **Paper 3**. The estimating equation then took the form

$$\begin{pmatrix} \hat{\hat{C}}_t^l \\ \hat{\hat{S}}_t^l \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^t \begin{pmatrix} \hat{\hat{S}}_{s-}^l & 0 \\ -\hat{\hat{S}}_{s-}^l & -\hat{\hat{S}}_{s-}^l \end{pmatrix} d \begin{pmatrix} \hat{\hat{A}}_s^{l,\text{Failure}} \\ \hat{\hat{A}}_s^{l,\text{Other death}} \end{pmatrix}.$$

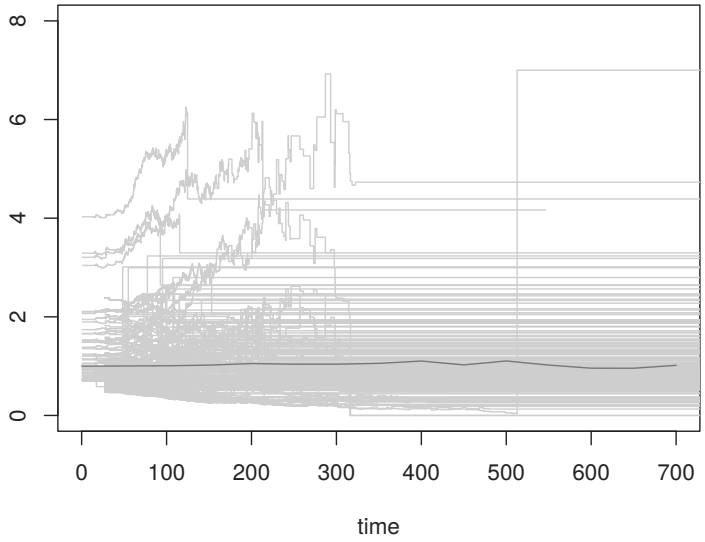
Due to the definition of the endpoints, the cumulative incidences could be very different shortly after diagnosis as many Failure events occurred early in the study period, e.g. after eight weeks and six months (see Section 7.4.2). On a longer time horizon, however, the cumulative incidences would depend less on details of the endpoint definition. When testing for differences in the treatment arms, we therefore put more emphasis on the events that happened a long time after diagnosis. This was done by choosing the function $K_t = t^{0.3}$, such that differences for larger t were emphasized in the Gray test statistic [31]

$$\hat{\hat{Z}}_{\mathcal{T}} = \int_0^{\mathcal{T}} K_t \cdot \Big(\frac{d\hat{\hat{C}}_t^{\text{RAD}}}{1 - \hat{\hat{C}}_{t-}^{\text{RAD}}} - \frac{d\hat{\hat{C}}_t^{\text{RP}}}{1 - \hat{\hat{C}}_{t-}^{\text{RP}}} \Big),$$

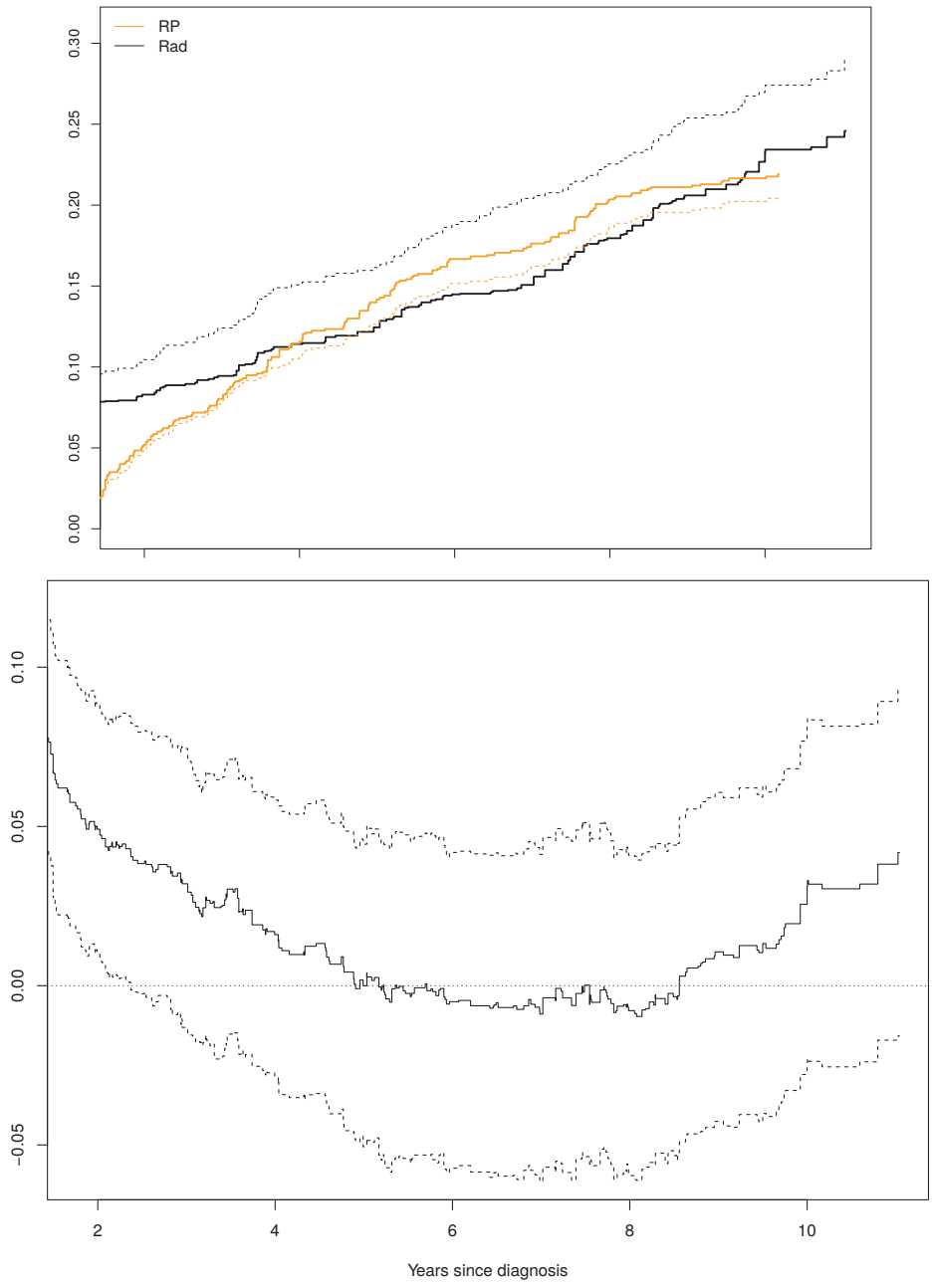
over the study period $[0, \mathcal{T}]$. We obtained p-values by bootstrapping the variance.

7.4.7. *Results.* We inspected the weighted and unweighted cumulative incidences, as well as the cumulative incidence difference along with 95% pointwise confidence intervals (shown in Thesis Figure 8). A visual inspection suggested that the groups were hard to distinguish. We performed a weighted Gray test to compare the cumulative incidences formally. Both the plot and the Gray test (p-value of 0.3310) indicated that the cumulative incidences were not different in the time period of interest. The weighted Gray test p-value was in contrast to the result we obtained from an analysis of the unweighted cumulative incidences. This naive analysis gave a Gray test p-value of 0.0083, which (misleadingly) indicated that the rates of Failure were different, and the Rad treated subjects had a significantly worse cumulative incidence.

7.4.8. *Clinical interpretation.* Our analysis suggests that the observed differences in rates of Failure are primarily due to spurious effects, i.e. systematic differences between the treatment groups at baseline and in the treatment initiation times. When accounting for these differences, we were unable to separate the marginal cumulative incidences. We take this to be the ideal marginal comparison.



THESIS FIGURE 7. A subset of the estimated weight trajectories $\{\hat{W}^i\}_i$. The average is shown in red. The time scale indicates days since diagnosis.



THESIS FIGURE 8. Upper panel: cumulative incidences of Failure under the two treatment regimens. The weighted analysis is shown in thick solid lines, while the unweighted naive analysis is shown in dotted lines. Lower panel: estimates of $G_t^{\mathrm{Rad}} - G_t^{\mathrm{RP}}$ from (19) with bootstrapped 95% confidence intervals based on a bootstrap sample of 3000.

8. Contributions, comments, and extensions

8.1. **Paper 1.**

8.1.1. *Contributions.* We describe a general method for modeling and inference on a range of parameters in survival analysis. To do this, we use hazard models as an intermediary step. Hazard models are well developed, but hazards can sometimes be difficult to interpret (see the discussion in Section 5.1). The methodology in this paper gives easy access to several other parameters in survival analysis and will be particularly useful in situations where hazards are difficult to interpret.

The methodology enables estimation of these parameters and their pointwise covariances using the general plugin procedure given by (10) and (12). These plugin estimators are implemented in a generic fashion in the `R` package `transform.hazards`; only the inputs $\hat{A}, \hat{X}_0, \hat{V}_0, F$, and the ∇F_j's are needed as input. Several worked examples that show how to use the code is found in Section 9, in the Supplementary material of **Paper 2**, and the package vignette.

8.1.2. *Comments.* The plugin estimation equations (10) and (12) can sometimes be solved explicitly, such that closed-form solutions can be expressed (this can e.g. be seen from the difference equations (13) and (14)). The true power of our estimation method is still in the inexplicit formulations since all parameters that solve (9) and their covariances can be estimated with the same procedure. Explicit solutions of (10) and (12), when they exist, can be calculated on a case by case basis only.

8.1.3. *Future work.* The ODE structure can be used further to interpret covariate-adjusted parameters; if the parameter X solves an ODE system driven by A, the covariate-adjusted parameter X^L that solves the same ODE system driven by the vector of cumulative hazards adjusted for L, which under the additive model reads

$$A = A^0 + LA^L.$$

Suppose the k'th element of X^L, $X^{L,k}$, is of interest, where L is a baseline covariate. Derived quantities such as $\Psi^L = X^{L+1,k}/X^{L,k}$ and $\Upsilon^L = X^{L+1,k} - X^{L,k}$ can be estimated jointly with X^L by augmenting the ODE system:

$$\begin{pmatrix} X_t^{L+1} \\ X_t^L \\ \Psi_t^L \\ \Upsilon_t^L \end{pmatrix} = \begin{pmatrix} X_0 \\ X_0 \\ 1 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} F(X_s^{L+1}) & (L+1)F(X_s^{L+1}) \\ F(X_s^L) & LF(X_s^L) \\ \frac{F^k(X_s^{L+1}) - F^k(X_s^L)\Psi_s^L}{X_s^{L,k}} & \frac{(L+1)F^k(X_s^{L+1}) - LF^k(X_s^L)\Psi_s^L}{X_s^{L,k}} \\ F^k(X_s^{L+1}) - F^k(X_s^L) & (L+1)F^k(X_s^{L+1}) - LF^k(X_s^L) \end{pmatrix} d \begin{pmatrix} A_s^0 \\ A_s^L \end{pmatrix},$$

where F^k is the k'th row of F. Ψ^l and Υ^l then give interpretations on the multiplicative and additive scale, of the impact one unit increase of L has on the parameter when it is adjusted to level $L = l$. Plugin estimators for Ψ^l and Υ^l can be implemented as part of the `transform.hazards` package.

One limitation of Theorem 1 and Theorem 2 is the assumption that F is Lipschitz on the image of X. This limits the study of parameters that solve (9), but when F is not Lipschitz. Examples of such parameters include ratios of cumulative hazards, the attributable fraction function [35], and the number needed to treat [36]. We are currently working on ideas for extending the plugin procedure in this paper, so that it can be used in several instances where (9) holds, but where the Lipschitz assumption is violated.

8.2. **Paper 2.**

8.2.1. *Contributions.* Standard hypothesis testing in survival analysis, such as the rank tests, are based on comparing hazards. Since hazards can be difficult to interpret (see the discussion in Section 5.1), this could lead to unfortunate consequences. Moreover, there is sometimes a mismatch between the null hypothesis and the actual research question. In particular, the rank tests are in practice sometimes used to test hypotheses that are different from the rank hypothesis.

We develop a general procedure for testing survival analysis parameters pointwise in time, and demonstrate that our method performs satisfactory using simulations. We also show using simulations that the rank tests have poor performance when testing our null hypothesis. An application on colon cancer data is considered in Section 6. The analysis is provided as a worked example in the Supplementary material, where we define the required ODE system, perform the analysis using the `transform.hazards` package, and calculate the test statistic (15). This may be useful for researchers that want to use the software.

Our testing procedure is derived from the general results in **Paper 1** and is therefore valid in many situations. The central requirement is that the ODE (9) is satisfied.

8.2.2. *Comments.* Rather than just performing the test at a pre-specified time point t_0, it is possible to plot the whole estimated trajectory with 95% pointwise confidence intervals. For one-dimensional parameters, one can plot $\hat{X}_t^1 - \hat{X}_t^2 \pm 1.96\sqrt{\hat{V}_t^{X^1-X^2}}$ for all t in the study period, where $\hat{V}^{X^1-X^2}$ is the plugin variance of $\hat{X}_t^1 - \hat{X}_t^2$. Such a plot would give a more detailed picture of $X_t^1 - X_t^2$ when $t \neq t_0$. Our pointwise hypothesis is then accepted at t_0 if zero is contained in the estimated confidence interval, and rejected otherwise. However, we must guard ourselves against cherry-picking, by choosing t_0 before such curves are plotted, and not afterward.

Another reservation should be made when it comes to cherry picking; one should not test several parameters using (15) and only report selected p-values. This can be avoided by reporting p-values from all the performed tests, or by employing our test statistic (15) to test differences between all parameters that are studied jointly.

We could also write down a test statistic that compares parameters at two or more time points, i.e. testing the hypothesis that $X_t^1 = X_t^2$ for $t \in \{t_0, t_1\}$, as we can write down plugin estimators for the covariance between $\hat{X}_{t_0}^1 - \hat{X}_{t_0}^2$ and $\hat{X}_{t_1}^1 - \hat{X}_{t_1}^2$. This test statistic would e.g. enable joint comparison of one and five year survival between two groups. We still think that the current test statistic (15) is general enough for many applications.

8.2.3. *Future work.* It is of general interest to estimate confidence bands for the plugin parameter estimators from **Paper 1**. Confidence bands would e.g. allow for testing over the full follow-up period, and not only for pre-specified time points. However,

development of such bands would require other techniques from the ones developed in **Paper 1**.

Extending on the comments in Section 8.1.3, we can also perform hypothesis testing on the quantities $\Psi^l_{t_0}$ and $\Upsilon^l_{t_0}$ at t_0, to check whether a one unit increase at level l is significant at t_0. Additionally, one can plot the plugin estimates of e.g. Υ^l along with 95% pointwise confidence intervals, to see how the impact of one unit increase of L under level $L = l$ varies over time. Similar tests and plots could be created with confidence bands if we had them.

8.3. **Paper 3.**

8.3.1. *Contributions.* We develop the continuous-time MSMs [10] in terms of

- Providing the consistent continuous-time weight estimator (16).
- Proving consistency results for weighted additive hazard regression that takes into account that the weights are estimated.
- Showing that the plugin estimation strategy from **Paper 1** can be used for estimating causal parameters that solve ODEs like (9).
- Describing software that calculates (16), and showing how a causal survival analysis can be performed in practice, explained through an example.

8.3.2. *Comments.* The stabilized inverse probability weights (6) that are used for estimating the discrete time MSMs (Section 6.3) can be viewed as approximations of the continuous-time weights (7) that are used for estimating the continuous-time MSMs (Section 6.4); as the time discretization is refined, the former will approach the latter. A heuristic argument is made below.

The discrete weight approximates the continuous-time likelihood ratio

We consider the time discretization $\{t_k\}_k$ from Section 6.3, and inspect the limit as the discretization times are refined. By performing algebraic manipulation of the logarithm of (6) divided by (7) we get

(20)
$$\log\Big(\frac{w_t^i}{R_t^i}\Big) = \sum_{k:t_k\leq t} \log\Big(\frac{q_k^i}{p_k^i}\Big) I(i,k) - \sum_{s\leq t} \log\Big(\frac{\tilde{\lambda}_s^i}{\lambda_s^i}\Big)\Delta N_s^i$$

(21)
$$+ \sum_{k:t_k\leq t} \Big(\log\big(1-q_k^i\big) - \log\big(1-p_k^i\big)\Big)(1-I(i,k)) - \int_0^t (\lambda_s^i - \tilde{\lambda}_s^i)ds,$$

where $q_k^i = P(A_k = 1|\bar{A}_{k-1} = \bar{a}_{k-1}^i), p_k^i = P(A_k = 1|\bar{A}_{k-1} = \bar{a}_{k-1}^i, \bar{L}_{k-1} = \bar{l}_{k-1}^i)$, and $I(i,k)$ is 1 if i is treated in $[t_{k-1}, t_k)$ and 0 otherwise.

We now introduce the equidistant "grid" $\{t_k^n\}_{k,n}$ so that $\Delta^n = t_k^n - t_{k-1}^n$ and $\lim_{n\longrightarrow\infty} \Delta^n = 0$. The probabilities $q^{i,n}, p^{i,n}$ will now depend on n, and the ratio $q^{i,n}/p^{i,n}$ approaches $\tilde{\lambda}^i/\lambda^i$, so that the right hand side of (20) vanishes in the limit. Furthermore, to leading order we have $\log(1-x) = -x$ when $x \approx 0$, and viewing the sum in (21) as a Riemann sum, we get that (21) also vanishes in the limit. This is because the sum approaches an integral and $q^{i,n}/\Delta^n, p^{i,n}/\Delta^n$ respectively approach $\tilde{\lambda}^i$ and λ^i.

The above argument suggests that one can use IPTW estimators based on (pooled) logistic regression, as in [9, eq. (17)], for estimating the continuous-time weights (7). Such IPTW estimators will be consistent if the number of discretization intervals grows as the sample size increases. The discretization interval width for the IPTWs may be viewed as a bandwidth parameter, similar to our bandwidth parameter κ from (16). From Thesis Figure 6 it can be seen that our weight estimator performs better when the underlying processes are time-continuous, at least concerning bias.

We have considered interventions where the hypothetical treatment intensity is given by $\tilde{\lambda}_t = E[\lambda_t | \mathcal{F}_{t-}^{\mathcal{V}_0}]$, i.e. when the processes and variables that are confounders are marginalized out. We have referred to them as "randomizing" interventions because the intensity $\tilde{\lambda}$ is not a function of the (time-dependent) confounders. However, different interventions could be of interest. One could e.g. be interested in hypothetical scenarios with a doubled treatment initiation rate. The results of this paper can be used to investigate several such scenarios without major changes.

We used a simple smoothing approach for estimating θ^i for our continuous-time weight estimator (16). The estimator for θ^i, (17), is a ratio between discrete derivatives, or kernel smoothed cumulative hazard estimates. This kernel smoothing could be performed in several ways, as long as the smoother is based on past information. Several well-known kernel smoothers could be used for this task, e.g. Gaussian, nearest neighbors, Epanechnikov, or local regression, all of which require specifications of at least one bandwidth parameter. When developing our estimator, we performed a range of simulations for the different smoothers and varying values of the bandwidth. None of the kernel smoothers were found to be noticeably better than the others for the tests we conducted, so we ended up using the simplest option.

8.3.3. *Future work.* A major result of **Paper 1** was that we were able to write down a general covariance estimator (12). It would be beneficial to also have an SDE estimator for the covariance when the cumulative hazard estimates are weighted. This would allow for estimation of confidence intervals and hypothesis testing pointwise in time for causal parameters that solve (9). It would moreover be of interest to find confidence bands when the cumulative hazard estimates are weighted. This would enable other kinds of hypothesis testing; see Section 8.2.3.

Our estimator (17) can run into problems in some situations, as the cumulative hazard estimates from the additive model can have negative increments. This could lead to numerical instability, especially in the case of multiple treatments, since numerous evaluations of the estimated θ^i's are needed in the estimator (16). We may want to consider other estimators for θ^i to improve the stability of our weight estimator.

8.4. **Paper 4.**

8.4.1. *Contributions.* We contribute to the debate [32,34] on the relative effectiveness of radical prostatectomy and radiation therapy among subjects with localized prostate cancer. Using an MSM, we find that the two treatment types have comparable failure rates up to eleven years after diagnosis.

We furthermore show how continuous-time MSMs can be utilized in practice. In the Supplementary material, we give details about model fitting for weight calculation, weight truncation, and choosing the bandwidth parameter. This information may be useful for other researchers that want to use the methodology.

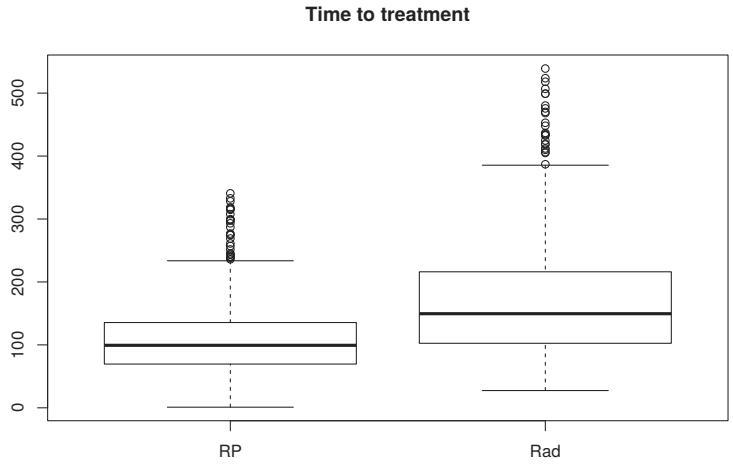
8.4.2. *Comments.* The results from our analysis shown in Sections 7.4.7 and 7.4.8 suggest that there is a large group of diagnosed subjects that should worry more about side effects, and less about prognosis when choosing treatment, as the failure rates for the two treatment types seem to be comparable. However, our targeted intervention is not entirely suitable for answering this query. This is because there is an actual difference in the treatment initiation times that depends on the treatment regimen. This can be seen in Thesis Figure 9; the average time to treatment is around 100 days in the RP group, and 150 days in the Rad group. The decision-theoretic problem "An average male was just diagnosed with prostate cancer, which treatment should he choose?" would more appropriately be answered by modifying the continuous-time intervention slightly; by enforcing the time to treatment initiation to be a random draw conditional on the assigned treatment group.

Formally, the Gray test is a rank test that compares subdistribution hazards obtained by modifying the at-risk set; subjects that experience the event of interest are in the at-risk set until they experience the event, while subjects that in reality experience the competing events remain in the modified at risk set indefinitely [31]. The Gray test therefore compares the hazards of a hypothetical population where the subjects are prevented from having events other than the one of interest. Using the Gray test may seem to be in conflict with one of the major points in **Paper 2**, which is that hypothesis testing should be performed on parameters that have clear interpretations. However, the Gray null hypothesis can be understood in another way, as there is a known relationship between the subdistribution hazards and the cumulative incidences; if C is the cumulative incidence, and γ is the subdistribution hazard we have $\gamma_t dt = dC_t/(1 - C_t)$. Therefore, the Gray null hypothesis is under mild conditions equivalent to equality of the cumulative incidences throughout the follow-up period. The Gray null hypothesis is therefore not harder to interpret than equality of the cumulative incidences throughout the follow-up period. Still, assessing equality at all times may be different from the hypothesis of primary interest.

The cumulative incidences will as mentioned in Section 7.4.6 be affected by details of the Failure endpoint definition for small times. These details could influence the Gray test outcome, because the test compares subdistribution hazards at all times,

including times shortly after diagnosis. A pointwise test after e.g. five years of follow-up would be less influenced by these details. Inspecting Thesis Figure 8 we see that such a pointwise test would also fail to detect significant differences since zero is contained in the confidence interval at all times after around two years of follow-up.

8.4.3. *Future work.* We want to use the continuous-time MSM methods on similar clinical problems. We would also like to make the software more streamlined so that our methods get more accessible to other researchers.



THESIS FIGURE 9. Boxplot of the treatment times for the two prostate cancer treatment groups. The average time to treatment for the whole cohort was 137 days.

9. Software

A short introduction to the main functions of the `R` packages `transform.hazards` and `ahw` developed during this thesis is provided. They are both freely available for anyone to use in the repository `github.com/palryalen`.

9.1. `transform.hazards`. Recall from **Paper 1** that we are interested in assessing parameters that solve differential equations driven by cumulative hazards A, i.e.

$$X_t = X_0 + \int_0^t F(X_s) dA_s,$$

where F is Lipschitz, two times continuously differentiable, and satisfies a linear growth bound. We write $F = (F_1, F_2, \cdots)$, so that F_j is the j'th column of F. The main function `pluginEstimate` in this package estimate such parameters nonparametrically using cumulative hazard estimates from Aalen's additive hazard model. `pluginEstimate` has the following input:

- `n`: the number of subjects.
- `hazMatrix`: the ordered increments of $\hat{A}$; see Section 7.1.4.
- `F`: the integrand function F.
- `JacobianList`: the Jacobian matrices of the columns of F, i.e. $\nabla F_1, \nabla F_2, \cdots$ as a list.
- `X0`: the initial values of X.
- `V0`: the initial values of V.
- `isLebesgue`: indices of X that correspond to regular dt integrals (optional).

We demonstrate how this package can be used through examples below.

The package can be downloaded and installed manually from the GitHub repository, but a simpler option is to use the `devtools` package. If `devtools` is installed, one can run the following command for installing the software:

```
devtools::install_github("palryalen/transform.hazards",
                         build_vignettes=TRUE)

library(transform.hazards)
```

In encourage the reader to inspect the package vignette. The vignette contains many worked examples that show how the analysis can be performed on up-to-date versions of the software. It is made accessible by the `build_vignettes` argument. Build and display the vignette with the command

```
browseVignettes("transform.hazards")
```

9.1.1. *Survival.* The survival function S solves a differential equation with initial value 1,

$$S_t = 1 - \int_0^t S_s dA_s,$$

where A is the cumulative hazard for death.

We estimate the survival function from a sample of $n = 300$ subjects, stored in the `data.frame` `data1`. An additive hazard model is fitted to the data by calling the `aalen` function. To use this function we load the `timereg` package. We can then extract the cumulative hazard increments:

```
library(timereg)
aaMod1 <- aalen(Surv(startTimes,stopTimes,censIndicator)~1,
                data = data1)

dA_est1 <- diff(c(0,aaMod1$cum[,2]))
hazMatrix <- matrix(dA_est1,nrow=1)
```

In this case, F takes the simple form $F(x) = -x$, and its Jacobian is therefore $\nabla F(x) = -1$. F and ∇F must be provided as matrix-valued functions along with the initial values X_0 and V_0. Note that ∇F must be provided in a list:

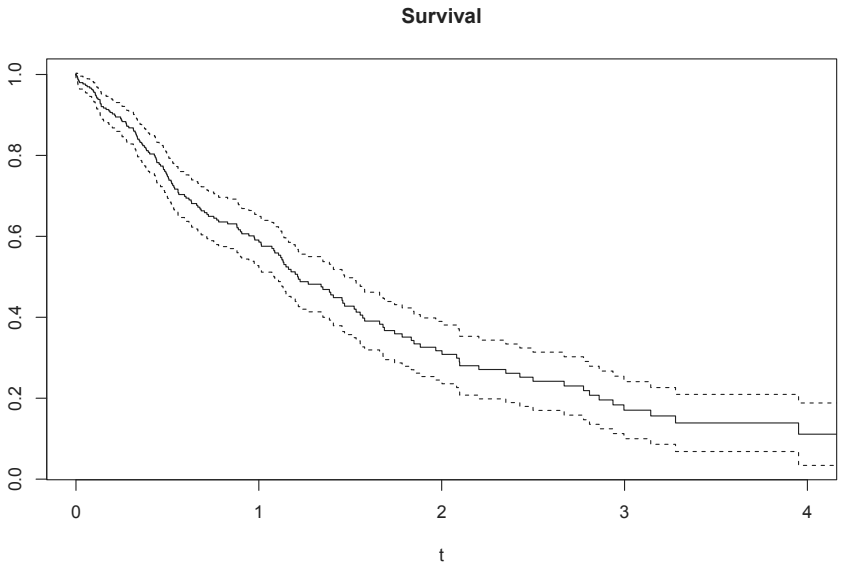
```
F_survival <- function(X)matrix(-X,nrow=1,ncol=1)
F_survival_JacobianList <- list(function(X)matrix(-1,nrow=1,ncol=1))
X0 <- matrix(1,nrow=1,ncol=1)
V0 <- matrix(0,nrow=1,ncol=1)
```

We obtain plugin estimates of S using the function call

```
param <- pluginEstimate(300, hazMatrix, F_survival,
                        F_survival_JacobianList, X0, V0)
```

The variable `param` is a list containing `X` (a matrix containing the plugin estimates of the parameter) and `covariance` (an array containing the plugin estimates of the covariance of $\hat{X}$). We can now plot the results with estimated 95 % confidence intervals:

```
t1 <- aaMod1$cum[,1]
plot(t1,param$X,type="s",xlim=c(0,4),xlab="t",ylab="",
     main="Survival")
lines(t1,param$X+1.96*sqrt(param$covariance[1,1,]),type="s",lty=2)
lines(t1,param$X-1.96*sqrt(param$covariance[1,1,]),type="s",lty=2)
```



THESIS FIGURE 10. Survival plugin estimate from Section 9.1.1, along with 95% pointwise confidence intervals.

9.1.2. *Restricted mean survival.* We estimate the restricted mean survival function R based on the same data set. It solves the system

$$\begin{pmatrix} S_t \\ R_t \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} -S_s & 0 \\ 0 & S_s \end{pmatrix} d \begin{pmatrix} A_s \\ s \end{pmatrix},$$

where S is the survival function. Here, the colums of F, F_1 and F_2, are given by

$$F_1(x_1, x_2) = \begin{pmatrix} -x_1 \\ 0 \end{pmatrix}, F_2(x_1, x_2) = \begin{pmatrix} 0 \\ x_1 \end{pmatrix}.$$

The Jacobian matrices are therefore respectively

$$\nabla F_1(x_1, x_2) = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \nabla F_2(x_1, x_2) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

We define these two functions along with initial values below.

```
F_restrict <- function(X)matrix(c(-X[1],0,0,X[1]),nrow=2)
F_restrict_JList <- list(function(X)matrix(c(-1,0,0,0),nrow=2),
                         function(X)matrix(c(0,1,0,0),nrow=2))
X0_restrict <- matrix(c(1,0),nrow=2)
V0_restrict <- matrix(0,nrow=2,ncol=2)
```

The restricted mean survival is a "regular" (i.e. Lebesgue) integral, and we must provide time increments (recall the discussion in Section 7.1.4). We choose the time interval $[0, 4]$ in 10^4 increments:

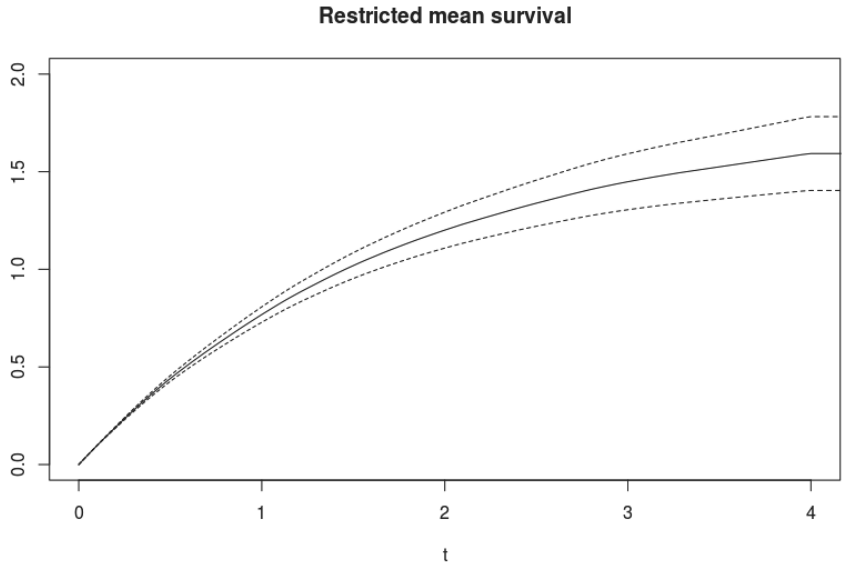
```
fineTimes <- seq(0,4,length.out = 1e4+1)
t2 <- sort(unique(c(fineTimes,t1)))
hazMatrix <- matrix(0,nrow=2,ncol=length(t2))
hazMatrix[1,match(t1,t2)] <- dA_est1
hazMatrix[2,] <- diff(c(0,t2))
```

We obtain plugin estimates using the call

```
param <- pluginEstimate(300, hazMatrix, F_restrict,
                        F_restrict_JList,X0_restrict,
                        V0_restrict,isLebesgue = 2)
```

Note that `param` also contains the plugin estimates of the survival S considered in Section 9.1.1. We plot the results, shown in Thesis Figure 11.

```
restrictCov <- param$covariance[2,2,]
plot(t2,param$X[2,],type="s",xlim=c(0,4),ylim=c(0,2),xlab="t",
    ylab="",main="Restricted mean survival")
lines(t2,param$X[2,]+1.96*sqrt(restrictCov),type="s",lty=2)
lines(t2,param$X[2,]-1.96*sqrt(restrictCov),type="s",lty=2)
```



THESIS FIGURE 11. Restricted mean survival plugin estimate from Section 9.1.2, along with 95% pointwise confidence intervals.

9.2. `ahw`. The main function in this package is `makeContWeights`, which estimates the weights (16) that are used in the weighted additive hazard regression (18). The function has the following input variables:

- `faFit`: the `aalen` fit for the observational hazard.
- `cfaFit`: the `aalen` fit for the hypothetical hazard.
- `dataFr`: `data.frame` or `data.table` on long format.
- `atRiskState`: at risk state(s) for treatment.
- `eventState`: treatment state.
- `stopTimeName`: name of the column with the transition times.
- `startStatusName`: name of the column with the starting states.
- `endStatusName`: name of the column with the end states.
- `idName`: name of column that identifies individuals.
- `b`: bandwidth parameter, equal to $1/\kappa$ in (16).
- `weightRange`: weight truncation interval (optional).
- `willPlotWeights`: weight trajectory plot indicator (optional).

If the `devtools` package is installed, one can run

```
devtools::install_github("palryalen/ahw")
library(ahw)
```

I also plan to make a GitHub vignette for this package to make the software more accessible to other researchers. When the vignette is complete, it can be loaded by adding the argument `build_vignettes == T` in the `install_github` call, and inspected using the command

```
browseVignettes("ahw")
```

9.2.1. *Randomizing treatment.* Consider subjects who are diagnosed with some disease, and receive treatment based on a binary baseline variable L. After being treated, the subjects are at risk of dying. We refer to this as the observational scenario.

Suppose we are interested in a hypothetical situation where, contrary to what we observe, the time to treatment is a random draw from the marginal treatment initiation times in the population. If α^A is the time to treatment hazard for the observational scenario, and $\tilde{\alpha}^A$ is the time to treatment hazard in the hypothetical

situation, the considered hazards take the form

$$\alpha_t^A = \alpha_t^0 + L\alpha_t^L$$
$$\tilde{\alpha}_t^A = \tilde{\alpha}_t^0,$$

where $\tilde{\alpha}_t^0 = E_L[\alpha_t^A]$. We simulate a data set from the observational scenario, and store it in the `data.frame fr1`. We display data for the first three subjects in the data set below

```
##   id      from        to L from.state to.state
## 1  1 0.0000000 0.0127410 1       diag    treat
## 2  1 0.0127410 0.7039590 1      treat    death
## 3  2 0.0000000 1.8553269 0       diag    treat
## 4  2 1.8553269 2.0091635 0      treat    death
## 5  3 0.0000000 0.4069482 1       diag    treat
## 6  3 0.4069482 0.6604173 1      treat    death
```

Subject 1 receives treatment at time 0.0127 and dies at time 0.704. Subject 2 receives treatment at time 1.855, and dies at time 2.009 and so on. We fit the observational and hypothetical treatment models below:

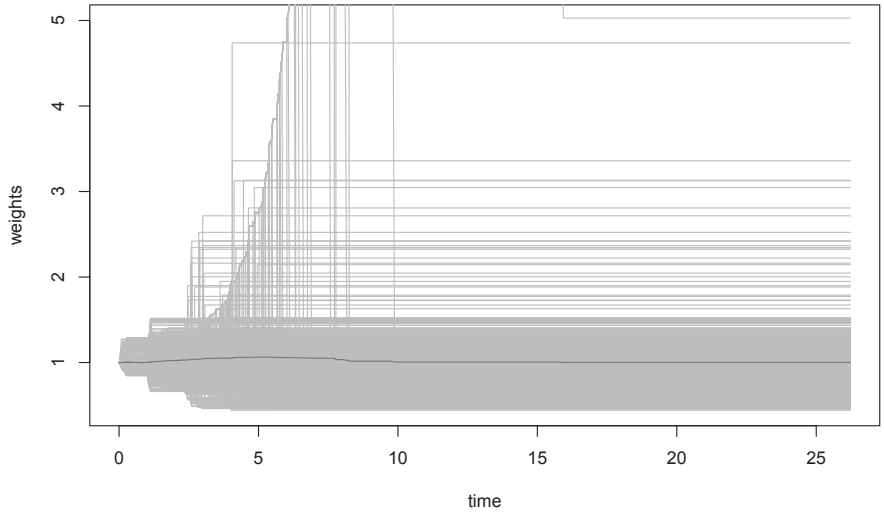
```
faFit <- aalen(Surv(from,to,to.state == "treat") ~ 1 + L,
               data=fr1[fr1$from.state=="diag",])
cfaFit <- aalen(Surv(from,to,to.state == "treat") ~ 1,
               data=fr1[fr1$from.state=="diag",])
```

We declare the input variables needed for using `makeContWeights`. The bandwidth parameter b is set to 0.3.

```
dataFr <- fr1
atRiskState <- "diag"
eventState <- "treat"
stopTimeName  <- "to"
startStatusName <- "from.state"
endStatusName <- "to.state"
idName <- "id"
b <- 0.3
```

Finally, we use the main function for weight estimation. We choose not to truncate weights, but want to plot the weight trajectories. The output is stored in `frame`.

```
frame <- makeContWeights(faFit, cfaFit, dataFr, atRiskState,
                         eventState,stopTimeName,startStatusName,
                         endStatusName,idName, b,
                         willPlotWeights = T)
```



THESIS FIGURE 12. Plot resulting from using the option `willPlotWeights = T`. The weight trajectories are well behaved in this simple example, with mean close to 1.

`frame` is an expanded `data.table`, where each subject has multiple rows as long as he is at risk of receiving treatment. We display the first 16 rows below:

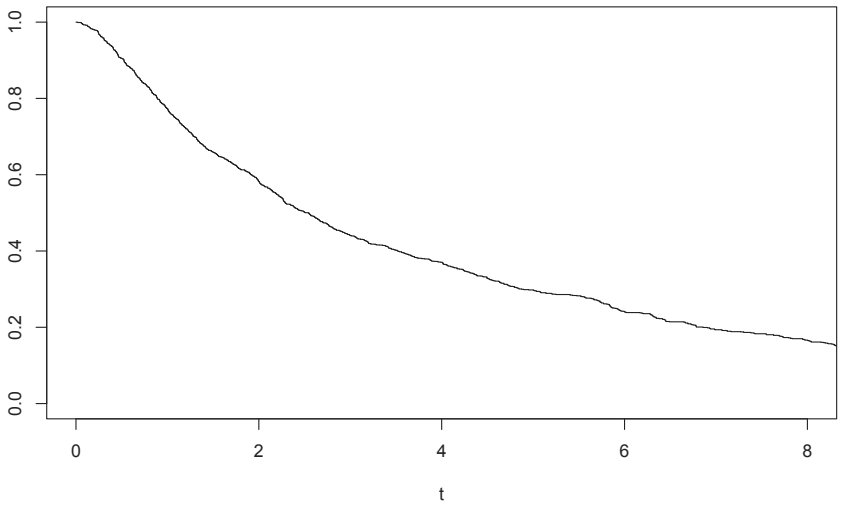
```
##    id        from          to L from.state to.state  weights
## 1   1 0.000000000 0.001265808 1       diag        0 1.000000
## 2   1 0.001265808 0.004741339 1       diag        0 0.999000
## 3   1 0.004741339 0.005543079 1       diag        0 0.999999
## 4   1 0.005543079 0.005807163 1       diag        0 0.998999
## 5   1 0.005807163 0.007286518 1       diag        0 0.998000
## 6   1 0.007286518 0.008172427 1       diag        0 0.997002
## 7   1 0.008172427 0.009478025 1       diag        0 0.997999
## 8   1 0.009478025 0.011068796 1       diag        0 0.997001
## 9   1 0.011068796 0.011178575 1       diag        0 0.997000
## 10  1 0.011178575 0.011324917 1       diag        0 0.997997
## 11  1 0.011324917 0.012741341 1       diag    treat 0.997997
## 12  1 0.012741341 0.703958983 1      treat    death 1.003302
## 13  2 0.000000000 0.001265808 0       diag        0 1.000000
## 14  2 0.001265808 0.004741339 0       diag        0 1.001000
## 15  2 0.004741339 0.005543079 0       diag        0 0.999999
## 16  2 0.005543079 0.005807163 0       diag        0 1.000999
```

We see that subject 1 is now recorded at several time points prior to him receiving treatment. These are the treatment times in the population that are smaller than subject 1's treatment time of 0.012741. In the `weights` column, there is a time-updated value of the weights just before the times in the `to` column, i.e. the treatment times in the population that are smaller than 0.012741. In particular, subject 1's weight is equal to 1 at 0.001266, the first treatment time in the population. Once the subjects are treated, they are no longer at risk of being treated, and the weight process is constant - see expression (7). There is therefore no need for a time-updated weight estimate after a subject has been treated.

We are now able to perform a weighted outcome regression to assess the hypothetical scenario. We estimate the marginal cumulative hazard for death we would see if, contrary to fact, the time to treatment were a random draw from $\tilde{\alpha}^A$, the marginal treatment initiation distribution in the sample as a whole. This is a weighted Nelson-Aalen estimator. We call a marginal `aalen` regression on the data set `frame`, and use the option `weights = frame$weights` to perform a weighted regression.

```
outMod <- aalen(Surv(from,to,to.state == "death")~1,data=frame,
        weights = frame$weights)
```

Having the weighted regression at hand, we can e.g. estimate the survival function under this hypothetical scenario using the `transform.hazards` package on the weighted cumulative hazard estimates from `outMod`. We do this, and plot the results in Thesis Figure 13. Note that the plugin variance should not be used, as we have not been able to account for the variance that is induced by weighting; see the discussion in Section 8.3.3.



THESIS FIGURE 13. Plot of the hypothetical survival function in example 9.2.1

References

[1] D. Cox, "Regression models and life tables," *Journal of the Royal Statistic Society*, vol. B, no. 34, pp. 187–202, 1972.

[2] O. Aalen, Ø. Borgan, and H. Gjessing, *Survival and Event History Analysis: A Process Point of View (Statistics for Biology and Health)*. Springer, 2008.

[3] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part i: Basic concepts and first analyses," *British Journal Of Cancer*, 2003.

[4] D. Cox and D. Oakes, "Analysis of survival data," *Chapman&Hall, London*, 1984.

[5] M. Hernán, "The hazards of hazard ratios," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 13, 2010.

[6] O. Aalen, R. Cook, and K. Røysland, "Does cox analysis of a randomized survival study yield a causal treatment effect?," *Lifetime data analysis*, vol. 21, no. 4, pp. 579–593, 2015.

[7] J. Robins and S. Greenland, "The probability of causation under a stochastic model for individual risk," *Biometrics*, pp. 1125–1138, 1989.

[8] M. Stensrud, M. Valberg, K. Røysland, and O. Aalen, "Exploring selection bias by causal frailty models: The magnitude matters," *Epidemiology*, vol. 28, no. 3, pp. 379–386, 2017.

[9] J. Robins, M. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," 2000.

[10] K. Røysland, "A martingale approach to continuous-time marginal structural models," *Bernoulli*, 2011.

[11] J. Jacod and A. Shiryaev, *Limit theorems for stochastic processes*, vol. 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag, second ed., 2003.

[12] E. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[13] H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, H. Skali, S. Solomon, S. Jacobus, M. Hughes, M. Packer, and L.-J. Wei, "Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis," *Journal of Clinical Oncology*, vol. 32, no. 22, pp. 2380–2385, 2014. PMID: 24982461.

[14] P. Royston and M. Parmar, "The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt," *Stat Med*, vol. 30, pp. 2409–2421, Aug 2011.

[15] M. Gassama, J. Benichou, L. Dartois, and A. Thiebaut, "Comparison of methods for estimating the attributable risk in the context of survival analysis," *BMC Med Res Methodol*, vol. 17, p. 10, 01 2017.

[16] O. Aalen, "A linear regression model for the analysis of life times.," *Statistics in Medicine*, 1989.

[17] O. Aalen, *Mathematical Statistics and Probability Theory: Proceedings, Sixth International Conference, Wisła (Poland), 1978 (Lecture Notes in Statistics)*. Springer, 1980.

[18] T. Martinussen, S. Vansteelandt, and P. Kragh Andersen, "Subtleties in the interpretation of hazard ratios," *arXiv e-prints*, p. arXiv:1810.09192, Oct. 2018.

[19] L. Sun and Z. Zhang, "A class of transformed mean residual life models with censored survival data," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 803–815, 2009.

[20] J. Pearl, *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press, 2nd ed., 2009.

[21] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[22] I. Shpitser and J. Pearl, "Identification of joint interventional distributions in recursive semi-markovian causal models," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pp. 1219–1226, AAAI Press, 2006.

[23] Y. Huang and M. Valtorta, "On the completeness of an identifiability algorithm for semi-markovian models," *Annals of Mathematics and Artificial Intelligence*, vol. 54, pp. 363–408, Dec 2008.

[24] J. Robins, "A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect.," *Mathematical modeling*, 1986.

[25] J. Robins, "The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies.," *Health Service Research Methodology: A Focus on AIDS*, 1989.

[26] S. Cole and M. Hernan, "Constructing inverse probability weights for marginal structural models," *Am. J. Epidemiol.*, vol. 168, pp. 656–664, Sep 2008.

[27] O. Aalen, K. Røysland, J. Gran, R. Kouyos, and T. Lange, "Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms," *Stat Methods Med Res*, vol. 25, pp. 2294–2314, 10 2016.

[28] J. Jacod, "Multivariate point processes: Predictable projection, radon-nikodym derivatives, representation of martingales.," *Probability Theory and Related Fields*, 1975.

[29] K. Røysland, O. Aalen, T. Lange, M. Nygaard, P. Ryalen, and V. Didelez, "Causal reasoning in survival analysis: Re-weighting and local independence graphs," 2018(in press).

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

[31] R. Gray, "A class of k-sample tests for comparing the cumulative incidence of a competing risk," *The Annals of Statistics*, vol. 16, no. 3, pp. 1141–1154, 1988.

[32] C. Wallis, A. Glaser, J. Hu, H. Huland, N. Lawrentschuk, D. Moon, D. Murphy, P. Nguyen, M. Resnick, and R. Nam, "Survival and complications following surgery and radiation for localized prostate cancer: An international collaborative review," *European Urology*, vol. 73, no. 1, pp. 11–20, 2017.

[33] A. Tree and D. Dearnaley, "Randomised controlled trials remain the key to progress in localised prostate cancer," *European Urology*, vol. 73, pp. 21—22, 2017.

[34] F. Hamdy, J. Donovan, J. Lane, M. Mason, C. Metcalfe, P. Holding, M. Davis, T. Peters, E. Turner, R. Martin, J. Oxley, M. Robinson, J. Staffurth, E. Walsh, P. Bollina, J. Catto, A. Doble, A. Doherty, D. Gillatt, R. Kockelbergh, H. Kynaston, A. Paul, P. Powell, S. Prescott, D. Rosario, E. Rowe, and D. Neal, "10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer," *New England Journal of Medicine*, vol. 375, no. 15, pp. 1415–1424, 2016.

[35] L. Chen, D. Lin, and D. Zeng, "Attributable fraction functions for censored event times," *Biometrika*, vol. 97, no. 3, pp. 713–726, 2010.

[36] D. Altman and P. Andersen, "Calculating the number needed to treat for trials where the outcome is time to an event," *BMJ*, vol. 319, no. 7223, pp. 1492–1495, 1999.

10. Papers

II

ON NULL HYPOTHESES IN SURVIVAL ANALYSIS.

MATS J. STENSRUD, KJETIL RØYSLAND AND PÅL C. RYALEN

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway

Abstract. The conventional nonparametric tests in survival analysis, such as the log-rank test, assess the null hypothesis that the hazards are equal at all times. However, hazards are hard to interpret causally, and other null hypotheses are more relevant in many scenarios with survival outcomes. To allow for a wider range of null hypotheses, we present a generic approach to define test statistics. This approach utilizes the fact that a wide range of common parameters in survival analysis can be expressed as solutions of differential equations. Thereby we can test hypotheses based on survival parameters that solve differential equations driven by cumulative hazards, and it is easy to implement the tests on a computer. We present simulations, suggesting that our tests perform well for several hypotheses in a range of scenarios. Finally, we use our tests to evaluate the effect of adjuvant chemotherapies in patients with colon cancer, using data from a randomised controlled trial.

KEY WORDS: Causal inference; Hazards; Hypothesis testing; Failure time analysis.

Date: January 29, 2019.

1. Introduction

The notion of hazards has been crucial for the development of modern survival analysis. Hazards are perhaps the most natural parameters to use when fitting statistical models to time-to-event data subject to censoring, and hazard functions were essential for the development of popular methods like Cox regression and rank tests, which are routinely used in practice.

In the field of causal inference, however, there is concern that many statisticians just do advanced 'curve fitting' without being careful about the interpretation of the parameters that are reported [1, 2, 3]. This criticism can be directed to several areas in statistics. In this spirit, we think that statisticians in general should pay particular attention to effect measures with clear-cut causal interpretations.

In survival analysis, it has been acknowledged that interpreting hazards as effect measures is delicate, see e.g. [4] and [5]. This contrasts the more traditional opinion, in which the proportional hazards model is motivated by the 'simple and easily understood interpretation' of hazard ratios [6, 4.3.a]. A key issue arises because the hazard, by definition, is conditioned on previous survival. If we consider causal diagrams [3, 2], it is clear that we condition on a 'collider' that opens a non-causal pathway from the exposure through any unobserved heterogeneity into the event of interest, see [4, 7, 8]. Since unobserved heterogeneity is present in most practical scenarios, even in randomized trials, the conditioning means that the hazards are fundamentally hard to interpret causally [4, 5].

Although we must be careful about assigning causal interpretations to hazards, we do not claim that hazards are worthless. On the contrary, hazards are key elements in the modelling of other parameters that are easier to interpret, serving as building blocks.

This point of view is also found in [2, 17.1]: "..., the survival analyses in this book privilege survival/risk over hazard. However, that does not mean that we should ignore hazards. The estimation of hazards is often a useful intermediate step for the estimation of survivals and risks." Indeed, we have recently suggested a generic method to estimate a range of effect measures in survival analysis, utilizing differential equations driven by cumulative hazards [9].

Nevertheless, the conventional hypothesis tests in survival analysis are still based on hazards. In particular the rank tests [10], including the log-rank test, are based on the null hypothesis

$$\mathbf{H}_0\colon\ \alpha_t^1 = \alpha_t^2 \text{ for all } t \in [0, \mathcal{T}], \tag{1}$$

where α_t^i is the hazard in group i. Formulating such hypotheses in a practical setting will often imply that we assign causal interpretations to these hazard functions. In the simplest survival setting this is not a problem, as there is a one-to-one relationship between hazards and the survival curves, and a null hypothesis comparing two or more survival curves is straightforward. In more advanced settings, e.g. scenarios with competing risks, hypotheses like (1) are less transparent, leading to issues with interpretation [11]. For example, in competing risks settings where competing events are treated as censoring events, the null hypothesis in (1) is based on cause-specific hazards, which are often not the target of inference [11].

We aimed to develop new hypothesis tests for time-to-event outcomes with two key characteristics: First, the tests should be rooted in explicit null hypotheses that are easy

to interpret. Second, the testing strategy should be generic, such that the scientist can apply the test to their estimand of interest.

Survival parameters as solutions of differential equations

We will consider survival parameters that are functions solving differential equations on the form

$$X_t = X_0 + \int_0^t F(X_s) dA_s, \tag{2}$$

where A is a q dimensional vector of cumulative hazards, and $F = (F_1, \cdots, F_q) : \mathbb{R}^p \longrightarrow \mathbb{R}^{p \times q}$ is Lipschitz continuous with bounded and continuous first and second derivatives, and satisfies a linear growth bound. The class of parameters also includes several quantities that are Lebesgue integrals, such that $dA_t^i = dt$ for some i. Here, X is a vector that includes our estimand of interest, but X may also contain additional nuisance parameters that are needed to formulate the estimand of interest.

Many parameters in survival analysis solve equations on the form (2). In particular, the survival function can be expressed on the form (2) as $S_t = 1 - \int_0^t S_s dA_s$, where A is the cumulative hazard for death. Other examples include the cumulative incidence function, the restricted mean survival function, and the prevalence function. We will present these parameters in detail in Section 3. Nonparametric plugin estimators have been thoroughly studied in [9]. The strategy assumes that A can be consistently estimated by

$$\hat{A}_t = \int_0^t G_{s-} dN_s, \tag{3}$$

where G is a $q \times l$ dimensional predictable process, and N is an l dimensional counting process. Furthermore, we assume that $\hat{A}$, the residuals $W^n = \sqrt{n}(\hat{A} - A)$, and its quadratic variation $[W^n]$, are so-called *predictably uniformly tight*. When the estimator is a counting process integral, a relatively simple condition ensures predictable uniformly tightness [9, Lemma 1]. Moreover, we suppose that $\sqrt{n}(\hat{A} - A)$ converges weakly (wrt the Skorohod metric) to a mean zero Gaussian martingale with independent increments, see [9, Lemma 1, Theorem 1 & 2] for details. Examples of estimators on the form (3) that satisfy these criteria are the Nelson-Aalen estimator, or more generally Aalen's additive hazard estimator; if Aalen's additive hazards model is a correct model for the hazard A, then Aalen's additive hazard model satisfy these criteria, in particular predictable uniformly tightness.

Our suggested plugin estimator of X is obtained by replacing A with $\hat{A}$, giving estimators that solve the system

$$\hat{X}_t = \hat{X}_0 + \int_0^t F(\hat{X}_{s-}) d\hat{A}_s, \tag{4}$$

where $\hat{X}_0$ is a consistent estimator of X_0. When the estimand is the survival function, this plugin estimator reduces to the Kaplan-Meier estimator. Ryalen et al [9] identified the asymptotic distribution of $\sqrt{n}(\hat{X}_t - X_t)$ to be a mean zero Gaussian martingale with covariance V solving a linear differential equation [9, eq. (17)]. The covariance V can also be consistently estimated by inserting the estimates $\hat{A}$, giving rise to the system

$$\begin{aligned} \hat{V}_t =& \hat{V}_0 + \sum_{j=1}^{q} \int_0^t \hat{V}_{s-} \nabla F_j(\hat{X}_{s-})^\intercal + \nabla F_j(\hat{X}_{s-}) \hat{V}_{s-} d\hat{A}_s^j \\ &+ n \int_0^t F(\hat{X}_{s-}) d[B]_s F(\hat{X}_{s-})^\intercal, \end{aligned} \tag{5}$$

where $\{\nabla F_j\}_{j=1}^q$ are the Jacobian matrices of the columns of $F = (F_1, \cdots, F_q)$ from (2), and $[B]_t$ is a $q \times q$ matrix defined by

$$\Big([B]_t\Big)_{i,j} = \begin{cases} 0, \text{ if } dA_t^i = dt \text{ or } dA_t^j = dt \\ \sum_{s \le t} \Delta \hat{A}_s^i \Delta \hat{A}_s^j, \text{ otherwise.} \end{cases}$$

The variance estimator (5), as well as the parameter estimator (4), can be expressed as difference equations, and therefore they are easy to calculate generically in computer programs. To be explicit, let $\tau_1, \tau_2, \ldots$ denote the ordered jump times of $\hat{A}$. Then, $\hat{X}_t = \hat{X}_{\tau_{k-1}} + F(\hat{X}_{\tau_{k-1}})\Delta \hat{A}_{\tau_k}$, as long as $\tau_k \le t < \tau_{k+1}$. Similarly, the plugin variance equation may be written as a difference equation,

$$\begin{aligned} \hat{V}_t =& \hat{V}_{\tau_{k-1}} + \sum_{j=1}^{q} \hat{V}_{\tau_{k-1}} \nabla F_j(\hat{X}_{\tau_{k-1}})^\intercal + \nabla F_j(\hat{X}_{\tau_{k-1}}) \hat{V}_{\tau_{k-1}} \Delta \hat{A}_{\tau_k}^j \\ &+ nF(\hat{X}_{\tau_{k-1}}) \Delta [B]_{\tau_k} F(\hat{X}_{\tau_{k-1}})^\intercal. \end{aligned}$$

2. Hypothesis testing

The null hypothesis is not explicitly expressed in many research reports. On the contrary, the null hypothesis is often stated informally, e.g. vaguely indicating that a difference between two groups is assessed. Even if the null hypothesis is perfectly clear to a statistician, this is a problem: the applied scientist, who frames the research question based on subject-matter knowledge, may not have the formal understanding of the null hypothesis.

In particular, we are not convinced that scientists faced with time-to-event outcomes profoundly understand how null hypotheses based on hazard functions. Hence, using null

hypotheses based on hazard functions, such as (1), may be elusive: in many scenarios, the scientist's primary interest is not to assess whether the hazard functions are equal *at all follow-up times*. Indeed, the research question is often more focused, and the scientist's main concern can be contrasts of other survival parameters at a prespecified t or in a prespecified time interval [12]. For example, we may aim to assess whether cancer survival differs between two treatment regimens five years after diagnosis. Or we may aim to assess whether a drug increases the average time to relapse in subjects with a recurring disease. We will highlight that the rank tests are often not suitable for such hypotheses.

Hence, instead of assessing hazards, let us study tests of (survival) parameters X^1 and X^2 in groups 1 and 2 at a prespecified time t_0. The null hypothesis is

$$\mathbf{H}_0^X\colon\ X_{t_0}^{1,i} = X_{t_0}^{2,i} \text{ for } i = 1, \cdots, p, \tag{6}$$

where p is the dimension of X. We emphasize that the null hypothesis in (6) is different from the null hypothesis in (1), as (6) is defined for any parameter X_{t_0} at a t_0. We will consider parameters X^1 and X^2 that solve (2); this is a broad class of important parameters, including (but not limited to) the survival function, the cumulative incidence function, the time dependent sensitivity and specificity functions, and the restricted mean survival function [9].

2.1. **Test statistics.** We consider two groups 1 and 2 with population sizes n_1, n_2 and let $n = n_1 + n_2$. We can estimate parameters X^1, X^2 and covariance matrices V^1, V^2 using the plugin method described in Section 1. The contrast $\sqrt{n}(\hat{X}_{t_0}^1 - \hat{X}_{t_0}^2)$ has an asymptotic mean zero normal distribution under the null hypothesis. If the groups are

independent, we may then use the statistic

$$(\hat{X}^1_{t_0} - \hat{X}^2_{t_0})\hat{V}^{-1}_{t_0}(\hat{X}^1_{t_0} - \hat{X}^2_{t_0})^\intercal, \tag{7}$$

to test for differences at t_0, where $\hat{V}_{t_0} = \hat{V}^1_{t_0}/n_1 + \hat{V}^2_{t_0}/n_2$, and where $\hat{V}^1_{t_0}$ and $\hat{V}^2$ are calculated using the covariance matrix estimator (5). Then, the quantity (7) is asymptotically χ^2 distributed with p degrees of freedom under the null hypothesis, which is a corollary of the results in [9], as we know from [9, Theorem 2] that $\sqrt{n_i}(\hat{X}^i - X^i)$ converges weakly to mean zero Gaussian martingale whose covariance matrix V^i can be consistently estimated using (5). Therefore, under the null hypothesis (6), the root n difference of the estimates, $\sqrt{n}(\hat{X}^1_{t_0} - \hat{X}^2_{t_0})$, will converge to a multivariate mean zero normal distribution with covariance matrix that can be estimated by $n\big(\hat{V}^1_{t_0}/n_1 + \hat{V}^2_{t_0}/n_2\big)$. Due to the continuous mapping theorem, the statistic (7) has an asymptotic χ^2 distribution.

Sometimes we may be interested in testing e.g. the r first components of X^1 and X^2, under the null hypothesis $X^{1,i}_{t_0} = X^{2,i}_{t_0}$ for $i = 1, \cdots, r < p$. It is straightforward to adjust the hypothesis (6) and the test statistic, yielding the same asymptotic distribution with r degrees of freedom.

3. Examples of test statistics

We derive test statistics for some common effect measures in survival and event history analysis. By expressing the test statistics explicitly, our tests may be compared with the tests based on conventional approaches.

3.1. **Survival at t_0.**

In clinical trials, the primary outcome may be survival at a prespecified t, e.g. cancer survival 5 years after diagnosis. Testing if survival at t is equal in

two independent groups can be done in several ways [12], e.g. by estimating the variance of Kaplan-Meier curves using Greenwood's formula. However, we will highlight that our generic tests also immediately deal with this scenario: it is straightforward to use the null hypothesis in (6), where S_t^1 and S_t^2 are the survival functions in group 1 and 2 at time t. Using the results in Section 2.1, we find that the plugin estimators of S^1 and S^2 are the standard Kaplan-Meier estimators. The plugin variance in group i solves

$$\hat{V}_t^i = \hat{V}_0^i - 2\int_0^t \hat{V}_{s-}^i d\hat{A}_s^i + n_i \int_0^t \Big(\frac{\hat{S}_{s-}^i}{Y_s^i}\Big)^2 dN_s^i, \tag{8}$$

for $i \in \{1, 2\}$, where Y_s^i is the number at risk in group i just before time s. Assuming that the groups are independent, the final variance estimator can be expressed as $\hat{V}_t = \hat{V}_t^1/n_1 + \hat{V}_t^2/n_2$, and the statistic (7) becomes $(\hat{S}_{t_0}^1 - \hat{S}_{t_0}^2)^2/\hat{V}_{t_0}$, which is approximately χ^2 distributed with 1 degree of freedom.

3.2. **Restricted mean survival until t_0.** As an alternative to the hazard ratio, the restricted mean survival has been advocated: it can be calculated without parametric assumptions and it has a clear causal interpretation [13, 14, 15]. The plugin estimator of the restricted mean survival difference between groups 1 and 2 is $\hat{R}_t^1 - \hat{R}_t^2 = \sum_{\tau_k \le t} \big(\hat{S}_{\tau_{k-1}}^1 - \hat{S}_{\tau_{k-1}}^2\big)\Delta\tau_k$, where $\Delta\tau_k = \tau_k - \tau_{k-1}$. The plugin estimator for the variance is

$$\begin{aligned}
\hat{V}_t^{R^i} &= \hat{V}_0^{R^i} + 2\sum_{\tau_k \le t} \hat{V}_{\tau_{k-1}}^{R^i, S^i} \Delta\tau_k \\
\hat{V}_t^{R^i, S^i} &= \hat{V}_0^{R^i, S^i} - \int_0^t \hat{V}_{s-}^{R^i, S^i} d\hat{A}_s^i + \sum_{\tau_k \le t} \hat{V}_{\tau_{k-1}}^{S^i} \Delta\tau_k,
\end{aligned}$$

where $\hat{V}^{S^i}$ is the plugin variance for $\sqrt{n_i}\hat{S}^i$, given in (8). The statistic (7) can be used to perform a test, with $\hat{V}_{t_0} = \hat{V}_{t_0}^{R^1}/n_1 + \hat{V}_{t_0}^{R^2}/n_2$.

3.3. **Cumulative incidence at t_0.** Many time-to-event outcomes are subject to competing risks. The Gray test is a competing risk analogue to the log-rank test: the null hypothesis is defined by subdistribution hazards λ_t, such that $\lambda_t = \frac{d}{dt}\log[1 - C_t]$ where C_t is the cumulative incidence of the event of interest, are equal at all t [16]. Analogous to the log-rank test, the Gray test has low power if the subdistribution hazard curves are crossing [17]. However, we are often interested in evaluating the cumulative incidence at a time t_0, without making assumptions about the subdistribution hazards, which are even harder to interpret causally than standard hazard functions. By expression the cumulative incidence on the form (2), we use our transformation procedure to obtain a test statistic for the cumulative incidence at t_0. The plugin estimator for the cumulative incidence difference is

$$\hat{C}_t^1 - \hat{C}_t^2 = \int_0^t \hat{S}_{s-}^1 d\hat{A}_s^{1,j} - \int_0^t \hat{S}_{s-}^2 d\hat{A}_s^{2,j},$$

where $A^{i,j}$ is the cumulative cause-specific hazard for the event j of interest, and $\hat{S}^i$ is the Kaplan-Meier estimate within group i. The groupwise plugin variances solve

$$\hat{V}_t^i = \hat{V}_0^i + 2\int_0^t \hat{V}_{s-}^i d\hat{A}_s^{i,j} + n_i \int_0^t \left(\frac{\hat{S}_{s-}^i}{Y_s^i}\right)^2 dN_s^{i,j},$$

where $N^{i,j}$ counts the event of interest.

3.4. **Frequency of recurrent events.** Many time-to-event outcomes are recurrent events. For example, time to hospitalization is a common outcome in medical studies, such as trials on cardiovascular disease. Often recurrent events are analysed with conventional methods, in particular the Cox model, restricting the analysis to only include the first event in each subject. A better solution may be to study the mean frequency function,

i.e. the marginal expected number of events until time t, acknowledging that the subject can not experience events after death [18]. We let $A^{i,E}$ and $A^{i,D}$ be the cumulative hazards for the recurrent event and death in group i, respectively, and let K^i and S^i be the mean frequency function and survival, respectively. Then, the plugin estimator of the difference is

$$\hat{K}_t^1 - \hat{K}_t^2 = \int_0^t \hat{S}_{s-}^1 d\hat{A}_s^{1,E} - \int_0^t \hat{S}_{s-}^2 d\hat{A}_s^{2,E}.$$

The plugin variances solve

$$\begin{aligned}
\hat{V}_t^{K^i} &= \hat{V}_0^{K^i} + \int_0^t \hat{V}_{s-}^{K^i,S^i} d(\hat{A}_s^{i,E} - \hat{A}_s^{i,D}) + n_i \int_0^t \Big(\frac{\hat{S}_{s-}^i}{Y_s^i}\Big)^2 dN_s^{i,E}, \\
\hat{V}_t^{K^i,S^i} &= \hat{V}_0^{K^i,S^i} - \int_0^t \hat{V}_{s-}^{K^i,S^i} d\hat{A}_s^{i,D} + \int_0^t \hat{V}_{s-}^{S^i} d\hat{A}_s^{i,E},
\end{aligned}$$

where $N^{i,E}$ counts the recurrent event, and where $\hat{V}^{S^i}$ is the survival plugin variance in group i, as displayed in (8).

3.5. **Prevalence in an illness-death model.** The prevalence denotes the number of individuals with a condition at a specific time, which is e.g. useful for evaluating the burden of a disease. We consider a simple Markovian illness-death model with three states: healthy:0, ill:1, dead:2. The population is assumed to be healthy initially, but individuals may get ill or die as time goes on. We aim to study the prevalence $P_t^{i,1}$ of the illness in group i as a function of time. Here, we assume that the illness is irreversible, but we could extend this to a scenario in which recovery from the illness is possible, similar to Bluhmki [19]. Let $A^{i,kj}$ be the cumulative hazard for transitioning from state k to j

in group i. Then, $P^{i,1}$ solves the system

$$\begin{pmatrix} P_t^{i,0} \\ P_t^{i,1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} -P_s^{i,0} & -P_s^{i,0} & 0 \\ P_s^{i,0} & 0 & -P_s^{i,1} \end{pmatrix} d \begin{pmatrix} A_s^{i,01} \\ A_s^{i,02} \\ A_s^{i,12} \end{pmatrix}.$$

The plugin estimator for the difference $P^{1,1} - P^{2,1}$ is

$$\hat{P}_t^{1,1} - \hat{P}_t^{2,1} = \int_0^t \hat{P}_{s-}^{1,0} d\hat{A}_s^{1,01} - \int_0^t \hat{P}_{s-}^{2,0} d\hat{A}_s^{2,01} - \int_0^t \hat{P}_{s-}^{1,1} d\hat{A}_s^{1,01} + \int_0^t \hat{P}_{s-}^{2,1} d\hat{A}_s^{2,01}.$$

The variance estimator for group i reads

$$\begin{aligned}
\hat{V}_t^{P^{i,1}} &= \hat{V}_0^{P^{i,1}} + 2\int_0^t \hat{V}_{s-}^{P^{i,0},P^{i,1}} d\hat{A}_s^{i,01} - 2\int_0^t \hat{V}_{s-}^{P^{i,1}} d\hat{A}_s^{i,12} \\
&\quad + n_i\Big(\int_0^t \Big(\frac{\hat{P}_{s-}^{i,0}}{Y_s^{i,0}}\Big)^2 dN_s^{i,01} + \int_0^t \Big(\frac{\hat{P}_{s-}^{i,1}}{Y_s^{i,1}}\Big)^2 dN_s^{i,12} \Big) \\
\hat{V}_t^{P^{i,0},P^{i,1}} &= \hat{V}_0^{P^{i,0},P^{i,1}} + \int_0^t \big(\hat{V}_{s-}^{P^{i,0}} - \hat{V}_{s-}^{P^{i,0},P^{i,1}}\big) d\hat{A}_s^{i,01} - \int_0^t \hat{V}_{s-}^{P^{i,0},P^{i,1}} d\hat{A}_s^{i,02} \\
&\quad - \int_0^t \hat{V}_{s-}^{P^{i,0},P^{i,1}} d\hat{A}_s^{i,12} - n_i \int_0^t \Big(\frac{\hat{P}_{s-}^{i,0}}{Y_s^{i,0}}\Big)^2 dN_s^{i,01} \\
\hat{V}_t^{P^{i,0}} &= \hat{V}_0^{P^{i,0}} - 2\int_0^t \hat{V}_{s-}^{P^{i,0}} d\hat{A}_s^{i,01} - 2\int_0^t \hat{V}_{s-}^{P^{i,0}} d\hat{A}_s^{i,02} \\
&\quad + n_i \int_0^t \Big(\frac{\hat{P}_{s-}^{i,0}}{Y_s^{i,0}}\Big)^2 d\big(N_s^{i,01} + N_s^{i,02}\big).
\end{aligned}$$

Here, $Y^{i,0}, Y^{i,1}$ are the number of individuals at risk in states 0 and 1, while $N^{i,kj}$ counts the transitions from state k to j in group i. By calculating $\hat{V}_t^{P^{i,1}}$ for $i \in \{1,2\}$, we can find the statistic (7). Here, the prevalence is measured as the proportion of affected individuals relative to the population at $t = 0$. We could use a similar approach to consider the proportion of affected individuals relative to the surviving population at t,

or we could record the cumulative prevalence until t to evaluate the cumulative disease burden.

4. Performance

In this section, we present power functions under several scenarios for the test statistics that were presented in Section 3. The scenarios were simulated by defining different relations between the hazard functions in the two exposure groups: (i) constant hazards in both groups, (ii) hazards that were linearly crossing, and (iii) hazards that were equal initially before diverging after a time t.

For each hazard relation (i)-(iii), we defined several κ's such that the true parameter difference was equal to κ at the prespecified time point t_0, i.e. $X^1_{t_0} - X^2_{t_0} = \kappa$. For each combination of target parameter, difference κ, and hazard scenario, we replicated the simulations m times to obtain m realizations of (7), and we artificially censored 10% of the subjects in each simulation. In the Supplementary material, we show additional simulations with different sample sizes (50, 100 and 500) and fractions of censoring (10%-40%). We have provided an overview of the simulations in figure 2, in which parameters of interest (solid lines) and hazard functions (dashed lines) are displayed in scenarios with fixed $\kappa = -0.05$ at $t_0 = 1.5$, i.e. $X^1_{1.5} - X^2_{1.5} = -0.05$.

In each scenario, we rejected $\mathbf{H}^X_0$ at the 5% confidence level. Thus, we obtained m Bernoulli trials in which the success probability is the power function evaluated at κ. The estimated power functions, i.e. the estimates of the Bernoulli probabilities, are displayed in figure 3 (solid lines). The power functions are not affected by the structure of the underlying hazards, as desired: our tests are only defined at t_0, and the particular parameterization of the hazard has minor impact on the power function.

The dashed lines in figure 3 show power functions of alternative nonparametric test statistics that are already found in the literature, tailor-made for the scenario of interest. In particular, for the survival at t, we obtained a test statistic using Greenwood's variance formula (and a cloglog transformation in the Supplementary Material) [12]. For the restricted mean survival at t, we used the statistic suggested in [15]. For the cumulative incidence at t, we used the statistic suggested in [20]. For the mean frequency function we used the estimators derived in [18], and in the prevalence example we used the variance formula in [21, p. 295], as implemented in the `etm` package in `R`. Our generic strategy gave similar power compared to the conventional methods for each particular scenario.

4.1. **Comparisons with the log-rank test.** We have argued that our tests are fundamentally different from the rank tests, as the null hypotheses are different. Nevertheless, since rank tests are widely used in practice, also when the primary interest seems to be other hypothesis than in (1), we aimed to compare the power of our test statistics with the log-rank test under different scenarios. In table 3, we compared the power of our test statistic and the rank test, using the scenarios in figure 2. In the first column, the proportional hazards assumption is satisfied (constant), and therefore the power of the log-rank test is expected to be optimal (assuming no competing risks). Our tests of the survival function and the restricted mean survival function show only slightly reduced power compared to the log rank test. For the cumulative incidence function at a time t_0, our test is less powerful than the Gray test and the log-rank test of the cause specific hazards. However, the cause specific hazard test have type one error rate is not nominal, which we will return to in the end of this section.

The second column displays power of tests under scenarios with crossing hazards. For the survival function, it may seem surprising that the log-rank test got higher power than our test, despite the crossing hazards. However, in this particular scenario the hazards are crossing close to the end of the study (dashed lines in Figure 2), and therefore the crossing has little impact on the power of the log-rank test. In contrast, the power of the log-rank test is considerably reduced in the scenarios where we study the restricted mean survival function and the cumulative incidence functions, in which the hazards are crossing at earlier points in time.

The third column shows the power under hazards that are deviating. For the survival function, our test provides higher power. Intuitively, the log-rank test has less power in this scenario because the hazards are equal or close to equal during a substantial fraction of the follow-up time. For the restricted mean survival, however, the log-rank has more power. This is not surprising [15], and it is due to the particular simulation scenario: Late events have relatively little impact on the restricted mean survival time, and in this scenario a major difference between the hazards was required to obtain κ. Since the log-rank test is sensitive to major differences in the hazards, it has more power in this scenario. For the cumulative incidence, in contrast, the power of the log-rank test is lower than the power of our test.

The results in table 3 illustrate that power depends strongly on the hazard scenario for the log-rank test, but this dependence is not found for our tests.

To highlight the basic difference between the log-rank test and our tests, we have studied scenarios where $\mathbf{H}_0^X$ in (6) is true (figure 4). That is, at $t_0 = 1.5$ the difference $X_{t_0}^1 - X_{t_0}^2 = 0$, but for earlier times the equality does not hold. In these scenarios, the log-rank test got high rates of type 1 errors. Heuristically, this is expected because the

hazards are different at most (if not all) times $t \in [0, t_0]$. Nevertheless, figure 4 confirms that the log-rank test does not have the correct type 1 error rate under null hypotheses as in (6), and should not be used for such tests, even if the power sometimes is adequate (as in table 3).

5. Example: Adjuvant chemotherapy in patients with colon cancer

To illustrate how our tests can be applied in practice, we assessed the effectiveness of two adjuvant chemotherapy regimes in patients with stage III colon cancer, using data that are available to anyone [22, 23]. The analysis is performed as a worked example in the supplementary data using the `R` package `transform.hazards`. After undergoing surgery, 929 patients were randomly assigned to observation only, levamisole (Lev) or levamisole plus fluorouracil (Lev+5FU) [22]. We restricted our analysis to the 614 subjects who got Lev or Lev+5FU. All-cause mortality and the cumulative incidence of cancer recurrence was lower in subjects receiving (Lev+5FU), as displayed in Figure 1.

We formally assessed the comparative effectivness of Lev and Lev+5FU after 1 and 5 years of follow-up, using the parameters from section 3. After 1 year, both the overall survival and the restricted mean survival were similar in the two treatment arms (Table 1). However, the cumulative incidence of recurrence was reduced in the Lev+5FU group, and the number of subjects alive with recurrent disease were lower in the Lev+5FU group. Also, the mean time spent alive and without recurrence was longer in the Lev+5FU group (Table 1, Restricted mean recurrence free survival). These results suggest that Lev+5FU has a beneficial effect on disease recurrence after 1 year of follow-up compared to Lev.

After 5 years of follow-up, overall survival and restricted mean survival was improved in the Lev+5FU group (Table 2). Furthermore, the cumulative incidence of recurrence was

reduced, and the prevalence of patients with recurrence was lower in the Lev+5FU group. These results suggest that Lev+5FU improves overall mortality and reduces recurrence after 1 year compared to Lev.

In conclusion, our analysis indicates that treatment with Lev+5FU improves several clinically relevant outcomes in patients with stage III colon cancer. We also emphasise that a conventional analysis using a proportional hazards model would not be ideal here, as the plots in Figure 1 indicate violations of the proportional hazards assumptions.

6. Covariate adjustments

Our approach allows us to conduct tests conditional on baseline covariates, using the additive hazard model; by letting the cumulative hazard integrand in (2) be conditional on specific covariates, we can test for conditional differences between groups, assuming that the underlying hazards are additive.

In more detail, we can test for differences between group 1 and 2 under the covariate level $Z = z_0$ by evaluating the cumulative hazards in each group at that level, yielding A^{1,z_0} and A^{2,z_0}. Estimates $\hat{A}^{1,z_0}$ and $\hat{A}^{2,z_0}$ can be found using standard software. This allows us to estimate parameters with covariances using (4) and (5), and test the null hypothesis of no group difference within covariate level z_0 using the test statistic (7), again assuming that the groups are independent.

7. Discussion

By expressing survival parameters as solutions of differential equations, we provide generic hypothesis tests for survival analysis. In contrast to the conventional approaches that are based on hazard functions [24, Section 3.3], our null hypotheses are defined with

respect to explicit parameters, defined at a time t_0. Our strategy also allows for covariate adjustment under additive hazard models.

We have presented some examples of parameters, and our simulations suggest that the test statistics are well-behaved in a range of scenarios. Indeed, for common parameters such as the survival function, the restricted mean survival function and the cumulative incidence function, our tests obtain similar power to conventional tests that are tailor made for a particular parameter. Importantly, our examples do not comprise a comprehensive set of relevant survival parameters, and several other effect measures for event histories may be described on the differential equation form (2), allowing for immediate implementation of hypothesis tests, for example using the `R` package `transforming.hazards` [9, 8]. The fact that our derivations are generic and easy to implement for customized parameters, is a major advantage.

Our tests differ from the rank tests, as the rank tests are based on assessing the equality of the hazards during the entire follow-up. However, our strategy is intended to be different: We aimed to provide tests that apply to scenarios where the null hypothesis of the rank tests is not the primary interests.

Restricting the primary parameter to a single time t_0 is sometimes considered to be a caveat. In particular, we ignore the time-dependent profile of the parameters before and after t_0. For some parameters, such as the survival function or the cumulative incidence function, this may be a valid objection in principle. However, even if our primary parameter is assessed at t_0, this parameter may account for the whole history of events until t_0. One example is the restricted mean survival, which considers the history of events until t_0. Indeed, the restricted mean survival has been suggested as an alternative effect measure to the hazard ratio, because it is easier to interpret causally,

and it does not rely on the proportional hazards assumption [14]. An empirical analysis of RCTs showed that tests of the restricted mean survival function yield results that are concordant with log-rank tests, under the conventional null hypothesis in (1) [14], and similar results were found in an extensive simulation study [15].

Furthermore, the time-dependent profile before t_0 is not our primary interest in many scenarios. In medicine, for example, we may be interested in comparing different treatment regimes, such as radiation and surgery for a particular cancer. Then, time to treatment failure is expected to differ in the shorter term due to the fundamental difference between the treatment regimes, but the study objective is to assess longer-term treatment differences [25]. Similarly, in studies of cancer screening, it is expected that more cancers are detected early in the screening arm, but the scientist's primary aim is often to assess long-term differences between screened and non-screened. In such scenarios, testing at a prespecified t_0 are more desirable than the null-hypothesis of the rank tests.

Nevertheless, we must assure that cherry picking of t_0 is avoided. In practice, there will often be a natural value of t_0. For example, t_0 (or multiple $t_1, t_2, \cdots, t_k$) can be prespecified in the protocol of clinical trials. In cancer studies, a common outcome is e.g. five year survival. Alternatively, t_0 can be selected based on when a certain proportion is lost to censoring. Furthermore, using confidence bands, rather than pointwise confidence intervals and p-values, is an appealing alternative when considering multiple points in time. There exist methods to estimate confidence bands based on wild bootstrap for competing risks settings [26], which were recently extended to reversible multistate models allowing for illness-death scenarios with recovery [19]. We aim to develop confidence bands for our estimators in future research.

Finally, we are often interested in assessing causal parameters using observational data, under explicit assumptions that ensure no confounding and no selection bias. Such parameters may be estimated after weighting the original data [8, 27]. Indeed, weighted point estimators are consistent when our approach is used [8], but we would also like to identify the asymptotic root n residual distribution, allowing us to estimate covariance matrices that are appropriate for the weighted parameters. We consider this to be an important direction for future work. Currently, such covariance matrices can only be obtained from bootstrap samples.

8. Software

We have implemented a generic procedure for estimating parameters and covariance matrices in an `R` package, available for anyone to use on `github.com/palryalen/transform.hazards`. It allows for hypothesis testing at prespecified time points using (6). Worked examples can be found the package vignette, or in the supplementary material here.

9. Funding

The authors were all supported by the research grant NFR239956/F20 - Analyzing clinical health registries: Improved software and mathematics of identifiability.

References

[1] Judea Pearl. The new challenge: From a century of statistics to the age of causation. In *Computing Science and Statistics*, pages 415–423. Citeseer, 1997.

[2] Miguel A Hernan and James M Robins. *Causal inference*. CRC Boca Raton, FL:, 2018 forthcomming.

[3] Judea Pearl. *Causality: Models, Reasoning and Inference 2nd Edition*. Cambridge University Press, 2000.

[4] Odd O Aalen, Richard J Cook, and Kjetil Røysland. Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime data analysis*, 21(4):579–593, 2015.

[5] Miguel A Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

[6] David R Cox and David Oakes. Analysis of survival data. 1984. *Chapman&Hall, London*, 1984.

[7] Mats J Stensrud, Morten Valberg, Kjetil Røysland, and Odd O Aalen. Exploring selection bias by causal frailty models: The magnitude matters. *Epidemiology*, 28(3):379–386, 2017.

[8] Pål C Ryalen, Mats J Stensrud, and Kjetil Røysland. The additive hazard estimator is consistent for continuous time marginal structural models. *arXiv preprint arXiv:1802.01946*, 2018.

[9] Pål C Ryalen, Mats J Stensrud, and Kjetil Røysland. Transforming cumulative hazard estimates. *accepted in Biometrika, arXiv preprint arXiv:1710.07422*, 2017.

[10] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

[11] Jessica G Young, Eric J Tchetgen Tchetgen, and Miguel A Hernán. The choice to define competing risk events as censoring events and implications for causal inference. *arXiv preprint arXiv:1806.06136*, 2018.

[12] John P Klein, Brent Logan, Mette Harhoff, and Per Kragh Andersen. Analyzing survival curves at a fixed point in time. *Statistics in medicine*, 26(24):4505–4519, 2007.

[13] Patrick Royston and Mahesh KB Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine*, 30(19):2409–2421, 2011.

[14] Ludovic Trinquart, Justine Jacot, Sarah C Conner, and Raphaël Porcher. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 34(15):1813–1819, 2016.

[15] Lu Tian, Haoda Fu, Stephen J Ruberg, Hajime Uno, and Lee-Jen Wei. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, 2017.

[16] Robert J Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154, 1988.

[17] A Latouche and R Porcher. Sample size calculations in the presence of competing risks. *Statistics in medicine*, 26(30):5370–5380, 2007.

[18] Debashis Ghosh and DY Lin. Nonparametric analysis of recurrent events and death. *Biometrics*, 56(2):554–562, 2000.

[19] Tobias Bluhmki, Claudia Schmoor, Dennis Dobler, Markus Pauly, Juergen Finke, Martin Schumacher, and Jan Beyersmann. A wild bootstrap approach for the aalen–johansen estimator. *Biometrics*, 2018.

[20] Mei-Jie Zhang and Jason Fine. Summarizing differences in cumulative incidence functions. *Statistics in Medicine*, 27(24):4939–4949, 2008.

[21] Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[22] Charles G Moertel, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Tangen, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of internal medicine*, 122(5):321–326, 1995.

[23] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[24] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[25] Pål C Ryalen, Mats J Stensrud, and Kjetil Røysland. Causal inference in continuous time: An example on prostate cancer therapy. *Accepted in Biostatistics*, 2018.

[26] DY Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in medicine*, 16(8):901–910, 1997.

[27] Miguel Angel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men, 2000.

TABLE 1. Estimates, 95% confidence intervals and p-values after 1 year

	Lev	Lev+5FU	p-value
Survival	0.91 (0.87,0.94)	0.92 (0.89,0.95)	0.62
Restricted mean survival	0.96 (0.95,0.98)	0.97 (0.95,0.98)	0.86
Cumulative incidence	0.27 (0.22,0.32)	0.15 (0.11,0.19)	0
Prevalence	0.19 (0.14,0.23)	0.09 (0.05,0.12)	0
Restricted mean recurrence free survival	0.85 (0.82,0.88)	0.90 (0.87,0.92)	0.01

TABLE 2. Estimates, 95% confidence intervals and p-values after 5 years

	Lev	Lev+5FU	p-value
Survival	0.54 (0.48,0.59)	0.63 (0.58,0.69)	0.01
Restricted mean survival	3.62 (3.44,3.81)	3.97 (3.79,4.15)	0.01
Cumulative incidence	0.47 (0.42,0.51)	0.34 (0.29,0.39)	0
Prevalence	0.07 (0.05,0.10)	0.03 (0.02,0.05)	0.02
Restricted mean recurrence free survival	2.29 (2.07,2.51)	2.95 (2.73,3.18)	0

TABLE 3. Power comparisons of our tests and rank tests (our/rank) for the scenarios displayed in figure 2, comparing two groups of 1500 individuals (based on 400 replications). In the lower row, we also display the power of Gray's test for competing risks (our/rank/Gray). The power of the rank tests is sensitive to the shape of the underlying hazards, while the power of our tests vary little across the scenarios. In particular, the power of the rank tests is very sensitive to the rate of change of the hazards when they are crossing or deviating; see also the third column of figure 2.

Parameter \ Hazard	Constant	Crossing	Deviating
Survival	0.81/0.88	0.79/0.96	0.83/0.70
Restricted mean survival	0.77/0.87	0.78/0.21	0.8/1
Cumulative incidence	0.85/0.94/0.88	0.86/0.80/0.70	0.86/0.83/0.76

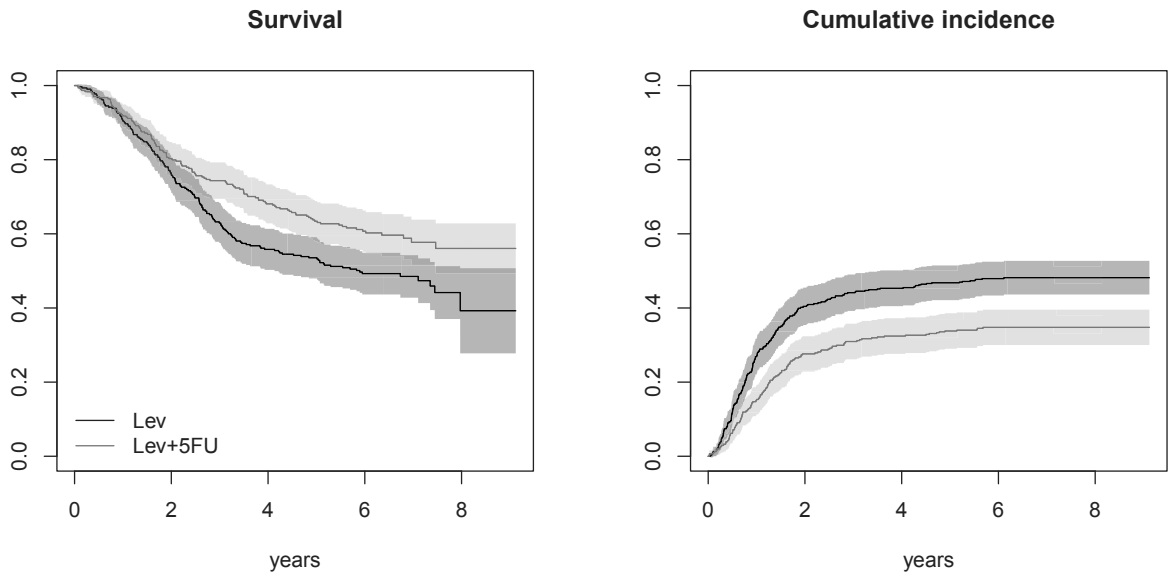


FIGURE 1. Survival curves (left) and the cumulative incidence of recurrence (right) along with 95% pointwise confidence intervals (shaded) from the colon cancer trial are displayed.

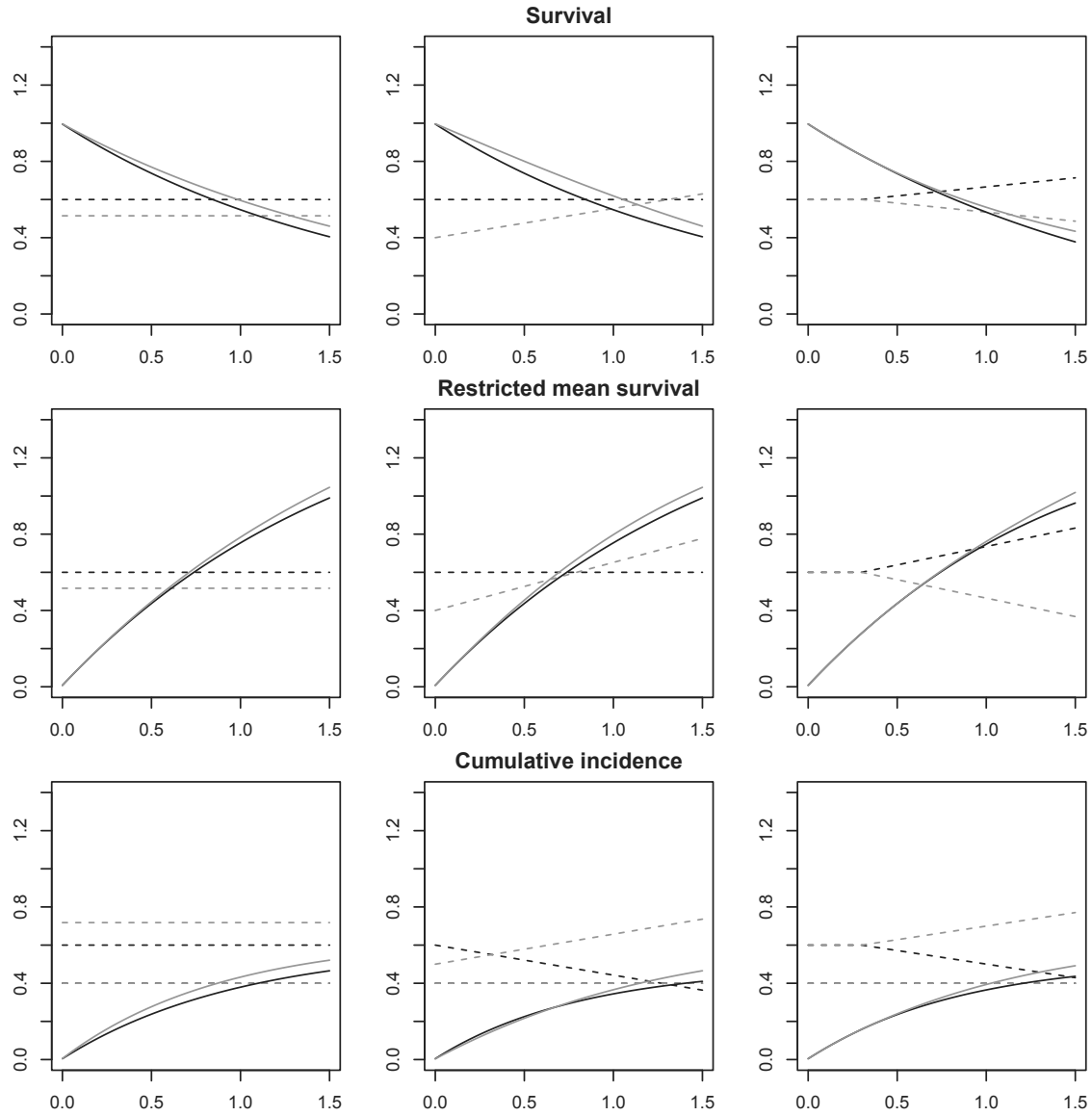


FIGURE 2. Simulation scenarios in which the true parameter difference was fixed to $\kappa = -0.05$ at $t_0 = 1.5$, i.e. $X^1_{1.5} - X^2_{1.5} = -0.05$. The upper row shows survival, the middle row shows restricted mean survival, and lower row shows cumulative incidences. The hazards are displayed as dotted lines; constant in the left column, linearly crossing in the middle column, and deviating in the right column. The X^1 parameters/hazards are black, and the X^2 parameters/hazards are green. See Table 3 for a power comparison of our tests and the rank tests for the scenarios that are displayed. The cumulative incidence panels: The cause specific hazards for the competing event are held constant equal to 0.4 at all times. We optimize the cause specific hazards for the event of interest so that $X^1_{1.5} - X^2_{1.5} = -0.05$ under the restrictions that they are constant (left), linearly crossing (middle), and equal before deviating (right).

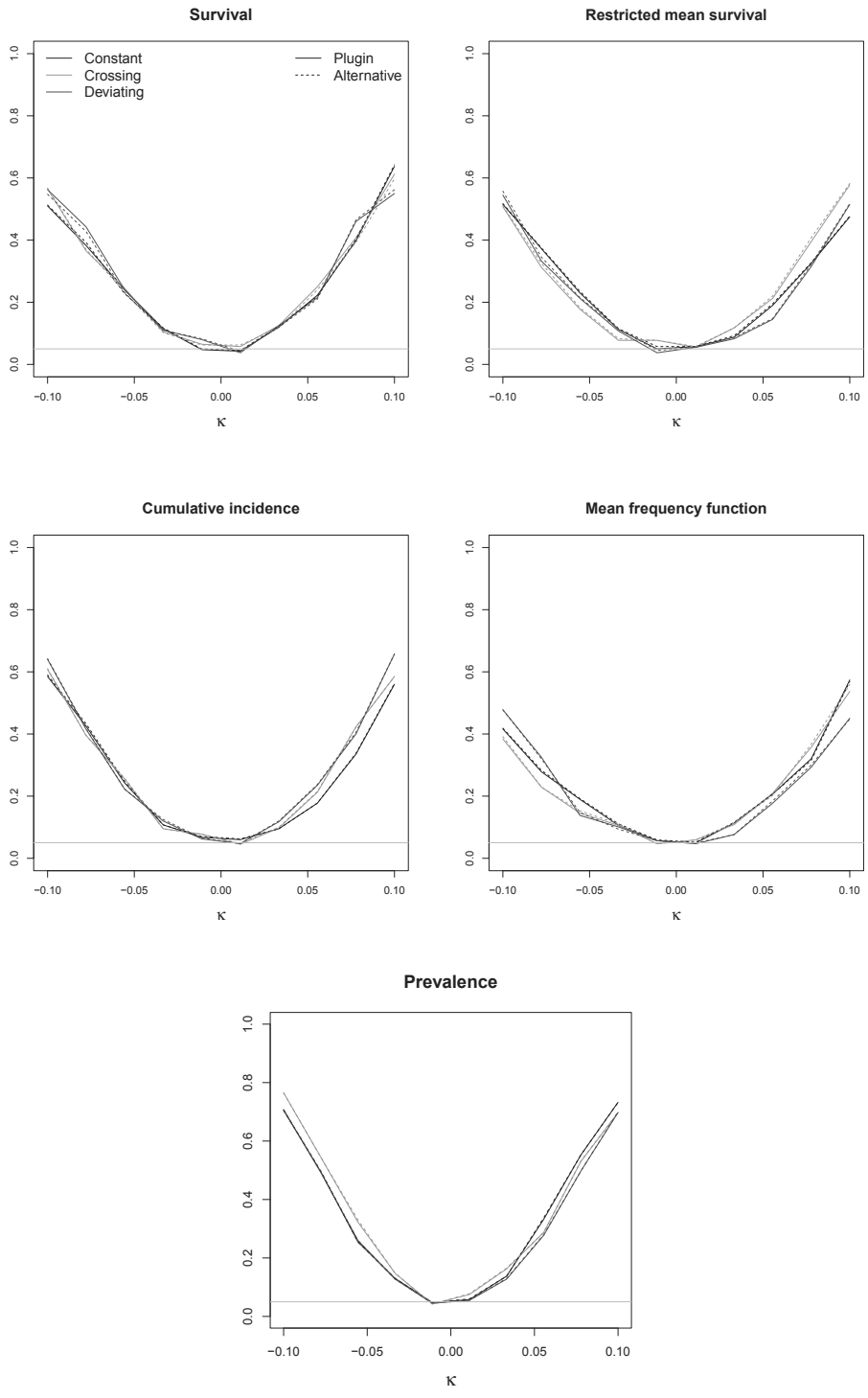


FIGURE 3. Estimated power functions for constant (black), crossing (green), and deviating (blue) hazards, based on 250 subjects with a replication number of 400. The dashed lines show test statistics derived from existing methods in the literature, that are tailor-made for the particular scenario. The confidence level is shown by the gray lines.

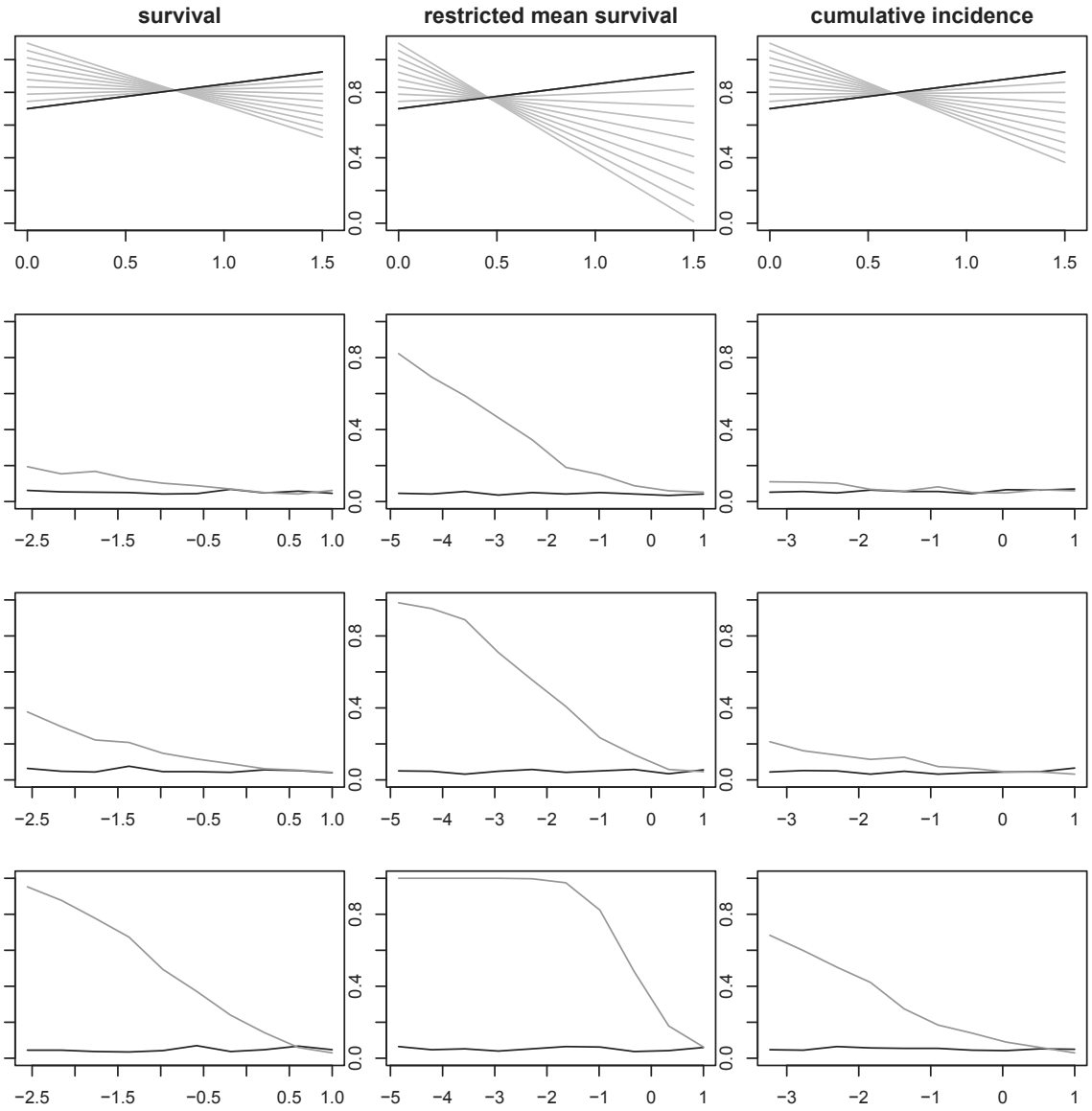


FIGURE 4. In the upper row, we display hazards functions in scenarios in which the hazard in group 1 is fixed (black line), and the hazards in group 2 varies (grey lines). The hazards are optimized such that the null hypothesis is true, i.e. $X^1_{t_0} = X^2_{t_0}$ for each combination of black/gray hazards at $t_0 = 1.5$. In the lower rows we show the estimated rejection rate as a function of the ratio of the hazard slopes (slope of gray/slope of black). This is done for sample sizes sample sizes 500 (row 2), 1000 (row 3), and 5000 (row 4) with a replication number of 500. The green curve shows the rejection rate of the log-rank test, while the black curve shows the rejection rate of our tests, which appear to be well calibrated along the 5% significance level. If the sample size is large, the rank tests can falsely reject the null hypothesis even when the hazards are crossing. The cumulative incidence panels: We only show the cause-specific hazards for the event of interest (which we compare using the rank test). The cause-specific hazard for the competing event is equal to 0.4 in both groups.

SUPPLEMENTARY MATERIAL: ON NULL HYPOTHESES IN SURVIVAL ANALYSIS.

MATS J. STENSRUD, KJETIL RØYSLAND AND PÅL C. RYALEN

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway

Date: January 25, 2019.

1. Colon data example

Here we provide a step-by-step explanation of the colon cancer data analysis in Section 5 of the main text. First, we create a system of integral equations, then we estimate the integrator, and finally we call `pluginEstimate` from the `transform.hazards` package for the plugin-estimation. We can then calculate the test statistics that compares the Lev and Lev+5FU groups after 1 and 5 years of follow-up.

1.1. **Setting up the system.** Before we do the analysis, we set up the ODE system. To to this, we let S be the survival function, and R be the restricted mean survival function. Furthermore we let C^R be the cumulative incidence of cancer recurrence, $S^{D\wedge R}$ be the recurrence free survival, and $R^{D\wedge R}$ be the restricted mean recurrence free survival, and P^R be the prevalence of patients with recurrence. We let A^D, A^R, and $A^{D|R}$ respectively be the marginal cumulative hazards for death, cancer recurrence, and death given cancer recurrence. We can thus consider

$$\begin{pmatrix} S \\ R \\ C^R \\ S^{D\wedge R} \\ P^R \\ R^{D\wedge R} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} -S & 0 & 0 & 0 \\ 0 & S & 0 & 0 \\ 0 & 0 & S^{D\wedge R} & 0 \\ -S^{D\wedge R} & 0 & -S^{D\wedge R} & 0 \\ 0 & 0 & S^{D\wedge R} & -P^R \\ 0 & S^{D\wedge R} & 0 & 0 \end{pmatrix} d \begin{pmatrix} A^D \\ s \\ A^R \\ A^{D|R} \end{pmatrix}. \tag{1}$$

The integrand function is therefore

$$F(x_1, x_2, x_3, x_4, x_5, x_6) = \begin{pmatrix} -x_1 & 0 & 0 & 0 \\ 0 & x_1 & 0 & 0 \\ 0 & 0 & x_4 & 0 \\ -x_4 & 0 & -x_4 & 0 \\ 0 & 0 & x_4 & -x_5 \\ 0 & x_4 & 0 & 0 \end{pmatrix}.$$

The fist column of the integrand function is $F_1(x_1, \cdots, x_6) = (-x_1, 0, 0, -x_4, 0, 0)^\top$. By taking derivatives we find that the Jacobian matrix is

$$\nabla F_1 = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = -e_{1,1} - e_{4,1}.$$

Similar calculations yield

$$\nabla F_2 = e_{2,1} + e_{6,4}, \qquad \nabla F_3 = e_{3,4} - e_{4,4} + e_{5,4}, \qquad \nabla F_4 = -e_{5,5}.$$

1.2. **Performing the analysis.** We install the `transform.hazards` package from GitHub using `devtools`. We also load the data set, and the library `timereg` for additive hazard modeling.

```
devtools::install_github("palryalen/transform.hazards")

library(transform.hazards)
library(timereg)
data(colon)
```

We define the input for the main function `pluginEstimate`; the integrand function, initial values, and the Jacobian matrices $\nabla F_1, \cdots, \nabla F_1$ (see Section 1.1) in a `list`:

```
F_fun <- function(X)matrix(c(-X[1],0,0,-X[4],0,0,
                             0,X[1],0,0,0,X[4],
                             0,0,X[4],-X[4],X[4],0,
                             0,0,0,0,-X[5],0),nrow=6)

J_F1 <- function(X)matrix(c(-1,0,0,-1,0,0,
                             rep(0,30)),nrow=6)
J_F2 <- function(X)matrix(c(0,1,rep(0,21),
                             1,rep(0,12)),nrow=6)
J_F3 <- function(X)matrix(c(rep(0,20),1,
                             -1,1,rep(0,13)),nrow=6)
J_F4 <- function(X)matrix(c(rep(0,28),-1,
                             rep(0,7)),nrow=6)

gradientList <- list(J_F1,J_F2,J_F3,J_F4)

X0 <- matrix(c(1,0,0,1,0,0),nrow=6)
V0 <- matrix(0,nrow=6,ncol=6)
```

We find the event times and the increments of the integrator for each group using the function `getHazMatrix` that can be found attached. The function returns a list containing the increments and the event times.

```
fr1 <- colon[colon$rx == "Lev",]
fr1$time <- fr1$time /365

fr2 <- colon[colon$rx == "Lev+5FU",]
fr2$time <- fr2$time /365

n_Lev <- nrow(fr1)/2
n_Lev_5FU <- nrow(fr2)/2
```

```
fineTimes <- seq(0,max(c(fr1$time,fr2$time)),
        length.out = 5e3)

Lev_list <- getHazMatrix(fr1)
Lev_5FU_list <- getHazMatrix(fr2)

tmsLev <- Lev_list[[2]]
hazLev <- Lev_list[[1]]

tmsLev_5FU <- Lev_5FU_list[[2]]
hazLev_5FU <- Lev_5FU_list[[1]]
```

We perform the analysis for the Lev group and the Lev+5FU group separately, calling the function `pluginEstimate` twice. Inspecting the system (1) we see that the second element in the integrator is a regular dt integral, so we use the input `isLebesge`$= 2$ to improve efficiency:

```
param_Lev <- pluginEstimate(n_Lev,hazLev,F_fun,
                gradientList,X0,V0,isLebesgue = 2)
param_Lev_5FU <- pluginEstimate(n_Lev_5FU,hazLev_5FU,
                F_fun,gradientList,X0,V0,isLebesgue = 2)
```

The calculations are done, and we can evaluate the test statistics after 1 and 5 years of follow-up:

```
t1_Lev <- max(which(tmsLev < 1))
t5_Lev <- max(which(tmsLev < 5))

t1_Lev_5FU <- max(which(tmsLev_5FU < 1))
t5_Lev_5FU <- max(which(tmsLev_5FU < 5))

X_Lev_t1 <- param_Lev$X[,t1_Lev]
X_Lev_5FU_t1 <- param_Lev_5FU$X[,t1_Lev_5FU]
V_t1 <- diag(param_Lev$covariance[,,t1_Lev] +
            param_Lev_5FU$covariance[,,t1_Lev_5FU])

X_Lev_t5 <- param_Lev$X[,t5_Lev]
X_Lev_5FU_t5 <- param_Lev_5FU$X[,t5_Lev_5FU]
V_t5 <- diag(param_Lev$covariance[,,t5_Lev] +
            param_Lev_5FU$covariance[,,t5_Lev_5FU])

# Test statistics
st1 <- (X_Lev_t1 - X_Lev_5FU_t1)^2/V_t1

st5 <- (X_Lev_t5 - X_Lev_5FU_t5)^2/V_t5
```

The vectors `st1` and `st5` contains the test statistics that compare $S, R, C^R, S^{D\wedge R}, P^R$ and $R^{D\wedge R}$, respectively, in the two groups after 1 and 5 years of follow-up.

2. Extended simulations

We expand the simulations in the main document for sample sizes $n = 50, 100$, and 500, as well as for 10% and 40% independent censoring. For each combination of sample size and censoring fraction, we replicated the number of simulations 500 times.

In Table 1 we compared the power of our test statistics with the rank tests, as in Section 4.1 in the main document. As expected, the power tends to be low when the amount of censoring is high, but that the difference is smaller when the sample size is small. Our tests have similar power for each combination of sample size and fraction of censoring. The rank tests give different results from our tests when the sample size is large, but the difference is small for small sample sizes.

In supplementary figure 1 - 5, we plot the estimated power functions from scenarios where the true parameter difference was set to $X^1_{t_0} - X^2_{t_0} = \kappa$ for $\kappa \in [-0.1, 0.1]$. For the smallest sample size of 50, we increased the interval width to $\kappa \in [-0.17, 0.17]$ to display scenarios with higher power. The performance of our test statistics appears to be similar to that of the alternative test statistics found in the literature for all the sample sizes and amounts of censoring we tried. In supplementary figure 1, we included two alternative test statistics; one based on the Greenwood estimator for the variance, and the other is based on a cloglog transformation of the Kaplan-Meier estimates [1, eq. (1) & (3)].

In supplementary figures 6 and 7 we plot the false rejection rates for our tests and the log-rank test in a range of crossing hazard scenarios under 10% and 40% censoring, respectively. The hazards were optimized such that our null hypothesis $X^1_{t_0} = X^2_{t_0}$ is true at $t_0 = 1.5$ for each parameter. Our tests and the rank tests perform similarly for the sample sizes 50 and 100 under both censoring values. For the largest sample sizes of 500 subjects in each group, we find that the rank tests have larger rejection rates even though the hazards are linearly crossing. We also see an interesting difference between these panels under 10% and 40% censoring: higher censoring rates yield larger rejection rates for two of the three parameters. This is due to a combination of two things; the hazards are crossing, and log-rank test depends on the number at risk Y^1_t and Y^2_t of both groups 1 and 2. For large amounts of censoring, Y^1_t and Y^2_t declines faster than for small amounts of censoring. The log-rank weight function $K_t = Y^1_t \cdot Y^2_t/(Y^1_t + Y^2_t)$ will therefore decline more quickly when censoring is increased, which will result in lower power in many situations. However, if the hazards are crossing, the fast decline may e.g. emphasize differences before the hazards have crossed, which will yield a larger log-rank test statistic. We see this for survival and cumulative incidence. Conversely, if the hazards are crossing early in the follow-up period, increasing the censoring

can make K_t emphasize the hazards both before and after they have crossed. Increased censoring may thus give a smaller test statistic, as we see for the restricted mean survival in the lowest panels of supplementary figure 6 and 7.

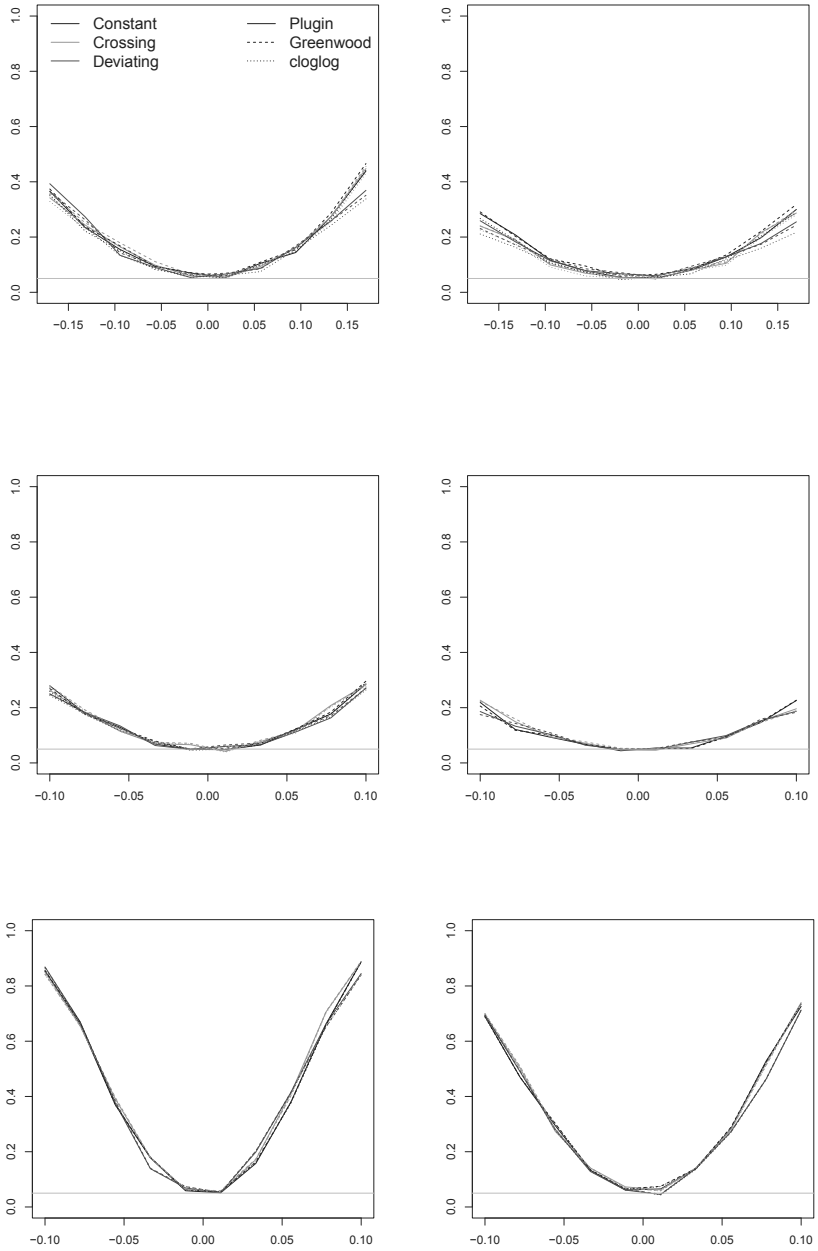
TABLE 1. Power comparisons of our tests and rank tests (our/rank). For the cumulative incidence rows, we have also added power based on Grays test (our/rank/Gray). The tests are performed at $t_0 = 1.5$, and our null hypothesis is slightly violated so that $X^1_{1.5} - X^2_{1.5} = -0.05$ in each scenario.

10% censoring

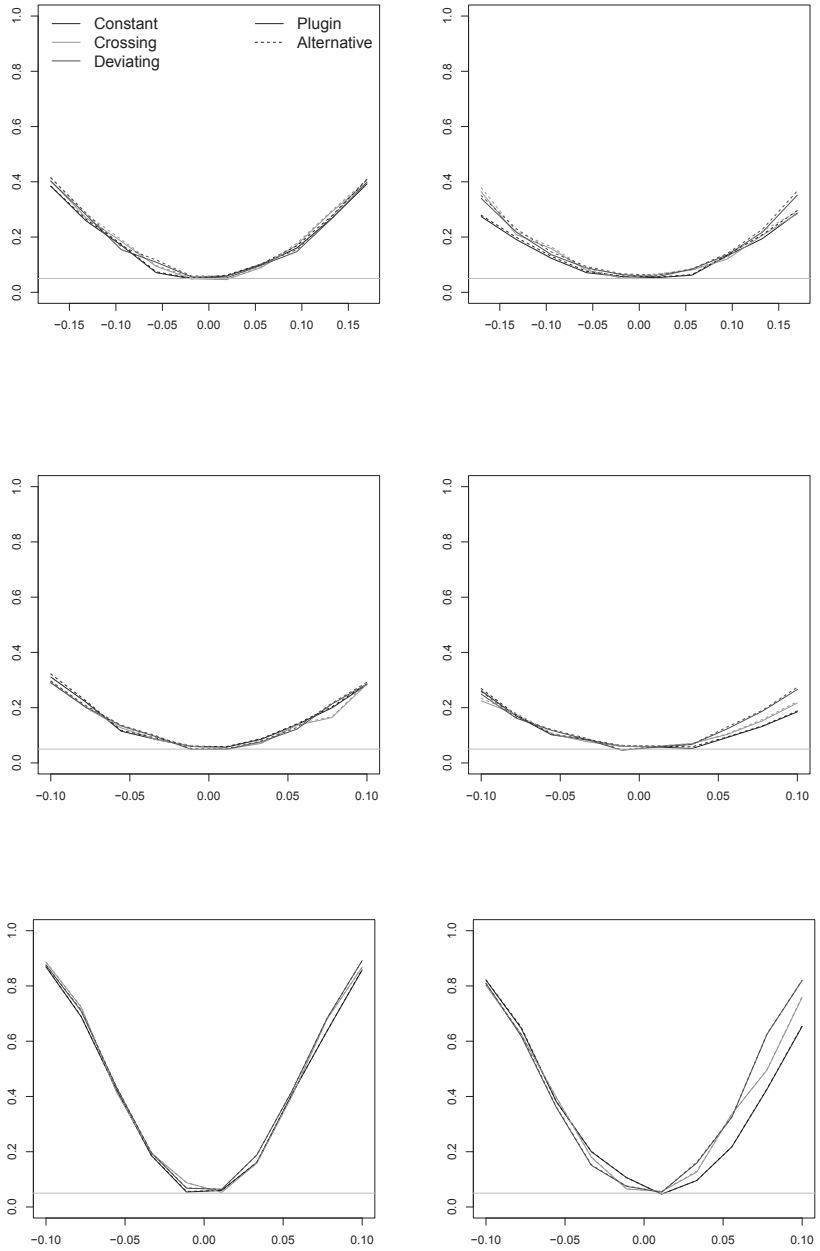
$n = 50$	Constant	Crossing	Deviating
Survival	0.08/0.10	0.08/0.09	0.07/0.08
Restricted mean survival	0.08/0.10	0.07/0.09	0.06/0.08
Cumulative incidence	0.08/0.10/0.10	0.10/0.09/0.09	0.08/0.08/0.07
$n = 100$			
Survival	0.12/0.13	0.12/0.10	12/0.14
Restricted mean survival	0.11/0.13	0.11/0.10	0.10/0.14
Cumulative incidence	0.12/0.13/0.14	0.12/0.10/0.09	0.12/0.14/0.12
$n = 500$			
Survival	0.41/0.49	0.38/0.36	0.39/0.38
Restricted mean survival	0.34/0.49	0.35/0.36	0.35/0.38
Cumulative incidence	0.40/0.49/0.42	0.44/0.36/0.31	0.43/0.38/0.32

40% censoring

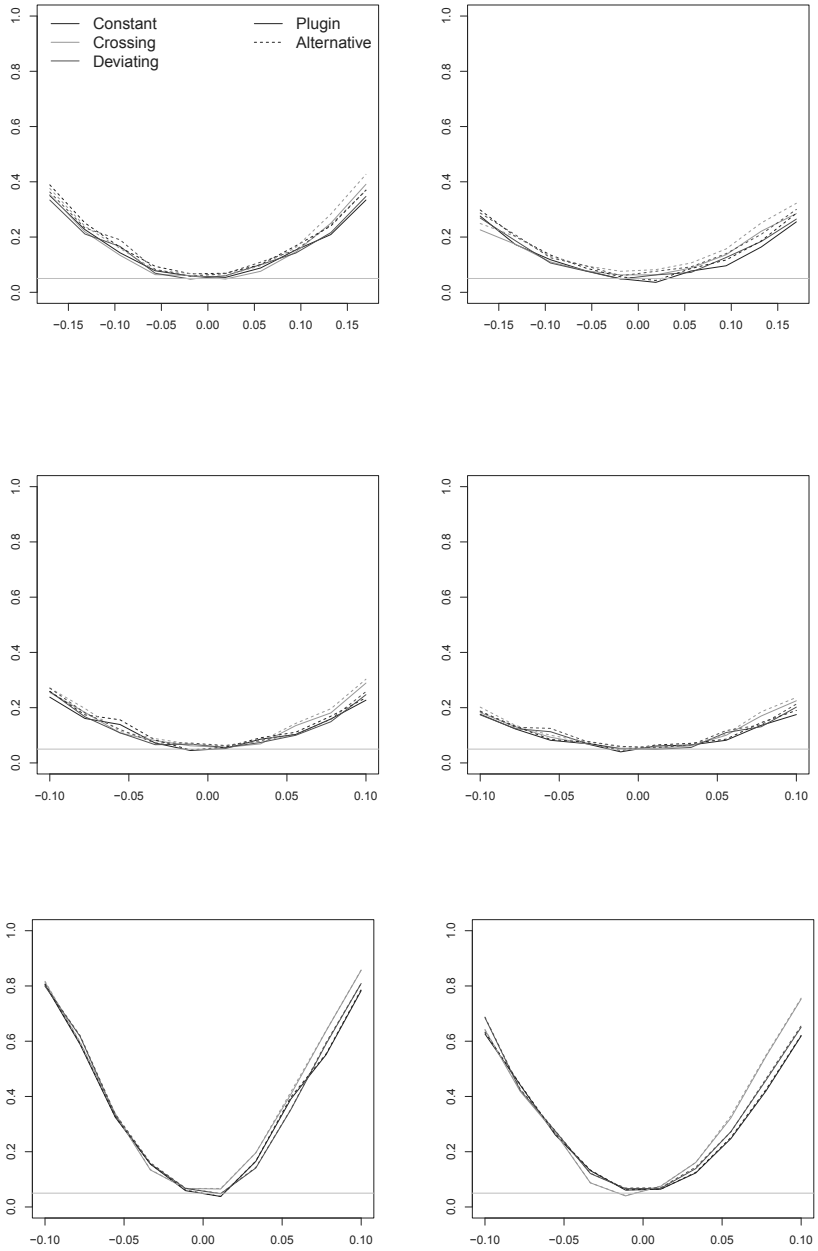
$n = 50$			
Survival	0.10/0.07	0.10/0.06	0.07/0.07
Restricted mean survival	0.07/0.07	0.07/0.06	0.07/0.07
Cumulative incidence	0.07/0.07/0.07	0.08/0.06/0.06	0.09/0.07/0.05
$n = 100$			
Survival	0.10/0.10	0.09/0.16	0.10/0.11
Restricted mean survival	0.11/0.11	0.09/0.07	0.09/0.17
Cumulative incidence	0.10/0.11/0.09	0.13/0.06/0.05	0.10/0.10/0.09
$n = 500$			
Survival	0.33/0.39	0.33/0.55	0.37/0.25
Restricted mean survival	0.32/0.35	0.35/0.16	0.30/0.67
Cumulative incidence	0.30/0.39/0.33	0.32/0.15/0.13	0.32/0.18/0.17



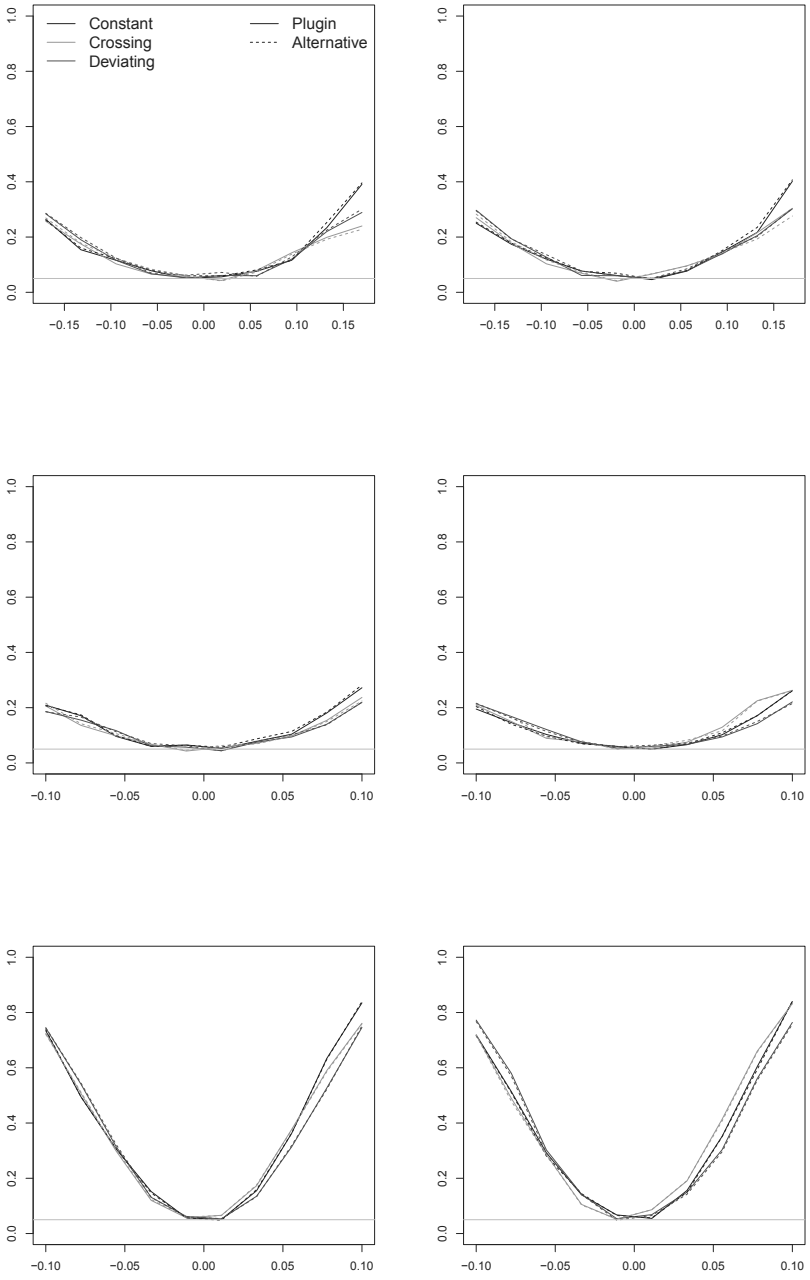
SUPPLEMENTARY FIGURE 1. Estimated survival power functions for constant (black), crossing (green), and deviating (blue) hazards. Different sample sizes (50, 100, and 500 from top to bottom) and amount of censoring (10% censoring to the left and 40% to the right). The dashed lines show the test statistics derived from the Greenwood variance. The dotted line shows the performance of a test statistic based on a cloglog transformation of the survival function. The confidence level is shown by the gray lines.



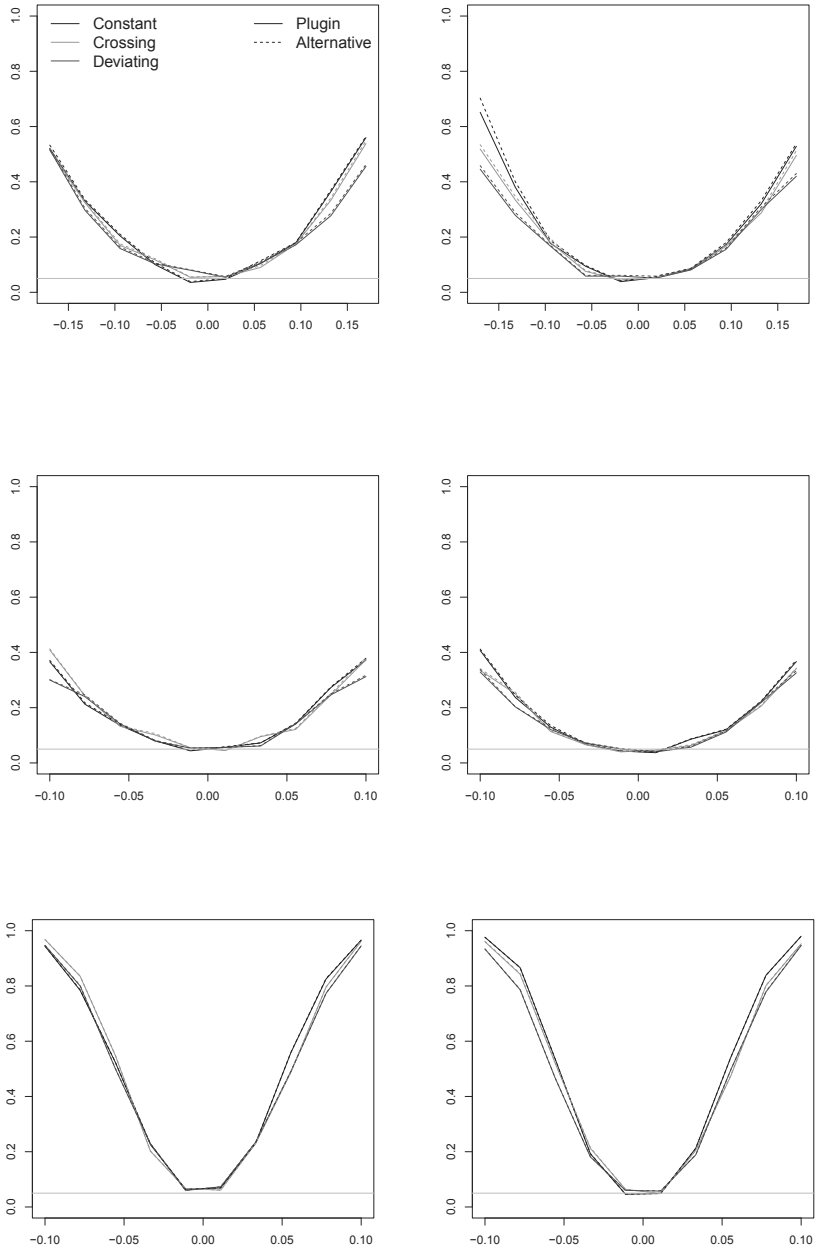
SUPPLEMENTARY FIGURE 2. Estimated cumulative incidence power functions for constant (black), crossing (green), and deviating (blue) hazards. Different sample sizes (50, 100, and 500 from top to bottom) and amount of censoring (10% censoring to the left and 40% to the right).



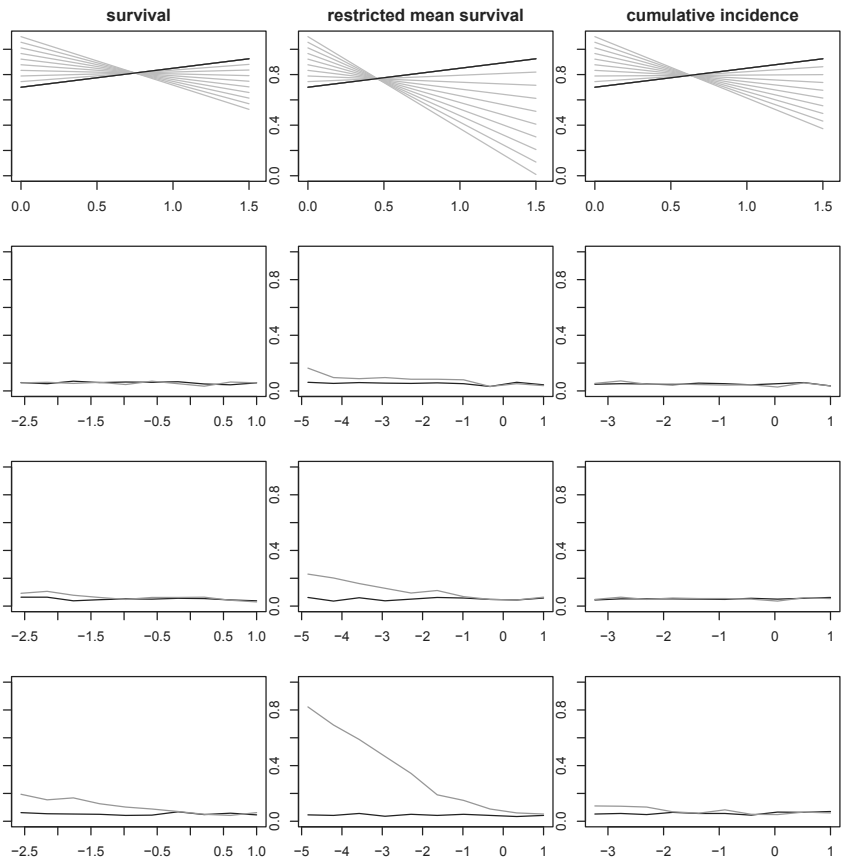
SUPPLEMENTARY FIGURE 3. Estimated restricted mean survival function power functions for constant (black), crossing (green), and deviating (blue) hazards. Different sample sizes (50, 100, and 500 from top to bottom) and amount of censoring (10% censoring to the left and 40% to the right).



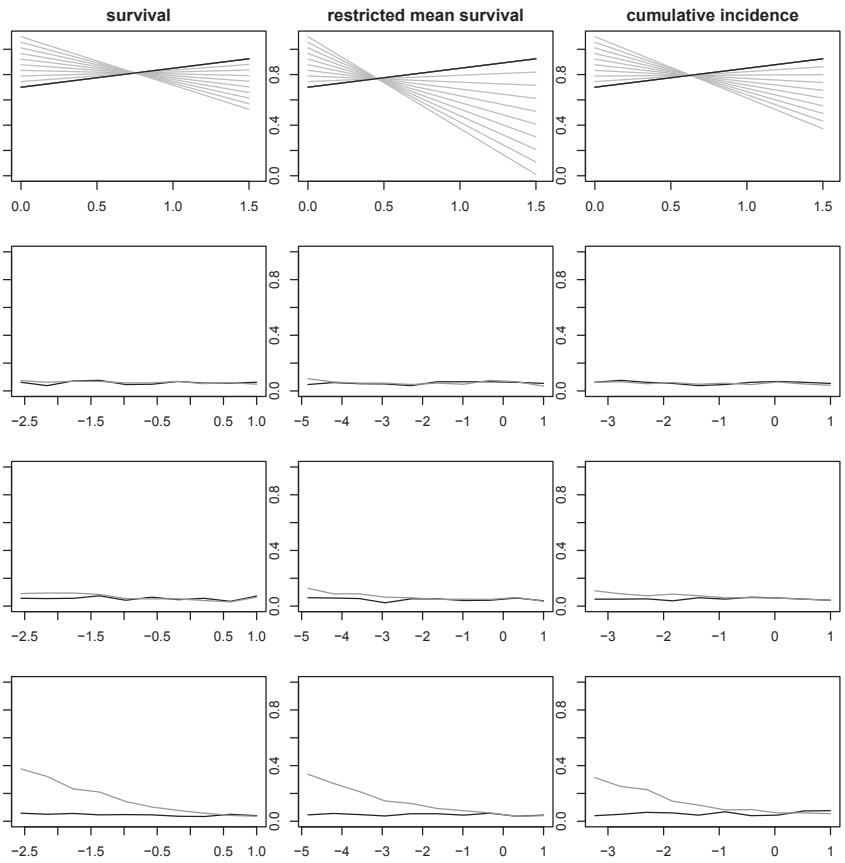
SUPPLEMENTARY FIGURE 4. Estimated power functions for the mean frequency function for constant (black), crossing (green), and deviating (blue) hazards. Different sample sizes (50, 100, and 500 from top to bottom) and amount of censoring (10% censoring to the left and 40% to the right).



SUPPLEMENTARY FIGURE 5. Estimated prevalence power functions for constant (black), crossing (green), and deviating (blue) hazards. Different sample sizes (50, 100, and 500 from top to bottom) and amount of censoring (10% censoring to the left and 40% to the right).



SUPPLEMENTARY FIGURE 6. False rejection rates of the rank tests (green) and our tests (black) for three parameters when the hazards are crossing. We have included sample sizes 50, 100, and 500 in rows 2-4, with 10% censoring in each panel. Our tests appear to be well calibrated, while the log-rank test may give unsatisfactory false rejection rates.



SUPPLEMENTARY FIGURE 7. False rejection rates of the rank tests (green) and our tests (black) for three parameters when the hazards are crossing. We have included sample sizes 50, 100, and 500 in rows 2-4, with 40% censoring in each panel. Our tests appear to be well calibrated, while the log-rank test may give unsatisfactory false rejection rates.

References

[1] John P Klein, Brent Logan, Mette Harhoff, and Per Kragh Andersen. Analyzing survival curves at a fixed point in time. *Statistics in medicine*, 26(24):4505–4519, 2007.

III

Noname manuscript No.
(will be inserted by the editor)

The additive hazard estimator is consistent for continuous-time marginal structural models

Pål C. Ryalen* · Mats J. Stensrud · Kjetil Røysland

Received: date / Accepted: date

Abstract Marginal structural models (MSMs) allow for causal analysis of longitudinal data. The standard MSM is based on discrete time models, but the continuous-time MSM is a conceptually appealing alternative for survival analysis. In applied analyses, it is often assumed that the theoretical treatment weights are known, but these weights are usually unknown and must be estimated from the data. Here we provide a sufficient condition for continuous-time MSM to be consistent even when the weights are estimated, and we show how additive hazard models can be used to estimate such weights. Our results suggest that continuous-time weights perform better than IPTW when the underlying process is continuous. Furthermore, we may wish to transform effect estimates of hazards to other scales that are easier to interpret causally. We show that a general transformation strategy can be used on weighted cumulative hazard estimates to obtain a range of other parameters in survival analysis, and explain how this strategy can be applied on data using our `R` packages `ahw` and `transform.hazards`.

Keywords Additive hazard models · Causal inference in survival analysis · Continuous time marginal structural models · Continuous time weights

* Corresponding author. E-mail: p.c.ryalen@medisin.uio.no

Address(es) of author(s) should be given

1 Introduction

MSMs can be used to obtain causal effect estimates in the presence of confounders, which e.g. may be time-dependent (Robins et al., 2000). The procedure is particularly appealing because it allows for a sharp distinction between confounder adjustment and model selection (Joffe et al., 2004): first, we adjust for observed confounders by weighing the observed data to obtain balanced pseudopopulations. Then, we calculate effect estimates from these pseudopopulations based on our structural model.

Traditional MSM techniques for survival analysis have considered time to be a discrete processes (Hernán et al., 2000b). In particular, inverse probability of treatment weights (IPTWs) are used to create the pseudopopulations, and then e.g. several subsequent logistic regressions are fitted for discrete time intervals to mimic a proportional hazards model.

However, time is naturally perceived as a continuous process, and it also seems natural to analyse time-to-event outcomes with continuous models. Inspired by the discrete time MSMs, Røysland (2011) suggested a continuous-time analogue to MSMs. Similar to the discrete MSMs, it has been shown that the continuous MSMs can be used to obtain consistent effect estimates when the theoretical treatment weights are known (Røysland, 2011). In particular, additive hazard regressions can be weighted with the theoretical continuous-time weights to yield consistent effect estimates. Nevertheless, the weights are usually unknown in real life and must be estimated from the data. To the best of our knowledge, the performance of MSM when the IPTW are estimated remains to be elucidated.

In this article, we show that continuous-time MSMs also perform desirable when the treatment weights are estimated from the data: we provide a sufficient condition to ensure that weighted additive hazard regressions are consistent. Furthermore, we show how such weighted hazard estimates can be consistently transformed to obtain other parameters that are easier to interpret causally. To do this, we use stability theory of SDEs, which allows us to target a range of parameters expressed as solutions of ordinary differential equations. Many examples of such parameters can be found in Ryalen et al. (2018b). This is immediately appealing for causal survival analysis: first, we can use hazard models, that are convenient for regression modeling, to obtain weights. Estimates on the hazard scale are hard to interpret causally per se (Robins and Greenland, 1989; Hernán, 2010; Aalen et al., 2015; Stensrud et al., 2017), but we present a generic method to consistently transform these effect estimates to several other scales that are easier to interpret.

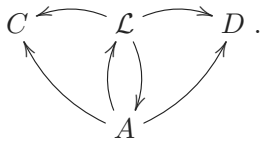
The continuous-time weights and the causal parameters can be estimated using the `R` package `ahw`. We show that this `ahw` weight estimator, which is based on additive hazard regression, is consistent in Theorem 2. We have implemented code for transforming cumulative hazard estimates in the package `transform.hazards`. These packages make continuous-time marginal structural modeling easier to implement for applied researchers.

2 Weighted additive hazard regression

2.1 Motivation

We will present a strategy for dealing with confounding and dependent censoring in continuous time. Confounding, which may be time-varying, will often be a problem when analysing observational data, e.g. coming from health registries. The underlying goal is to assess the effect a treatment strategy has on an outcome.

We can describe processes in continuous time using local (in)dependence relations, and we can use local independence graphs to visualise these relations. A precise description of local independence can be found in Røysland (2011). The local independence graph we will focus on is



Heuristically, the time-dependent confounders $\mathcal{L}$ and the exposure A can influence the censoring process C and the event of interest D. Moreover, the time-dependent confounders can both influence and be influenced by the exposure process. We include baseline variables, some of which may be confounders, in Section 2.2.

The above graph can e.g. describe a follow-up study of HIV-infected subjects, where the initiation and adjustment of HIV treatment depend on CD4 count measurements over time (Hernán et al., 2000a). The CD4 count is a predictor of future survival, and it is also a diagnostic factor that informs initiation of zidovudine treatment; a CD4 count below a certain threshold indicates that treatment is needed. The CD4 count will, in turn, tend to increase in response to treatment, and is monitored over time to inform the future treatment strategy. Hence, it is a time-dependent confounder. In most follow-up studies there is a possibility for subjects to be censored, and we allow the censoring to depend on the covariate and treatment history, as long as subjects are alive.

In Ryalen et al. (2018a) we analysed a cohort of Norwegian males diagnosed with prostate cancer, using the theory from this article to compare treatment effectiveness of radiation and surgery, even though time-dependent confounding were thought to be a minor issue. The continuous-time MSMs allowed us to estimate causal cumulative incidences on the desired time-scale, starting from the time of diagnosis. This example shows that (continuous-time) MSMs can also be a preferable choice in the absence of time-dependent confounding.

2.2 Hypothetical scenarios and likelihood ratios

We consider observational event-history data where n i.i.d. subjects are followed over the study period $[0, T]$. Let $N^{i,A}$ and $N^{i,D}$ respectively be counting

processes that jump when treatment A and outcome D of interest occur for subject i. Furthermore, let $Y^{i,A}, Y^{i,D}$ be the at-risk processes for A and D. We let $\mathcal{V}_0$ be the collection of baseline variables that are not confounders, as well as the treatment and outcome processes. $\mathcal{L}$ are the (time-dependent) confounders. For now, we assume independent censoring, but we will show how our methods can be applied in some scenarios with dependent censoring in Section 6.

Let $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$ denote the filtration that is generated by all the observable events for individual i. Moreover, let P^i denote the probability measure on $\mathcal{F}_T^{i,\mathcal{V}_0\cup\mathcal{L}}$ that governs the frequency of observations of these events, and $\lambda_t^{i,D}$ denote the intensity for $N^{i,D}$ with respect to P^i and the filtration $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$.

We aim to estimate the outcome in a hypothetical situation where a treatment intervention is made according to a specified strategy. Suppose that the frequency of observations we would have seen in this hypothetical scenario is described by another probability measure $\tilde{P}^i$ on $\mathcal{F}_T^{i,\mathcal{V}_0\cup\mathcal{L}}$. Furthermore, assume that all the individuals are also i.i.d. in the hypothetical scenario and that $\tilde{P}^i \ll P^i$, i.e. that there exists a likelihood ratio

$$R_t^i := \frac{d\tilde{P}^i|_{\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}}}{dP^i|_{\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}}}$$

for each time t. We will later describe how an explicit form of $\{R^i\}_i$ can be obtained. It relies on the assumption that the underlying model is causal, a concept we define in Section 3. For the moment we will not require this, but only assume that $\lambda_t^{i,D}$ defines the intensity with respect to $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$ for both P^i and $\tilde{P}^i$; that is, the functional form of $\lambda_t^{i,D}$ is identical under both P^i and $\tilde{P}^i$.

Suppose that $N^{i,D}$ has an additive hazard with respect to $\tilde{P}^i$ and the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$ that is generated by the components of $\mathcal{V}_0$. We stress that we consider the intensity process marginalised over $\mathcal{L}$, and it is thereby defined with respect to $\mathcal{F}_t^{i,\mathcal{V}_0}$, and not $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$. In other words, we assume that the hazard for event D with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$ is additive, and can be written as

$$\boldsymbol{X}_{t-}^{i\intercal}\boldsymbol{b}_t, \tag{1}$$

where $\boldsymbol{b}_t$ is a bounded and continuous vector-valued function, and the components of $\boldsymbol{X}^i$ are covariate processes or baseline variables from $\mathcal{V}_0$.

2.3 Re-weighted additive hazard regression

Our main goal is to estimate the cumulative coefficient function in (1), i.e.

$$\boldsymbol{B}_t := \int_0^t \boldsymbol{b}_s ds \tag{2}$$

from the observational data distributed according to $P = P^1 \otimes \cdots \otimes P^n$. If we had known all the true likelihood ratios, we could try to estimate (2) by re-weighting each individual in Aalen's additive hazard regression (Andersen et al., 1993, VII.4) according to its likelihood ratio. However, the true weights are unlikely to be known, even if the model is causal. In real-life situations, we can only hope to have consistent estimators for these weights. We therefore consider $\mathcal{F}_t^{1,\mathcal{V}_0\cup\mathcal{L}} \otimes \cdots \otimes \mathcal{F}_t^{n,\mathcal{V}_0\cup\mathcal{L}}$-adapted estimates $\{R_t^{(i,n)}\}_n$ that converge to R_t^i under relatively weak assumptions, such that Aalen's additive hazard regression for the outcome re-weighted according to $\{R_t^{(i,n)}\}$ gives consistent estimates of the causal cumulative hazard. The estimator we will consider is defined as follows: let $\boldsymbol{N}^{(n)}$ be the vector of counting processes and $\boldsymbol{X}^{(n)}$ the matrix containing the $\boldsymbol{X}^i$'s, that is,

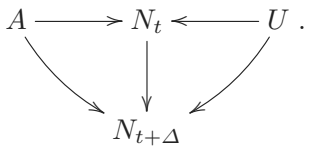
$$\boldsymbol{N}_t^{(n)} := \begin{pmatrix} N_t^{1,D} \\ \vdots \\ N_t^{n,D} \end{pmatrix} \text{ and } \boldsymbol{X}_s^{(n)} := \begin{pmatrix} X_s^{1,1} & \dots & X_s^{1,p} \\ \vdots & & \vdots \\ X_s^{n,1} & \dots & X_s^{n,p} \end{pmatrix}, \tag{3}$$

and let $\boldsymbol{Y}_s^{(n),D}$ denote the $n \times n$-dimensional diagonal matrix where the i'th diagonal element is $Y_s^{i,D} \cdot R_{s-}^{(i,n)}$. The weighted additive hazard regression is given by:

$$\boldsymbol{B}_t^{(n)} := \int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)})^{-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} d\boldsymbol{N}_s^{(n)}. \tag{4}$$

2.3.1 Parameters that are transformations of cumulative hazards

It has recently been emphasised that the common interpretation of hazards in survival analysis as the causal risk of death during $(t, t+\Delta]$ for an individual that is alive at t, is often not appropriate; see e.g. Hernán (2010). An example in Aalen et al. (2015) shows that this can also be a problem in RCTs; if N is a counting process that jumps at the time of the event of interest, A is a randomised treatment, and U is an unobserved frailty, the following causal diagram describes such a situation:



If we consider the probability of an event before $N_{t+\Delta}$, conditioning on no event at time t, we condition on a collider that opens a non-causal path from A to the outcome. This could potentially have dramatic consequences since much of survival analysis is based on the causal interpretation of hazards, e.g. hazard ratios.

In Ryalen et al. (2018b), we have suggested a strategy to handle this situation: even if it is difficult to interpret hazard estimates causally per se, we can

use hazard models to obtain other parameters that have more straightforward interpretations. Population based measures such as the survival function, the cumulative incidence functions, and the restrictive mean survival function, do not condition on survival and will therefore not be subject to the selection bias. Moreover, these measures, and many others (see Ryalen et al. (2018b); Stensrud et al. (2018) for examples), solve differential equations driven by cumulative hazards, i.e. they are functions $\boldsymbol{\eta}_t$ that can be written on the form

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_0 + \int_0^t F(\boldsymbol{\eta}_{s-}) d\boldsymbol{B}_s, \tag{5}$$

where $\boldsymbol{B}$ are cumulative hazard coefficients, and F is a Lipschitz continuous matrix-valued function. In Ryalen et al. (2018b), we showed how to estimate $\boldsymbol{\eta}$ by replacing the integrator in (5) with an estimator $\boldsymbol{B}^{(n)}$ that can be written as a counting process integral. Examples of such $\boldsymbol{B}^{(n)}$ include the Nelson-Aalen, or more generally Aalen's additive hazard estimator. This gave rise to the stochastic differential equation

$$\boldsymbol{\eta}_t^{(n)} = \boldsymbol{\eta}_0^{(n)} + \int_0^t F(\boldsymbol{\eta}_{s-}^{(n)}) d\boldsymbol{B}_s^{(n)}, \tag{6}$$

that is easy to solve on a computer; it is a piecewise constant, recursive equation that jumps whenever the integrator $\boldsymbol{B}^{(n)}$ jumps. Hence, (6) can be solved using a `for` loop over the jump times of $\boldsymbol{B}^{(n)}$, i.e. the survival times of the population.

A simple example of a parameter on the form (5) is the survival function, which reads $S_t = 1 - \int_0^t S_s dB_s$, where B is the cumulative hazard for death. In this case, the estimation strategy (6) yields the Kaplan-Meier estimator. Nevertheless, some commonly studied parameters cannot be written on the form (5), such as the median survival, and the hazard ratio.

In Ryalen et al. (2018b) we showed that $\boldsymbol{\eta}^{(n)}$ provides a consistent estimator of $\boldsymbol{\eta}$ if

- $\lim_{n\to\infty} P(\sup_{t\leq T} |\boldsymbol{B}_t^{(n)} - \boldsymbol{B}_t| \geq \epsilon) = 0$ for every $\epsilon > 0$, i.e. the cumulative hazard estimator is consistent, and
- the estimator $\boldsymbol{B}^{(n)}$ is predictably uniformly tight, abbreviated P-UT.

The additive hazard estimator satisfies both these criteria, and additive hazard regression can thus be used as an intermediate step for flexible estimation of several parameters, such as the survival, the restricted mean survival, and the cumulative incidence functions (Ryalen et al., 2018b). In Theorem 1, we show that also the re-weighted additive hazard regression satisfies these properties, which is a major result in this article. Thus, we can calculate causal cumulative hazard coefficients, and transform them to estimate MSMs that solve ordinary differential equations consistently. In Section 4.4 we illustrate how such estimation can be done, by including an example of a marginal structural relative survival model on simulated data.

A mathematically precise definition of P-UT is given in Jacod and Shiryaev (2003, VI.6a). We will not need the full generality of this definition here.

Rather, we will use Ryalen et al. (2018b, Lemma 1) to determine if processes are P-UT. The Lemma states that whenever $\{\boldsymbol{J}_t^{(n)}\}_n$ is a sequence of semi-martingales on $[0,T]$ with Doob-Meyer decompositions

$$\boldsymbol{J}_t^{(n)} = \int_0^t \boldsymbol{\rho}_s^{(n)} ds + \boldsymbol{M}_t^{(n)},$$

where $\{\boldsymbol{M}^{(n)}\}_n$ are square integrable local martingales and $\{\boldsymbol{\rho}^{(n)}\}_n$ are predictable processes such that

$$\lim_{a\to\infty} \sup_n P\bigg(\sup_s |\boldsymbol{\rho}_s^{(n)}|_1 \geq a \bigg) = 0 \text{ and} \tag{7}$$

$$\lim_{a\to\infty} \sup_n P\bigg(\operatorname{Tr}\langle \boldsymbol{M}^{(n)} \rangle_T \geq a \bigg) = 0, \tag{8}$$

then $\{\boldsymbol{J}_t^{(n)}\}_n$ is P-UT. Here, Tr is the trace function, and $\langle\cdot\rangle$ is the predictable variation.

2.4 Consistency and P-UT property

The consistency and P-UT property of $\boldsymbol{B}^{(n)}$ introduced in Section 2.3 is stated as a Theorem below. A proof can be found in the Appendix.

Theorem 1 (Consistency of weighted additive hazard regression) *Suppose that*

I) The conditional density of $R_t^{(i,n)}$ given $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$ does not depend on i,

II)

$$E_P[\sup_{t\leq T} |\lambda_t^{1,D}|^2] < \infty \text{ and } E_P[\sup_{t\leq T} |\boldsymbol{X}_t^1|^2] < \infty$$

III) Let

$$\boldsymbol{\Gamma}_t^{(n)} := \left(\frac{1}{n} \boldsymbol{X}_{t-}^{(n)\intercal} \boldsymbol{Y}_t^{(n),D} \boldsymbol{X}_{t-}^{(n)} \right) = \left(\tfrac{1}{n} \textstyle\sum_{k=1}^n R_{t-}^{(k,n)} X_{t-}^{k,i} Y_t^{k,D} X_{t-}^{k,j} \right)_{i,j},$$

and suppose that

$$\lim_{a\to\infty} \inf_n P\bigg(\sup_{t\leq T} \operatorname{Tr}\left(\boldsymbol{\Gamma}_t^{(n)-1} \right) > a \bigg) = 0,$$

IV) Suppose that $\{R^i\}_i$ and $\{R^{(i,n)}\}_{i,n}$ are uniformly bounded and

$$\lim_{n\longrightarrow\infty} P\big(\big| R_t^{(i,n)} - R_t^i \big| > \delta \big) = 0 \tag{9}$$

for every i, $\delta > 0$ and t.

Then $\{\boldsymbol{B}^{(n)}\}_n$ *is P-UT and*

$$\lim_{n\longrightarrow\infty} P\Big(\sup_{t\leq T}\big|\boldsymbol{B}_t^{(n)} - \boldsymbol{B}_t\big| \geq \delta\Big) = 0, \tag{10}$$

for every $\delta > 0$.

Heuristically, condition I) states that if we know individual i's realisation of the variables and processes in $\mathcal{V}_0 \cup \mathcal{L}$ up to time t, no other information on individual i is used for estimating her weight at t. Condition II) ensures that the number of outcome events will not blow up, or suddenly grow by an extreme amount. Condition III) implies that there can be no collinearity among the covariates, or more precisely that the inverse matrix of $\big(E[X_t^{1,i}X_t^{1,j}]\big)_{i,j}$ is uniformly bounded in t. Condition IV) states that the weight estimator converges to the theoretical weights R_t^i, in a not very strong sense. The uniform boundedness of $\{R^i\}_i$ is a positivity condition similar to the positivity condition required for standard inverse probability weighting.

3 Causal validity and a consistent estimator for the individual likelihood ratios

We can model the individual likelihood ratio in many settings where the underlying model is causal. To do this, we assume that each subject is represented by the outcomes of r baseline variables $Q_1, \dots Q_r$, and d counting processes $N^1, \dots, N^d$. Moreover, we let $\mathcal{F}_t$ denote the filtration that is generated by all their possible events before t.

Suppose that $\lambda^1, \dots, \lambda^d$ are the intensities of the counting processes $N^1, \dots, N^d$ with respect to the filtration $\mathcal{F}_t$ and the observational probability P. Now, by Jacod (1975), $P|_{\mathcal{F}_T}$ is uniquely determined by all the intensities and the conditional densities at baseline of the form $dP\big(Q^k|Q^{k-1}, \dots, Q^1\big)$, because the joint density at baseline factorises as a product of conditional densities.

Suppose that the observational scenario, where the frequency of events are described by P, is subject to an intervention on the component represented by N^j. Our model is said to be **causal** if such an intervention would not change the 'local characteristics' of the remaining nodes. More precisely this means that

- The functional form of the intensities on which we do not intervene coincide under P and the intervened scenario $\tilde{P}$, i.e. λ^k would also define the intensity for N^k with respect to $\tilde{P}$ when $k \neq j$, and
- The conditional density of each Q^k, given $Q^{k-1}, \dots, Q^1$ would be the same with respect to both P and $\tilde{P}$, i.e.

$$dP\big(Q^k|Q^{k-1}, \dots, Q^1\big) = d\tilde{P}\big(Q^k|Q^{k-1}, \dots, Q^1\big)$$

for $k = 1, \cdots r$.

If the intervention instead were targeted at a baseline variable, say Q^j, and this intervention would replace $dP\big(Q^k|Q^{k-1},\dots,Q^1\big)$ by $d\tilde{P}\big(Q^k|Q^{k-1},\dots,Q^1\big)$, for $k=1,\cdots r$, the model is said to be causal if

- The intensity process for N^k with respect to P and $\tilde{P}$ coincide for all $k=1,\cdots p$, and
- The remaining conditional densities at baseline coincide, i.e.

$$dP\big(Q^k|Q^{k-1},\dots,Q^1\big) = d\tilde{P}\big(Q^k|Q^{k-1},\dots,Q^1\big),$$

for $k \neq j$.

Note that the latter is in agreement with Pearl's definition of a causal model (Pearl, 2000).

This notion of causal validity leads to an explicit formula for the likelihood ratio. If the intervention is aimed at N^j, changing the intensity from λ^j to $\tilde{\lambda}^j$, then the likelihood ratio takes the form

$$R_t = \big(\prod_{s\leq t} \theta_s^{\Delta N_s^j}\big) \exp\big(\int_0^t \lambda_s^j - \tilde{\lambda}_s^j ds\big), \tag{11}$$

where $\theta_t := \frac{\tilde{\lambda}_t^j}{\lambda_t^j}$, see Røysland (2011) and Jacod (1975).

If the intervention is targeted at a baseline variable, the likelihood ratio corresponds to the ordinary propensity score

$$R_0 := \frac{d\tilde{P}\big(Q^j|Q^{j-1},\dots,Q^1\big)}{dP\big(Q^j|Q^{j-1},\dots,Q^1\big)}. \tag{12}$$

Interventions on several nodes yield a likelihood ratio that is a product of terms on the form (11) and (12). The terms in the product could correspond to baseline interventions, time-dependent treatment interventions, or interventions on the censoring intensity. It is natural to estimate the likelihood ratio, or weight process by a product of baseline weights, treatment weights, and censoring weights.

We want, of course, to identify the likelihood ratio that corresponds to $\tilde{P}$, as this is our strategy to assess the desired randomised trial. Following equations (11) and (12), we see that the intervened intensities and baseline variables must be modeled correctly, and specifically that a sufficient set of confounders must be included when modeling the treatment intensity. Additionally, the MSM for the outcome must be correctly specified. An important consequence of the results in this paper is that a class of MSM parameters that solve ODEs driven by cumulative hazards can be estimated consistently.

As long as the intervention acts on a counting process or a baseline variable, the same formula would hold in much more general situations where the remaining covariates are represented by quite general stochastic processes. The assumption of 'coinciding intensities' must then be replaced by the assumption that the 'characteristic triples', a generalisation of intensities to more general processes, coincides for P and $\tilde{P}$; see Jacod and Shiryaev (2003, II.2).

3.1 Estimation of continuous-time weights using additive hazard regression

Suppose we have a causal model as described in the beginning of Section 3, allowing us to obtain a known form of the likelihood ratio R^i. To model the hypothetical scenario, we need to rely on estimates of the likelihood ratio. In the following, we will only focus on a causal model where we replace the intensity of treatment by $\tilde{\lambda}^{i,A}$, the intensity of $N^{i,A}$ with respect to P and the subfiltration $\mathcal{F}_t^{\mathcal{V}_0}$. It is a consequence of the innovation theorem (Andersen et al., 1993) that $E[\lambda_t^{i,A}|\mathcal{F}_{t-}^{\mathcal{V}_0}] = \tilde{\lambda}_t^{i,A}$. Moreover, an exercise in asymptotics of stochastic processes shows that if we discretise time, the associated marginal model structural weights from Robins et al. (2000) approximate (11) gradually as the time-resolution increases.

We will not follow the route of Robins et al. (2000) to estimate R^i. Instead, we will use that (11) is the unique solution to the stochastic differential equation

$$\begin{aligned} R_t^i &= R_0^i + \int_0^t R_{s-}^i dK_s^i \\ K_t^i &= \int_0^t (\theta_s^i - 1) dN_s^{i,A} + \int_0^t \lambda_s^{i,A} ds - \int_0^t \tilde{\lambda}_s^{i,A} ds, \end{aligned}$$

with $\theta^i = \frac{\tilde{\lambda}^{i,A}}{\lambda^{i,A}}$. To proceed, we assume that $\lambda^{i,A}$ and $\tilde{\lambda}^{i,A}$ satisfy the additive hazard model, i.e. that there are vector valued functions $\boldsymbol{h}_t$ and $\tilde{\boldsymbol{h}}_t$, and covariate processes $\boldsymbol{Z}_t$ and $\tilde{\boldsymbol{Z}}_t$ that are adapted to $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$ and $\mathcal{F}_t^{i,\mathcal{V}_0}$ respectively, and

$$\lambda_t^{i,A} = Y_t^{i,A} \boldsymbol{Z}_t^{i\intercal} \boldsymbol{h}_t \text{ and } \tilde{\lambda}_t^{i,A} = Y_t^{i,A} \tilde{\boldsymbol{Z}}_t^{i\intercal} \tilde{\boldsymbol{h}}_t. \tag{13}$$

The previous equation translates into the following:

$$\begin{aligned} R_t^i &= R_0^i + \int_0^t R_{s-}^i dK_s^i \\ K_t^i &= \int_0^t (\theta_s^i - 1) dN_s^{A,i} + \int_0^t Y_s^{i,A} \boldsymbol{Z}_s^{i\intercal} d\boldsymbol{H}_s - \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_s^{i\intercal} d\tilde{\boldsymbol{H}}_s, \end{aligned}$$

where $\boldsymbol{H}_t = \int_0^t \boldsymbol{h}_s ds$ and $\tilde{\boldsymbol{h}}_t = \int_0^t \tilde{\boldsymbol{h}}_s ds$.

Our strategy is to replace R_0^i, $\boldsymbol{H}$, $\tilde{\boldsymbol{H}}$ and θ^i by estimators. This gives the following stochastic differential equation:

$$\begin{aligned} R_t^{(i,n)} &= R_0^{(i,n)} + \int_0^t R_{s-}^{(i,n)} dK_s^{(i,n)} \\ K_t^{(i,n)} &= \int_0^t (\theta_{s-}^{(i,n)} - 1) dN_s^{i,A} + \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} - \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}, \end{aligned} \tag{14}$$

where the quantity $R_0^{(i,n)}$ is assumed to be a consistent estimator of R_0^i. We will use the additive hazard regression estimators $\boldsymbol{H}^{(n)}$ and $\tilde{\boldsymbol{H}}^{(n)}$ for estimating

$\boldsymbol{H}$ and $\tilde{\boldsymbol{H}}$ (Andersen et al., 1993). Moreover, suppose that $\theta_0^{(i,n)}$ is a consistent estimator of θ_0^i, the intensity ratio evaluated at zero. Our candidate for $\theta_t^{(i,n)}$ when $t > 0$ depends on the choice of an increasing sequence $\{\kappa_n\}_n$ with $\lim_{n\longrightarrow\infty} \kappa_n = \infty$ such that $\sup_n \frac{\kappa_n}{\sqrt{n}} < \infty$. This estimator takes the form

$$\theta_t^{(i,n)} = \begin{cases} \theta_0^{(i,n)}, 0 \leq t < 1/\kappa_n \\ \frac{\int_{t-1/\kappa_n}^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}}{\int_{t-1/\kappa_n}^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)}}, 1/\kappa_n \leq t \leq T. \end{cases} \tag{15}$$

κ_n can thus be interpreted as a smoothing parameter. We let $\boldsymbol{Y}^{(n),A}$ be the diagonal matrix where the i'th diagonal element is $Y^{i,A}$. The following Theorem says that the above strategy works out.

Theorem 2 *Suppose that*

a. Each θ^i is uniformly bounded, and right-continuous at $t = 0$.
b. For each i,

$$\lim_{\delta\to 0} P\big(\inf_{t\leq T} |\tilde{\boldsymbol{Z}}_t^{i\intercal} \tilde{\boldsymbol{h}}_t| \leq \delta\big) = 0, \tag{16}$$

c. $E\big[\sup_{s\leq T} |\boldsymbol{Z}_s^i|_3^3\big] < \infty$ and $E\big[\sup_{s\leq T} |\tilde{\boldsymbol{Z}}_s^i|_3^3\big] < \infty$ for every i
d.

$$\lim_{a\to\infty} \sup_n P\bigg(\sup_{s\leq T} \mathrm{Tr}\left(\big(\frac{1}{n} \boldsymbol{Z}_s^{(n)\intercal} \boldsymbol{Y}_s^{(n),A} \boldsymbol{Z}_s^{(n)}\big)^{-1}\right) \geq a\bigg) = 0$$

and

$$\lim_{a\to\infty} \sup_n P\bigg(\sup_{s\leq T} \mathrm{Tr}\left(\big(\frac{1}{n} \tilde{\boldsymbol{Z}}_s^{(n)\intercal} \boldsymbol{Y}_s^{(n),A} \tilde{\boldsymbol{Z}}_s^{(n)}\big)^{-1}\right) \geq a\bigg) = 0$$

Then we have that

$$\lim_{n\to\infty} P\bigg(\sup_{t\leq T} |R_t^{(i,n)} - R_t^i| > \delta\bigg) = 0 \tag{17}$$

for every $\delta > 0$ and i.

For Theorem 1 to apply we need that our additive hazard weight estimator and the likelihood ratio are uniformly bounded. The latter will for instance be the case if both $\lambda^{i,A} - \tilde{\lambda}^{i,A}$ and $\tilde{\lambda}^{i,A}/\lambda^{i,A}$ are uniformly bounded. We will, however, only assume that the theoretical weights R^i are uniformly bounded. In that case we can also make our weight estimator $R^{(i,n)}$ uniformly bounded, by merely truncating trajectories that are too large.

4 Example

4.1 Software

We have developed `R` software for estimation of continuous-time MSMs that solve ordinary differential equations, in which additive hazard models are used

to model both the time to treatment and the time to the outcome of interest. Our procedure involves two steps: first, we estimate continuous-time weights using fitted values of the treatment model. These weights can be used to reweight the sample for estimating the outcome model. Second, we take the cumulative hazard coefficients of the weighted (or causal) outcome model and transform them to estimate ODE parameters that have a more appealing interpretation than cumulative hazards. The two steps can be performed using the `R` packages `ahw` and `transform.hazards`, both of which are available in the repository `github.com/palryalen`. Below, we show an example on how to use the packages on simulated data.

4.2 A simulation study

We simulate an observational study where individuals may experience a terminating event D, so that the hazard for D depends additively on the treatment A and a covariate process L. A and L are counting processes that jump from 0 to 1 for an individual at the instant treatment is initiated or the covariate changes, respectively. The subjects receive treatment depending on L, such that L is a time-dependent confounder. The subjects in the $L = 1$ group can move into treatment, while the subjects in the $L = 0$ group may receive treatment or move to the $L = 1$ group in any order. All subjects are at risk of experiencing the terminating event. The following data generating hazards for D, A, and L are utilised:

$$\alpha_t^D = \alpha_t^{D|0} + \alpha_t^{D|A} A_{t-} + \alpha_t^{D|L} L_{t-} + \alpha_t^{D|A,L} A_{t-} L_{t-} \tag{18}$$

$$\alpha_t^A = \alpha_t^{A|0} + \alpha_t^{A|L} L_{t-} \tag{19}$$

$$\alpha_t^L = \alpha_t^{L|0} + \alpha_t^{L|A} A_{t-}.$$

We want to assess the effect of A on D we would see if A were randomised, i.e. if treatment initiation did not depend on L. To find the effect A has on D we perform a weighted analysis.

We remark that this scenario could be made more complicated by e.g. allowing the subjects to move in and out of treatment, or have recurrent treatments. We could also have included a dependent censoring process, and reweighted to a hypothetical scenario in which censoring were randomised (see Section 6).

4.3 Weight calculation using additive hazard models

We assume that the longitudinal data is organised such that each individual has multiple time-ordered rows; one row for each time either A, L or D changes.

Our goal is to convert the data to a format suitable for weighted additive hazard regression. Heuristically, the additive hazard estimates are cumulative sums of least square estimations evaluated at the event times in the sample.

The main function will therefore need to do two jobs; a) the data must be expanded such that every individual, as long as he is still at risk of D, has a row for each time D occurs in the population, and b) each of those rows must have an estimate of his weight process evaluated just before that event time.

Our software relies on the `aalen` function from the `timereg` package. We fit two additive hazard models for the transition from untreated to treated. The first model assesses the transitions that we observe, i.e. where treatment is influenced by a subjects realisation of L. Here, we use (19), i.e. the true data generating hazard model for treatment initiation; an additive hazard model with intercept and L as a covariate. The second model describes the transitions under the hypothetical randomised trial in which each individual's treatment initiation time is a random draw of the treatment initiation times in the population as a whole. The treatment regime in our hypothetical trial is given by the marginal treatment initiation hazard of the study population, which is the hazard obtained by integrating out L from (19). We estimate the cumulative hazard using the Nelson-Aalen estimator for the time to treatment initiation, by calling a marginal `aalen` regression.

In this way we obtain a factual and a hypothetical `aalen` object that are used as inputs in our `makeContWeights` function. Other input variables include the bandwidth parameter used in (15), weight truncation options, and an option to plot the weight trajectories.

The output of the `makeContWeights` function is an expanded data frame where each individual has a row for every event time in the population, with an additional `weight` column containing time-updated weight estimates. To do a weighted additive hazard regression for the outcome, we will use the `aalen` function once again. Weighted regression is performed on the expanded data frame by setting the weights argument equal to the weight column.

When the weighted cumulative hazard estimates are at hand, we can transform our cumulative hazard estimates as suggested in Section 2.3.1, to obtain effect measures that are easier to interpret. This step can be performed using the `transform.hazards` package; see the GitHub vignette for several worked examples.

4.4 A marginal structural model

We now suppose the intervention that imposes a marginal treatment initiation rate is causally valid. This implies that the intensity for the event D has the same form under the randomised scenario $\tilde{P}$, i.e. that the hazard for D under $\tilde{P}$ for the filtration $\mathcal{F}_t^{A\cup D\cup L}$, generated by A, D, and L, takes the same functional form as (18). We are, however, interested in the hazard with respect to $\tilde{P}$ and the subfiltration $\mathcal{F}_t^{A\cup D}$, the filtration generated by A and D (note that $\mathcal{F}_t^{A\cup D\cup L}$ and $\mathcal{F}_t^{A\cup D}$ respectively correspond to $\mathcal{F}_t^{\mathcal{V}_0\cup\mathcal{L}}$ and $\mathcal{F}_t^{\mathcal{V}_0}$ from Section 2.2). By the innovation theorem the hazard with respect to $\tilde{P}$ and $\mathcal{F}_t^{A\cup D}$ takes the form

$$\beta(t|A) = \beta_t^0 + \beta_t^A A_{t-}.$$

A straightforward regression analysis of the observational data cannot yield causal estimates. Using the ideas from Section 2, we can estimate the cumulative coefficients $B_t^{A=0} = \int_0^t \beta_s^0 ds$ and $B_t^{A=1} - B_t^{A=0} = \int_0^t \beta_s^A ds$ consistently by performing a weighted additive hazard regression.

Cumulative hazards, however, are not easy to interpret. We therefore assess effects on the survival scale, using a marginal structural relative survival model. In this example, our marginal structural relative survival RS^A solves

$$RS_t^{A=a} = 1 + \int_0^t \left(-RS_s^{A=a}\ RS_s^{A=a}\right) d \begin{pmatrix} B_s^{A=a} \\ B_s^{A=0} \end{pmatrix}. \tag{20}$$

The quantity $RS^{A=1}$ can be understood as the survival probability a subject would have if he were exposed at time 0, relative to the survival probability he would have if he were never exposed. Our suggested plugin-estimator is obtained by inserting the estimated causal cumulative coefficients, i.e. the weighted estimates $\hat{B}^{A=a}$ and $\hat{B}^{A=0}$:

$$\hat{RS}_t^{A=a} = 1 + \int_0^t \left(-\hat{RS}_{s-}^{A=a}\ \hat{RS}_{s-}^{A=a}\right) d \begin{pmatrix} \hat{B}_s^{A=a} \\ \hat{B}_s^{A=0} \end{pmatrix}.$$

4.5 Simulation details and results

We simulate subjects, none of which are treated at baseline. Initially, all the patients start with $L = 0$, and the hazards for transitioning from one state to another is constant. As described in Section 4.3, we fit additive hazard models for the time to treatment initiation, one for the observed treatment scenario, i.e. (19), and one for the hypothetical randomised scenario. These models are inserted into `makeContWeights` to obtain weight estimates. Finally, we estimate the additive hazard model by calling the `aalen` function where the `weights` option is set equal to the weight column in the expanded data set.

We make comparisons to the discrete-time, stabilised IPTWs, calculated using pooled logistic regressions. To do this, we discretise the study period $[0, 10]$ into K equidistant subintervals, and include the time intervals as categorical variables in the regressions. We fit two logistic regressions; one for the weight numerator, regressing only on the intercept and the categorical time variables, and a covariate-dependent model for the weight denominator, regressing on the intercept, the categorical time variables, and the time-updated covariate process. We then calculate IPTWs by extracting the predicted probabilities of the two logistic regression model fits, and inserting them into the cumulative product formula (Robins et al., 2000, eq. (17)).

In the upper three rows of Figure 1 we display estimates of the causal cumulative hazard coefficient, i.e. estimates of $B^{A=1} - B^{A=0}$, for a range of sample sises. We include estimates weighted according to our estimator (14), the IPTW estimator, and the theoretical weights, i.e. the true likelihood ratios $\{R^i\}_i$. Compared to the discrete weight estimators, our continuous-time weight estimator (14) gives better approximations to the curves that are estimated

with the theoretical weights. In the lowest row of Figure 1 we plot $\hat{RS}^{A=1}$, i.e. transformed estimates of the cumulative hazard coefficients re-weighted according to the different weight estimators. We used the `transform.hazards` package to perform the plugin-estimation.

5 Performance

In Figure 2 we plot mean weight estimates based on aggregated simulations of the set-up in Section 4. The plot suggests that the discrete weights gradually approximate the continuous likelihood ratio as the time discretisation is refined. However, the continuous-time weights (14) are closer to the expected value of 1 at all times t, indicating less bias.

Choosing the bandwidth parameter will influence the weight estimator and weighted additive hazard estimator in a bias-variance tradeoff; a small κ_n will yield estimates with large bias and small variance, while a large κ_n will give rise to small bias but large variance. It is difficult to provide an exact recipe for choosing the bandwidth parameter, since a good choice depends on several factors, such as the sample size, the distribution of the treatment times, as well as the form and complexity of the true treatment model: if the true treatment hazard is constant, a small κ_n is often appropriate. If the treatment hazard is highly time-varying, κ_n should be chosen to be large, depending on the sample size. Heuristically, several treatment times in the interval $[t - 1/\kappa_n, t]$ for each t would be desirable, but this is not possible in every situation, e.g. when the treatment time distribution is skewed. Such distributions can lead to instable, and possibly large weights for some subjects, even if the chosen bandwidth parameter is a good choice for most other subjects. One option is to truncate weights that are larger than a specified threshold, at the cost of introducing bias. We can assess sensitivity concerning the choice of the bandwidth by performing an analysis for several bandwidth values, truncating weights if necessary, and comparing the resulting weighted estimators. This approach was taken in (Ryalen et al., 2018a, see e.g. Supplementary Figure 4), where no noticeable difference was found for four values of κ_n.

We inspect the bias and variance of our weight estimator for sample sizes n under four bandwidth choices κ_n^z, $z = 1, 2, 3, 4$ at a specified time t_0. By aggregating estimates of k samples for each n we get precise estimates of the bias and variance as a function of n for each choice. The bandwidth functions are scaled such that they are identical at the smallest sample n_0, with $\kappa_{n_0}^1 = \kappa_{n_0}^2 = \kappa_{n_0}^3 = \kappa_{n_0}^4 = 1/t_0$. Otherwise they satisfy $\kappa_n^1 \propto n^{1/2}, \kappa_n^2 \propto n^{1/3}, \kappa_n^3 \propto n^{1/5}$, and $\kappa_n^4 \propto n^{1/10}$.

We simulate a simple scenario where time to treatment initiation depends on a binary baseline variable, such that $\lambda_t^{i,A} = Y_t^{i,A}(\alpha_t^0 + \alpha_t^A x^i)$ for individual i with at-risk indicator $Y^{i,A}$ and binary variable x^i. We calculate weights that re-weight to a scenario where the baseline variable has been marginalised out, i.e. where the treatment initiation intensity is marginal. Utilising the fact that the true likelihood ratio R^i has a constant mean equal to 1, we can find precise

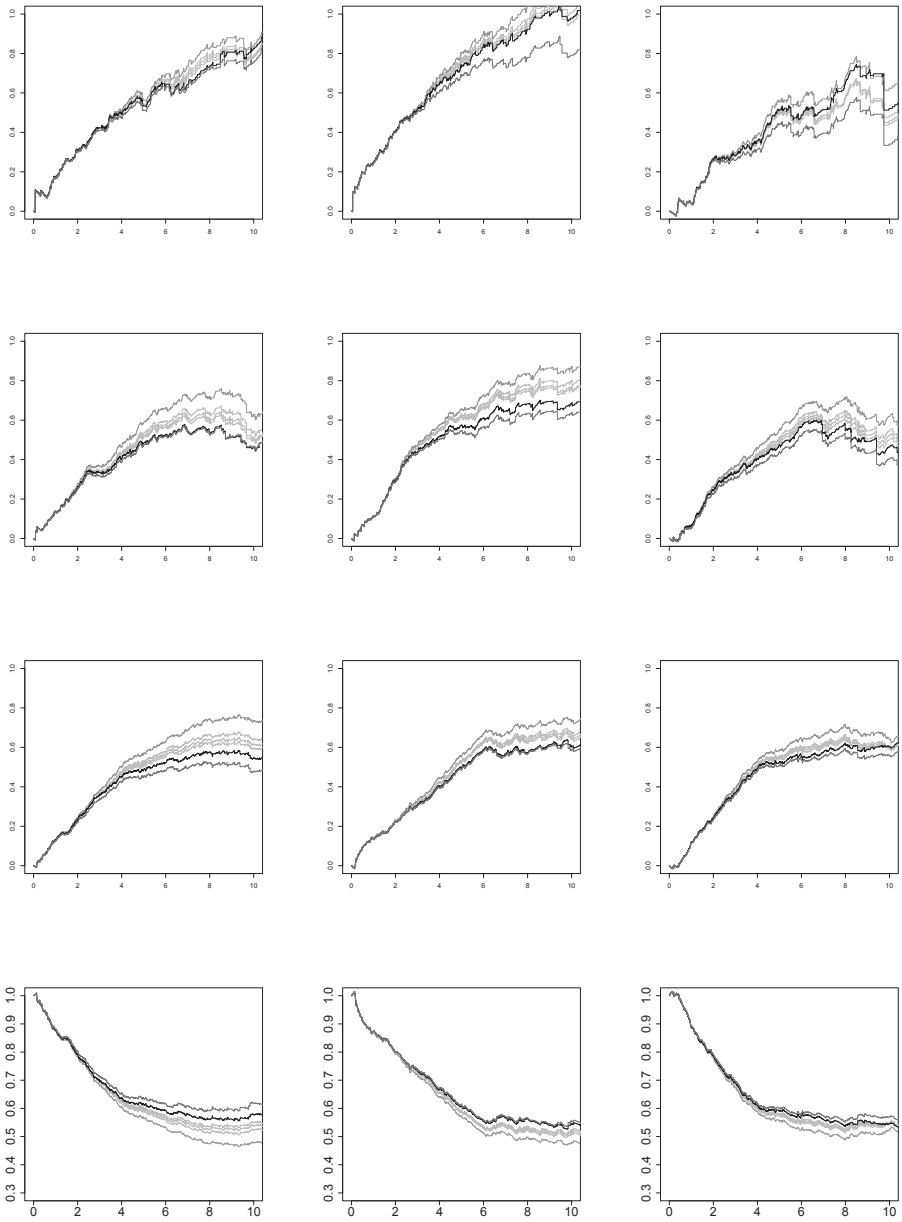


Fig. 1: The upper three rows: three realisations of the cumulative treatment effect estimates for the same scenario, with $n = 500$, 1000, and 2000 from top to bottom. A red line based on estimates re-weighted with the true R^i's is included for reference. The green line shows the unweighted estimates, the gray lines are obtained using the IPTW estimates, while the black line is obtained using our additive hazard weight estimates. The discrete weights were estimated using pooled logistic regressions based on $K = 4$, 8, and 16 time intervals. Increasing the number of intervals moved the curves closer to the red curve. The lowermost row: estimated causal effect of being treated at $t = 0$ versus never being treated according to the relative survival MSM, based on the $n = 2000$ sample.

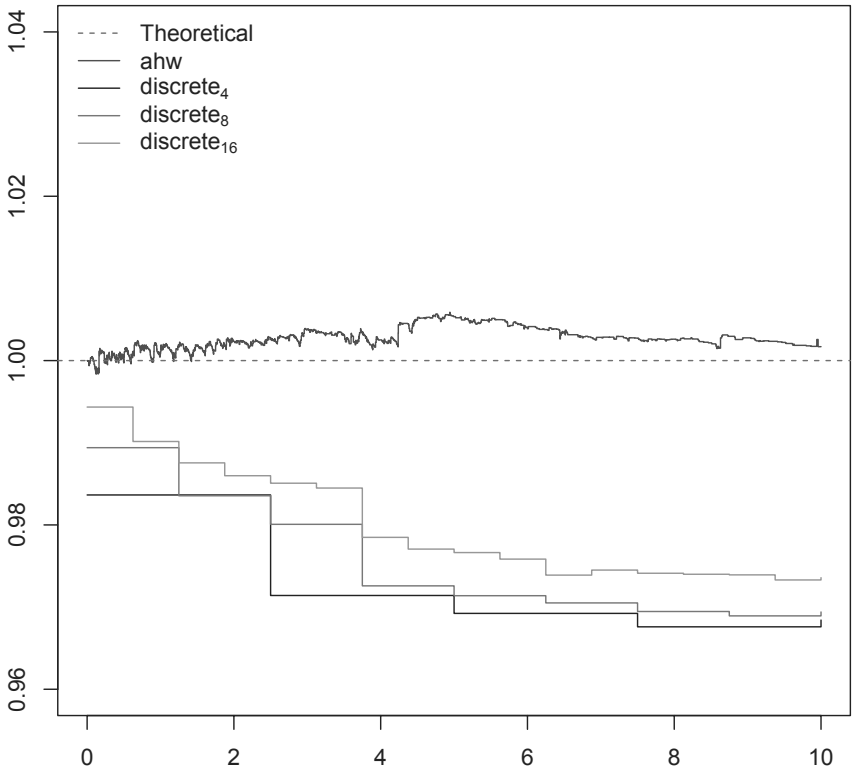


Fig. 2: Average weights based on a sample size of 3000. The theoretical weights have expected value 1. Included are our additive hazard weights, as well as IPTW with $K = 4$, 8, and 16 time intervals. We see that the discrete weights are biased approximations of the theoretical likelihood ratio, while our additive hazard weight estimator appears to be less biased.

estimates of the bias and variance of the additive hazard weight estimator (14) at time t_0.

We plot the bias and variance of the weight estimator as a function of n under the strategies $\kappa_n^1, \kappa_n^2, \kappa_n^3$ and κ_n^4 in Figure 3. We see that the convergence strategy κ_n^1 yields a faster relative decline in bias, but a higher variance as the sample size increases. Meanwhile, the strategy κ_n^4 has a slower decline in bias, but a smaller variance than the other strategies. Finally, the strategies κ_n^2 and κ_n^3 lie mostly between κ_n^1 and κ_n^4 both concerning bias and variance, as a function of the sample size. We also see empirical justification for the

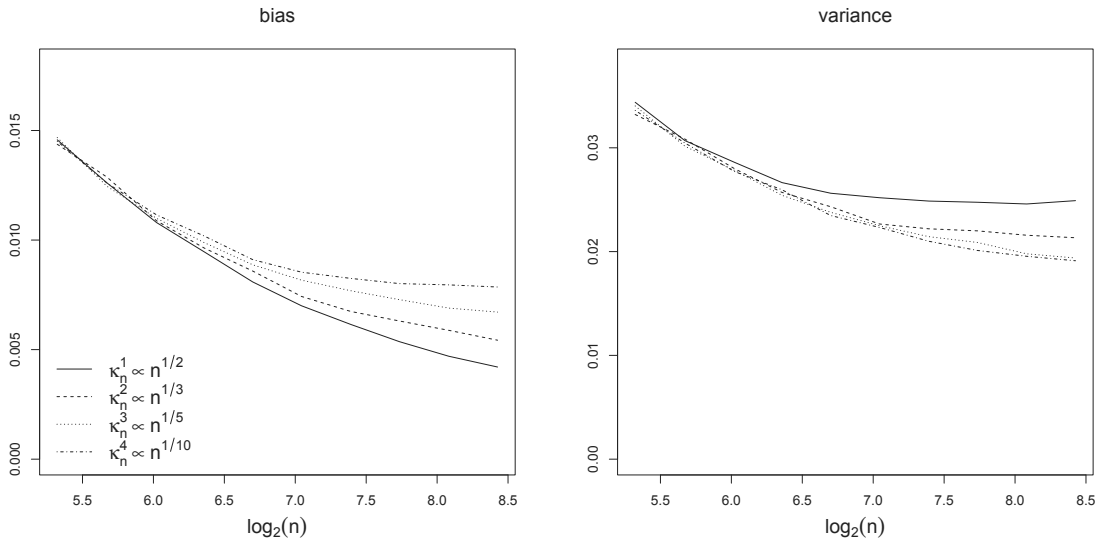


Fig. 3: Bias and variance as a function of n, for four bandwidth refinement strategies.

requirement $\sup_n \kappa_n/n^{1/2} < \infty$, as the variance under the strategy κ_n^1 declines very slowly as n is increased.

6 Censoring weights

Most standard martingale-based estimators in survival analysis are consistent when we have independent censoring, see Andersen et al. (1993, III.2.1). We have assumed independent censoring when conditioning on $\mathcal{V}_0$. A likely situation where this is violated is when we have independent censoring when conditioned on $\mathcal{L} \cup \mathcal{V}_0$, but have dependent censoring if we only condition on $\mathcal{V}_0$. If the model is causal with respect to an intervention that randomises censoring sufficiently, we can model the scenario where this intervention had been applied, and censoring is independent when conditioning on $\mathcal{V}_0$. This means that many estimators that are common in survival analysis will be consistent. Suppose that $N^{i,c}$ is a counting process that jumps when individual i is censored. Moreover, let $\lambda_t^{i,c}$ denote the intensity of $N^{i,c}$ with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$, and let $\tilde{\lambda}_t^{i,c}$ denote its intensity of with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$.

Suppose that there is a meaningful intervention that would give a scenario with frequencies that are governed by $\tilde{P}$ and its intensity for censoring with respect to $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$, is replaced by $\tilde{\lambda}_t^{i,c}$. If the model is causal with respect to

this intervention, the corresponding likelihood ratio process is given by

$$R_t^{i,c} = \prod_{s \leq t} \Big(\frac{\tilde{\lambda}_s^{i,c}}{\lambda_s^{i,c}} \Big)^{\Delta N_s^{i,c}} \exp\big(- \int_0^t \tilde{\lambda}_s^{i,c} - \lambda_s^{i,c} ds \big). \tag{21}$$

However, as we only need to apply weights to observations strictly before the time of censoring, we only need to consider

$$R_t^{i,c} = \exp\big(- \int_0^t \tilde{\lambda}_s^{i,c} - \lambda_s^{i,c} ds \big). \tag{22}$$

This process is a solution to the equation

$$R_t^{i,c} = 1 + \int_0^t R_s^{i,c} (\lambda_s^{i,c} - \tilde{\lambda}_s^{i,c}) ds. \tag{23}$$

Furthermore, we assume additive hazard models, i.e. that

$$\lambda_t^c = Y_t^{i,c} \boldsymbol{U}_{t-}^{i\intercal} \boldsymbol{g}_t \text{ and } \tilde{\lambda}_t^{i,c} = Y_t^{i,c} \tilde{\boldsymbol{U}}_{t-}^{i\intercal} \tilde{\boldsymbol{g}}_t, \tag{24}$$

for an $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$-adapted covariate process $\boldsymbol{U}^i$, and an $\mathcal{F}_t^{i,\mathcal{V}_0}$-adapted covariate process $\tilde{\boldsymbol{U}}^i$, and vector valued functions $\boldsymbol{g}$ and $\tilde{\boldsymbol{g}}$. Following Theorem 2, we see that these weights are consistently estimated by $R^{(i,n,c)}$ defined by the equation:

$$\begin{aligned} R_t^{(i,n,c)} &= 1 + \int_0^t R_{s-}^{(i,n,c)} dK_s^{(i,n,c)} \\ K_t^{(i,n,c)} &= \int_0^t Y_s^{i,c} \boldsymbol{U}_{s-}^{i\intercal} d\boldsymbol{G}_s^{(n)} - \int_0^t Y_s^{i,c} \tilde{\boldsymbol{U}}_{s-}^{i\intercal} d\tilde{\boldsymbol{G}}_s^{(n)}, \end{aligned}$$

where $\boldsymbol{G}^{(n)}$ and $\tilde{\boldsymbol{G}}^{(n)}$ are the usual additive hazards estimates of $\int_0^{\cdot} \boldsymbol{g}_s ds$ and $\int_0^{\cdot} \tilde{\boldsymbol{g}}_s ds$.

7 Discussion

Marginal structural modeling is an appealing concept for causal survival analysis. Here we have developed theory for continuous-time MSMs that may motivate the approach for practical research. Indeed, we show that the continuous-time MSMs yield consistent effect estimates, even if the treatment weights are estimated from the data. Our continuous-time weights seem to perform better than the discrete time weights when we study processes that develop in continuous time. Furthermore, our weights can be estimated using additive hazard regressions, which are easy to fit in practice. Importantly, we also show that causal effect estimates on the hazard scale, e.g. weighted cumulative hazard estimates, can be transformed consistently to estimate other parameters that are easier to interpret causally. We thereby offer a broad strategy to obtain causal effect estimates for time-to-event outcomes. Previously, Huffer and

McKeague (1991) and McKeague (1987) derived results on weighted additive hazard regression, but they do not cover our needs, as our weights are estimates of likelihood ratios with respect to filtrations that are larger than the filtration for the additive hazard that we want to estimate.

Estimators of IPTWs may be unstable and inefficient, e.g. when there are strong predictors of the treatment allocation. In practice, applied researchers will often face a bias-variance tradeoff when considering confounder control and efficient weight estimation. This bias-variance tradeoff has been discussed in the literature, and weight truncation has been suggested to reduce the variance, at the cost of introducing bias; see e.g. Cole and Hernán (2008). Similar to IPTWs, and for the same reasons, our continuous-time weight estimator may be instable, and proper weight estimation requires a delicate balance between confounder control and precision in most practical situations.

We have considered the treatment process A to be a time-to-event variable, but our strategy can be generalised to handle recurrent, or piecewise constant exposures. If A is allowed to have multiple jumps, the estimation procedure becomes more complex, but the same estimators (4) and (14) can be used with few modifications. We think, however, that many important applications can be explored assuming that A is the time to an event.

A different approach that accounts for time-dependent confounding is the structural nested model, which parameterises treatment effects directly in a structural model (Robins, 2014). While this procedure avoids weighting, and will often be more stable and efficient, it relies on other parametric assumptions and can be harder to implement (Vansteelandt and Sjolander, 2016).

We conjecture that there is a similar consistency result as Theorem 1 when the outcome model is a weighted Cox regression. However, using a Cox model in the hypothetical scenario after marginalisation leads to restrictions on the data generating mechanisms that are not properly understood, see e.g. Havercroft and Didelez (2012). This issue is related to the non-collapsibility of the Cox model, and it is a problem regardless of the weights being used are continuous or discrete.

8 Funding

The authors were all supported by The Research Council of Norway, grant NFR239956/F20 - Analyzing clinical health registries: Improved software and mathematics of identifiability.

Appendix: proofs

We need some lemmas to prove Theorem 1.

Lemma 1 *Suppose that* $\{V^i\}_i$ *are processes on* $[0,T]$ *such that* $\sup_i E\big[\sup_s |V_s^i|\big] < \infty$, *then*

$$\lim_{a\to\infty} \sup_n P\bigg(\sup_s \big|\frac{1}{n}\sum_{i=1}^n V_s^i\big| \geq a\bigg) = 0. \tag{25}$$

Proof By Markov's inequality, we have for every $a > 0$ *that*

$$P\bigg(\sup_s \big|\frac{1}{n}\sum_{i=1}^n V_s^i\big| \geq a\bigg) \leq \frac{1}{na}\sum_{i=1}^n E_P\bigg[\sup_s \big|V_s^i\big|\bigg],$$

which proves the claim.

Lemma 2 (A perturbed law of large numbers) *Suppose*

I) $p^{-1} + q^{-1} = 1$, $p < \infty$,

II) $\{V_i\}_i \subset L^p(P)$, $\{S_i\}_i \subset L^q(P)$ *such that* $\{(V_i, S_i)\}_i$ *is i.i.d., and* V_i, S_i *are measurable with respect to a* σ*-algebra* $\mathcal{F}_i$,

III) Triangular array $\{S_{(i,n)}\}_{n,i\leq n}$ *such that*

$$\lim_{n\longrightarrow\infty} P\big(|S_{(1,n)} - S_1| \geq \epsilon\big) = 0 \tag{26}$$

for every $\epsilon > 0$, *and there exists a* $\tilde{S} \in L^q(P)$ *such that* $\tilde{S} \geq |S_{(1,n)}|$ *for every* n,

IV) The conditional density of $S_{(i,n)}$ *given* $\mathcal{F}_i$ *does not depend on* i.

This implies that

$$\lim_{n\longrightarrow\infty} E\bigg[\bigg|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - E_P[S_1V_1]\bigg|\bigg] = 0. \tag{27}$$

Proof From the triangle inequality and condition IV) we have that

$$\begin{aligned} &E\bigg[\bigg|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - \frac{1}{n}\sum_{i=1}^n S_iV_i\bigg|\bigg] \leq \frac{1}{n}\sum_{i=1}^n E\big[\big|\big(S_{(i,n)} - S_i\big)V_i\big|\big] \\ =&E\big[\big|\big(S_{(1,n)} - S_1\big)V_1\big|\big]. \end{aligned}$$

The dominated convergence theorem implies that the last term converges to 0. *Finally, the weak law of large numbers and the triangle inequality yields*

$$\begin{aligned} &\lim_{n\longrightarrow\infty} E\bigg[\bigg|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - E_P[S_1V_1]\bigg|\bigg] \\ \leq &\lim_{n\longrightarrow\infty} E\bigg[\bigg|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - \frac{1}{n}\sum_{i=1}^n S_iV_i\bigg|\bigg] + E\bigg[\bigg|\frac{1}{n}\sum_{i=1}^n S_iV_i - E[S_1V_1]\bigg|\bigg] = 0. \end{aligned}$$

Lemma 3 $\{V_i\}_i$ *i.i.d. non-negative variables in* $L^2(P)$, *then*

$$\lim_{n\to\infty} P\bigg(\frac{1}{n}\max_{i\leq n} V_i \geq \epsilon\bigg) = 0 \tag{28}$$

for every $\epsilon > 0$.

Proof Note that

$$P\Big(\frac{1}{n}\max_{i\leq n} V_i > \epsilon\Big) = 1 - P\Big(\max_{i\leq n} V_i \leq \epsilon n\Big) = 1 - P\Big(V_1 \leq \epsilon n\Big)^n$$
$$= 1 - \Big(1 - P\big(V_1 > \epsilon n\big)\Big)^n$$

If $n > \|V_1\|_2 \epsilon^{-1}$*, we therefore have by Chebyshev's inequality that*

$$P\Big(\frac{1}{n}\max_{i\leq n} V_i > \epsilon\Big) \leq 1 - \Big(1 - \frac{E[V_1^2]}{n^2\epsilon^2}\Big)^n,$$

where the last term converges to 0 *when* $n \to \infty$ *since* $\lim_{n\longrightarrow\infty} n \log\big(1 - \frac{E[V_1^2]}{n^2\epsilon^2}\big) = 0$ *for every* $\epsilon > 0$*.*

Lemma 4 *Define* $\gamma_s^i := Y_s^{i,D} \boldsymbol{X}_s^i \boldsymbol{b}_s$*, where* $\boldsymbol{X}_s^i$ *is the i'th row of* $\boldsymbol{X}_s^{(n)}$*. If the assumptions of Theorem 1 are satisfied, then*

$$\lim_{n\longrightarrow\infty} P\Big(\sup_t \Big| \int_0^t \boldsymbol{\Gamma}^{(n)-1} \frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)} \boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D} - \gamma_s^i) ds \Big| \geq \delta\Big) = 0 \qquad (29)$$

for every $\delta > 0$*.*

Proof Assumption III) from Theorem 1 and Lemma 1 implies that

$$\lim_{J\to\infty} \inf_n P\Big(\sup_t \big|\Gamma_t^{(n)-1} \frac{1}{n}\sum_{i=1}^n R_{t-}^{(i,n)} \boldsymbol{X}_{t-}^{i\intercal}(\lambda_t^{i,D} - \gamma_t^i)\big| > J\Big) = 0. \qquad (30)$$

Moreover, Lemma 2 implies that

$$\frac{1}{n}\sum_{i=1}^n R_{t-}^{(i,n)} \boldsymbol{X}_{t-}^{i\intercal}(\lambda_t^{i,D} - \gamma_t^i)$$

converges in probability to

$$E_P\big[R_{t-}^1 \boldsymbol{X}_{t-}^{1\intercal}(\lambda_t^{1,D} - \gamma_t^1)\big]$$

However, from the innovation theorem we have that this equals

$$E_{\tilde{P}}\big[\boldsymbol{X}_{t-}^{1\intercal}(\lambda_t^{1,D} - \gamma_t^1)\big] = E_{\tilde{P}}\big[\boldsymbol{X}_{t-}^{1\intercal}(E_{\tilde{P}}[\lambda_t^{1,D}|\mathcal{F}_{t-}^{1,\mathcal{V}_0}] - \gamma_t^1)\big] = 0,$$

since $\boldsymbol{X}_{t-}^1$ *and* γ_t^1 *are* $\mathcal{F}_{t-}^{1,\mathcal{V}_0}$ *measurable. This and* (30) *enables us to apply Andersen et al. (1993, Lemma II.5.3) to obtain* (29)*.*

Lemma 5 *Suppose that II) and III) from Theorem 1 are satisfied and let* $\boldsymbol{M}_t^{(n)} := \left(N_t^{1,D} - \int_0^t \lambda_s^{1,D} ds, \ldots, N_t^{n,D} - \int_0^t \lambda_s^{n,D} ds\right)^\intercal$. *Then*

$$\boldsymbol{\Xi}_t^{(n)} := \frac{1}{n} \int_0^t \boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} d\boldsymbol{M}_s^{(n)} \tag{31}$$

defines a square integrable local martingale with respect to the filtration $\mathcal{F}_s^{1,\mathcal{V}_0\cup\mathcal{L}} \otimes \cdots \otimes \mathcal{F}_s^{n,\mathcal{V}_0\cup\mathcal{L}}$ *and*

$$\lim_{n\to\infty} P\bigg(\operatorname{Tr}(\langle \boldsymbol{\Xi}^{(n)} \rangle_T) \geq \delta \bigg) = 0 \tag{32}$$

for every $\delta > 0$.

Proof Writing $\boldsymbol{\lambda}^{(n)}$ *for the diagonal matrix with i'th diagonal element equal to* $\lambda^{i,D}$, *we have that*

$$\operatorname{Tr}(\langle \boldsymbol{\Xi}^{(n)} \rangle_T) = \int_0^T \frac{1}{n^2} \operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{\lambda}_s^{(n)} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)} \boldsymbol{\Gamma}_s^{(n)-1} \right) ds. \tag{33}$$

Moreover,

$$\begin{aligned}
&\frac{1}{n^2} \operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{\lambda}_s^{(n)} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)} \boldsymbol{\Gamma}_s^{(n)-1} \right) && (34)\\
\leq &\frac{1}{n^2} \operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)} \boldsymbol{\Gamma}_s^{(n)-1} \right) \max_{i\leq n} Y_s^{i,D} R_{s-}^{(i,n)} \lambda_s^{i,D} && (35)\\
\leq &\operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \right) \left(\frac{1}{n} \max_{i\leq n} \lambda_s^{i,D}\right) \|R^{(i,n)}\|_\infty && (36)\\
\leq &\operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \right) \left(\frac{1}{n} \sum_{i\leq n} \lambda_s^{i,D}\right) \|R^{(i,n)}\|_\infty && (37)
\end{aligned}$$

Now, III), (37) and Lemma 1 implies that

$$\lim_{a\to\infty} \inf_n P\bigg(\sup_s \frac{1}{n^2} \operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{\lambda}_s^{(n)} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)} \boldsymbol{\Gamma}_s^{(n)-1} \right) \geq a \bigg) = 0.$$

On the other hand, Lemma 3, (36) and III) gives us that

$$\lim_{n\to\infty} P\bigg(\frac{1}{n^2} \operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{\lambda}_s^{(n)} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)} \boldsymbol{\Gamma}_s^{(n)-1} \right) \geq \delta \bigg) = 0$$

for every s and $\delta > 0$, *so Andersen et al. (1993, Propositon II.5.3) implies that (31) also holds.*

Proof (Theorem 1) We have the following decomposition:

$$\begin{aligned}\boldsymbol{B}_t^{(n)} - \boldsymbol{B}_t &= \int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)})^{-1}\big(\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{\lambda}_s^{(n)} - \boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)}b_s\big)ds \\ &\quad + \int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)})^{-1}\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}d\boldsymbol{M}_s^{(n)} \\ &= \int_0^t \boldsymbol{\Gamma}^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D} - \gamma_s^i)ds + \boldsymbol{\Xi}_t^{(n)}.\end{aligned}$$

Leglarts inequality (Jacod and Shiryaev, 2003, Lemma I.3.30) together with Lemma 5 implies that $\boldsymbol{\Xi}^{(n)}$ converges uniformly in probability to 0. Moreover, Lemma 4 implies that $\int_0^{\cdot}\boldsymbol{\Gamma}^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D} - \gamma_s^i)ds$ converges in same sense to 0, which proves the consistency.

To see that $\boldsymbol{B}^{(n)}$ is P-UT, note that it coincides with the sum of $\boldsymbol{B}_t$, $\boldsymbol{\Xi}^{(n)}$ and $\int_0^{\cdot}\boldsymbol{\Gamma}_s^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^i - \gamma_s^i)ds$. According to Ryalen et al. (2017, Lemma 1), the latter is P-UT since III) and Lemma 1 implies (7). Moreover, $\boldsymbol{B}_t = \int_0^{\cdot}\boldsymbol{b}_s ds$ is clearly P-UT, since $\boldsymbol{b}_t$ is uniformly bounded. $\boldsymbol{\Xi}^{(n)}$ is also P-UT since Lemma 5 implies that (8) is satisfied. Finally, as $\boldsymbol{B}^{(n)}$ is a sum of three processes that are P-UT, it is necessarily P-UT itself.

Proof of Theorem 2

Lemma 6 *Suppose that c. and d. from Theorem 2 are satisfied, and that*

I)

$$\lim_{a\to\infty}\sup_n P\bigg(\sup_t\big|\theta_t^{(i,n)}\big| \geq a\bigg) = 0,$$

II) $\theta_{t-}^{(i,n)}$ *converges to* θ_t^i *in probability for each* i *and* t.

Then we have that $K^{(i,n)}$ *is predictably uniformly tight (P-UT) and*

$$\lim_n P\bigg(\sup_t\big|K_t^{(i,n)} - K_t^i\big| \geq \epsilon\bigg) = 0 \tag{38}$$

for every i *and* $\epsilon > 0$.

Proof Note that

$$K_t^{(i,n)} - K_t^i = \int_0^t(\theta_{s-}^{(i,n)} - \theta_s^i)dN_s^{i,A} + n^{-1/2}\int_0^t Y_s^i\boldsymbol{Z}_{s-}^{i\intercal}d\boldsymbol{W}_s^{(n)} - n^{-1/2}\int_0^t Y_s^{i,A}\tilde{\boldsymbol{Z}}_{s-}^{i\intercal}d\tilde{\boldsymbol{W}}_s^{(n)}, \tag{39}$$

where $\boldsymbol{W}_t^{(n)} := n^{1/2}(\boldsymbol{H}_t^{(n)} - \boldsymbol{H}_t)$ *and* $\tilde{\boldsymbol{W}}_t^{(n)} := n^{1/2}(\tilde{\boldsymbol{H}}_t^{(n)} - \tilde{\boldsymbol{H}}_t)$ *are square-integrable martingales with respect to* $\mathcal{F}_t^{1,\mathcal{V}_0\cup\mathcal{L}}\otimes\cdots\otimes\mathcal{F}_t^{n,\mathcal{V}_0\cup\mathcal{L}}$ *and* $\mathcal{F}_t^{1,\mathcal{V}_0}\otimes\cdots\otimes\mathcal{F}_t^{n,\mathcal{V}_0}$ *respectively.*

Let τ be an optional stopping time and note that

$$E\left[\left|\int_0^\tau (\theta_{s-}^{(i,n)} - \theta_s^i) dN_s^{i,A}\right|\right] \leq E\left[\int_0^\tau \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| dN_s^{i,A}\right] = E\left[\int_0^\tau \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| \lambda_s^{i,A} ds\right],$$

so by Lenglarts inequality, (Jacod and Shiryaev, 2003, I.3.30), we see that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T}\left|\int_0^t (\theta_{s-}^{(i,n)} - \theta_s^i) dN_s^{i,A}\right| \geq \epsilon\right) = 0 \tag{40}$$

for every $\epsilon > 0$ if

$$\lim_{n\longrightarrow\infty} P\left(\int_0^T \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| \lambda_s^{i,A} ds \geq \epsilon\right) = 0, \tag{41}$$

for every $\epsilon > 0$. The latter property holds due to I), II) and Andersen et al. (1993, Proposition II.5.3).

Since $\{\int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}\}_n$ converges in the skorokhod topology, we have that $\{\sup_{t\leq T} |\int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}|\}_n$ is tight (Jacod and Shiryaev, 2003, Theorem VI.3.21). Therefore, we also get that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} |n^{-1/2} \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}| \geq \epsilon\right) = 0 \tag{42}$$

for every $\epsilon > 0$. For the same reason we also have

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} |n^{-1/2} \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{W}}_s^{(n)}| \geq \epsilon\right) = 0. \tag{43}$$

By combining (42),(43) and (40), we obtain that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} |K_t^{(i,n)} - K_t^i| \geq \epsilon\right) = 0 \tag{44}$$

for every $\epsilon > 0$.

To see that $K^{(i,n)}$ is P-UT, note that the compensator of $\int_0^\cdot (\theta_{s-}^{(i,n)} - 1) dN_s^{i,A}$ equals $\int_0^\cdot (\theta_{s-}^{(i,n)} - 1)\lambda_s^{i,A} ds$ and

$$\langle \int_0^\cdot (\theta_{s-}^{(i,n)} - 1) dN_s^{i,A} - \int_0^\cdot (\theta_{s-}^{(i,n)} - 1)\lambda_s^{i,A} ds\rangle_T = \int_0^T (\theta_{s-}^{(i,n)} - 1)^2 \lambda_s^{i,A} ds.$$

The assumptions I) in this Lemma and c) together with Ryalen et al. (2017, Lemma 1) therefore imply that $\int_0^\cdot (\theta_{s-}^{(i,n)} - 1) dN_s^{i,A}$ is P-UT.

To see that $\int_0^\cdot Y_s^i \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}$ is P-UT, note that

$$\int_0^\cdot Y_s^i \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)} = n^{-1/2} \int_0^\cdot Y_s^i \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{W}}_s^{(n)} + \int_0^\cdot Y_s^i \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s. \tag{45}$$

An analogous decompositon yields that $\int_0^\cdot Y_s^i \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)}$ is P-UT. This means that $K^{(i,n)}$ is a sum of three processes that are P-UT, and must therefore be P-UT itself.

Lemma 7 *Suppose that*

I) $\{\kappa_n\}_n$ *increasing sequence of positive numbers such that*

$$\lim_{n\longrightarrow\infty}\kappa_n=\infty \text{ and } \sup_n\frac{\kappa_n}{\sqrt{n}}<\infty,$$

II) $\boldsymbol{h}_t$ *is a bounded and continuous vector valued function,*
III) $\boldsymbol{Z}^i$ *is caglad with* $E[\sup_{t\leq T}|\boldsymbol{Z}^i_t|^3_3]<\infty$,
IV)

$$\lim_{J\to\infty}\sup_n P\bigg(\text{Tr}\left((\frac{1}{n}\boldsymbol{Z}^{(n)\intercal}_{t-}\boldsymbol{Y}^{(n),A}_t\boldsymbol{Z}^{(n)}_{t-})^{-1}\right)\geq J\bigg)=0 \tag{46}$$

V) $Y^{i,A}\boldsymbol{Z}^{i\intercal}_{\cdot-}\boldsymbol{h}$ *defines the intensity for* $N^{i,A}$ *with respect to* P *and* $\mathcal{F}^{i,\mathcal{V}_0}_\cdot$. *Now,*

$$\lim_{n\longrightarrow\infty}P\bigg(\sup_{1/\kappa_n\leq t\leq T}\Big|\kappa_n\int_{t-1/\kappa_n}^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{H}^{(n)}_s-Y^{i,A}_t\boldsymbol{Z}^{i\intercal}_{t-}\boldsymbol{h}_t\Big|\geq\epsilon\bigg)=0. \tag{47}$$

Proof Note that

$$\kappa_n\int_{t-1/\kappa_n}^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{H}^{(n)}_s-Y^{i,A}_t\boldsymbol{Z}^{i\intercal}_{t-}\boldsymbol{h}_t \tag{48}$$

$$=\frac{\kappa_n}{\sqrt{n}}\int_0^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{W}^{(n)}_s-\frac{\kappa_n}{\sqrt{n}}\int_0^{t-1/\kappa_n}Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{W}^{(n)}_s \tag{49}$$

$$+\kappa_n\int_{t-1/\kappa_n}^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}\boldsymbol{h}_sds-Y^{i,A}_t\boldsymbol{Z}^{i\intercal}_{t-}\boldsymbol{h}_t. \tag{50}$$

The martingale central limit theorem implies that $\{\boldsymbol{W}^{(n)}\}$ *is a sequence of martingales that converges in law to a continuous Gaussian processes with independent increments, see Andersen et al. (1993). Moreover, Ryalen et al. (2017, Proposition 1) says that* $\{\boldsymbol{W}^{(n)}\}_n$ *is P-UT.*

Therefore Jacod and Shiryaev (2003, Theorem VI 6.22) implies that $\int_0^\cdot Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{W}^{(n)}_s$ *converges in law to a continuous process, so it is C-tight. Moreover, from Jacod and Shiryaev (2003, Proposition VI.3.26) we have that*

$$\lim_{n\longrightarrow\infty}P\bigg(\sup_{1/\kappa_n\leq t\leq T}\Big|\int_0^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{W}^{(n)}_s-\int_0^{t-1/\kappa_n}Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}d\boldsymbol{W}^{(n)}_s\Big|\geq\epsilon\bigg)=0 \tag{51}$$

for every $\epsilon>0$. *The mean value theorem of elementary calculus implies that*

$$\lim_{n\longrightarrow\infty}\sup_{1/\kappa_n\leq t\leq T}\Big|\kappa_n\int_{t-1/\kappa_n}^t Y^{i,A}_s\boldsymbol{Z}^{i\intercal}_{s-}\boldsymbol{h}_sds-Y^{i,A}_t\boldsymbol{Z}^{i\intercal}_{t-}\boldsymbol{h}_t\Big|=0 \tag{52}$$

P *a.s. Combining* (51) *and* (52) *yields the claim.*

Proof (Proof of Theorem 2)

Combining (16) and the decomposition in the proof of Lemma 7, we see that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n\leq t\leq T}\left|\kappa_n\int_{t-1/\kappa_n}^{t} Y_s^{i,A}\tilde{\boldsymbol{Z}}_{s-}^{i\intercal}d\tilde{\boldsymbol{H}}_s^{(n)}/\tilde{\lambda}_t^{i,A}-1\right|\geq\epsilon\right)=0. \quad (53)$$

Combining (16) and a. we also have

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n\leq t\leq T}\left|\kappa_n\int_{t-1/\kappa_n}^{t} Y_s^{i,A}\boldsymbol{Z}_{s-}^{i\intercal}d\boldsymbol{H}_s^{(n)}/\lambda_t^{i,A}-1\right|\geq\epsilon\right)=0. \quad (54)$$

Whenever $t\geq 1/\kappa_n$, we have that by the continuous mapping theorem that

$$\begin{aligned}
&\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n\leq t\leq T}\left|\theta_t^{(i,n)}-\theta_t^i\right|\geq\epsilon\right)\\
=&\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n\leq t\leq T}\left|\theta_t^i\left(\frac{\kappa_n\int_{t-1/\kappa_n}^{t} Y_s^{i,A}\tilde{\boldsymbol{Z}}_{s-}^{i\intercal}d\tilde{\boldsymbol{H}}_s^{(n)}/\tilde{\lambda}_t^{i,A}}{\kappa_n\int_{t-1/\kappa_n}^{t} Y_s^{i,A}\boldsymbol{Z}_{s-}^{i\intercal}d\boldsymbol{H}_s^{(n)}/\lambda_t^{i,A}}-1\right)\right|\geq\epsilon\right)\\
=&0.
\end{aligned}$$

Since θ^i is right-continuous at $t=0$, we have that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{0\leq t\leq T}\left|\theta_t^{(i,n)}-\theta_t^i\right|\geq\epsilon\right)=0. \quad (55)$$

Finally, Jacod and Shiryaev (2003, Corollary VI 3.33) implies that $\{(R_0^{(i,n)},K^{(i,n)})\}_n$ converges to (R_0^i,K^i) in probability. Since $K^{(i,n)}$ is P-UT,

$$R_t^{(i,n)}=1+\int_0^t R_{s-}^{(i,n)}dK_s^{(i,n)}$$

and

$$R_t^i=1+\int_0^t R_{s-}^i dK_s^i$$

Jacod and Shiryaev (2003, Theorem IX 6.9) implies that $R^{(i,n)}$ converges to R^i in probability.

References

O. Aalen, R. Cook, and K. Røysland. Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime data analysis*, 21 (4):579–593, 2015.

P. Andersen, Ø. Borgan, R. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.

S. Cole and M. Hernán. Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.*, 168(6):656–664, Sep 2008.

W. Havercroft and V. Didelez. Simulating from marginal structural models with time-dependent confounding. *Statistics in medicine*, 31(30):4190–4206, 2012.

M. Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

M. Hernán, B. Brumback, and J. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5):561–570, 2000a. ISSN 10443983. URL `http://www.jstor.org/stable/3703998`.

M. Hernán, B. Brumback, and J. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men, 2000b.

F. Huffer and I. McKeague. Weighted least squares estimation for aalen's additive risk model. *Journal of the American Statistical Association*, 86(413): 114–129, 1991. ISSN 01621459. URL `http://www.jstor.org/stable/2289721`.

J. Jacod. Multivariate point processes: Predictable projection, radon-nikodym derivatives, representation of martingales. *Probability Theory and Related Fields*, 1975.

J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43932-3.

M. Joffe, T. Ten Have, H. Feldman, and S. Kimmel. Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279, 2004.

I. McKeague. Asymptotic theory for weighted least squares estimators in aalen's additive risk model, 1987.

J. Pearl. *Causality: Models, Reasoning and Inference 2nd Edition*. Cambridge University Press, 2000.

J. Robins. Structural nested failure time models. *Wiley StatsRef: Statistics Reference Online*, 2014.

J. Robins and S. Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, pages 1125–1138, 1989.

J. Robins, M. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

K. Røysland. A martingale approach to continuous-time marginal structural models. *Bernoulli*, 2011.

P. Ryalen, M. Stensrud, and K. Røysland. Transforming cumulative hazard estimates. *arXiv preprint arXiv:1710.07422*, 2017.

P. Ryalen, M. Stensrud, S. Fosså, and K. Røysland. Causal inference in continuous time: an example on prostate cancer therapy. *Biostatistics*, page kxy036, 2018a. doi: 10.1093/biostatistics/kxy036. URL `http://dx.doi.org/10.1093/biostatistics/kxy036`.

P. Ryalen, M. Stensrud, and K. Røysland. Transforming cumulative hazard estimates. *Biometrika*, page asy035, 2018b. doi: 10.1093/biomet/asy035. URL http://dx.doi.org/10.1093/biomet/asy035.

M. Stensrud, M. Valberg, K. Røysland, and O. Aalen. Exploring selection bias by causal frailty models: The magnitude matters. *Epidemiology*, 28(3): 379–386, 2017.

M. Stensrud, K. Røysland, and P. Ryalen. On null hypotheses in survival analysis. *ArXiv e-prints*, July 2018.

S. Vansteelandt and A. Sjolander. Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods*, 5(1):37–56, 2016.