



# On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects

Paul Frédéric Blanche<sup>1,2</sup> · Anders Holt<sup>2</sup> · Thomas Scheike<sup>1</sup> 

Received: 20 September 2021 / Accepted: 9 June 2022 / Published online: 7 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

Simple logistic regression can be adapted to deal with right-censoring by inverse probability of censoring weighting (IPCW). We here compare two such IPCW approaches, one based on weighting the outcome, the other based on weighting the estimating equations. We study the large sample properties of the two approaches and show that which of the two weighting methods is the most efficient depends on the censoring distribution. We show by theoretical computations that the methods can be surprisingly different in realistic settings. We further show how to use the two weighting approaches for logistic regression to estimate causal treatment effects, for both observational studies and randomized clinical trials (RCT). Several estimators for observational studies are compared and we present an application to registry data. We also revisit interesting robustness properties of logistic regression in the context of RCTs, with a particular focus on the IPCW weighting. We find that these robustness properties still hold when the censoring weights are correctly specified, but not necessarily otherwise.

**Keywords** Average treatment effect · Competing risks · Ipcw adjustment · Logistic regression

## 1 Introduction

To handle right censored data when fitting logistic regression models, it has been suggested to weight the estimating equations, see. e.g. Zheng et al. (2006); Uno et al. (2007), or to weight the outcome, see e.g. Scheike et al. (2008). In this manuscript

---

✉ Thomas Scheike  
ts@biostat.ku.dk

<sup>1</sup> Section of Biostatistics, Department of Public Health, University of Copenhagen, Øster Farimagsgade 5B, P.O.B. 2099, 1014 Copenhagen K, Denmark

<sup>2</sup> Department of Cardiology, Copenhagen University Hospital–Herlev and Gentofte, Copenhagen, Denmark

we review these two approaches and we study and compare their large sample properties. We further present their potential to estimate meaningful treatment effects in observational studies and randomized clinical trials.

The two weighting approaches can be used very similarly with or without competing risks, which is convenient. As pointed out by e.g. Schumacher et al. (2016), competing risks are very common in medical research. The two weighting approaches have the advantage that they directly extend the usual logistic regression method for binary data to both survival data and competing risks data.

Logistic regression is appealing both for its simplicity and its direct interpretation when the interest lies in estimating a risk or survival probability at a fixed time-point. We refer to this as a  $t$ -year risk or survival, in what follows. Unlike alternative methods based on hazards regression, the parameters of the model lead to a direct interpretation for the  $t$ -year risk or survival (probability) being modeled. The modeling assumptions are also easier to communicate and check, in our opinion. This is in contrast to hazard-based regression, where model parameters do not have a direct interpretation for the  $t$ -year risk or survival, and are commonly misunderstood by clinicians. This is true without competing risks (Sutradhar and Austin 2018), and even more so with competing risks. Standard approaches to model hazards involve structural assumptions, typically proportional hazards, whose consequences for the risk of interest are typically not clear. In addition, it can be challenging to assess the validity of these structural assumptions. Stensrud and Hernán (2020) also mentioned that treatment effects calculated from  $t$ -year risks are often more helpful for clinical decision-making and more easily understood by patients than hazard ratios.

The rest of the manuscript is organized as follows. In Sect. 2, we introduce the estimating equations to fit a logistic regression model and we study the large sample properties of the two different weighting approaches. To derive the large sample results, we use only “simple” arguments based on Taylor expansions and “standard” martingales theory (Aalen et al. 2008, Sec. 2.2). We avoid the use of projection arguments as in e.g. Tsiatis (2006), which we hope facilitates the understanding of the origins of the key results. Building on the comparison of the asymptotic properties of each weighting approach, in Sect. 2.8, we show that the two approaches sometimes have very different performances, in terms of efficiency. In Sect. 3, we further show how to use logistic regression to build G-computation (standardized) or double robust estimators to account for confounding in observational studies, when estimating treatment effects. Here we focus on the situation with a baseline treatment and baseline confounders. We do not cover the more complex settings with time-dependent treatment and/or time-dependent confounding. As compared to similar approaches building on hazard-based regression, as that of Ozenne et al. (2020), here logistic regression enables a direct parametrization of the outcome model. This can facilitate modeling choices and the discussion of their strengths and limitations. In Sect. 4, we present an application to Danish registry data (Holt et al. 2021), where we compared the 33-months risk of cardiovascular death among patients who have initiated a beta-blocker treatment to those who have not. In Sect. 5, we revisit interesting robustness properties of logistic regression fitted to randomized clinical trials (RCT) data. We find that these robustness properties still hold with the two weighting approaches, when the censoring weights are correctly specified. Finally, Sect. 6 presents a discussion.

## 2 Logistic regression with censored data

### 2.1 Observed data

We assume to observe  $(\tilde{T}_i, \Delta_i, \tilde{\eta}_i, \mathbf{L}_i, A_i), i = 1, \dots, n$ , i.i.d. copies of  $(\tilde{T}, \Delta, \tilde{\eta}, \mathbf{L}, A)$ . Here  $T$  is the time-to-event with cause of death  $\eta \in \{1, 2\}$ , that is observed subject to right censoring. Specifically, because of a censoring time  $C$ , we observe  $\tilde{T} = \min(T, C)$ , the right censored time to event,  $\Delta = \mathbb{1}\{T \leq C\}$ , the event (or non-censoring) indicator and the observed cause of death  $\tilde{\eta} = \Delta\eta$ . In addition to  $(\tilde{T}, \Delta, \tilde{\eta})$ , we also observe baseline covariates  $\mathbf{L}$  and a binary treatment group  $A$ .

In the worked example we consider in Sect. 4,  $T$  is the time from start of follow-up to death, with  $\eta \in \{1, 2\}$  indicating either a cardiovascular death, when  $\eta = 1$ , or a non-cardiovascular death, when  $\eta = 2$ . The censoring time  $C$  is the time from start of follow-up to either administrative censoring at end of follow-up (e.g. 31 December 2018) or loss of follow-up due to other reasons (here only emigration), whichever comes first. The variable  $A$  indicates whether a patient has initiated a beta-blocker treatment ( $A = 1$ ) or not ( $A = 0$ ) within 3 months before start of follow-up. As follow-up starts 3 months after a myocardial infarction (MI), the variable  $A$  indicates whether a patient has initiated the treatment within 3 months after MI. Here  $\mathbf{L}$  denotes a vector of many baseline (pre-treatment) variables, such as age and diabetes history (see Sect. 4 for the full list).

Finally, note that the case without competing risk simply corresponds to the case where  $\eta = 1$  with probability 1. In that case, the clinical interpretation of the statistical results will often be simpler. The fact that we assume only two competing events is also without loss of generality as all competing events can always be grouped in a single one.

### 2.2 Additional notations

Further, for a given time horizon  $t$  of interest (e.g.  $t=1$  year) we also define the binary indicator of experiencing a main event within  $t$ -year as  $D(t) = \mathbb{1}\{T \leq t, \eta = 1\}$ . Because of censoring,  $D(t)$  is not always fully observed, but we do observe  $\tilde{D}(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{\eta} = 1\}$ . These two indicators are related via the equation  $\tilde{D}(t) = D(t) \cdot \Delta(t)$ , where  $\Delta(t) = \mathbb{1}\{t \wedge T \leq C\} = 1 - \mathbb{1}\{\tilde{T} \leq t\} \cdot (1 - \Delta)$  is the indicator of not observing a censored time before  $t$ . We use the notation  $x \wedge y$  to indicate the minimum of  $x$  and  $y$ . To present technical results, we further introduce the notation  $\lambda_c(t, A, \mathbf{L})$  for the hazard rate of  $C$  given the covariates  $(\mathbf{L}, A)$  and we let  $M^c(t, A, \mathbf{L}) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 0\} - \int_0^t Y(v) \lambda_c(v, A, \mathbf{L}) dv$  be the related standard martingale, where  $Y(v) = \mathbb{1}\{\tilde{T} \geq v\}$ . When the hazard rate  $\lambda_c(t, A, \mathbf{L})$  is assumed to not depend on either  $\mathbf{L}$  or  $A$  (or both), we will omit them in the notation, for conciseness. We will also use the notation  $\mathbf{V}^2 = \mathbf{V}\mathbf{V}^T$ , for any vector  $\mathbf{V}$ , and  $\text{Var}(\mathbf{V})$  to denote the variance-covariance matrix of  $\mathbf{V}$ . We also use the “expit” and “logit” functions, defined as  $\text{expit}(x) = \exp(x)/\{1 + \exp(x)\}$  and  $\text{logit}(x) = \text{expit}^{-1}(x) = \log\{x/(1 - x)\}$ .

### 2.3 Assumptions about the censoring

Throughout the manuscript, we assume the usual independent censoring assumption that  $C$  is independent of the outcome  $(T, \eta)$  given  $(L, A)$ . This allows some dependency between  $C$  and  $(L, A)$  that is present in many practical situations such as the study of cardiovascular death presented in Sect. 4. In some sections, we will make the stronger assumption that  $C$  is independent of  $T$  to derive specific results. This assumption is more restrictive but nevertheless still often plausible and used in practice. For identifiability, we further assume that not all subjects of any subgroup defined at baseline are systematically censored before  $t$ , i.e.,  $P\{C > t | L = l, A = a\} > 0$  for all  $l$  and  $a = 0, 1$ .

### 2.4 Logistic regression & estimating equations

In this section, we aim to estimate a conditional  $t$ -year risk of event, often known as the *absolute risk* (Pfeiffer and Gail 2017) and defined as  $F_1(t, a, l) = P\{D(t) = 1 | A = a, L = l\}$ . Within our context presented in Sect. 2.1, it has a simple interpretation as the proportion of patients who are expected to die from a cardiovascular death within the 33 months of follow-up, among patients from treatment group  $A = a$ , with baseline characteristics  $L = l$ . Both Geskus (2016) and Young et al. (2020) discuss the use of this quantity as well as various other relevant quantities. We assume that the absolute risk  $F_1(t, a, l)$  can be modeled by a logistic model, that is,

$$Q(t, \mathbf{x}, \boldsymbol{\beta}) = \text{expit}(\mathbf{x}^T \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta})},$$

where  $\mathbf{x}$  is a pre-specified vector of  $p$  random variables constructed from  $(l, a)$  and  $\mathbf{x}^T \boldsymbol{\beta}$  denotes the linear predictor associated to  $X = \mathbf{x}$ . For example,  $X = (1, A, L)^T$  is used in the simplest model and  $X = (1, A, L, AL)^T$  can be used to model interactions for heterogeneous treatment effects. Note that we implicitly assume that the parameter vector  $\boldsymbol{\beta}$  depends on  $t$ . For ease of notation we will sometimes write  $Q(t, \mathbf{x})$  for  $Q(t, \mathbf{x}, \boldsymbol{\beta})$ , without emphasizing that it depends on the parameter  $\boldsymbol{\beta}$ . Whenever convenient, we will also use  $F_1(t, \mathbf{x})$  to denote the conditional risk  $P\{D(t) = 1 | X = \mathbf{x}\}$ .

For uncensored data, i.e. assuming  $P(\Delta = 1) = 1$ , the binary variable  $D_i(t)$  is observed for all subjects  $i = 1, \dots, n$ . In that case, we can simply fit a usual generalized linear model (glm) and it is well known that maximizing the likelihood leads to the following  $p$  estimating equations, see e.g. McCullagh and Nelder (1989),

$$U_{glm}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n X_i \left\{ D_i(t) - Q(t, X_i, \boldsymbol{\beta}) \right\} = \mathbf{0}. \tag{1}$$

For right censored data,  $D_i(t)$  is not observed for all subjects, but only for subjects  $i$  such that  $\Delta_i(t) = 1$ . Two modifications of the estimating equations have been proposed to obtain consistent estimators via inverse probability of censoring weighting. The first consists of weighting each individual contribution in (1) as follows,

$$\widehat{U}_{ipcw-glm}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \widehat{W}_i(t) \cdot \mathbf{X}_i \left\{ D_i(t) - Q(t, \mathbf{X}_i, \boldsymbol{\beta}) \right\} = \mathbf{0} \tag{2}$$

using subject specific weight

$$\widehat{W}_i(t) = \frac{\Delta_i(t)}{\widehat{G}_c(t \wedge \widetilde{T}_i, \mathbf{X}_i)} = \frac{\mathbb{1}\{\widetilde{T}_i \leq t\} \cdot \Delta_i}{\widehat{G}_c(\widetilde{T}_i, \mathbf{X}_i)} + \frac{\mathbb{1}\{\widetilde{T}_i > t\}}{\widehat{G}_c(t, \mathbf{X}_i)},$$

where  $\widehat{G}_c(u, \mathbf{x}_i)$  denotes a consistent estimator of  $G_c(u, \mathbf{x}_i) = P(C_i > u | \mathbf{X}_i = \mathbf{x}_i)$  for all  $u \leq t$ . We refer to this approach below as IPCW-GLM. Although  $D_i(t)$  is not observed for all subjects, we note that (2) can be computed since  $\widehat{W}_i(t) \cdot \widetilde{D}_i(t) = \widehat{W}_i(t) \cdot D_i(t)$ .

Often, a Kaplan-Meier estimator is used to compute  $\widehat{G}_c$ . It may be computed from the entire dataset, when the censoring is independent of the covariates, or within strata defined from  $\mathbf{x}_i$  (e.g. in each treatment group) when this is not the case. Alternatively, a Cox model can be used to model the censoring distribution. This approach has been used by Zheng et al. (2006) and Uno et al. (2007) in a context without competing risks and by Azarang et al. (2017) in presence of competing risks.

A second approach consists of weighting each individual outcome in (1) as follows,

$$\widehat{U}_{oipcw}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \left\{ D_i(t) \cdot \widehat{W}_i(t) - Q(t, \mathbf{X}_i, \boldsymbol{\beta}) \right\} = \mathbf{0} . \tag{3}$$

Here again, we could equivalently write  $D_i(t)$  instead of  $\widetilde{D}_i(t)$  in (3) to clarify that the computation can be done using the observed censored data. We refer to this approach as the outcome weighted IPCW approach or the OIPCW below. This approach has been used by Scheike et al. (2008), among others, in a context with competing risks.

### 2.5 Properties and comparison of the two approaches

In this section, we study and compare some properties of the two different IPCW adjustments given by (2) or (3). Surprisingly, the two weighting approaches can lead to very different performance in some settings, as it will be illustrated in Sect. 2.8. But, first, we start by noting a situation in which the two approaches are identical. This happens when  $\widehat{G}_c(t, \mathbf{x}_i)$  is fully non-parametric, that is, when for all values of  $\mathbf{x}_i$ , the survival probability  $G_c(t, \mathbf{x}_i)$  can be estimated with the Kaplan-Meier estimator. Indeed, for each strata  $s(\mathbf{x}_i)$  of  $n_{s(\mathbf{x}_i)}$  subjects having this covariate value  $\mathbf{x}_i$ , we note that  $\sum_{j \in s(\mathbf{x}_i)} \widehat{W}_j(t) = n_{s(\mathbf{x}_i)}$  for all  $t$  such that  $\widehat{G}_c(t, \mathbf{x}_i) > 0$ , see e.g. Appendix A of Cortese et al. (2017). Therefore it follows that

$$\sum_{j \in s(\mathbf{x}_i)} \widehat{W}_j(t) \mathbf{x}_j Q(t, \mathbf{x}_j) = n_{s(\mathbf{x}_i)} \mathbf{x}_i Q(t, \mathbf{x}_i) ,$$

which makes the two set of estimating Eqs. (2) and (3) equivalent and give the same estimator.

Second, we study the large sample properties of the two IPCW adjustments given by (2) or (3). In the remainder of this section, we will restrict our attention to the case in which a simple Kaplan-Meier estimator is used to estimate the censoring distribution and the distribution of  $C$  does not depend on the covariates  $(A, \mathbf{L})$ . We will not necessarily assume that the logistic model is correctly specified. That is, we allow for  $Q(t, \mathbf{x}, \boldsymbol{\beta}) \neq F_1(t, \mathbf{x})$ . In that case,  $\boldsymbol{\beta}$  denotes the limit of the solution to the estimating Eq. (1), when the sample size  $n$  tends to infinity. Such a  $\boldsymbol{\beta}$  exists under mild conditions, see e.g. Uno et al. (2007). Following similar lines as those of Bang and Tsiatis (2000), in Appendix A.1, we provide i.i.d. decompositions for the score equations, in Theorem 1 below.

**Theorem 1** *The scores  $\widehat{U}_{oipcw}(\boldsymbol{\beta})$  and  $\widehat{U}_{ipcw-glm}(\boldsymbol{\beta})$  have the i.i.d. representation*

$$\sqrt{n} \widehat{U}_m(\boldsymbol{\beta}) = n^{-1/2} \sum_{i=1}^n \Phi_m\{X_i, D_i(t), \widetilde{T}_i, \widetilde{\eta}_i, t\} + o_p(1),$$

where the subscript  $m$  denotes either “oipcw” or “ipcw-glm”, with

$$\Phi_m\{X, D(t), \widetilde{T}, \widetilde{\eta}, t\} = X\{D(t) - Q(t, X)\} - \int_0^t \varphi_m(X, D(t), s) \frac{dM^c(s)}{G_c(s)}, \quad (4)$$

where

$$\begin{aligned} \varphi_{oipcw}(X, D(t), s) &= XD(t) - E[XD(t) \mid T \geq s], \\ \varphi_{ipcw-glm}(X, D(t), s) &= \varphi_{oipcw}(X, D(t), s) \\ &\quad - \left\{ XQ(t, X) - E[XQ(t, X) \mid T \geq s] \right\}. \end{aligned}$$

From the above Theorem 1, the asymptotic normality of the estimators follows and the asymptotic variances are calculated following standard martingale theory (Aalen et al. 2008, Sec. 2.2). Note that similar results have been given before in Scheike et al. (2008) for the OIPCW case and by Azarang et al. (2017) for the IPCW-GLM case, although the studied estimation equations were slightly different.

**Corollary 1** *Let  $\widehat{\boldsymbol{\beta}}_m$  be the estimator that solves  $\widehat{U}_m(\boldsymbol{\beta}) = \mathbf{0}$ . Asymptotically,  $\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_m)$ , with the usual sandwich form of the variance,  $\boldsymbol{\Sigma}_m = \mathcal{I}^{-1} \boldsymbol{\Omega}_m \mathcal{I}^{-1}$ , with “bread”  $\mathcal{I} = E[X^2 Q(t, X)\{1 - Q(t, X)\}]$  and “meat”  $\boldsymbol{\Omega}_m = \mathcal{J} + \int_0^t f_m(s) f_c(s) ds$ , where  $f_c(s) = \lambda_c(s)/G_c(s)$ ,  $\mathcal{J} = E([\mathbf{X}\{D(t) - Q(t, X)\}]^2)$  and  $f_m(s) = E[\{\varphi_m(X, D(t), s)\}^2 \mathbb{1}\{T \geq s\}]$ . A consistent estimator  $\widehat{\boldsymbol{\Sigma}}_m$  of  $\boldsymbol{\Sigma}_m$  is provided in Appendix A.2. Further, if the model is well-specified, that is, if  $Q(t, \mathbf{x}, \boldsymbol{\beta}) = F_1(t, \mathbf{x})$  for all  $\mathbf{x}$ , then  $\mathcal{J} = \mathcal{I}$ .*

**Proof** Following standard martingale theory (Aalen et al. 2008, Sec. 2.2), the second term of the two in the right-hand side of (4) is a martingale relative to the filtration  $\{\mathcal{F}_u\}$ , where  $\mathcal{F}_u$  is the  $\sigma$ -algebra generated by  $\{\mathbb{1}\{C \leq t\}, t \leq u; (T, \eta, X)\}$ . Further,

the two terms in the right-hand side of (4) therefore are uncorrelated as the first term is  $\mathcal{F}(0)$  measurable. With notation  $\mathbf{\Omega}_m = \text{Var}\left[\Phi_m\{X, D(t), \tilde{T}, \tilde{\eta}, t\}\right]$ , as the two terms are also 0 mean, it follows

$$\mathbf{\Omega}_m = E\left([X\{D(t) - Q(t, X)\}]^2\right) + E\left(\left[\int_0^t \varphi_m(X, D(t), s) \frac{dM^c(s)}{G_c(s)}\right]^2\right).$$

By definition  $E\left([X\{D(t) - Q(t, X)\}]^2\right) = \mathcal{J}$  and we note that that  $\mathcal{J} = \mathcal{I}$  follows from the law of iterated expectations, when  $Q(t, \mathbf{x}, \beta) = F_1(t, \mathbf{x})$  for all  $\mathbf{x}$ . Again following standard variance calculation for martingales (Aalen et al. 2008, Sec. 2.2) and the law of iterated expectations, it further follows

$$\begin{aligned} \mathbf{\Omega}_m &= \mathcal{J} + \int_0^t E\left(\left[\varphi_m(X, D(t), s) \frac{Y(s)}{G_c(s)}\right]^2\right) \lambda_c(s) ds, \\ &= \mathcal{J} + \int_0^t E\left[\left\{\varphi_m(X, D(t), s)\right\}^2 \mathbb{1}\{T \geq s\} \frac{E\{\mathbb{1}\{C \geq s\} | X, T, \eta\}}{G_c(s)^2}\right] \lambda_c(s) ds, \\ &= \mathcal{J} + \int_0^t f_m(s) f_c(s) ds, \end{aligned} \tag{5}$$

where  $f_m(s) = E\left[\left\{\varphi_m(X, D(t), s)\right\}^2 \mathbb{1}\{T \geq s\}\right]$  and  $f_c(s) = \lambda_c(s)/G_c(s)$ .

Finally, the delta-method implies that, asymptotically, the variance of the estimator  $\hat{\beta}_m$  solving  $\hat{U}_m(\beta) = \mathbf{0}$  will be proportional, in the sample size, to the usual sandwich form of the variance,  $\mathbf{\Sigma}_m = \mathcal{I}^{-1} \mathbf{\Omega}_m \mathcal{I}^{-1}$ , since  $\mathcal{I}$  is equal to the mean of the derivative of the estimating Eqs. (1), (2) and (3), see e.g., Stefanski and Boos (2002). That is, asymptotically,  $\sqrt{n}(\hat{\beta}_m - \beta) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_m)$ .  $\square$

We now compare the asymptotic efficiency of the two IPCW adjustments given by (2) or (3), when the model is well specified, and show that which of the two is the most efficient depends on the censoring time distribution. Specifically, there exists a function  $\mathbf{g}$  such that  $\mathbf{\Omega}_{ipcw-glm} - \mathbf{\Omega}_{oipcw} = \int_0^t \mathbf{g}(s) f_c(s) ds$  and the average  $\int_0^t \mathbf{g}(s) f_c(s) ds$  can be either positive definite or negative definite, depending on the shape of  $f_c = \lambda_c/G_c$ , and hence of the density of the censoring time  $C$  within  $[0, t]$ . Therefore, for some censoring time distributions the estimator  $\hat{\beta}_{ipcw-glm}$  will be more efficient than  $\hat{\beta}_{oipcw}$ , but for other distributions it will be the other way around. Hence, a general recommendation about which method to prefer between OIPCW and IPCW-GLM cannot be made based on general efficiency criteria. To the best of our knowledge, this result, which we encapsulated in Corollary below, is new. For illustration, in Fig. 1 of Sect. 2.8 we show the diagonal of  $\mathcal{I}^{-1} \mathbf{g}(s) \mathcal{I}^{-1}$  versus  $s \in [0, t]$ , from which originates the difference in asymptotic variance between  $\hat{\beta}_{ipcw-glm}$  and  $\hat{\beta}_{oipcw}$ , in a specific setting with two binary covariates.

**Corollary 2** *When the model is well-specified, the difference of the asymptotic variance of the two scores  $\hat{U}_{oipcw}(\beta)$  and  $\hat{U}_{ipcw-glm}(\beta)$ , that is,  $\mathbf{\Omega}_{ipcw-glm} - \mathbf{\Omega}_{oipcw}$ , can be*

either positive definite or negative definite, depending on the shape of the density of the censoring time within the interval  $[0, t]$ . This means that for some distributions of the censoring time  $\widehat{\beta}_{oipcw}$  is more efficient than  $\widehat{\beta}_{ipcw-glm}$ , but for some others it is the other way around.

**Proof** In the following, we will make extensive use of the law of iterated expectations, without explicitly mentioning it each time. We first note that  $E[XD(t)\mathbb{1}\{T \geq s\}] = E[XH_1(s, t, X)]$ , with  $H_1(s, t, X) = F_1(t, X) - F_1(s, X)$ . Further, it follows

$$\begin{aligned} f_{oipcw}(s) &= E \left[ \left\{ XD(t) - E[XD(t) | T \geq s] \right\}^2 \mathbb{1}\{T \geq s\} \right] \\ &= E \left[ XD(t)\mathbb{1}\{T \geq s\} - \frac{\mathbb{1}\{T \geq s\}}{S(s)} E[XH_1(s, t, X)] \right]^2 \\ &= E \left[ X^2 H_1(s, t, X) \right] - \{E[XH_1(s, t, X)]\}^2 / S(s). \end{aligned}$$

Similarly, we obtain

$$\begin{aligned} f_{ipcw-glm}(s) &= f_{oipcw}(s) + E \left[ \left\{ XQ(t, X) - E[XQ(t, X) | T \geq s] \right\}^2 \mathbb{1}\{T \geq s\} \right] \\ &\quad - E \left[ XD(t) \left\{ XQ(t, X) - E[XQ(t, X) | T \geq s] \right\}^T \mathbb{1}\{T \geq s\} \right] \\ &\quad - E \left[ \left\{ XQ(t, X) - E[XQ(t, X) | T \geq s] \right\} X^T D(t) \mathbb{1}\{T \geq s\} \right] \\ &= f_{oipcw}(s) + g(s) \end{aligned}$$

with,

$$\begin{aligned} g(s) &= E \left[ X^2 F_1^2(t, X) S(s, X) \right] - \frac{1}{S(s)} \left\{ E[XF_1(t, X)S(s, X)] \right\}^2 \\ &\quad - 2E \left[ X^2 H_1(s, t, X) F_1(t, X) \right] \\ &\quad + \frac{1}{S(s)} E \left[ XF_1(t, X)S(s, X) \right] E \left[ XH_1(s, t, X) \right]^T \\ &\quad + \frac{1}{S(s)} E \left[ XH_1(s, t, X) \right] E \left[ XF_1(t, X)S(s, X) \right]^T, \end{aligned}$$

where  $S(s, X) = P(T \geq s|X)$ , since we assumed  $Q(t, X) = F_1(t, X)$ . Interestingly, because  $S(0) = S(0, X) = 1$  and  $H_1(0, t, X) = F_1(t, X)$ , we note that the function  $g : \mathbb{R} \rightarrow \mathbb{R}^{p \times p}$  starts being negative definite in  $s = 0$ , where

$$g(0) = - \left\{ E \left[ X^2 F_1^2(t, X) \right] - (E[XF_1(t, X)])^2 \right\} = -\text{Var}(XF_1(t, X))$$



and because  $H_1(t, t, X) = 0$ , it ends up being positive definite in  $s = t$ , with

$$\begin{aligned} \mathbf{g}(t) &= E\left[X^2 F_1^2(t, X) S(t, X)\right] - \frac{1}{S(t)} \left\{E\left[X F_1(t, X) S(t, X)\right]\right\}^2 \\ &= E\left[\left\{X F_1(t, X) - E\left[X F_1(t, X) \mid T \geq t\right]\right\}^2 \mathbb{1}\{T > t\}\right]. \end{aligned}$$

The consequence of this is that the average  $\int_0^t \mathbf{g}(s) f_c(s) ds$  can be either positive definite or negative definite, depending on the shape of  $f_c(\cdot)$  within the interval  $[0, t]$ . This means that which set of estimating equations to prefer, between (2) and (3), depends on how the censoring times are distributed.  $\square$

### 2.6 On naive standard error computation obtained using standard software

Most statistical software for fitting logistic models have a “weight” option. That is for instance the case of the function `glm` in R and of procedure `LOGISTIC` in SAS, which can be used to solve  $\widehat{U}_{ipcw-glm}(\beta) = \mathbf{0}$ , using the weights  $\widehat{W}_i(t), i = 1, \dots, n$ . Other software can solve  $\widehat{U}_{oipcw}(\beta) = \mathbf{0}$  by solving the usual score equations of a logistic model for the weighted outcomes  $\widehat{W}_i(t) \widetilde{D}_i(t), i = 1, \dots, n$ . That is the case of the function `geese` in R (from the `geepack` package) and of procedure `GENMOD` in SAS. Standard software can therefore be used to compute  $\widehat{\beta}_m$  by solving  $\widehat{U}_m(\beta) = \mathbf{0}$ , for  $m = oipcw$  and  $ipcw - glm$ , once the weights have been computed in a previous step, e.g. using Kaplan-Meier.

From a practical point of view, it is of interest to know how the “naive” standard errors, which the software will compute by default in that case, will compare to the correct values given by the diagonal of  $\widehat{\Sigma}_m$  in Corollary 1. In short, we expect them to be different as the software will ignore the fact that the weights  $\widehat{W}_i(t)$  are estimated from the data and therefore are not known in advance. We show by simple arguments that the “naive” estimator of the variance of  $\widehat{\beta}$  is conservative. As it will be illustrated in Sect. 2.8, and perhaps surprisingly, it can be very conservative. This emphasizes that using  $\widehat{\Sigma}_m$  of Corollary 1 or Bootstrapping is important for correct standard error computation.

In the rest of this section, we restrict our attention to the case in which the censoring weights have been computed by simple Kaplan-Meier estimates, thus assuming  $C$  does not depend on any of the covariates. We first note that the “naive” computation of the software will estimate the variance of  $\sqrt{n}(\widehat{\beta}_m - \beta)$  by a variance that converges towards  $\mathcal{I}^{-1} \Omega'_m \mathcal{I}^{-1}$ , where

$$\begin{aligned} \Omega'_{oipcw} &= \text{Var}\left(X_i \left\{W_i(t) D_i(t) - Q(t, X_i, \beta)\right\}\right) \\ \text{and } \Omega'_{ipcw-glm} &= \text{Var}\left(W_i(t) X_i \left\{D_i(t) - Q(t, X_i, \beta)\right\}\right), \end{aligned}$$

where  $W(t) = \Delta(t) / G_c(t \wedge \widetilde{T})$ . This follows from usual theory of estimating equations, see. e.g. Stefanski and Boos (2002), when a robust standard error is computed by the software. The fact that the “naive” robust standard errors are (asymptotically)

systematically computed too large is therefore a consequence of the following proposition.

**Proposition 1** *The difference  $\Omega_m - \Omega'_m$  is negative definite for both  $m = oipcw$  and  $ipcw - glm$ .*

**Proof** First, note that  $\text{Var} \{n^{-1/2}U_m(\beta)\} = \Omega'_m$ , where  $U_m(\beta)$  is defined by replacing  $\widehat{W}_i(t)$  by  $W_i(t)$  in  $\widehat{U}_m(\beta)$ . Second, we note that  $\sqrt{n}U_m(\beta)$  can be re-written as the same i.i.d. representation as that of  $\sqrt{n}\widehat{U}_m(\beta)$  except for the absence of the conditional expectations terms in  $\varphi_m(X, D(t), s)$ , that is, without the term  $E(XD(t)|T > t)$  for the OIPCW estimator or the term  $E(X[D(t) - Q(t, X)]|T > t)$  for the IPCW-GLM estimator. This follows from calculations already presented in Appendix A.1 and especially the identity  $W_i(t) = 1 - \int_0^t \{1/G_c(u)\}dM_i^c(u)$ . Consequently, following the same martingale arguments as in the proof of Corollary 1, it follows that

$$\begin{aligned} \Omega_m - \Omega'_m &= \int_0^t E \left[ \left\{ \mathbf{h}_m(X) - E[\mathbf{h}_m(X) | T \geq s] \right\}^2 \mathbb{1}\{T \geq s\} \right] f_c(s) ds \\ &\quad - \int_0^t E \left[ \left\{ \mathbf{h}_m(X) \right\}^2 \mathbb{1}\{T \geq s\} \right] f_c(s) ds \\ &= \int_0^t \left( -\frac{1}{S(s)} \left\{ E \left[ \mathbf{h}_m(X) \mathbb{1}\{T \geq s\} \right] \right\}^2 \right) f_c(s) ds, \end{aligned}$$

with  $\mathbf{h}_{oipcw}(X) = XD(t)$  and  $\mathbf{h}_{ipcw-glm}(X) = X\{D(t) - Q(t, X)\}$ . This shows that  $\Omega_m - \Omega'_m$  is negative definite. □

### 2.7 Augmented estimators

Following ideas from Robins and Rotnitzky (1992) and Bang and Tsiatis (2000), we now construct modifications of estimating Eqs. (2) and (3) which could lead to (asymptotically) more efficient estimators. Here we will again restrict our attention to the case where the Kaplan-Meier estimator is used to estimate the weights and  $C$  is independent of the covariates  $(A, L)$ , as in the two previous sections. To do this, we first define augmented equations  $\widehat{U}_m^{Aug}(\beta)$  of the form

$$\widehat{U}_m^{Aug}(\beta) = \widehat{U}_m(\beta) + \frac{1}{n} \sum_{i=1}^n \int_0^t \mathbf{e}_m(X_i, s) \frac{dM_i^c(s)}{G_c(s)} = \mathbf{0},$$

where  $\mathbf{e}_m(X, s)$  is an arbitrary function of  $X$  and  $s$ . These modified equations still have mean zero (asymptotically) and thus any choice of  $\mathbf{e}_m(X, s)$  will provide a consistent estimator of  $\beta$ . A natural question is therefore which choices lead to the most efficient estimators. The following proposition answers this question and shows that the optimal choice of  $\mathbf{e}_m(X, s)$  also makes the two augmented equations for OIPCW and IPCW-GLM become asymptotically equivalent.

**Proposition 2** Let  $\widehat{\beta}_m^{Aug}$  be the solution of  $\widehat{U}_m^{Aug}(\beta)$ . Then, the asymptotic variance of  $\widehat{\beta}_m^{Aug}$  attains its minimum when

$$e_{oipcw}(X, s) = E[XD(t) \mid T \geq s] - E[XD(t) \mid X, T \geq s]$$

and  $e_{ipcw-glm}(X, s) = e_{oipcw}(X, s) + XQ(t, X) - E[XQ(t, X) \mid T \geq s]$ .

In that case,  $\sqrt{n} \widehat{U}_{oipcw}^{Aug}(\beta) = \sqrt{n} \widehat{U}_{ipcw-glm}^{Aug}(\beta) + o_p(1)$  and the asymptotic variances of  $\widehat{\beta}_{oipcw}^{Aug}$  and  $\widehat{\beta}_{ipcw-glm}^{Aug}$  are equal.

**Proof** From the i.i.d representation of  $\sqrt{n} \widehat{U}_m(\beta)$  given in Theorem 1, we get that of the augmented score, that is,  $\sqrt{n} \widehat{U}_m^{Aug}(\beta) = n^{-1/2} \sum_{i=1}^n \Phi_m^{Aug}\{X_i, D_i(t), \widetilde{T}_i, \widetilde{\eta}_i, t\} + o_p(1)$ , with influence function

$$\Phi_m^{Aug}\{X, D(t), \widetilde{T}, \widetilde{\eta}, t\} = X\{D(t) - Q(t, X)\} - \int_0^t \varphi_m^{Aug}(X, D(t), s) \frac{dM^c(s)}{G_c(s)},$$

with  $\varphi_m^{Aug}(X, D(t), s) = e_m(X, s) + \varphi_m(X, D(t), s)$ . Similarly to (5) in the Proof of Corollary 1, it follows

$$\text{Var}\left[\Phi_m^{Aug}\{X, D(t), \widetilde{T}, \widetilde{\eta}, t\}\right] = \mathcal{J} + \int_0^t f_m^{Aug}(s) f_c(s) ds,$$

where

$$\begin{aligned} f_m^{Aug}(s) &= E\left[\{e_m(X, s) + \varphi_m(X, D(t), s)\}^2 \mathbb{1}\{T \geq s\}\right] \\ &= E\left[\{e_m(X, s) + \varphi_m(X, D(t), s)\}^2 \mid T \geq s\right] S(s) \\ &= E\left\{E\left[\{e_m(X, s) + \varphi_m(X, D(t), s)\}^2 \mid X, T \geq s\right] \mid T \geq s\right\} S(s) \end{aligned}$$

which achieves its minimum for

$$e_m(X, s) = -E\left[\varphi_m(X, D(t), s) \mid X, T \geq s\right].$$

Therefore, we have

$$\begin{aligned} e_{oipcw}(X, s) &= -E[XD(t) \mid X, T \geq s] + E[XD(t) \mid T \geq s] \\ \text{and } e_{ipcw-glm}(X, s) &= e_{oipcw}(X, s) \\ &\quad + E\left[XQ(t, X) - E[XQ(t, X) \mid T \geq s] \mid X, T \geq s\right] \\ &= e_{oipcw}(X, s) + XQ(t, X) - E[XQ(t, X) \mid T \geq s], \end{aligned}$$

hence the first result of the proposition. As  $\boldsymbol{\varphi}_m^{Aug}(\mathbf{X}, D(t), s) = \mathbf{e}_m(\mathbf{X}, s) + \boldsymbol{\varphi}_m(\mathbf{X}, D(t), s)$ , it further follows

$$\begin{aligned} \boldsymbol{\varphi}_{oipcw}^{Aug}(\mathbf{X}, D(t), s) &= \mathbf{X}D(t) - E[\mathbf{X}D(t) \mid \mathbf{X}, T \geq s] \\ \text{and } \boldsymbol{\varphi}_{oipcw-glm}^{Aug}(\mathbf{X}, D(t), s) &= \boldsymbol{\varphi}_{oipcw}^{Aug}(\mathbf{X}, D(t), s) \\ &\quad - \left\{ \mathbf{X}Q(t, \mathbf{X}) - E[\mathbf{X}Q(t, \mathbf{X}) \mid T \geq s] \right\} \\ &\quad + \mathbf{X}Q(t, \mathbf{X}) - E[\mathbf{X}Q(t, \mathbf{X}) \mid T \geq s] \\ &= \boldsymbol{\varphi}_{oipcw}^{Aug}(\mathbf{X}, D(t), s). \end{aligned}$$

This shows that  $\sqrt{n} \widehat{U}_{oipcw}^{Aug}(\boldsymbol{\beta}) = \sqrt{n} \widehat{U}_{oipcw-glm}^{Aug}(\boldsymbol{\beta}) + o_p(1)$  and therefore also that the asymptotic variances of  $\widehat{\boldsymbol{\beta}}_{oipcw}^{Aug}$  and  $\widehat{\boldsymbol{\beta}}_{oipcw-glm}^{Aug}$  are equal.  $\square$

When using the optimal choice of  $\mathbf{e}_m(\mathbf{X}, s)$  of Proposition 2, the augmented equations become

$$\begin{aligned} \widehat{U}_m^{Aug}(\boldsymbol{\beta}) &= \widehat{U}_m(\boldsymbol{\beta}) + \frac{1}{n} \sum_{i=1}^n \int_0^t \left\{ E[\mathbf{X}_i D_i(t) \mid \mathbf{X}_i, T_i \geq s] \right. \\ &\quad \left. - E[\mathbf{X}D(t) \mid T \geq s] \right\} \frac{dM_i^c(s)}{G_c(s)} = \mathbf{0}. \end{aligned} \tag{6}$$

One cannot solve  $\widehat{U}^{Aug}(\boldsymbol{\beta}) = \mathbf{0}$  in practice, since the augmentation term depends on unknown parameters. However, one can instead solve a similar equation after replacing  $M_i^c(s)$  and  $G_c(s)$  by  $\widehat{M}_i^c(s)$  and  $\widehat{G}_c(s)$  (as defined in Appendix A.2) and the two terms in the brackets  $\{ \}$  above by consistent estimators of each. While the second term  $E[\mathbf{X}D(t) \mid T \geq s]$  can be estimated nonparametrically in a simple fashion (as described in Appendix A.2), this is not necessarily the case for the first term because of the conditioning on  $\mathbf{X}_i$ , especially if the dimension of  $\mathbf{X}_i$  is not small and/or some components of  $\mathbf{X}_i$  are continuous covariates. In that case, one will often need to model  $F_1(s, \mathbf{X})$  for all  $s \in [0, t]$ , build on the equality  $E[\mathbf{X}D(t) \mid \mathbf{X}, T \geq s] = \mathbf{X}H_1(s, t, \mathbf{X})/S(s, \mathbf{X})$  and instead solve

$$\begin{aligned} \widehat{U}_m^{Aug}(\boldsymbol{\beta}) &= \widehat{U}_m(\boldsymbol{\beta}) + \frac{1}{n} \sum_{i=1}^n \int_0^t \left\{ \frac{\mathbf{X}_i \widehat{H}_1(s, t, \mathbf{X}_i)}{\widehat{S}(s, \mathbf{X}_i)} - \widehat{E}[\mathbf{X}D(t) \mid T \geq s] \right\} \frac{d\widehat{M}_i^c(s)}{\widehat{G}_c(s)} \\ &= \mathbf{0}, \end{aligned} \tag{7}$$

where  $\widehat{H}_1(s, t, \mathbf{X}) = \widehat{F}_1(t, \mathbf{X}) - \widehat{F}_1(s, \mathbf{X})$  and  $\widehat{S}(s, \mathbf{X})$  is an estimator of  $S(s, \mathbf{X}) = P(T > s \mid \mathbf{X})$ . Popular methods to estimate  $F_1(s, \mathbf{X})$  and  $S(s, \mathbf{X})$  simultaneously over all time  $s \in [0, t]$  are based on cause-specific hazards modeling, see e.g. Ozenne et al. (2017). Although  $\widehat{E}[\mathbf{X}D(t) \mid T \geq s]$  can be computed non-parametrically, it might be more natural to compute it as  $\left\{ \sum_{i=1}^n Y_i(s) \mathbf{X}_i \widehat{H}_1(s, t, \mathbf{X}_i) / \widehat{S}(s, \mathbf{X}_i) \right\} / \left\{ \sum_{i=1}^n Y_i(s) \right\}$

in that case. From a computational point of view, it is also interesting to note that  $d\widehat{M}_i^c(s)$  in (7) can be replaced by  $dN_i^c(s)$ , where  $N_i^c(s) = \mathbb{1}\{\widetilde{T}_i \leq s, \Delta_i = 0\}$ .

Following Robins and Rotnitzky (1992), one could rigorously show that Eq. (7) provide the best augmented estimators and that this will provide more efficient estimators than (2) and (3), if the plugged-in estimates  $\widehat{F}_1(s, \mathbf{X})$  and  $\widehat{S}(s, \mathbf{X})$  are close enough to their counterpart for all  $s \in [0, t]$ .

On one hand, this efficiency result makes the augmented estimating Eq. (7) appealing. This is especially the case when  $F_1(s, \mathbf{X})$  and  $S(s, \mathbf{X})$  do not appear particularly difficult to estimate consistently for all  $s \in [0, t]$ . On the other hand, in practice it might be challenging to estimate all these quantities precisely. This might often happen when the dimension of  $\mathbf{X}$  is not small and/or some components of  $\mathbf{X}$  are continuous covariates. The prespecification of modeling choices, which is often deemed important in medical research (FDA 2021; Loder et al. 2010), might further complicate the task. This should be kept in mind when considering the use of the augmented approach in practice, since  $\widehat{\beta}_m^{Aug}$  might be less efficient than  $\widehat{\beta}_m$  when  $F_1(s, \mathbf{X})$  and  $S(s, \mathbf{X})$  are not estimated consistently for all  $s \in [0, t]$ .

Finally, the Proposition 3 below provides a simple computational trick to solve estimating equations asymptotically equivalent to the augmented Eq. (6), using standard software, when  $\mathbf{X}$  is a vector of a few categorical variables. The results and its proof might also help to get some intuition for more general results, which state that efficiency gains can generally be obtained by using inverse probability of censoring weights that depends on  $\mathbf{X}$ , even when the censoring time  $C$  is independent of  $\mathbf{X}$  (Robins and Rotnitzky 1992).

**Proposition 3** *Assume that  $\mathbf{X}$  is a vector of categorical variables defining  $K$  strata and let  $s(\mathbf{x}) \in \{s_1, \dots, s_K\}$  denote the strata corresponding to  $\mathbf{x}$ . Further assume that a stratified Kaplan-Meier estimator, stratified on  $s(\mathbf{X})$ , is used to compute the inverse probability of censoring weights. In that case, solving the equations  $\widehat{U}_m(\beta) = \mathbf{0}$  (given by (2) and (3)) is asymptotically equivalent to solving the augmented augmented  $\widehat{U}_m^{Aug}(\beta)$  (given by (6)), for which a marginal Kaplan-Meier estimator is used to compute the weights. That is,  $\sqrt{n} \widehat{U}_m(\beta) = \sqrt{n} \widehat{U}_m^{Aug}(\beta) + o_p(1)$ .*

**Proof** First, as pointed out before at the beginning of Sect. 2.5, we note that  $\widehat{U}_m(\beta)$  is identical for both choices of  $m$  in that case. Second, using that  $\sum_{i=1}^n = \sum_{k=1}^K \sum_{i:s(\mathbf{X}_i)=s_k}$  and proceeding as in Appendix A.1 for each sum  $\sum_{i:s(\mathbf{X}_i)=s_k}$ , we obtain a similar i.i.d decomposition as in Theorem 1 and it follows that  $\sqrt{n} \widehat{U}_m(\beta) = n^{-1/2} \sum_{i=1}^n \Theta_m\{\mathbf{X}_i, D_i(t), \widetilde{T}_i, \widetilde{\eta}_i, t\} + o_p(1)$  with

$$\begin{aligned} \Theta_m\{\mathbf{X}, D(t), \widetilde{T}, \widetilde{\eta}, t\} &= \mathbf{X}\{D(t) - Q(t, \mathbf{X})\} \\ &\quad - \int_0^t \left\{ XD(t) - E\{XD(t) \mid s(\mathbf{X}), T \geq s\} \right\} \\ &\quad \frac{dM^c(s, s(\mathbf{X}))}{G_c(s, s(\mathbf{X}))}. \end{aligned}$$

We now note that conditioning on  $X$  or  $s(X)$  is equivalent and that assuming  $C$  independent of  $X$  implies  $M^c(s, s(X)) = M^c(s)$  and  $G_c(s, s(X)) = G_c(s)$ . This shows that  $\sqrt{n} \widehat{U}_m(\beta) = \sqrt{n} \widehat{U}_m^{Aug}(\beta) + o_p(1)$ . □

### 2.8 Theoretical computation and comparison of efficiency

In this section, we compare the asymptotic variances of  $\widehat{\beta}_{ipcw-glm}$  and  $\widehat{\beta}_{oipcw}$  in specific settings, to illustrate their theoretical differences detailed above. For a simple illustration, we consider  $X = (X_1, X_2)$  where  $X_1$  and  $X_2$  are two independent binary covariates, with  $P(X_1 = -1) = P(X_1 = 1) = 0.5$  and  $P(X_2 = 0) = P(X_2 = 1) = 0.5$ . We assume that  $F_1(t, X) = \text{expit}(\beta_0(t) + \beta_1 X_1 + \beta_2 X_2)$  with  $\beta_0(t) = \log[\rho_1(1 - e^{-t})]$ ,  $\beta_1 = 0.5$  and  $\beta_2 = -0.5$ . For the absolute risk of the competing event, i.e.  $F_2(t, X) = P(T \leq t, \eta = 2|X)$ , we assume  $F_2(t, X) = \text{expit}(\mu(t) - 0.5 \cdot X_1 + 0.5 \cdot X_2) \cdot \{1 - F_1(6, X)\}$  with  $\mu(t) = \log[\rho_2(1 - e^{-t})]$ . This parametrization satisfies the constraint  $F_1(t, x) + F_2(t, x) \leq 1$  for all  $x = (x_1, x_2)$  and  $t \in [0, 6]$ . To generate independent censoring, we used a constant hazard  $\lambda_c(t, X) = r_c$ .

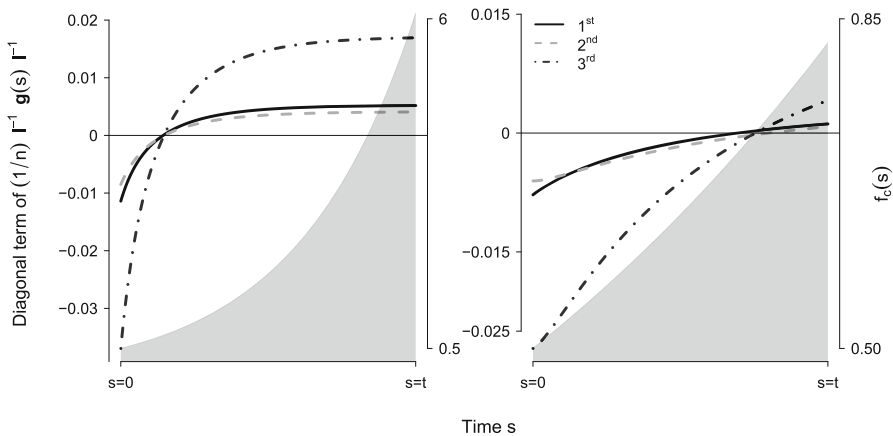
To illustrate several scenarios, we use two values of each parameter  $\rho_1, \rho_2$  and  $r_c$  and two time horizons  $t = 1$  and  $5$ . This leads to 16 scenarios with a marginal risk  $F_1(t) = E\{F_1(t, X)\}$  ranging from 5% to 42% and a proportion of censored observations before time  $t$ , i.e.  $P(\widetilde{T} \leq t, \widetilde{\eta} = 0)$ , ranging from 4% to 80% (Table 1). We consider the estimation of the well-specified logistic model for  $F_1(t, X)$  at time  $t = 1$  and  $5$  and we compute the corresponding asymptotic variances, when using OIPCW and IPCW-GLM. The computation was made using the theoretical (asymptotic) formulae derived in Sect. 2.5. In addition, we performed simulations that led to similar results with sample sizes  $n = 400$  and  $n = 1600$  (not shown), which confirmed the results presented below. In the simulation results, the gain for  $n = 400$  for the OIPCW approach were slightly better than in the theoretical (asymptotic) computations.

The theoretical computations, presented in Table 1, reveal which type of IPCW adjustment gave the smallest asymptotic variances in our settings. Overall, OIPCW appears to be preferable as the gain using this method is sometimes considerable. For a high risk  $F_1(t) = 42\%$  and a low risk of competing event ( $\rho_2 = 0.1$ , which corresponds to 8% competing risk) the parameters  $\beta_0(t), \beta_1$  and  $\beta_2$  are estimated with a variance that is between 2.17 and 2.75 times larger when using IPCW-GLM instead of OIPCW.

In Fig. 1 we show the 3 diagonal components of the matrix  $(1/n)\mathcal{I}^{-1}g(s)\mathcal{I}^{-1}$ , together with  $f_c(s)$ , for  $s \in [0, t]$ , for the two settings in which we saw the largest difference in asymptotic variance between the two methods favoring OIPCW or IPCW-GLM, respectively. This provides insights on the origin of the variance differences. As seen in Sect. 2.5, the differences in asymptotic variance, for each parameter estimator, are given by the diagonal of  $(1/n)\mathcal{I}^{-1}\{\int_0^t g(s)f_c(s)ds\}\mathcal{I}^{-1}$ , where  $f_c$  depends only on the censoring time distribution. In our settings,  $f_c(s) = r_c e^{r_c s}$ , which increases fast with  $s$ . This makes the values of  $g(s)$  for late times  $s$  particularly influential in the variance differences. This explains why the left plot corresponds to large variance differences whereas the right plot corresponds to smaller differences. On the left plot of Fig. 1, there is a large interval in which the diagonal terms of  $g(s)$  are positive and

**Table 1** Ratio of variances for the estimators of each parameter of the logistic model, when using IPCW-GLM relative to OIPCW. A symbol “-” means that the value is similar to that just above. “Cens” displays the proportion of censored observations before time  $t$ , i.e.  $P(\tilde{T} \leq t, \tilde{\eta} = 0)$ . See text for the definition of the other parameters

$t$	$\rho_1$	$F_1(t)$	$\rho_2$	$r_c$	Cens	$\beta_0(t)$	$\beta_1$	$\beta_2$
1	0.1	5%	0.1	0.1	9%	1.00	1.00	1.00
-	-	-	-	0.5	36%	1.00	1.00	1.01
-	-	-	5	0.1	4%	1.00	1.00	1.00
-	-	-	-	0.5	16%	1.00	0.99	0.99
-	0.9	32%	0.1	0.1	7%	1.00	1.00	1.00
-	-	-	-	0.5	30%	1.02	1.02	1.03
-	-	-	5	0.1	4%	0.99	0.98	0.98
-	-	-	-	0.5	17%	0.97	0.93	0.93
5	0.1	8%	0.1	0.1	33%	1.02	1.02	1.03
-	-	-	-	0.5	80%	1.23	1.31	1.41
-	-	-	5	0.1	8%	1.00	0.99	1.00
-	-	-	-	0.5	24%	1.02	1.00	1.03
-	0.9	42%	0.1	0.1	23%	1.09	1.13	1.14
-	-	-	-	0.5	59%	2.17	2.68	2.75
-	-	-	5	0.1	7%	0.99	0.96	0.97
-	-	-	-	0.5	23%	1.11	1.03	1.09



**Fig. 1** Diagonal terms for  $(1/n)\mathcal{I}^{-1}\mathbf{g}(s)\mathcal{I}^{-1}$ , for  $s \in [0, t]$ , from which originates the difference in asymptotic variance between each parameter estimator, when using either OIPCW or IPCW-GLM (with  $n = 400$ ). The 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> diagonal terms relate to the differences in variance for estimators of  $\beta_0(t)$ ,  $\beta_1$  and  $\beta_2$ , respectively. The gray area and the right axis additionally show  $f_c(s)$  for  $s \in [0, t]$ , which weights the contribution of  $(1/n)\mathcal{I}^{-1}\mathbf{g}(s)\mathcal{I}^{-1}$ , at each time  $s \in [0, t]$ , to the variance differences given by  $(1/n)\mathcal{I}^{-1}\{\int_0^t \mathbf{g}(s)f_c(s)ds\}\mathcal{I}^{-1}$ . The right and left plots correspond to situations with data generating parameters  $\rho_1, \rho_2, r_c$  and  $t$  of lines 14 and 8 of Table 1, respectively

$f_c(s)$  is large. By contrast, on the right plot, the large interval for which the diagonal terms of  $\mathbf{g}(s)$  are negative contains a wide range of time  $s$  for which  $f_c(s)$  is rather small.

**Table 2** Ratio of naive-variance to true variance, for scenarios with censoring  $r_c = 0.5$  and  $t = 5$ , when using either a marginal or a stratified Kaplan-Meier estimator to compute the IPCW weights

$\rho_1$	$\rho_2$	Marginal K-M			Stratified K-M		
		$\beta_0(t)$	$\beta_1$	$\beta_2$	$\beta_0(t)$	$\beta_1$	$\beta_2$
OIPCW							
0.1	0.1	1.01	1.00	1.00	1.02	1.01	1.02
–	5	1.03	1.00	1.00	1.05	1.05	1.05
0.9	0.1	1.06	1.00	1.00	1.13	1.11	1.12
–	5	1.13	1.00	1.00	1.28	1.25	1.26
IPCW-GLM							
0.1	0.1	1.18	1.00	1.00	1.47	1.33	1.43
–	5	1.05	1.01	1.00	1.09	1.06	1.09
0.9	0.1	1.45	1.00	1.00	3.23	2.91	3.00
–	5	1.18	1.07	1.01	1.52	1.38	1.41

Figure 1 also illustrates the result that which method is asymptotically the most efficient depends on the censoring time distribution. Graphically, which method is most efficient depends on the location of the gray area and different distributions will lead to different shapes of the gray area. For instance, if the censoring times were all early with respect to the time-point  $t$  of interest, then the IPCW-GLM would be more efficient, as in that case  $f_c(s)$  would be non zero only for early times  $s$ . Conversely, if they were all located just before  $t$ , then OIPCW would be more efficient.

Finally, we compare the “naive” variance, which standard software can estimate by default, to the true variance (asymptotically). This exemplifies the asymptotic overestimation of the variance showed in Sect. 2.6. The results are provided in Table 2, for all scenarios with time  $t = 5$ . The table provides the results when using a simple (marginal) Kaplan-Meier estimator or a (fully) stratified Kaplan-Meier estimator to compute the weights. We note that for all settings and both estimators the variance is overestimated for  $\beta_0(t)$ , and this is particularly so for the IPCW-GLM method. For instance, this will make the confidence interval for the predicted risk in the reference group too wide. The variance of covariate effect estimators are well estimated for the OIPCW using the simple naive variance estimator, when using a simple Kaplan-Meier to compute the weight. However, when using the stratified Kaplan-Meier, we see differences. The IPCW-GLM “naive” standard errors were considerably off when compared to the truth when using the stratified Kaplan-Meier weights. We conclude that it is important to use appropriate methods to compute standard errors.

Comparing the results obtained using the stratified versus those using the simpler Kaplan-Meier estimator also illustrates the efficiency gain that can be obtained using the augmented estimators. This is because using the (fully) stratified Kaplan-Meier is asymptotically equivalent to using the augmented estimator, as seen in Proposition 3.

### 3 Average causal treatment effect and confounding

We now present estimators of the average causal treatment effect. We use potential outcome notations, as e.g. in Ozenne et al. (2020) or Young et al. (2020), to define the



associated target parameters. Let  $(T^a, \eta^a)$  and  $D^a(t)$ , for  $a = 0, 1$ , denote the potential outcome variables  $(T, \eta)$  and  $D(t)$  that are, or would have been observed, under the treatment value  $a$ . Averaged causal risks are defined by  $F_1(t, a) = P\{D^a(t) = 1\}$ , for  $a = 0$  and  $1$ , where the term ‘‘average’’ emphasizes that we average over baseline covariates, i.e.  $F_1(t, a) = E[P\{D^a(t) = 1|L\}]$ . The term ‘‘causal’’ emphasizes that  $F_1(t, a)$  relates to the potential outcome  $D^a(t)$ . Within our context presented in Sect. 2.1, the interpretation of  $F_1(t, a)$  is that of the proportion of patients that would have been expected to die from a cardiovascular death within the 33 months of follow-up, had all the patients initiated a beta-blocker treatment, when  $a = 1$ , or had none of them been treated with beta-blocker, when  $a = 0$ . The corresponding risk difference  $\text{Diff}(t) = F_1(t, 1) - F_1(t, 0)$ , and risk ratio  $\text{RR}(t) = F_1(t, 1)/F_1(t, 0)$ , therefore each defines an average treatment effect.

Specific assumptions, in addition of those of Sect. 2.3, are required to make valid causal inference. In the rest of the manuscript, we will assume the usual ‘‘no unmeasured confounding’’, ‘‘consistency’’ and ‘‘positivity’’ assumptions. No unmeasured confounding means that  $(T^a, \eta^a)$ , and thus also  $D^a(t)$ , are independent of  $A$  conditionally on  $L$ . Positivity means that  $\pi(a, l) > 0$  for all  $l$  and  $a = 0, 1$ , where  $\pi(a, l) = P\{A = a|L = l\}$ . Consistency means that  $(T, \eta) = (1 - A) \cdot (T^0, \eta^0) + A \cdot (T^1, \eta^1)$ . See, e.g., Hernan and Robins (2020) for more details on these standard assumptions and counterfactual notations.

### 3.1 G-computation

As in Ozenne et al. (2020) and Zhang and Zhang (2011), for  $a = 0, 1$ , we can estimate  $F_1(t, a)$  using the G-computation estimator

$$\widehat{F}_1^g(t, a) = \frac{1}{n} \sum_{i=1}^n \widehat{Q}(t, a, L_i) \tag{8}$$

where  $\widehat{Q}(t, a, L_i) = Q(t, a, L_i, \widehat{\beta})$ , using any consistent estimator  $\widehat{\beta}$  of  $\beta$  seen above, e.g.,  $\widehat{\beta}_{oipcw}$  or  $\widehat{\beta}_{oipcw}$ . With a slight abuse of notations, we will often write  $Q(t, A, L, \beta)$  for  $Q(t, X, \beta)$  in what follows, to emphasize that it depends on both  $A$  and  $L$ . Similarly, we will also write  $Q(t, a, L, \beta)$  in the case  $A$  is either observed with, or set to, the value  $a$ . The rationale for  $\widehat{F}_1^g(t, a)$  is that, under the above assumptions,  $F_1(t, a) = E\{P\{D(t) = 1|A = a, L\}\}$ . The g-computation estimators of  $\text{Diff}(t)$  and  $\text{RR}(t)$  are then obtained as the difference or the ratio of  $\widehat{F}_1^g(t, 1)$  and  $\widehat{F}_1^g(t, 0)$ . Instead of using a logistic model, Ozenne et al. (2020) previously suggested to combine two cause-specific Cox models to compute  $\widehat{F}_1^g(t, a)$  for  $a = 0, 1$ , whereas Zhang and Zhang (2011) suggested to use a Fine and Gray (1999) model.

Many asymptotic properties of  $\widehat{F}_1^g(t, a)$ , and of the corresponding estimators of  $\text{Diff}(t)$  and  $\text{RR}(t)$ , can be derived directly from Taylor-expansions and the results of the previous sections. We especially note the following result, which provides insights to compare the asymptotic efficiency of different versions of the estimator  $\widehat{F}_1^g(t, a)$ , defined by plugging-in different estimators  $\widehat{\beta}$  of  $\beta$ . Maybe unsurprisingly, this result

implies that using the most efficient estimator  $\widehat{\beta}$  will lead to the most efficient estimator  $\widehat{F}_1^g(t, a)$ , when the model is well-specified.

**Proposition 4** *Assume that the model is well-specified, that is,  $Q(t, a, \mathbf{I}, \beta) = F_1(t, a, \mathbf{I})$  for all  $\mathbf{I}$ , and that  $C$  does not depend on the covariates  $(A, \mathbf{L})$ . Asymptotically,  $\sqrt{n} \left\{ \widehat{F}_1^g(t, a) - F_1(t, a) \right\}$  is normally distributed with variance*

$$\text{Var} \left\{ Q(t, a, \mathbf{L}, \beta) \right\} + n \left\{ \mathbf{B}_\beta(a) \right\} \text{Var}(\widehat{\beta}) \left\{ \mathbf{B}_\beta(a) \right\}^T,$$

where  $\mathbf{B}_\beta(a) = E \left\{ \partial / \partial \theta^T Q(t, a, \mathbf{L}, \theta) |_{\theta = \beta} \right\}$ . Consequently, if we consider any two versions of  $\widehat{F}_1^g(t, a)$ , each defined by plugging-in a different estimator  $\widehat{\beta}$  of  $\beta$ , the most efficient of the two is obtained by plugging-in the most efficient estimator  $\widehat{\beta}$  (asymptotically).

**Proof** A proof can be found in Bartlett (2018, Appendix A.2), in a slightly different context. We repeat the main arguments in our context in Appendix A.3, for completeness. □

### 3.2 Double robust estimating equations

The above g-computation estimator requires the model for  $F_1(t, a, \mathbf{I})$  to be well-specified to be consistent (unless specific additional assumptions hold, as in Sect. 5). To relax this assumption, so called double-robust (DR) estimators have been proposed. These estimators build on estimators for both  $F_1(t, a, \mathbf{I})$  (the “outcome model”) and  $\pi(a, \mathbf{I})$  (the “propensity score model”) and requires only one of the two to be consistent for the DR estimator to be consistent. See e.g. Hernán and Robins (2020). Let  $\widehat{\pi}(a, \mathbf{L}_i)$  denote an estimator of  $\pi(a, \mathbf{L}_i)$ . For binary (uncensored data), the DR estimators of  $F_1(t, a)$  is

$$\widehat{F}_1^{dr}(t, a) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}\{A_i = a\}}{\widehat{\pi}(a, \mathbf{L}_i)} \left( D_i(t) - \widehat{Q}(t, a, \mathbf{L}_i) \right) + \frac{1}{n} \sum_{i=1}^n \widehat{Q}(t, a, \mathbf{L}_i). \tag{9}$$

In short, the double robustness comes from the following properties. If the outcome model is correctly specified, the law of iterated expectations and the law of large number together imply that the first average on the right-hand side of Eq. (9) converges to zero. Specifically, the first average in the right-hand side of (9) will converge to

$$\begin{aligned} & E \left\{ E \left[ \frac{\mathbb{1}\{A_i = a\}}{\widehat{\pi}(a, \mathbf{L}_i)} \left( D_i(t) - \widehat{Q}(t, a, \mathbf{L}_i) \right) \middle| A_i, \mathbf{L}_i \right] \right\} \\ & \approx E \left\{ \frac{\mathbb{1}\{A_i = a\}}{\widehat{\pi}(a, \mathbf{L}_i)} \left( E[D_i(t) | A_i, \mathbf{L}_i] - \widehat{Q}(t, A_i, \mathbf{L}_i) \right) \right\}, \end{aligned}$$

which will also converge to zero when  $\widehat{Q}(t, A_i, \mathbf{L}_i)$  converges to  $E[D_i(t) | A_i, \mathbf{L}_i] = F_1(t, A_i, \mathbf{L}_i)$ . Hence, in that case the DR-estimator has asymptotically the same mean

as the g-computation estimator. If the propensity model is correctly specified, then the two ways of averaging, using the empirical average  $(1/n) \sum_i$  or the weighted average  $(1/n) \sum_i \frac{\mathbb{1}\{A_i=a\}}{\hat{\pi}(a, L_i)}$ , are both consistent for the population average. Hence the weighted and empirical average of  $\widehat{Q}(t, a, L_i)$  cancel each other out (in the limit) and the DR estimator converges to the expectation of  $(1/n) \sum_{i=1}^n \frac{\mathbb{1}\{A_i=a\}}{\hat{\pi}(a, L_i)} D_i(t)$ , which is indeed  $F_1(t, a)$ .

To define an estimator with right censored data, we note that only the first of the two averages on the right-hand side of Eq. (9) needs to be modified. Further, the formula of this first average resembles very much that of the estimating equation for binary data of Eq. (1). Hence we can build on the same idea: we can either weight the individual contribution to the average (as in (2)) or the censored outcome (as in (3)). The first option leads to

$$\begin{aligned} \widehat{F}_1^{dr,1}(t, a) &= \frac{1}{n} \sum_{i=1}^n \frac{\widehat{W}_i(t) \mathbb{1}\{A_i = a\}}{\hat{\pi}(a, L_i)} \left( D_i(t) - \widehat{Q}(t, a, L_i) \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \widehat{Q}(t, a, L_i) \end{aligned} \tag{10}$$

the second to

$$\begin{aligned} \widehat{F}_1^{dr,2}(t, a) &= \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}\{A_i = a\}}{\hat{\pi}(a, L_i)} \left( \widehat{W}_i(t) \cdot D_i(t) - \widehat{Q}(t, a, L_i) \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \widehat{Q}(t, a, L_i) . \end{aligned} \tag{11}$$

This second approach is similar to a DR estimator suggested by Ozenne et al. (2020). The only difference is that we suggest to use a logistic model for the outcome model  $Q(t, a, L_i)$ , whereas they instead suggest a model based on two cause-specific Cox regressions. Estimators of  $\widehat{\text{Diff}}(t)$  and  $\widehat{RR}(t)$  are then obtained as the difference or the ratio of the estimator for each group.

Standard errors may be computed either by bootstrapping or via asymptotic decompositions, which follow after Taylor expanding in the direction of the different parameters, similarly to what we did for the G-computation estimator in Appendix A.3.

### 3.3 Regression based DR estimators via clever covariates

Building on the work of Scharfstein et al. (1999, p. 1140–1441), Bang and Robins (2005) pointed out that a DR estimator closely related to that of Eq. (9) can be computed via simple averages of predicted risks similar to that of the G-computation estimator in (8). They suggested to proceed as follows. First, fit a logistic model using, in addition to the covariates  $(A, L)$ , the two constructed covariates defined by  $h_a\{A, L\} = \mathbb{1}\{A =$

$a\}/\widehat{\pi}(a, \mathbf{L})$ , for  $a = 0, 1$ , which Moore and van der Laan (2009) referred to as ‘‘clever covariates’’. That is, fit the logistic regression model

$$Q'(t, A, \mathbf{L}) = \text{expit}\left[\text{logit}\{Q(t, A, \mathbf{L})\} + \beta_{h_0}h_0\{A, \mathbf{L}\} + \beta_{h_1}h_1\{A, \mathbf{L}\}\right],$$

where  $Q(t, A, \mathbf{L})$  is the same logistic model as before. Then, instead of (8), compute,

$$\widehat{F}_1^{rdr}(t, a) = \frac{1}{n} \sum_{i=1}^n \widehat{Q}'(t, a, \mathbf{L}_i). \tag{12}$$

Below, we refer to this estimator as the regression-based DR estimator. The rationale for this approach is that  $\widehat{F}_1^{rdr}(t, a)$  can be re-written almost exactly as the right-hand side of Eq. (9), with the only difference that  $\widehat{Q}$  should be replaced by  $\widehat{Q}'$ . Indeed,

$$\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}\{A_i = a\}}{\widehat{\pi}(a, \mathbf{L}_i)} \left[ D_i(t) - \widehat{Q}'(t, a, \mathbf{L}_i) \right] = 0, \tag{13}$$

when fitting the logistic model with the clever covariates, using estimating Eq. (1). This is because the equation corresponding to the clever covariate  $h_a\{A, \mathbf{L}\}$ , for  $a = 0, 1$ , is nothing else than (13), by definition of  $h_a\{A_i, \mathbf{L}_i\}$  and because  $h_a\{A_i, \mathbf{L}_i\} \cdot \widehat{Q}(t, A_i, \mathbf{L}) = h_a\{A_i, \mathbf{L}_i\} \cdot \widehat{Q}'(t, a, \mathbf{L})$ .

Of course, the minor difference between  $\widehat{F}_1^{rdr}(t, a)$  and  $\widehat{F}_1^{dr}(t, a)$  has no consequence for the double robustness property. Indeed, if the outcome model  $Q(t, A, \mathbf{L})$  is well specified, then  $Q'(t, A, \mathbf{L})$  is also well-specified and adding the clever covariates or not into the model does not matter much, because estimators of parameter  $\beta_{h_0}$  and  $\beta_{h_1}$  will converge towards zero, as the sample size  $n$  increases. If the propensity score model is well specified, then the outcome model does not need to be correctly specified and thus it does not matter either for consistency.

Scharfstein et al. (1999) and Bang and Robins (2005) also note that if we only aim to estimate the causal difference  $\text{Diff}(t)$  and not each of the two risks, then we can use a single clever covariate instead of two, when fitting the logistic model. That is, one can fit the logistic regression model

$$Q''(t, A, \mathbf{L}) = \text{expit}\left[\text{logit}\{Q(t, A, \mathbf{L})\} + \beta_{h_{\{1-0\}}}h_{\{1-0\}}\{A, \mathbf{L}\}\right],$$

with the single clever covariate  $h_{\{1-0\}}\{A, \mathbf{L}\} = h_1\{A, \mathbf{L}\} - h_0\{A, \mathbf{L}\}$ . The risk difference can then be estimated by a simple average of differences,

$$\frac{1}{n} \sum_{i=1}^n \left\{ \widehat{Q}''(t, 1, \mathbf{L}_i) - \widehat{Q}''(t, 0, \mathbf{L}_i) \right\}.$$

The rationale is similar to that of using two clever covariates. This becomes clear after noticing that

$$\begin{aligned} \widehat{F}_1^{dr}(t, 1) - \widehat{F}_1^{dr}(t, 0) &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\mathbb{1}\{A_i = 1\}}{\widehat{\pi}(1, \mathbf{L}_i)} - \frac{\mathbb{1}\{A_i = 0\}}{\widehat{\pi}(0, \mathbf{L}_i)} \right\} \left\{ D_i(t) - \widehat{Q}(t, A_i, \mathbf{L}_i) \right\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{Q}(t, 1, \mathbf{L}_i) - \widehat{Q}(t, 0, \mathbf{L}_i) \right\}, \end{aligned}$$

which holds because  $\mathbb{1}\{A_i = a\} \cdot \widehat{Q}(t, A_i, \mathbf{L}_i) = \mathbb{1}\{A_i = a\} \cdot \widehat{Q}(t, a, \mathbf{L}_i)$ . Scharfstein et al. (1999) pointed out that this approach can be slightly more efficient than that based on the two clever covariates. However, we believe that often we do not aim to estimate only the risk difference but also the risk in each group. Hence, using two clever covariates is probably often the most attractive option to compute the three estimates in a unified way.

To define regression-based DR estimators for right censored data, it is sufficient to use either the estimating Eqs. (2) or (3) to fit the logistic model and otherwise proceed as in the binary uncensored case described above. The rationale is exactly the same: the first average on the right-hand side of Eqs. (10) and (11) are similar to the estimation equation induced by the clever covariates in either (2) or (3). Finally, note that the DR estimators may further be censoring augmented along the lines of Sect. 2.7, and this has been done in Ozenne et al. (2020).

### 3.4 tMLE-like estimators

Moore and van der Laan (2009) suggested a very similar yet different approach to the regression-based DR approach. The regression models used for the outcome, i.e.,  $Q'(t, A, \mathbf{L})$ , and the propensity score, i.e.,  $\pi(A, \mathbf{L})$ , are the same. The estimator used for the propensity score, and thus to construct the clever covariates, are the same and the averaging step (12) is also exactly the same. However, the procedure to estimate the parameters of the outcome model  $Q'(t, A, \mathbf{L})$ , and hence to compute  $\widehat{Q}'(t, a, \mathbf{L}_i)$ , is slightly different. Moore and van der Laan (2009) proposed to first estimate  $F_1(t, A_i, \mathbf{L}_i)$  by fitting the model without the clever covariates, that is,  $Q(t, A, \mathbf{L})$ . Then, in a second step, the model is updated by regressing on the clever covariates while including  $\logit\{\widehat{Q}(t, A, \mathbf{L})\}$ , obtained from the previous step, as an offset in the model. Hence, the final estimate of the individual predicted risk  $\widehat{Q}'(t, a, \mathbf{L}_i)$  closely resembles that of the regression based DR, except for the fact that not all regression parameters have been estimated simultaneously. Instead, a two steps approach was used. First the parameters in  $Q(t, A_i, \mathbf{L})$  were estimated, and then the parameters  $\beta_{h_0}$  and  $\beta_{h_1}$ . A similar two steps approach is also suggested to compute  $\widehat{Q}''(t, a, \mathbf{L}_i)$ , when using one clever covariate instead of two.

The double robustness property is of course preserved with this approach. If the outcome regression model is well specified, then consistency follows from the same argument as for the regression-based DR approach. If the propensity score model is consistent, here again the consistency follows because, in the second step, the estimating equations induced by the clever covariates are similar to the first average on the right-hand side of Eq. (9).

This estimator is referred to as the tMLE estimator by Moore and van der Laan (2009). This stands for “targeted maximum likelihood estimation”. Overviews about tMLE, in general, can be found in Van Der Laan and Rubin (2006) and Van der Laan and Rose (2011). The computation of the tMLE estimator described above is presented in details in a tutorial by Luque-Fernandez et al. (2018).

Interestingly, with right censored data, a similar approach to that of Moore and van der Laan (2009) can easily be used. One can just use the estimating Eqs. (2) or (3) to fit the logistic models and otherwise proceed as in the binary uncensored case.

## 4 Application to Danish registry data

We re-analyze data from Holt et al. (2021), who investigated the long-term cardio-protective effect associated with beta-blocker treatment in stable, optimally treated, myocardial infarction (MI) patients without heart failure. The motivation for this study was that the evidence of a protective beta-blocker effect relies on data from old randomized trials, which were conducted before catheter-based reperfusion became standard care during hospital admission for MI. Consequently, the evidence relates to a rather different patient population than the current population of MI patients. This has led to increasing skepticism about the benefit of initiating a beta-blocker treatment and to increasingly less systematic treatment initiations (Rossello et al. 2015).

As in Holt et al. (2021), in our analysis follow-up starts 3 months after MI and we compare the  $t = 33$ -months risk of cardiovascular death among patients who have initiated a beta-blocker treatment ( $A = 1$ ) to those who have not ( $A = 0$ ), among patients alive 3 months after MI. The data were collected from nationwide registers and we included Danish patients with first-time MI discharged between 2003 and 2018. Following inclusion criteria detailed in Holt et al. (2021), this results into 24,770 patients included in the treated group ( $A = 1$ ) and 5,407 in the untreated group ( $A = 0$ ). Follow-up ended at death, emigration, 33 months after inclusion or 31 December 2018, whichever came first. Here non-cardiovascular death is a competing risk. The definition of each variable  $T$ ,  $\eta$ ,  $A$  and  $C$  within this context was already presented in Sect. 2.1.

We assume that the no unmeasured confounding assumption holds when adjusting on the following baseline variables  $L$ : age group (30-60, 60-70, 70-80, 80-85), sex, calendar year at inclusion (2003-2008, 2009-2013, 2014-2018), educational level (3 levels), procedure during MI hospital admission (3 types), diabetes, history of stroke, hypertension, peripheral arterial disease, kidney and liver disease, gastrointestinal bleeding and cancer.

Another important assumption of the methods discussed in this paper is the consistency of the estimator of the censoring probabilities. That is, the consistency of  $\widehat{G}_c(u, \mathbf{x})$ , for all  $u \leq t$  and  $\mathbf{x}$ . This means that the choice of estimator matters and we therefore illustrate four choices, referred to as C1, C2, C3 and C4, in what follows.

## 4.1 Considered estimators for the censoring distribution

First, we consider the simplest and most common choice, that is, the use of a (marginal) Kaplan-Meier estimator (C1). Here the covariate values  $x$  do not play any role. Second, we use a Kaplan-Meier estimator stratified by treatment and age group (C2). The rationale for stratifying by treatment group is the growing skepticism about the benefit of initiating beta-blocker treatment during the study period (2003–2018). Hence, patients included recently, e.g. in 2010's, are more likely to be non treated than those included less recently, say in the 2000's. The rationale for further stratifying by age group is to try to increase efficiency, see e.g. Malani (1995) and related remarks in Sect. 2.7. The third choice consists to use a Kaplan-Meier estimator stratified by group of inclusion year, before or after 1st of January 2015 (C3). The rationale for this choice is that patients included before 2015 can be lost of follow-up within 33-months only if they leave Denmark, which is extremely rare in the study population. By contrast, those included after March 2016 will all be censored within 33-months due to study end on 31 December 2018. The arbitrary cutoff in 2015 was chosen to strike a good balance between two conflicting objectives. The first is to define a strata containing only patients similar enough to those included after March 2016. The second is to include enough patients in this strata to obtain reliable estimates. These patients need to be similar enough with respect to standard health care received, including the propensity to receive a beta-blocker treatment. The fourth choice is similar to the third, except that we further stratified by age and treatment group in the strata of patients included in 2015–2018 (C4). Here the rationale is to both adjust for the increasing skepticism and increase efficiency.

Note that choices C3 and C4 closely relate to the work of Rotnitzky et al. (2007) and Lok et al. (2018), who modeled two competing risks of censoring, to rely on weaker assumptions when estimating the IPCW weights. In our context, two competing risks of censoring exist: one due to study end, the other due to emigration. With choices C3 and C4, we approximately modeled these two competing risks of censoring, by stratifying on the inclusion year (before or after 2015). We therefore believe that choices C3 and C4 are more suitable than C1 and C2. The choice C4 makes less assumptions than C3, hence we believe that it leads to the most trustworthy results.

## 4.2 Logistic models for the outcome and propensity models

All baseline variables described above were included in the logistic model for the 33-months risk of CV death. To capture potentially different treatment effect among patients of different groups of age, year of inclusion, procedure during MI hospital admission and sex, interaction terms between each of these variables and the treatment variable were added. The same model choices were used in Holt et al. (2021), up to minor differences in the definition of some groups. For the double robust estimators, modeling choices are also needed for the probably of initiating a beta-blocker treatment. We used a logistic model and all baseline variables were also included, without any interaction term.

### 4.3 Compared methods

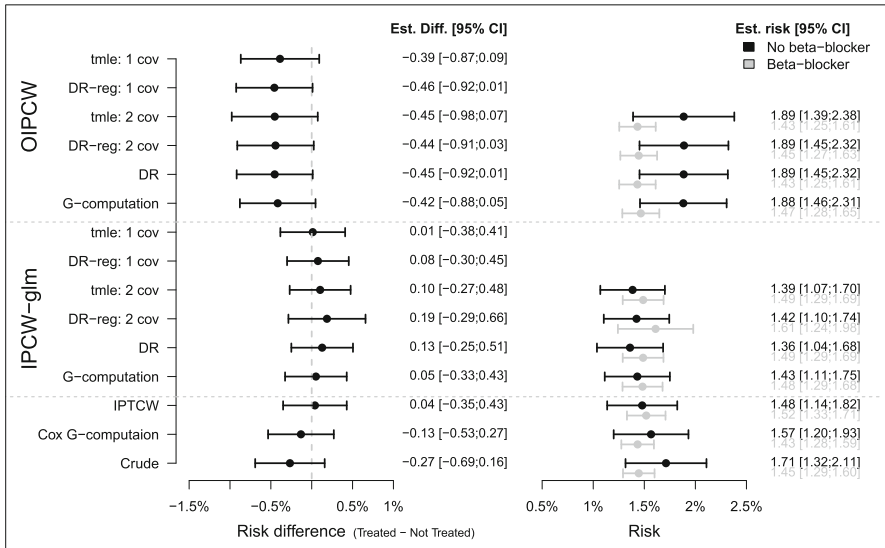
We present results obtained for all estimators introduced in Sect. 3. For comparison purpose, we also present results obtained using three other approaches to estimate the risk  $F_1(t, a)$ , for  $a = 0, 1$ , and the risk difference  $F_1(t, 1) - F_1(t, 0)$ . The first method is to estimate the risk in each treatment group using the Aalen-Johanssen estimator, as an unadjusted (crude) analysis. Note that this is algebraically equivalent to using a logistic model that only includes the treatment variable, when fitting the model using a Kaplan-Meier estimator stratified by treatment groups to compute the weights (Scheike et al. 2008; Geskus 2016). The second approach consists in using G-computation based on Cox regression for each cause-specific hazard, instead of logistic regression, as detailed in Ozenne et al. (2020). Here we used the same variables and interaction terms as in the logistic model, to model each cause specific hazard. Finally, the third approach uses inverse probability of treatment and censoring weighting, to estimate each risk  $F_1(t, a)$  as

$$\widehat{F}_1^{iptcw}(t, a) = \frac{\sum_{i=1}^n \widehat{\omega}_i(t, a, \mathbf{L}_i) D_i(t)}{\sum_{i=1}^n \widehat{\omega}_i(t, a, \mathbf{L}_i)}, \quad \text{with } \widehat{\omega}_i(t, a, \mathbf{L}_i) = \frac{\mathbb{1}\{A_i = a\} \widehat{W}_i(t)}{\widehat{\pi}(a, \mathbf{L}_i)}.$$

By contrast to the approaches described in Sect. 3, the first of these approaches does not account for confounding and it is expected to lead to biased results. The second adjusts for confounding, but it relies on a correct model specification of the cause-specific hazards of the two competing risks, at any time within the 33 month of follow-up. However, it does not rely on any specific modeling assumption for the censoring distribution or the propensity score. Only the assumption of independent censoring is needed, which allows the censoring time to arbitrarily depend on the treatment variable  $A$  as well as baseline variables  $\mathbf{L}$ . The third relies on a correct model specification of the propensity score and of the censoring distribution, but it does not rely on any modeling assumption about the conditional risk  $F_1(t, A, \mathbf{L})$ .

Note that for the initial (main) analysis of these data, whose results are detailed in Holt et al. (2021), we chose to use G-computation via logistic regression. We will now briefly explained why we preferred this approach to alternatives, in this specific context. Despite the large sample size ( $n = 24,770 + 5,407$ ), the number of observed cardiovascular deaths was not very large ( $333 + 72$ ) and the prevalence of some of the many important baseline characteristic we needed to adjust for were rare (see Table 1 in Holt et al. (2021)). This seemed to rule out the use of very flexible modeling approaches, which generally need more data in each subgroup. Therefore, parametric modeling assumptions seemed needed and important to pre-specify carefully, using background knowledge, to obtain reliable results. We found the background clinical knowledge easier to translate into reliable parametric assumptions for the 33-months absolute risk of CV death than for the two cause-specific hazards at any time within the follow-up. This was especially the case when considering parametric assumptions related to the inclusion of interaction terms, which can be important. Consequently, we chose to pre-specify a G-computation approach based on logistic regression of the absolute risk over an alternative based on regressions of the two cause-specific hazards. That is, we chose the approach presented in Sect. 3.1 over that presented in Ozenne





**Fig. 2** Estimates risks per treatment group and risk differences, with 95% confidence intervals. Here the IPCW weights are based on choice C4. The labels on the Y-axis indicate the method used to obtain the results: “Crude”: unadjusted (crude) analysis using a stratified Aalen-Johanssen estimator; “Cox G-computation”: G-computation based on Cox regression for each cause-specific hazard; “IPTCW”: inverse probability of censoring and treatment weighting ( $\widehat{F}_1^{iptcw}(t, a)$ ); “G-computation”: G-computation based on logistic regression for the 33-month risk of CV death ( $\widehat{F}_1^g(t, a)$ ); “DR”: DR estimator of Sect. 3.2 ( $\widehat{F}_1^{dr}(t, a)$ ); “DR-reg: 1 cov” and “DR-reg: 2 cov”: regression based DR estimators of Sect. 3.3 with one, respectively two, clever covariates; “tmle: 1 cov” and “tmle: 2 cov”: tMLE-like estimators of Sect. 3.4 with one, respectively two, clever covariates

et al. (2020). Pre-specifying a reliable model for the likelihood of initiating a Beta-Blocker treatment appeared also rather challenging in our context, using background knowledge. That is why we pre-specified the choice of an analysis that did not use propensity scores either.

Finally, also for similar reasons, it would have been challenging to use the augmented estimators of Sect. 2.7 in a meaningful way. Therefore, we did not pursue the use of these interesting estimators in this application. All results building on logistic regressions that we present are based on estimators solving the estimating Eqs. (2) or (3).

### 4.4 Results

Figure 2 displays all results obtained from the different methods, with censoring modeling choice C4, which we believe to be the best. Standard errors (se) were computed by Bootstrap (400 samples) and 95% confidence intervals (CI) for the risk differences were computed as  $CI = estimate \pm 1.96 \cdot se$ . Similar figures for choices C1, C2 and C3 are provided in the supplementary material.

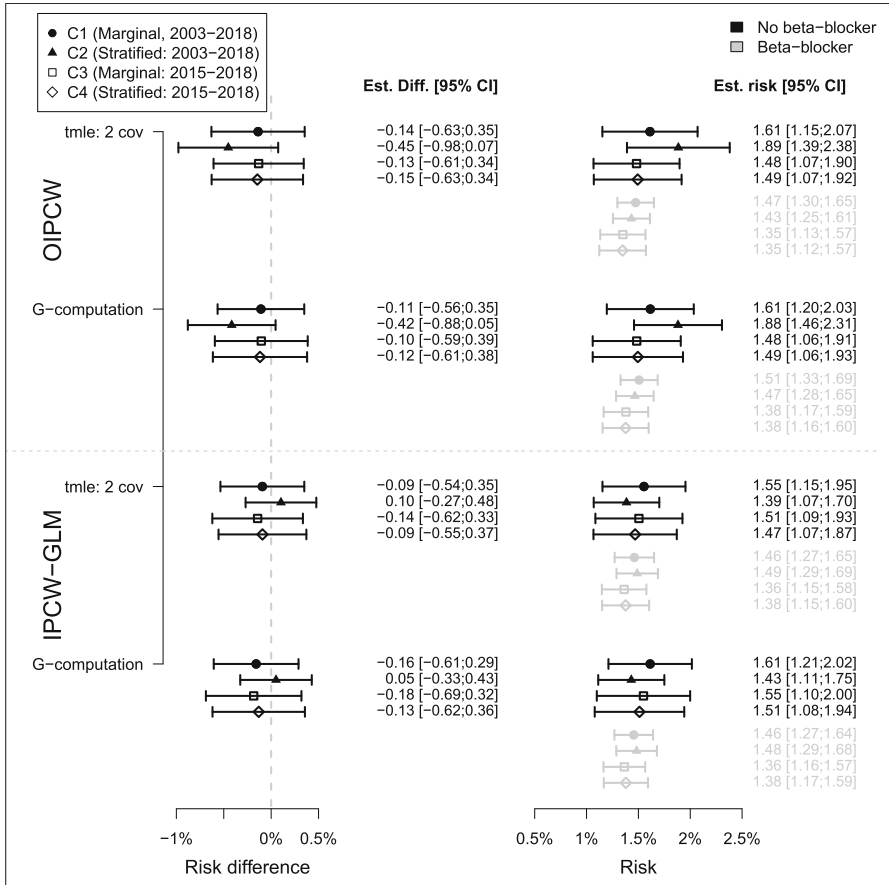


Fig. 3 Sensitivity of the results to the modeling choices to compute the IPCW weights

Figure 2 does not show large differences between the results obtained when using OIPCW or IPCW-GLM. The results of the crude analysis is surprisingly close to the results obtained with the other methods that adjust for confounding. The reason is probably that the distribution of the baseline covariates  $L$  is actually quite similar in the two treatment groups, which is what Table 1 in Holt et al. (2021) suggests. That might also partly explain why all the other results are also quite similar. We note that the narrowest confidence interval is obtained when using G-computation based on Cox regressions for the two cause-specific hazards. This is not surprising, since this approach makes stronger parametric assumptions. However, the difference is modest and we believe that it has a negligible impact on the clinical interpretation of the results.

Figure 3 provides an overview of the sensitivity of the results to choices C1-C4, when using the G-computation and tMLE-like estimators of Sects. 3.1 and 3.4, using either IPCW-GLM or OIPCW. Unlike in Fig. 2, here we show the results for only one of the five double robust estimators. This is because the results are quite similar

for the others (see Figures in the supplementary material). Interestingly, Fig. 3 and choices C1-C4 illustrate that incorrectly modeling some truly existing associations between censoring and baseline covariates can lead to more bias than to not model them. Indeed, if we consider the choice C4 as the most rational in our context, we see that choice C2 leads to results further away from those of choice C4 than choice C1. This is the case although choice C2 accounts for the strong dependence between censoring and treatment in a flexible way, whereas choice C1 does not account for any dependence between censoring and baseline covariates. Figure 3 also illustrates that using OIPCW or IPCW-GLM can give substantially different results. For instance, the difference is substantial with censoring adjustment C2.

## 5 Efficient and robust analysis of RCT

When the data come from a randomized trial, one can estimate  $F_1(t, a)$ , for  $a = 0$  and 1, directly by an unadjusted logistic regression. This is because unlike in observational studies,  $A$  is independent of  $L$  in RCTs, because of randomization. Hence,  $F_1(t, a) = P(D(t) = 1|A = a)$ .

Note that also with censored data, when using a stratified Kaplan-Meier estimator, stratified on treatment  $A$ , to compute the IPCW weights, fitting the unadjusted logistic regression model is algebraically equivalent to a widespread approach. Indeed, it reduces to using a stratified Kaplan-Meier estimator, with survival data, or a stratified Aalen-Johanssen estimator, with competing risks data, to estimate  $F_1(t, a)$ , for  $a = 0, 1$  (Geskus 2016).

However, instead of using an unadjusted analysis, one can gain efficiency by leveraging the information contained in the baseline covariates  $L$ . See, e.g., Robinson and Jewell (1991); Zhang et al. (2008). The simplest way to do that is to add the baseline covariates  $L$  into the model, in addition to  $A$ , to model and estimate  $F_1(t, a, L)$ . Interestingly, even if this is done using an incorrect model, that is, even if  $Q(t, L, a) \neq F_1(t, L, a)$ , the type-I error rate for testing a treatment effect will be controlled, as long as the Wald-type test statistic is computed with a robust standard error. Furthermore, the g-computation estimator  $\widehat{F}_1^g(t, a)$  defined in (8) will still be consistent under arbitrary model misspecification (Moore and van der Laan 2009; Rosenblum and Steingrimsson 2016). Consequently, adjusting for baseline covariates  $L$  when analyzing RCTs is recommended by regulatory agencies such as the Food and Drug Administration (FDA 2021) and the European Medicine Agency (EMA 2015). The latter states that “*Variables known a priori to be strongly, or at least moderately, associated with the primary outcome and/or variables for which there is a strong clinical rationale for such an association should also be considered as covariates in the primary analysis.*”

In the next section, we start by briefly reviewing these robustness results, in the binary uncensored case. We then provide related results for the right censored case, when an additional censoring adjustment is made based on our IPCW weighted logistic regression models.

To the best of our knowledge, previous work on similar robustness properties with right-censored data focused on hazards modeling, not on risk modeling as we do below

(Struthers and Kalbfleisch 1986; Lin and Wei 1989; DiRienzo and Lagakos 2001; Lu and Tsiatis 2008; Kim 2013; Vansteelandt et al. 2014).

### 5.1 Robustness for binary uncensored data

We will assume the use a simple logistic model, where  $X = (1, A, L)$ . That is, the risk of  $D(t) = 1$  given  $\mathbf{x} = (1, a, l)$  being modeled by

$$Q(t, a, l, \beta) = \frac{\exp(\beta_0 + \beta_{Aa} + \beta_L l)}{1 + \exp(\beta_0 + \beta_{Aa} + \beta_L l)}. \quad (14)$$

Further, we assume that, as expected by design, randomization makes the treatment variable  $A$  independent of the baseline covariates  $L$ . With binary uncensored data and when fitting the above model (14) via usual maximum likelihood, i.e. by solving the estimating Eq. (1), then Rosenblum and Steingrimsson (2016) established the interesting robustness properties that follow. First,  $\beta_A$  converges in probability towards 0 if and only if the null hypothesis of no treatment effect,  $H_0 : F_1(t, 0) = F_1(t, 1)$ , holds, under any model misspecification. Furthermore, the G-computation estimator  $\widehat{F}_1^g(t, a) = (1/n) \sum_{i=1}^n Q(t, a, L_i, \widehat{\beta})$  converges in probability towards  $F_1(t, a)$ , for  $a = 0, 1$ , also under any model misspecification. The estimated treatment effects  $\widehat{F}_1^g(t, 1) - \widehat{F}_1^g(t, 0)$  and  $\widehat{F}_1^g(t, 1)/\widehat{F}_1^g(t, 0)$  are therefore also consistent.

As pointed out by Rosenblum and Steingrimsson (2016), the robustness property of the G-computation estimator is very useful in practice. It provides a simple and transparent approach to leverage the information contained in baseline variables which i) does not rely on any modeling assumptions and ii) estimates the same average treatment effect as that estimated by an unadjusted analysis. The use of this G-computation is therefore now promoted by the most recent FDA guidelines (FDA 2021).

For computing confidence intervals and p-values, Rosenblum and Steingrimsson (2016) note that one can compute robust standard errors via nonparametric bootstrap and exploit the asymptotic normal distribution of the estimators. Finally, note that Rosenblum and Steingrimsson (2016) also showed that the two hypothesis tests for  $H_0$ , defined via the G-computation estimator  $\widehat{F}_1^g(t, 1) - \widehat{F}_1^g(t, 0)$  and via the maximum likelihood estimator  $\widehat{\beta}_A$  of  $\beta_A$ , do not only both control the type-I error rate, but they are also equally powerful, asymptotically, under any model misspecification.

### 5.2 Results for right-censored data

With right-censored data, the robustness properties still hold under additional conditions, but not necessarily without. This the topic of the following theorems, for which we sketch the proofs in Appendix B.

**Theorem 2** *For both approaches, OPICW and IPCW-GLM, when the censoring adjustment model is correct, then the G-computation estimator  $\widehat{F}_1^g(t, a)$  is consistent, for  $a = 0, 1$ . Further,  $\widehat{\beta}_A$  converges to 0 if and only if the null hypothesis of no treatment effect,  $H_0 : F_1(t, 0) = F_1(t, 1)$ , is true. The two results hold under arbi-*

rary misspecification of the working model  $Q(t, a, \mathbf{I})$ , in randomized trials where  $A$  is independent of  $\mathbf{L}$ .

Consequently, hypothesis testing for a treatment effect based on either the average risk difference estimator  $\widehat{F}_1^g(t, 1) - \widehat{F}_1^g(t, 0)$  or  $\widehat{\beta}_A$  will control the type-I error, asymptotically, as long as an appropriate (robust) standard error is used in the Wald-type test statistic. However, the estimated value  $\widehat{\beta}_a$  will usually not have a meaningful interpretation under model misspecification, unlike  $\widehat{F}_1^g(t, 1) - \widehat{F}_1^g(t, 0)$ . Indeed, if  $F_1(t, a, \mathbf{I}) \neq Q(t, a, \mathbf{I})$ , the limit of  $\widehat{\beta}_a$  does no longer have a simple interpretation as the logarithm of a conditional odds ratio.

For many clinical trials, the censoring distribution is not expected to depend on  $(A, \mathbf{L})$ . This is for instance the case when only administrative censoring occurs. Consequently, using a Kaplan-Meier estimator, possibly stratified by treatment group, should be sufficient to ensure a correct censoring adjustment and thus guarantee the above robustness properties.

Without a correct censoring model, some of the robustness properties still hold under specific assumptions, when using a stratified Kaplan-Meier estimator to compute the IPCW weights. The next theorem identifies two such assumptions. The first is that the hazard of the censoring time  $C$  has an additive structure. The second is that there is no conditional treatment effect given baseline covariates, neither on the risk of the event of interest nor on the competing risk. Without these assumptions, it can be seen from the proof of the theorem (or simple simulations) that the robustness properties will generally no longer hold.

**Theorem 3** Assume that i)  $\lambda_c(t, A, \mathbf{L}) = \lambda_0(t) + A\lambda_1(t) + \lambda_2(t, \mathbf{L})$ , where  $\lambda_0(t)$  and  $\lambda_1(t)$  do not depend on  $(A, \mathbf{L})$  and  $\lambda_2(t, \mathbf{L})$  does not depend on  $A$ , and ii)  $F_k(s, 0, \mathbf{I}) = F_k(s, 1, \mathbf{I})$  for all  $\mathbf{I}$ ,  $s \leq t$  and  $k = 1, 2$ , where  $F_k(s, a, \mathbf{I}) = P(T \leq s, \eta = k | A = a, \mathbf{L} = \mathbf{I})$ . Further, assume that a stratified Kaplan-Meier estimator is used to compute the IPCW weights, stratified on  $A$ . Then, for both approaches, OPICW and IPCW-GLM, both the  $G$ -computation estimator of the risk difference,  $\widehat{F}_1^g(t, 1) - \widehat{F}_1^g(t, 0)$ , and  $\widehat{\beta}_{|A}$ , converge to 0. The results hold under arbitrary misspecification of the working model  $Q(t, a, \mathbf{I})$ , in randomized trials where  $A$  is independent of  $\mathbf{L}$ .

Group sequential trials are increasingly used, especially in the pharmaceutical industry. Interestingly, recent results from Martens and Logan (2020) show that  $\widehat{\beta}_{opicw}$  has the asymptotic properties required by the common stopping boundary specification methods, to design group sequential trials. The above robustness results are therefore also useful for this type of trials.

## 6 Discussion

We have presented two approaches to fit logistic regression models with right censored data. One method deals with the censored data by weighting the outcome (OIPCW), the other by weighting the estimating equations (IPCW-GLM). We have studied the large sample properties of the two approaches and this revealed that which approach is the most efficient depends on the distribution of the censoring times. We have further

illustrated that there can be surprisingly big differences between the two methods. In a concrete setting, we saw that a parameter estimated using OIPCW had a variance almost three times smaller than when using IPCW-GLM. In most other settings, however, the two methods gave similar results.

We have also reviewed the estimation of causal average treatment effects in observational studies, via G-computation and double-robust estimation, both building on logistic regression. Then, we showed that, here again, OIPCW and IPCW-GLM can be used to deal with right censored data in simple ways. In our application to Danish registry data, all methods gave similar results with negligible difference in terms of the clinical interpretation.

As pointed out by a reviewer, the causal interpretation of the results in our specific application might not be as simple as it appears at first reading. This is because the patients included in our study either have initiated the treatment at any time within 3 months after MI ( $A = 1$ ) or have had the opportunity but did not ( $A = 0$ ). Hence, all patients included in our study need to have survived 3 months (and meet other inclusion criteria) to be included in the analysis. It is difficult to proceed differently with our registry data, but this introduces a selection mechanism that can be a consequence of the treatment. This complicates the interpretation of the results as it will usually make the two treatment groups no longer comparable at start of follow-up, 3 months after MI. This selection mechanism is similar to that involved in the definition of hazard ratios (Martinussen et al. 2020) and estimands based on functional outcomes in RCTs with high mortality (Colantuoni et al. 2018). One could think about our observational study as an attempt to emulate a RCT, when considering an outcome similar to those “truncated due to death” discussed in Colantuoni et al. (2018). Thus, the results should probably not be interpreted on their own, but in conjunction with others, e.g., with the probability of a patient being included in the analysis, for each treatment group. We refer to Colantuoni et al. (2018) for excellent discussions about closely related topics.

Finally, we have reviewed important robustness properties of logistic regression models fitted to RCT data. These properties support recently updated guidelines from regulatory agencies, which promote adjustment on baseline variables and the use of G-computation via logistic regression (FDA 2021). We showed that these properties are preserved when using OIPCW or IPCW-GLM to deal with right censored data, under correct censoring adjustment. This makes logistic regression and the derived G-computation estimator appealing to analyze time-to-event endpoints in RCTs. Especially when considering estimands, G-computation may be a relevant alternative to the use of more traditional hazard-based regression models (Rufibach 2019).

Our robustness results have been shown assuming simple randomization, not stratified randomization. Recently, Wang et al. (2021) showed that for uncensored data, the robustness results are preserved under stratified randomization. We think that this is also the case when using OIPCW and IPCW-GLM to deal with right censored data, under correct censoring adjustment. This would be interesting to confirm in future work.

Software have been made available to facilitate the use of both the opicw and the IPCW-GLM approach. They provide asymptotically correct standard error computation based on large sample expansions such as those of Theorem 1. See functions `binreg` and `logitIPCW` in the R package `metS` (Holst and Scheike 2021), to fit a

logistic regression model with OICPW and IPCW-GLM, respectively. See functions `binregATE` and `logitATE` to compute causal risks and risk differences. These functions implement the G-computation and double robust estimators of Sects. 3.1 and 3.2, with OICPW and IPCW-GLM, respectively.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10985-022-09564-6>.

**Acknowledgements** We are grateful to the editors as well as two referees for their constructive and useful comments.

## Appendix A

### Appendix A.1: Proof of Theorem 1

Overall, we follow similar lines as those of Bang and Tsiatis (2000). First, note that for all  $s \in [0, t]$ ,

$$W(s) = \frac{\Delta(s)}{G_c(s \wedge \tilde{T})} = 1 - \int_0^s \frac{dM^c(u)}{G_c(u)}, \tag{15}$$

with  $\Delta(s) = \mathbb{I}\{s \wedge T \leq C\}$ ,  $Y(u) = \mathbb{I}\{\tilde{T} \geq u\}$  and  $M^c(u) = \mathbb{I}\{\tilde{T} \leq u, \Delta = 0\} - \int_0^u Y(v)\lambda_c(v)dv$ . The second equality in (15) has been pointed out by Robins and Rotnitzky (1992) for  $s = \infty$  and it can be shown as follows for any  $s \in [0, t]$  as follows.

$$\begin{aligned} 1 - \int_0^s \frac{dM^c(u)}{G_c(u)} &= 1 - \left\{ \frac{(1 - \Delta)\mathbb{I}\{\tilde{T} \leq s\}}{G_c(\tilde{T})} - \int_0^{\tilde{T} \wedge s} \frac{\lambda_c(u)}{G_c(u)} du \right\} \\ &= 1 - \left\{ \frac{(1 - \Delta)\mathbb{I}\{\tilde{T} \leq s\}}{G_c(\tilde{T})} - \left[ \frac{1}{G_c(\tilde{T} \wedge s)} - \frac{1}{G_c(0)} \right] \right\}, \end{aligned}$$

because

$$\frac{\lambda_c(u)}{G_c(u)} du = -\frac{dG_c(u)}{\{G_c(u)\}^2} \quad \text{and} \quad \frac{d}{du} \left( \frac{1}{G_c(u)} \right) = -\frac{dG_c(u)}{\{G_c(u)\}^2}.$$

Since  $G_c(0) = 1$ , it follows

$$1 - \int_0^s \frac{dM^c(u)}{G_c(u)} = \begin{cases} \Delta/G_c(\tilde{T}) & \text{if } \tilde{T} \leq s \\ 1/G_c(s) & \text{if } \tilde{T} > s \end{cases} = W(s).$$

Second, recall a well-known martingale integral representation for the Kaplan-Meier estimator, see e.g. Andersen et al. (1993, Sec. IV.3), for all  $s \in [0, t]$ ,

$$\frac{\widehat{G}_c(s) - G_c(s)}{G_c(s)} = - \sum_{i=1}^n \int_0^s \frac{\widehat{G}_c(u-)}{Y_{\bullet}(u)} \frac{dM_i^c(u)}{G_c(u)}, \tag{16}$$

where  $Y_{\bullet}(u) = \sum_{i=1}^n \mathbb{1}\{\widetilde{T}_i \geq u\}$ . Third, note that

$$n^{-1}Y_{\bullet}(u) = \widehat{G}_c(u-)\widehat{S}(u-), \tag{17}$$

where  $\widehat{S}(u-)$  is the Kaplan-Meier estimator for  $S(u-) = P(T > u-)$ . With the notations

$$\begin{aligned} \widehat{U}_{oipcw}(\boldsymbol{\beta}) &= n^{-1} \sum_{i=1}^n X_i \left\{ \widehat{W}_i(t) D_i(t) - Q(t, X_i, \boldsymbol{\beta}) \right\} \\ U_{oipcw}(\boldsymbol{\beta}) &= n^{-1} \sum_{i=1}^n X_i \left\{ W_i(t) D_i(t) - Q(t, X_i, \boldsymbol{\beta}) \right\} \\ \text{and } U_{glm}(\boldsymbol{\beta}) &= n^{-1} \sum_{i=1}^n X_i \left\{ D_i(t) - Q(t, X_i, \boldsymbol{\beta}) \right\}, \end{aligned}$$

we first note that

$$\begin{aligned} n^{1/2} \widehat{U}_{oipcw}(\boldsymbol{\beta}) &= \underbrace{n^{1/2} U_{oipcw}(\boldsymbol{\beta})}_{(*)} \\ &+ \underbrace{n^{-1/2} \sum_{j=1}^n X_j D_j(t) \frac{\Delta_j(t)}{\widehat{G}_c(t \wedge \widetilde{T}_j)} \left\{ \frac{G_c(t \wedge \widetilde{T}_j) - \widehat{G}_c(t \wedge \widetilde{T}_j)}{G_c(t \wedge \widetilde{T}_j)} \right\}}_{(**)}. \end{aligned}$$

From (15), we get

$$(*) = n^{1/2} U_{glm}(\boldsymbol{\beta}) - n^{-1/2} \sum_{i=1}^n \int_0^t X_i D_i(t) \frac{dM_i^c(u)}{G_c(u)},$$

and from (16), we get

$$\begin{aligned} (**) &= n^{-1/2} \sum_{j=1}^n X_j D_j(t) \frac{\Delta_j(t)}{\widehat{G}_c(t \wedge \widetilde{T}_j)} \left\{ - \sum_{i=1}^n \int_0^{t \wedge \widetilde{T}_j} \frac{\widehat{G}_c(u-)}{Y_{\bullet}(u)} \frac{dM_i^c(u)}{G_c(u)} \right\} \\ &= -n^{-1/2} \sum_{i=1}^n \int_0^t \left[ \frac{\widehat{G}_c(u-)}{Y_{\bullet}(u)} \sum_{j=1}^n X_j D_j(t) \frac{\Delta_j(t)}{\widehat{G}_c(t \wedge \widetilde{T}_j)} \mathbb{1}\{\widetilde{T}_j \geq u\} \right] \frac{dM_i^c(u)}{G_c(u)} \end{aligned}$$



and from (17), we get

$$(**) = -n^{-1/2} \sum_{i=1}^n \int_0^t \widehat{E}[XD(t)|T \geq u] \frac{dM_i^c(u)}{G_c(u)}$$

where

$$\begin{aligned} \widehat{E}[XD(t)|T > u] &= \frac{1}{n\widehat{S}(u-)} \sum_{i=1}^n \frac{\Delta_i(t) \mathbb{1}\{\widetilde{T}_i \geq u\} X_i D_i(t)}{\widehat{G}_c(t \wedge \widetilde{T}_i)} \\ &= E[XD(t) | T \geq u] + o_p(1), \end{aligned}$$

because of independent censoring and the uniform convergence of the Kaplan-Meier estimator  $\widehat{G}_c(\cdot)$  in  $[0, t]$ . Consequently,

$$\begin{aligned} n^{1/2} \widehat{U}_{oipcw}(\boldsymbol{\beta}) &= n^{1/2} U_{glm}(\boldsymbol{\beta}) - n^{-1/2} \\ &\quad \sum_{i=1}^n \int_0^t \left\{ X_i D_i(t) - \widehat{E}[XD(t) | T \geq u] \right\} \frac{dM_i^c(u)}{G_c(u)} \\ &= n^{1/2} U_{glm}(\boldsymbol{\beta}) - n^{-1/2} \\ &\quad \sum_{i=1}^n \int_0^t \left\{ X_i D_i(t) - E[XD(t) | T \geq u] \right\} \frac{dM_i^c(u)}{G_c(u)} \\ &\quad + o_p(1) \end{aligned}$$

For  $n^{1/2} \widehat{U}_{ipcw-glm}(\boldsymbol{\beta})$ , the result follows similarly, since the calculation is similar except from  $X_i D_i(t)$  being replaced with  $X_i \{D_i(t) - Q(t, X_i)\}$  in (\*) and (\*\*).

### Appendix A.2: variance estimator $\widehat{\boldsymbol{\Sigma}}_m$

From the above derivations, the same calculations without using

$$n^{1/2} U_{oipcw}(\boldsymbol{\beta}) = n^{1/2} U_{glm}(\boldsymbol{\beta}) - n^{-1/2} \sum_{i=1}^n \int_0^t X_i D_i(t) \frac{dM_i^c(u)}{G_c(u)}$$

lead to  $n^{1/2} \widehat{U}_{oipcw}(\boldsymbol{\beta}) = n^{-1/2} \sum_{i=1}^n \boldsymbol{\epsilon}_i + o_p(1)$  where

$$\boldsymbol{\epsilon}_i = X_i \left\{ W_i(t) D_i(t) - Q(t, X_i, \boldsymbol{\beta}) \right\} + \int_0^t E[XD(t) | T \geq u] \frac{dM_i^c(u)}{G_c(u)}.$$

One can consistently estimate  $\boldsymbol{\epsilon}_i$  by

$$\widehat{\boldsymbol{\epsilon}}_i = X_i \left\{ \widehat{W}_i(t) \widetilde{D}_i(t) - Q(t, X_i, \widehat{\boldsymbol{\beta}}_{oipcw}) \right\} + \int_0^t \widehat{E}[XD(t) | T \geq u] \frac{d\widehat{M}_i^c(u)}{\widehat{G}_c(u)},$$

where  $\widehat{M}_i^c(s) = \mathbb{1}\{\widetilde{T}_i \leq s, \Delta_i = 0\} - \int_0^s Y_i(v)d\widehat{\Lambda}_c(v)$ , where  $\widehat{\Lambda}_c(s)$  is the Nelson-Aalen estimator of the cumulative hazard of  $C$  at time  $s$  and

$$\widehat{E}[XD(t)|T > u] = \frac{1}{n\widehat{S}(u-)} \sum_{i=1}^n \frac{\Delta_i(t)\mathbb{1}\{\widetilde{T}_i \geq u\}X_i\widetilde{D}_i(t)}{\widehat{G}_c(t \wedge \widetilde{T}_i)}.$$

Consequently, one can consistently estimate  $\Omega_{oipcw}$  by  $\widehat{\Omega}_{oipcw} = (1/n) \sum_{i=1}^n \widehat{\epsilon}_i \widehat{\epsilon}_i^T$ . Similarly, to estimate  $\Omega_{ipcw-glm}$  we can use  $\widehat{\Omega}_{ipcw-glm} = (1/n) \sum_{i=1}^n \widehat{\omega}_i \widehat{\omega}_i^T$  with

$$\begin{aligned} \widehat{\omega}_i &= \widehat{W}_i(t)X_i \left\{ \widetilde{D}_i(t) - Q(t, X_i, \widehat{\beta}_{ipcw-glm}) \right\} \\ &+ \int_0^t \widehat{E}[X\{D(t) - Q(t, X_i, \beta)\} | T \geq u] \frac{d\widehat{M}_i^c(u)}{\widehat{G}_c(u)} \end{aligned}$$

where

$$\begin{aligned} &\widehat{E}[X\{D(t) - Q(t, X_i, \beta)\} | T \geq u] \\ &= \frac{1}{n\widehat{S}(u-)} \sum_{i=1}^n \frac{\Delta_i(t)\mathbb{1}\{\widetilde{T}_i \geq u\}X_i \left\{ \widetilde{D}_i(t) - Q(t, X_i, \widehat{\beta}_{ipcw-glm}) \right\}}{\widehat{G}_c(t \wedge \widetilde{T}_i)}. \end{aligned}$$

Consistent estimators of  $\mathcal{I}$  are  $\widehat{\mathcal{I}}_m = n^{-1} \sum_{i=1}^n [X_i^2 Q(t, X_i, \widehat{\beta}_m) \{1 - Q(t, X_i, \widehat{\beta}_m)\}]$ , for  $m = oipcw$  or  $ipcw - glm$ . Finally,  $\Sigma_m = \mathcal{I}^{-1}\Omega_m\mathcal{I}^{-1}$  can be estimated by  $\widehat{\Sigma}_m = \widehat{\mathcal{I}}_m^{-1}\widehat{\Omega}_m\widehat{\mathcal{I}}_m^{-1}$ .

**Appendix A.3: Proof of Proposition 4**

A proof was provided by Bartlett (2018, Appendix A.2), in a slightly different context. We repeat the main arguments here for completeness. First, we note that a Taylor-expansion and the results from Sects. 2.5 and 2.7 imply that

$$\sqrt{n} \left\{ \widehat{F}_1^g(t, a) - F_1(t, a) \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ Q(t, a, L_i) - F_1(t, a) + \mathbf{B}_\beta(a)\psi_i \right\} + o_p(1)$$

with  $\psi_i = \mathcal{I}^{-1}\Phi\{X_i, D_i(t), \widetilde{T}_i, \widetilde{\eta}_i, t\}$ . Here,  $\Phi$  denotes  $\Phi_m$  or  $\Phi_m^{Aug}$ , for  $m = oipcw$  or  $ipcw - glm$ , depending on which estimator  $\widehat{\beta}$  we plugged-in to define  $\widehat{F}_1^g(t, a)$ . Formulas for the different versions of  $\Phi$  can be found in Theorem 1 or in the proof of Proposition 2. Hence, it only remains to prove that  $\text{Cov}\{Q(t, a, L), \mathbf{B}_\beta(a)\psi\} = 0$ . This can be done by expanding the covariance, using the conditional covariance formula, as

$$\text{Cov}\left\{E[Q(t, a, L) | X], E[\mathbf{B}_\beta(a)\psi | X]\right\} + E\left[\text{Cov}\{Q(t, a, L), \mathbf{B}_\beta(a)\psi | X\}\right]$$

The second term is 0, because the first component is constant, conditional on  $X$ . We now explain why the first term is also 0. First note, that

$$E\left[\mathbf{B}_{\beta(a)}\psi \mid X\right] = \mathbf{B}_{\beta(a)}\mathcal{I}^{-1}E\left[\Phi\{\dots\} \mid X\right]$$

and

$$E\left[\Phi\{\dots\} \mid X\right] = E\left[X\{D(t) - Q(t, X)\} - \int_0^t \varphi(X, D(t), s) \frac{dM^c(s)}{G_c(s)} \mid X\right].$$

Second, note that  $E[X\{D(t) - Q(t, X)\} \mid X] = X\{F_1(t, X) - Q(t, X)\} = \mathbf{0}$  because the model is assumed to be well specified. Third,  $E\left[\int_0^t \varphi(X, D(t), s) \frac{dM^c(s)}{G_c(s)} \mid X\right] = 0$  follows by independent censoring and standard martingale theory (Aalen et al. 2008, Sec. 2.2).

## Appendix B

### 7.1 Sketch of Proof for Theorem 2

A proof was provided by Rosenblum and Steingrimsson (2016) for the uncensored case. We here essentially repeat their main arguments, which also apply in our case, up to minor differences introduced by the IPCW weights. First, note that  $\hat{\beta}_{\text{ipcw}}$  is consistent for  $\beta$ , where  $\beta$  is the solution to

$$E\left[X\left\{D(t) \cdot W(t, X) - Q(t, X, \beta)\right\}\right] = \mathbf{0}, \tag{18}$$

with  $X = (1, A, L)^T$  and  $W(t, X) = \Delta(t)/G_C(t \wedge \tilde{T}, X)$ , using a notation that emphasizes the (potential) dependence of the IPCW weight on  $X$ . Second, note that the first equation in (18), which corresponds to the first component of  $X = (1, A, L)^T$ , is

$$\begin{aligned} E[D(t)W(t, X)] &= E[Q(t, X, \beta)] \\ &= \sum_{a \in \{0,1\}} E\left[Q(t, (1, A, L)^T, \beta) \mid A = a\right] \cdot \pi(a) \\ &= E[Q(t, (1, 1, L)^T, \beta)] \cdot \pi(1) + E\left[Q(t, (1, 0, L)^T, \beta)\right] \cdot \pi(0) \end{aligned} \tag{19}$$

where  $\pi(a) = P(A = a)$ , for  $a = 0, 1$ , and where the last equality follows because  $A$  is assumed independent of  $L$  (randomization), hence the conditioning disappears in the expectations. Similarly, the second equation in (18), which corresponds to the second component of  $X = (1, A, L)^T$ , is

$$E[AD(t)W(t, X)] = E[AQ(t, X, \beta)]$$

$$= E[Q(t, (1, 1, L)^T, \beta)] \cdot \pi(1). \tag{20}$$

Furthermore,

$$E[D(t)W(t, X)] = E[AD(t)W(t, X)] + E[(1 - A)D(t)W(t, X)]. \tag{21}$$

By plugging-in (19) and (20) into (21), we find

$$E[(1 - A)D(t)W(t, X)] = E[Q(t, (1, 0, L)^T, \beta)] \cdot \pi(0). \tag{22}$$

Further note that we also have the identities

$$E[AD(t)W(t, X)] = E[D(t)W(t, X)|A = 1] \cdot \pi(1) \tag{23}$$

and  $E[(1 - A)D(t)W(t, X)] = E[D(t)W(t, X)|A = 0] \cdot \pi(0). \tag{24}$

Hence, Eqs. (23) and (20) give

$$E[D(t)W(t, X) | A = 1] = E[Q(t, (1, 1, L)^T, \beta)], \tag{25}$$

and Eqs. (24) and (22) give

$$E[D(t)W(t, X) | A = 0] = E[Q(t, (1, 0, L)^T, \beta)]. \tag{26}$$

The robustness result for the G-computation estimator, that is,  $E[Q(t, (1, a, L), \beta)] = F_1(t, a)$ , for  $a = 0, 1$ , therefore follows for OIPCW because  $E[D(t)W(t, X)|A = a] = F_1(t, a)$ , when the censoring adjustment is correct.

To show that  $\widehat{\beta}_A$  converges in probability towards zero if and only if  $F_1(t, 0) = F_1(t, 1)$ , we here repeat the argument of Rosenblum and Steingrimsson (2016). As, we have just shown that

$$F_1(t, 0) - F_1(t, 1) = E[\text{expit}(\beta_0 + \beta_L L) - \text{expit}(\beta_0 + \beta_A + \beta_L L)]$$

and since  $x \mapsto \text{expit}(x)$  is monotonically increasing, then it follows that  $F_1(t, 0) - F_1(t, 1) = 0$  if and only if  $\beta_A = 0$ . This completes the proof for the OIPCW approach.

The same arguments apply for IPCW-GLM too. Briefly, instead of the solution to (18), we consider the solution to

$$E[XW(t, X)\{D(t) - Q(t, X, \beta)\}] = \mathbf{0}.$$

Instead of (25) and (26), this leads to

$$E[D(t)W(t, X) | A = 0] = E[W(t, X)Q(t, (1, 0, L)^T, \beta) | A = 0] \tag{27}$$

$$E[D(t)W(t, X) | A = 1] = E[W(t, X)Q(t, (1, 1, L)^T, \beta) | A = 1]. \tag{28}$$

Note that in the expectations in (27) and (28) the conditioning does not vanish as in (26) and (25), because of the weight  $W(t, \mathbf{X})$ . If the censoring adjustment is correct, then  $E[W(t, \mathbf{X}) \mid A = a, \mathbf{X}] = 1$  for  $a = 0, 1$ , which together with the law of iterated expectations implies that the right-hand side of (27) and (28) are the same as that of (25) and (26). Hence (27) and (28) are equivalent to (26) and (25) and the rest of the proof is identical to that of OIPCW.

### 7.2 Sketch of Proof for Theorem 3

The proof exploits the three key equalities (29), (30) and (31) below. They are proven at the end of this section, after the proof of the theorem which exploits them. First, the stratified Kaplan-Meier estimator of  $P(C > s \mid A = a)$ , for  $s \in [0, t]$ , will converge to  $\tilde{G}_c(s, A) = \exp\{-\int_0^s \tilde{\lambda}_c(u, A)du\}$ , where

$$\tilde{\lambda}_c(u, A) = \lim_{h \rightarrow 0} \frac{1}{h} P(C \in [u, u + h] \mid C > u, T > u, A) = \tilde{\lambda}_0(u) + A\lambda_1(u), \tag{29}$$

where  $\tilde{\lambda}_0(u)$  does not depends on  $(A, \mathbf{L})$ . Second,

$$E\left[\tilde{W}(t, A)D(t) \mid A = 0\right] = E\left[\tilde{W}(t, A)D(t) \mid A = 1\right], \tag{30}$$

where  $\tilde{W}(t, A) = \Delta(t)/\tilde{G}_c(t \wedge T, A)$ , where the notation  $\tilde{W}(t, A)$  emphasizes both the misspecification of the censoring adjustment and the dependence on  $A$ . Third, for  $a = 0, 1$ ,

$$E\left[\tilde{W}(t, A)Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = a\right] = E\left[r_c(T \wedge t, \mathbf{L})Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta})\right], \tag{31}$$

where  $r_c(T \wedge t, \mathbf{L}) \geq 0$  depends on  $(\mathbf{L}, T)$  but not on  $A$  (see Sect. 7.2.3 for details).

Other useful results are “new versions” of (25) and (26) (for OIPCW) and (27) and (28) (for IPCW-GLM). The equations of (25), (26), (27) and (28) still hold under the weaker assumptions of Theorem 2, up to the minor difference that  $W(t, \mathbf{X})$  should now be replaced by  $\tilde{W}(t, A)$ . Indeed, (25), (26), (27) and (28) were derived without making any assumptions on the censoring adjustment. They were only consequences of the estimating equations. Below we refer to these “new versions” when we cite (25), (26), (27) or (28).

Let us now consider the OIPCW case. Equations (30), (25) and (26), imply

$$E[Q(t, (1, 1, \mathbf{L})^T, \boldsymbol{\beta})] = E[Q(t, (1, 0, \mathbf{L})^T, \boldsymbol{\beta})]$$

which proves that  $\hat{F}_1^g(t, 1) - \hat{F}_1^g(t, 0)$  convergences to 0, for OIPCW. Using the same monotonicity argument as in the Proof of Theorem 2, this implies that  $\hat{\beta}_A$  also convergences to 0.

Let us now consider the IPCW-GLM case. Equations (30), (27) and (28), imply

$$E\left[\tilde{W}(t, A)Q(t, (1, 1, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = 1\right] = E\left[\tilde{W}(t, A)Q(t, (1, 0, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = 0\right]$$

which, using (31), further leads to

$$E\left[r_c(T \wedge t, \mathbf{L})\left\{\text{expit}(\beta_0 + \boldsymbol{\beta}_L \mathbf{L}) - \text{expit}(\beta_0 + \beta_A + \boldsymbol{\beta}_L \mathbf{L})\right\}\right] = 0.$$

Because  $r_c(T \wedge t, \mathbf{L}) > 0$  and  $x \mapsto \text{expit}(x)$  is monotonically increasing, then it follows that  $\beta_A = 0$ , which proves that  $\hat{\beta}_A$  converges to 0. This further implies that  $E[Q(t, (1, 1, \mathbf{L})^T, \boldsymbol{\beta})] = E[Q(t, (1, 0, \mathbf{L})^T, \boldsymbol{\beta})]$ , which proves that  $\hat{F}_1^g(t, 1) - \hat{F}_1^g(t, 0)$  converges to 0.

### 7.2.1 Proof of (29)

By construction, the stratified Kaplan-Meier estimator of  $P(C > s \mid A = a)$ , for  $s \in [0, t]$ , will converge to  $\tilde{G}_c(s, A) = \exp\left\{-\int_0^s \tilde{\lambda}_c(u, A)du\right\}$ , where

$$\begin{aligned} \tilde{\lambda}_c(u, A) &= \lim_{h \rightarrow 0} \frac{1}{h} P(C \in [u, u + h] \mid C > u, T > u, A) \\ &= \frac{\lim_{h \rightarrow 0} \frac{1}{h} P(C \in [u, u + h], T > u \mid A)}{P(C > u, T > u \mid A)} \end{aligned}$$

First, let's look at the numerator,

$$\begin{aligned} &\lim_{h \rightarrow 0} \frac{1}{h} P(C \in [u, u + h], T > u \mid A) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} E\left[\mathbb{1}\{C \in [u, u + h]\} \cdot \mathbb{1}\{T > u\} \mid A\right] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} E\left[E\left(\mathbb{1}\{C \in [u, u + h]\} \cdot \mathbb{1}\{T > u\} \mid A, \mathbf{L}\right) \mid A\right] \\ &= E\left[\lambda_c(u, A, \mathbf{L})G_c(u, A, \mathbf{L})S(u, \mathbf{L}) \mid A\right] \text{ as } C \perp\!\!\!\perp T \mid (A, \mathbf{L}), \end{aligned}$$

where  $G_c(u, A, \mathbf{L}) = P(C > u \mid A, \mathbf{L})$  and  $S(u, \mathbf{L}) = 1 - F(u, A, \mathbf{L}) - F_2(u, A, \mathbf{L})$ , which depends on  $\mathbf{L}$  but not on  $A$ , by the assumption of no conditional treatment effects. This further leads to

$$\begin{aligned} &\lim_{h \rightarrow 0} \frac{1}{h} P(C \in [u, u + h], T > u \mid A) \\ &= E\left[\{\lambda_0(t) + A\lambda_1(t) + \lambda_2(t, \mathbf{L})\}e^{-\int_0^u \{\lambda_0(s) + A\lambda_1(s) + \lambda_2(s, \mathbf{L})\}ds} S(u, \mathbf{L}) \mid A\right] \\ &= \tau(A, u)\left[\{\lambda_0(t) + A\lambda_1(t)\}\gamma(u) + \xi(u)\right] \end{aligned}$$

with

$$\begin{aligned} \tau(A, u) &= e^{-\int_0^u \{\lambda_0(s) + A\lambda_1(s)\} ds} \\ \gamma(u) &= E\left[ e^{-\int_0^u \lambda_2(s, \mathbf{L}) ds} S(u, \mathbf{L}) \right] \\ \xi(u) &= E\left[ e^{-\int_0^u \lambda_2(s, \mathbf{L}) ds} S(u, \mathbf{L}) \lambda_2(t, \mathbf{L}) \right] \end{aligned}$$

where the conditioning on  $A$  in the expectations vanishes because we assumed  $A \perp\!\!\!\perp L$  (randomization). Similarly, it follows  $P(C > u, T > u | A) = \tau(A, u)\gamma(u)$ . Hence,  $\tilde{\lambda}_c(u, A) = \tilde{\lambda}_0(u) + A\lambda_1(u)$  with  $\tilde{\lambda}_0(u) = \lambda_0(u) + \xi(u)/\gamma(u)$ .

### 7.2.2 Proof of (30)

Using result (29) and the above definition of  $\tilde{G}_c(t, A)$ , we now show (30). First, we note that result (29) implies that the ratio  $r_c(t, \mathbf{L}) = G_c(t, A, \mathbf{L})/\tilde{G}_c(t, A) = \exp[-\int_0^t \{\lambda_0(s) + \lambda_2(s, \mathbf{L}) - \tilde{\lambda}_0(s)\} ds]$  does not depend on  $A$ , but on  $\mathbf{L}$  and  $t$  (hence the notation  $r_c(t, \mathbf{L})$ ). Therefore, we have

$$\begin{aligned} E\left[\tilde{W}(t, A)D(t) \mid A = 0\right] &= E\left[E\left\{\frac{\mathbb{1}\{T \wedge t \leq C\}}{\tilde{G}_c(t \wedge T, A)} D(t) \mid T, \eta, \mathbf{L}, A = 0\right\} \mid A = 0\right] \\ &= E\left[\frac{E\left\{\mathbb{1}\{t \wedge T \leq C\} \mid T, \eta, \mathbf{L}, A = 0\right\}}{\tilde{G}_c(T \wedge t, 0)} D(t) \mid A = 0\right] \\ &= E\left[r_c(T \wedge t, \mathbf{L})D(t) \mid A = 0\right] \quad \text{as } C \perp\!\!\!\perp (T, \eta) \mid (A, \mathbf{L}) \\ &= E\left[E\left\{\int_0^t r_c(s, \mathbf{L}) dD(s) \mid \mathbf{L}, A = 0\right\} \mid A = 0\right] \\ &= E\left[\int_0^t r_c(s, \mathbf{L}) E\left\{dD(s) \mid \mathbf{L}, A = 0\right\} \mid A = 0\right] \\ &= E\left[\int_0^t r_c(s, \mathbf{L}) dF_1(t, 0, \mathbf{L}) \mid A = 0\right] \\ &= E\left[\int_0^t r_c(s, \mathbf{L}) dF_1(t, 1, \mathbf{L}) \mid A = 0\right] \\ &\quad \text{(no conditional treatment effect)} \\ &= E\left[\int_0^t r_c(s, \mathbf{L}) dF_1(t, 1, \mathbf{L}) \mid A = 1\right] \\ &\quad \text{(randomization, i.e., } A \perp\!\!\!\perp \mathbf{L}\text{)} \\ &= E\left[\tilde{W}(t, A)D(t) \mid A = 1\right] \end{aligned}$$

### 7.2.3 Proof of (31)

As in the Proof of (30), here again we use that (29) implies that  $r_c(t, \mathbf{L}) = G_c(t, A, \mathbf{L})/\tilde{G}_C(t, A)$  does not depend on  $A$ , but on  $\mathbf{L}$  and  $t$ . Therefore,

$$\begin{aligned} & E \left[ \tilde{W}(t, A) Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = a \right] \\ &= E \left[ E \left\{ \frac{\mathbb{1}\{T \wedge t \leq C\}}{\tilde{G}_c(t \wedge T, A)} Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \mid T, \mathbf{L}, A \right\} \mid A = a \right] \\ &= E \left[ \frac{E \left\{ \mathbb{1}\{t \wedge T \leq C\} \mid T, \mathbf{L}, A \right\}}{\tilde{G}_C(T \wedge t, A)} Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = a \right] \\ &= E \left[ r_c(T \wedge t, \mathbf{L}) Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \mid A = a \right] \quad \text{as } C \perp\!\!\!\perp T \mid (A, \mathbf{L}) \\ &= E \left[ r_c(T \wedge t, \mathbf{L}) Q(t, (1, a, \mathbf{L})^T, \boldsymbol{\beta}) \right], \end{aligned}$$

where the last equality follows from the randomization assumption  $A \perp\!\!\!\perp \mathbf{L}$  and the assumption of no conditional treatment effect, which further implies  $A \perp\!\!\!\perp T$ .

## References

- Aalen O, Borgan Ø, Gjessing HK, Gjessing S (2008) Survival and event history analysis: a process point of view. Springer, Berlin
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical Models Based on Counting Processes. Springer, New York
- Azarang L, Scheike T, de Uña-Álvarez J (2017) Direct modeling of regression effects for transition probabilities in the progressive illness-death model. *Stat Med* 36(12):1964–1976
- Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973
- Bang H, Tsiatis AA (2000) Estimating medical costs with censored data. *Biometrika* 87(2):329–343
- Bartlett JW (2018) Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical stat* 17(5):648–666
- Colantuoni E, Scharfstein DO, Wang C, Hashem MD, Leroux A, Needham DM, Girard TD (2018) Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*, 360
- Cortese G, Holmboe SA, Scheike TH (2017) Regression models for the restricted residual mean life for right-censored and left-truncated data. *Stat Med* 36(11):1803–1822
- DiRienzo A, Lagakos S (2001) Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *J Royal Stat Society: Series B (Statistical Methodology)* 63(4):745–757
- EMA (2015). Guideline on adjustment for baseline covariates in clinical trials. [https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials\\_en.pdf](https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf)
- FDA (2021). Adjusting for covariates in randomized clinical trials for drugs and biological products guidance for industry. <https://www.fda.gov/media/148910/download>
- Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. *J Amer Stat Assoc* 94(446):496–509
- Geskus RB (2016) *Data analysis with competing risks and intermediate states*. CRC Press Boca Raton
- Hernán M, Robins J (2020) Causal Inference: What If. Chapman & Hall/CRC, Boca Raton
- Holst KK, Scheike T (2021) *mets: Analysis of Multivariate Event Times*. R package version 1(2):9
- Holt A, Blanche P, Zareini B, Rajan D, El-Sheikh M, Schjerning A-M, Schou M, Torp-Pedersen C, McGeehan P, Gislason GH et al (2021) Effect of long-term beta-blocker treatment following myocardial



- infarction among stable, optimally treated patients without heart failure in the reperfusion era: a Danish, nationwide cohort study. *Eur Heart J* 42(9):907–914
- Kim JP (2013) A Note on Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models. *Biometrics* 69(1):282–289
- Lin DY, Wei L-J (1989) The robust inference for the Cox proportional hazards model. *J Amer stat Assoc* 84(408):1074–1078
- Loder E, Groves T, MacAuley D (2010) Registration of observational studies
- Lok JJ, Yang S, Sharkey B, Hughes MD (2018) Estimation of the cumulative incidence function under multiple dependent and independent censoring mechanisms. *Lifetime Data Anal* 24(2):201–223
- Lu X, Tsiatis AA (2008) Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* 95(3):679–694
- Luque-Fernandez MA, Schomaker M, Racht B, Schnitzer ME (2018) Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Stat Med* 37(16):2530–2546
- Malani HM (1995) A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 82(3):515–526
- Martens MJ, Logan BR (2020) Group sequential tests for treatment effect on survival and cumulative incidence at a fixed time point. *Lifetime Data Anal* 26(3):603–623
- Martinussen T, Vansteelandt S, Andersen PK (2020) Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal* 26(4):833–855
- McCullagh P, Nelder JA (1989) *Generalized Linear Models*. Chapman and Hall, London, 2nd edition
- Moore KL, van der Laan MJ (2009) Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med* 28(1):39–64
- Ozenne B, Sørensen AL, Scheike T, Torp-Pedersen C, Gerds TA (2017) riskRegression: predicting the risk of an event using Cox regression models. *The R J* 9(2):440–460
- Ozenne BMH, Scheike TH, Stærk L, Gerds TA (2020) On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *Biom J* 62(3):751–763
- Pfeiffer RM, Gail MH (2017) *Absolute Risk: Methods and Applications in Clinical Management and Public Health*. CRC Press
- Robins JM, Rotnitzky A (1992) In: Recovery of information and adjustment of dependent censoring using surrogate markers. *AIDS Epidemiology-Methodological Issues*. Birkhäuser, Boston, pp 24–33
- Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, pp. 227–240
- Rosenblum M, Steingrimsson JA (2016) Matching the Efficiency Gains of the Logistic Regression Estimator While Avoiding its Interpretability Problems, in *Randomized Trials*. Technical report
- Rossello X, Pocock SJ, Julian DG (2015) Long-term use of cardiovascular drugs: challenges for research and for patient care. *J Amer Coll Cardiology* 66(11):1273–1285
- Rotnitzky A, Farall A, Bergesio A, Scharfstein D (2007) Analysis of failure time data under competing censoring mechanisms. *J Royal Stat Soc: Series B (Stat Methodol)* 69(3):307–327
- Rufibach K (2019) Treatment effect quantification for time-to-event endpoints-Estimands, analysis strategies, and beyond. *Pharmaceutical stat* 18(2):145–165
- Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120 (with Rejoinder, 1135–1146)
- Scheike T, Zhang M, Gerds T (2008) Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 95:205–220
- Schumacher M, Ohneberg K, Beyersmann J (2016) Competing risk bias was common in a prominent medical journal. *Journal of Clinical Epidemiology* 80:135–136
- Stefanski LA, Boos DD (2002) The calculus of M-estimation. *The Amer Stat* 56(1):29–38
- Stensrud MJ, Hernán MA (2020) Why test for proportional hazards? *Jama* 323(14):1401–1402
- Struthers CA, Kalbfleisch JD (1986) Misspecified Proportional Hazard Models. *Biometrika* 73:363–369
- Sutradhar R, Austin PC (2018) Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Ann epidemiology* 28(1):54–57
- Tsiatis A (2006) *Semiparametric theory and missing data*. Springer Science & Business Media
- Uno H, Cai T, Tian L, Wei L (2007) Evaluating prediction rules for t-year survivors with censored regression models. *J Amer Stat Assoc* 102(478):527–537
- Van der Laan MJ, Rose S (2011) *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media

- Van Der Laan MJ, Rubin D (2006) Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1)
- Vansteelandt S, Martinussen T, Tchetgen ET (2014) On adjustment for auxiliary covariates in additive hazard models for the analysis of randomized experiments. *Biometrika* 101(1):237–244
- Wang B, Susukida R, Mojtabai R, Amin-Esmaeili M, Rosenblum M (2021) Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, pp. 1–12
- Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA (2020) A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med* 39(8):1199–1236
- Zhang M, Tsiatis AA, Davidian M (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64(3):707–715
- Zhang X, Zhang M-J (2011) SAS macros for estimation of direct adjusted cumulative incidence curves under proportional subdistribution hazards models. *Comput methods programs in biomedicine* 101(1):87–93
- Zheng Y, Cai T, Feng Z (2006) Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* 62(1):279–287

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.