The Survivor Average Causal Effect for Outcomes Truncated by Death in RCTs

Master Thesis in Biostatistics (STA495)

by

Chiara Vanetta

10-741-122

supervised by

Prof. Dr. Leonhard Held

Dr. Stefanie von Felten

University of Zurich^UZH

Zurich, January 2020

Contents

Preface

Abstract

Researchers are increasingly evaluating non-mortality outcomes such as cognition, physical function and quality of life in RCTs of critically ill or older patients (Colantuoni *et al.*, 2018). Outcomes truncated by death are particularly common in such studies, and present a statistical challenge since they are not missing in the usual sense. In their presence, a complete case analysis of the data could undermine the benefit of randomization and produce misleading results, while an imputation of the data would create values that do not exist and is thus inappropriate. A useful approach to handle them, yet not so well-known among clinical researchers and epidemiologists, is principal stratification (Frangakis and Rubin, 2002) and the concept of survivor average causal effect (SACE, Rubin, 2006).

We implemented in `R` the SACE estimator proposed by Hayden *et al.* (2005) and we used it to reanalyze a completed RCT on the effect of early prophylactic high-dose recombinant human erythropoietin in very preterm infants on neurodevelopment at 2 years of age (Epo trial, Natalucci *et al.*, 2016). The results obtained by SACE approach confirmed those reported in the original publication, which were obtained by complete case analysis and by single imputation of the worst observed outcome.

In addition, we conducted a simulation study to compare Hayden's method with complete case analysis and multiple imputation analysis under different scenarios. We evaluated the performance of the three methods with respect to their three targeted estimands. In scenarios where survival was not affected by treatment, the three methods yielded similar results and the estimates were unbiased with respect to all estimands. In these circumstances, complete case analysis may be used to estimate the SACE. However, in scenarios where survival was affected by treatment, the estimates derived by complete case analysis were biased with respect to the estimands targeted by Hayden's method and multiple imputation analysis, and/or vice versa. Although the results gained by multiple imputation were similar to those obtained by SACE approach, multiple imputation should not be used to analyze studies with outcomes truncated by death, unless inference about an hypothetical population without deaths is desired.

Our findings highlighted the importance of aligning the choice of the statistical method to use with the study research question, the targeted estimand and the expected scenario. In particular, possible post-randomization events such as death, together with strategies to address them, should be explicitly defined at the planning stage of the study and determine choices about study design, data collection and statistical analysis. (European Medicines Agency, 2017; US Food and Drug Administration, 2017).

Acknowledgments

Writing this thesis has been an exciting and rewarding experience, for which I desire to thank several people. First, I would like to thank my supervisors: Prof. Dr. Leonhard Held, for providing a topic aligned with my interests, for his inspiring ideas and his guidance, and Dr. Stefanie von Felten, for conceiving this interesting project, for her valuable insights and her availability. I have learned a lot from them, and I am also grateful for the opportunity they gave me to participate to the XXXIst ROeS conference of the International Biometric Society. Secondly, I would like to thank my fellow students for making the study environment so enjoyable: Xijin Chen, Uriah Daugaard, Seraphina Kissling, Lucas Kook, Mei-Yee Ng, Katrin Petermann, and especially Natalia Popova, who shared the whole path with me, giving me support and a constant exchange of ideas. Thanks also to all the people involved in the Master Program in Biostatistics and in particular to Dr. Eva Furrer, for providing a great study program and for seeking the best for students. Last but not least, I would like to thank my mom, my sister, my grandparents and Olaf, for their continuous support and for always believing in me.

Chiara Vanetta
January 2020

Chapter 1

Introduction

Missing outcome measurements occur in most RCTs. In their presence, a complete case analysis of the data may undermine the benefit of randomization and produce misleading results. Multiple imputation is usually the preferred approach to handle them (e.g. Vickers and Altman, 2013), and with the availability of software to generate imputations and to pool estimates of effect sizes, this method has become increasingly popular.

Particularly challenging are outcomes truncated by death, which occur when some subjects die after randomization, before their outcome of interest can be measured. These are common, for example, in therapeutic trials of advanced-stage or rapidly fatal diseases and in studies to compare non-mortality outcomes in older individuals, but similar truncations are also present in other fields, such as economics (Rubin, 2006). Outcomes truncated by death are not missing in the usual sense, i.e. outcomes not measured which could have been measured, since they do not exist and could never be observed. Imputation of such data is not appropriate because it would generate data that are not defined. Moreover, public-health decision makers may be more interested in knowing the effective impact of a treatment in the real population, rather than in a population that exists only statistically (Chaix *et al.*, 2012).

Kurland *et al.* (2009) present an overview of the statistical models and estimands that have been proposed to analyze longitudinal data with follow-up truncated by death. The issue seems well recognized in the field of longitudinal data analysis, but less recognized and less often applied in analyses of cross-sectional outcomes in RCTs.

One approach that can deal with outcomes truncated by death is principal stratification (Frangakis and Rubin, 2002), and the concept of survivor average causal effect (SACE), introduced by Rubin (2006). The SACE is defined as the average causal effect in the subgroup of patients that would have survived under both treatment assignments. Because this subgroup is defined at the baseline and is not affected by post-treatment events such as death, the benefit of randomization is preserved and no non-existent data are created.

Unfortunately, the SACE is not identifiable without further assumptions (Zhang and Rubin, 2003). Several approaches have been proposed to enable the SACE identification and estimation. We focus on the specific estimator proposed by Hayden *et al.* (2005), which exploits the baseline covariates and makes the so-called *explainable nonrandom survival* assumption.

In particular, the purpose of our work is threefold. In the first place, to implement Hayden's SACE estimator in `R` (R Core Team, 2019). Secondly, to apply it to data from a completed placebo-controlled, double-blind RCT on the effect of early prophylactic high-dose recombinant human erythropoietin (rhEPO) in very preterm infants on neurodevelopment at 2 years of age (Epo trial, Natalucci *et al.*, 2016), and to compare the SACE estimate with the estimates reported in the original publication. Lastly, to conduct a simulation study to compare Hayden's method with complete case analysis and analysis using multiple imputation under different scenarios. Although multiple imputation is inappropriate to analyze outcomes truncated by death, we aim to assess how results gained with this method differ from those of the other methods,

and whether the differences are large or actually quite small. In order to have clear objectives and a well-designed study, we produce a protocol prior to the simulation study. Since the compared methods estimate different quantities, we discuss the results of the simulation study with reference to the topic of estimands.
The motivation for the present work came from a currently ongoing, similar RCT (EpoRepair trial, Rüegger *et al.*, 2015) to evaluate the effect of high-dose rhEPO on long-term neurocognitive outcomes of very preterm infants suffering from intraventricular hemorrhage. The primary outcome is IQ at 5 years of age. Since mortality up to term equivalent age of this vulnerable population is around 15%, a relevant proportion of outcomes truncated by death is expected at 5 years of age. Moreover, there is currently limited awareness of the fact that outcomes truncated by death are not missing data in the usual sense. With this work we hope to promote awareness of the problem and methodological knowledge of how it could be dealt with.
The thesis is organized as follows. In the following sections, the approach of principal stratification and the concept of survivor average causal effect are introduced, followed by an overview of the different SACE estimators that have been proposed. Chapter 2 begins with a discussion about the characteristics of the statistical methods considered in our work. Then, Hayden's method is presented, and the implementations in R of Hayden's method and of multiple imputation are described. Finally, the methods used for the analysis of the Epo trial are reported, and the protocol of the simulation study is presented. In Chapter 3, the results of the Epo trial analysis and the results of the simulation study are shown. The thesis ends with a discussion of the results (Chapter 4), which comprises the interpretation, the limitations and the generalizability of the findings.

1.1 Principal stratification framework

Principal stratification (Frangakis and Rubin, 2002) is a general framework to adjust for post-randomization variables when comparing treatments. The key idea is to stratify patients with respect to the joint potential values of the post-randomization variable under each of the compared treatments, and then to compute causal effects only within the obtained strata.
In studies with outcomes truncated by death, the post-randomization variable is survival. To illustrate the application of principal stratification to these studies, we consider the randomized experiment described in Rubin (2006), which compares an active treatment, which we denote as treatment 1, with the control treatment, which we denote as treatment 0. The primary outcome is quality of life 2 years post-randomization and is denoted by Y. Many patients do not reach the 2 years post-randomization endpoint, their outcome Y is thus truncated by death.
In this case, we stratify patients with respect to 2 years survival, but not on the *observed* survival, which is generally affected by the treatment received, rather on the *bivariate* survival: survival if assigned to the control treatment and survival if assigned to the active treatment. Because the stratification with respect to the bivariate survival is not affected by the treatment received, even though which of the two outcomes is actually observed is affected by the treatment received, the randomization is preserved.
The four resulting strata are called principal strata and are defined as follows:

- Individuals who would live under either treatment assignment, LL
- Individuals who would die under control but live under the active treatment, DL
- Individuals who would die under either treatment assignment, DD
- Individuals who would live under control but die under the active treatment, LD

The strata represent different types of people. The LL subjects may be considered as the most robust, the DL subjects to be of typical health status, the DD subjects as frail and the LD

subjects as acutely ill patients who may be more susceptible to previously unrecognized adverse effects of the treatment (Colantuoni *et al.*, 2018), or as patients that feel so much better under treatment (even if it does not affect their disease progression) that they "overdo" it (Rubin, 2006).

A causal effect must be a comparison of treatment potential outcomes and control potential outcomes on a common subset of units (Rubin, 2006). We first define it at the individual level, as the difference between the potential outcome under assignment to the active treatment, $Y(1)$, and the potential outcome under assignment to the control treatment, $Y(0)$. The value $Y(1)-Y(0)$ is well-defined only for the subjects where both $Y(1)$ and $Y(0)$ are defined, i.e. for the subject that would survive up to 2 years post-randomization under both treatment assignments. Since a well-defined value for the causal effect exists only for the subjects of the LL group, the computation of causal effects is restricted to this group. The average causal effect in this stratum is obtained averaging the individual causal effects and is thus called the survivor average causal effect.

The specific artificial case provided in Rubin (2006) is displayed in Table 1.1. For the LL people, the average outcome $\bar{Y}$ would be 900 if all were treated and 700 if no one was treated. Therefore, the survivor average causal effect of the treatment is $900 - 700 = 200$.

Table 1.1: Principal strata defined by potential outcomes, and average causal effect. Asterisks represent undefined values.

Principal	Control treatment		Active treatment		Average causal
stratum	S	$\bar{Y}$	S	$\bar{Y}$	effect on Y
LL	1	700	1	900	200
DL	0	*	1	600	*
DD	0	*	0	*	*
LD	1	800	0	*	*

However, we do not get to observe all the potential outcomes and thus all the values in Table 1.1. For patients who survived, we can only observe the outcome under the treatment they were assigned to. Thus, we do not know the principal stratum to which each patient actually belongs: if we observe that one patient did not survive, we know that he cannot be in the LL stratum, but if we observe that he survived, we do not know whether he is in the LL stratum because we do not know if he would have survived under the treatment he was not assigned to.

Therefore, we can not simply compute the difference between the average outcomes of the subjects in the LL stratum. Some modelling assumptions are required to identify the patients in the LL stratum and to estimate the SACE.

1.2 SACE estimators

In order to have an overview of the studies that have been conducted after the publication of Hayden's method, we collected the papers citing Hayden *et al.* (2005) from the Web of Science and PubMed. We classified each of them as *methodological*, *application* or *study protocol*, depending on whether the article discussed statistical methods, applied Hayden's estimator or another SACE estimator to analyze a study, or was a study protocol (Table A.1), respectively. Then, we mainly examined the methodological studies and we used them to find out about the different SACE estimators developed.

As observed by Jemiai *et al.* (2007), two approaches have been proposed to identify the SACE. The first approach calculates bounds on the SACE (Zhang and Rubin, 2003; Chiba, 2012; Yang and Small, 2016). The second approach makes assumptions that allow to identify and estimate the SACE, and/or conducts sensitivity analyses around assumptions about the distribution of the outcome or survival (Hayden *et al.*, 2005; Egleston *et al.*, 2006; Shepherd *et al.*, 2006; Jemiai *et al.*, 2007; Chiba and VanderWeele, 2011; Wang *et al.*, 2017b). For example, Yang and Small (2016) propose a set of assumptions that make use of survival information before and after the measurement of the outcome to narrow the bounds on the SACE, and develop a programming approach to obtain the closed form for the bounds under these assumptions. Wang *et al.* (2017b) use a substitution variable in place of the latent membership to the LL group to enable the SACE identification. They need some identification conditions for the substitution variable, which are conceptually similar to conditions for a conditional instrumental variable. Chiba and VanderWeele (2011) propose a method that is particularly simple to implement in practice and does not require special statistical programming. They express the SACE as the difference between the crude comparison of the outcomes among survivors and a sensitivity parameter, which has to be set by the investigator according to what is thought plausible, and can be varied to examine how conclusions vary under different values of the parameter. A limitation of this method is that it gives a range of corrected values and not a single point estimate (Merchant *et al.*, 2018).

As previously discussed, the SACE is not identifiable without further assumptions. Common assumptions are the *stable unit treatment value* assumption, the *ranked average scores* assumption and the *monotonicity* assumption. Each of these may be more plausible in some contexts and less in others. The stable unit treatment value assumption, developed by Rubin (Rubin, 1980, 1986), states that there are no different forms or versions of the same treatment, and that the outcome of an individual is unaffected by the treatment assignment to the other individuals. The ranked average scores assumption is made e.g. by Zhang and Rubin (2003), Chiba (2012) and Yang and Small (2016), and states that the LL patients have on average a better outcome than the DL patients. The monotonicity assumption is made by many authors (Zhang and Rubin, 2003; Egleston *et al.*, 2006; Shepherd *et al.*, 2006; Jemiai *et al.*, 2007; Chiba and VanderWeele, 2011; Chiba, 2012; Yang and Small, 2016) and states that survival under the active treatment is at least as good as survival under the control treatment, or equivalently, that no LD patients are present. As pointed out by Wang *et al.* (2017b), this assumption may be plausible in some observational studies, for example in studies evaluating the effect of smoking on memory decline in an aged population, where non-smoking is considered as the treatment, since it is commonly believed that smoking is always bad for overall health and hence overall survival. However, it is questionable in RCTs with acute diseases, because if researchers believe that one treatment benefits survival a priori, a clinical trial would be unethical. Moreover, the monotonicity assumption may not be appropriate if, instead of a treatment and placebo, two experimental treatments are compared (Jemiai *et al.*, 2007). Hayden *et al.* (2005) do neither impose monotonicity nor ranked average scores, but assumes stable unit treatment value and explainable nonrandom survival, which is described in Section 2.2.

Chapter 2

Methods

2.1 Characteristics of the statistical methods considered

Complete case analysis, multiple imputation analysis and single imputation analysis are commonly used approaches to analyze RCTs with missing outcome measurements. In the following, we discuss their characteristics, in general and with respect to their application to studies with outcomes truncated by death.

Complete case analysis is the simplest method. In general, complete case analysis does not conform with the intention-to-treat (ITT) principle, since only a subset of the randomized patients is analyzed. The restriction to a subset of the randomized patients also decreases the precision of the treatment effect estimate and the power of the study. Moreover, it is based on a post-randomization event, the observation of the outcome, and thus undermines the benefit of the randomization. If the missingness of the outcome is directly or indirectly related to the baseline covariates as well as to the treatment group, a baseline imbalance among individuals with observed outcomes is created, which may result in an incorrect estimate of treatment effect (Groenwold *et al.*, 2014). The dependence of the outcome missingness on the baseline covariates is usually the case when the post-randomization event is survival, as in the case of studies with outcomes truncated by death. Thus, if outcomes truncated by death are dependent on treatment, the complete case analysis results are biased. For example, if the treatment benefits survival of less healthy individuals, then patients who survive in the placebo group may be healthier than the patients who survive in the treatment group (Colantuoni *et al.*, 2018) and the treatment effect may be underestimated. However, if survival is not affected by the treatment, then the randomization is preserved and complete case analysis estimates the same quantity as the SACE approach, as no DL and LD patients (only LL and DD) are present.

Multiple imputation is an ideal method to deal with missing data, especially when their amount is extensive, since all the randomized patients can be analyzed, the randomization is preserved and the analysis conforms with the ITT principle. The uncertainty associated with the missing data can be taken into account using Rubin's rules (Rubin, 1987) to combine the estimates from the imputed data sets. However, when applied to analyze outcomes truncated by death, creates data that do not exist. These are data that could not be observed and are not defined, which is what makes this method inappropriate in our context.

Single imputation analysis is frequently used to deal with missing data, and imputation of the "best" or "worst" outcome is often used as sensitivity analysis in RCTs (Sterne *et al.*, 2009). Like multiple imputation, it preserves the initial randomization and conforms with the ITT principle, but, differently from multiple imputation, it fails to account for the uncertainty about the missing values and thus the estimated standard errors are often too small.

The SACE approach prevents the potential distortion of results of complete case analysis and is particularly appropriate in case of truncation by death. In fact, in this case, it may be more clinically relevant than the previously discussed methods as it compares patients of the same type (LL vs. LL, instead of $LL + DL$ vs. $LL + DL$ as complete case analysis) and does not assume that the dead patients could still be alive (as multiple imputation and single imputation do). Moreover, it estimates a causal effect, since it compares potential outcomes on a common subset of units. The subgroup of patients analyzed is defined at the baseline and it is not affected by post-randomization events, therefore the SACE approach preserves the initial randomization. However, since a subset of the randomized patients is analyzed, it does not conform with the ITT principle and loses some precision and power compared to the other methods. Also, the SACE estimate is not identifiable from the data alone and requires strong assumptions that are not testable on the data. On the other hand, some investigators have argued that the use of relatively strong assumptions to identify the principal strata are justifiable (Kurland *et al.*, 2009). The characteristics of the discussed statistical methods, when used to analyze studies with outcomes truncated by death, are summarized in Table 2.1.

Table 2.1: Summary of the characteristics of some methods when used to analyze RCTs with outcomes truncated by death.

Complete case analysis	Multiple imputation analysis	Single imputation analysis	SACE approach
+ Simple - Randomization may not be preserved - Does not conform with ITT principle - Less precision and power	+ Randomization preserved + Conforms with ITT principle - Creates data that could not be observed	+ Randomization preserved + Conforms with ITT principle - Creates data that could not be observed (without accounting for their uncertainty)	+ Randomization preserved + Causal effect - Does not conform with ITT principle - Less precision and power - Non-testable assumptions

2.2 Hayden's method

In the following, the derivation of the SACE estimator made by Hayden *et al.* (2005) is shown. Before going into the details of the method, we briefly introduce the setting considered and some notation.

Let us assume that we want to compare the effect of two treatments, denoted by $z = 0, 1$, on an outcome of interest, denoted by Y, which is assessed at a later time. Suppose that patients are randomly assigned to receive one of the treatments. For patients on treatment z:

- $S(z)$ is the indicator of survival up to follow-up assessment under treatment z;
- $Y(z)$ is the outcome of interest under treatment z assessed at follow-up, defined and observable only if $S(z) = 1$;
- $S(1 - z)$ is the indicator of survival up to follow-up assessment under treatment $1 - z$, not observable;
- $Y(1 - z)$ is the outcome of interest under treatment $1 - z$ assessed at follow-up, defined only if $S(1 - z) = 1$, not observable.

The SACE estimand is the outcome difference in the patients who would have survived under both treatments. Using the notation above, it can be expressed as

$$\begin{aligned}\mu &= \frac{\mathrm{E}\{[Y(1)-Y(0)]S(0)S(1)\}}{\mathrm{E}\{S(0)S(1)\}} \\ &= \frac{\mathrm{E}\{Y(1)S(1)S(0)\}}{\mathrm{E}\{S(1)S(0)\}} - \frac{\mathrm{E}\{Y(0)S(0)S(1)\}}{\mathrm{E}\{S(0)S(1)\}},\end{aligned} \tag{2.1}$$

where $S(0)S(1) = 1$ if and only if a patient would have survived under both treatments, and is 0 otherwise. The quantity $Y(z)S(z)$ is always defined: its value is equal to the value of $Y(z)$ if $S(z) = 1$ and is equal to 0 if $S(z) = 0$.

However, since each patient receives only one treatment, we only know if the patient survived under that treatment and we do not know whether he would have survived under the other treatment. Therefore, we can not observe the joint distributions in the numerator and denominator of (2.1) and we can not directly identify μ from the data. Hayden *et al.* (2005) make use of the baseline covariates, denoted by X, and make the following independence assumptions, which are referred as explainable nonrandom survival:

A1) $S(z) \perp\!\!\!\perp S(1-z) \mid X$

A2) $S(z) \perp\!\!\!\perp Y(1-z) \mid X, \{S(1-z) = 1\}$

The first assumption states that, conditional on the baseline covariates, the survival status of subjects under treatment z is independent of their survival status under treatment $1-z$. The second assumption states that, conditional on surviving when assigned to treatment $1-z$, and on the baseline covariates, the survival status of subjects under treatment z is independent of their outcome under treatment $1-z$. These assumptions essentially mean that no unmeasured confounders are present and, unfortunately, are not testable on the data. Thus, in order to make them more plausible, one should collect and use baseline covariates which are strongly predictive for survival, and ideally incorporate an analysis of sensitivity of results to departures from A1) and A2).

Let $p(z) = \mathrm{E}\{S(z) \mid X\}$. Then, if A1) and A2) hold, we have

$$\begin{aligned}\mathrm{E}\{Y(z)S(z)S(1-z) \mid X\} &= \mathrm{E}\{Y(z)S(z) \mid X\}\, \mathrm{E}\{S(1-z) \mid X\} \\ &= \mathrm{E}\{Y(z)S(z) \mid X\}\, p(1-z) \\ &= \mathrm{E}\{Y(z)S(z)\, p(1-z) \mid X\}.\end{aligned} \tag{2.2}$$

At this point, the idea to estimate the SACE is simple. We want to compute the outcome mean in those patients assigned to treatment 1 who survived and who we think are quite likely to have survived had they been assigned to treatment 0, and the outcome mean in those patients assigned to treatment 0 who survived and who we think are quite likely to have survived had they been assigned to treatment 1.

To estimate the survival probability of each subject under the treatment not assigned, we fit two logistic regression models for survival depending on the baseline covariates: one to the patients assigned to treatment 1 and the other to the patients assigned to treatment 0. The model fitted to patients assigned to treatment 1 is used to estimate the survival probability under treatment 1 of patients assigned to treatment 0. Similarly, the model fitted to patients assigned to treatment 0 is used to estimate the survival probability under treatment 0 of patients assigned to treatment 1.

Finally, the outcome mean under treatment 1 is computed in patients who survived under treatment 1, weighting their outcome by their survival probability under treatment 0. Similarly, the outcome mean under treatment 0 is computed in patients who survived under treatment 0, weighting their outcome by their survival probability under treatment 1.

Formally, from (2.2), conditional on X, and with a consistent estimator $\hat{p}(z)$ of $p(z)$ for $z = 0, 1$, a SACE estimator is given by the difference in the following weighted means:

$$\hat{\mu} = \frac{\sum\limits_i Y_i(1)S_i(1)\hat{p}_i(0)}{\sum\limits_i S_i(1)\hat{p}_i(0)} - \frac{\sum\limits_j Y_j(0)S_j(0)\hat{p}_j(1)}{\sum\limits_j S_j(0)\hat{p}_j(1)}, \tag{2.3}$$

where i indexes over patients assigned to arm $z = 1$ and j indexes over patients assigned to arm $z = 0$.

The variance of (2.3) is calculated using an asymptotic approximation to $\hat{p}_i(z)$ and by application of the Delta method to the variance-covariance matrix of $(\mu_{1n}, \mu_{1d}, \mu_{0n}, \mu_{0d})$, where $\mu_{zn} = \mathrm{E}\{Y(z)S(z)S(1-z)\}$, $\mu_{zd} = \mathrm{E}\{S(z)S(1-z)\}$, and thus $\mu = \frac{\mu_{1n}}{\mu_{1d}} - \frac{\mu_{0n}}{\mu_{0d}}$.

In particular, the following expressions for $(\mu_{1n}, \mu_{1d}, \mu_{0n}, \mu_{0d})$ are given:

$$\mu_{1n} = \sum_{i=1}^{n_1} S_i(1)Y_i(1)p_i(0) + \sum_{i=1}^{n_0} \left\{ \sum_{k=1}^{n_1} S_k(1)Y_k(1)p_k(0)(1-p_k(0))\mathbf{X}_k^T \right\} \times \mathbf{I}_1^{-1}\mathbf{X}_i(S_i(0) - p_i(0))$$

$$\mu_{1d} = \sum_{i=1}^{n_1} S_i(1)p_i(0) + \sum_{i=1}^{n_0} \left\{ \sum_{k=1}^{n_1} S_k(1)p_k(0)(1-p_k(0))\mathbf{X}_k^T \right\} \times \mathbf{I}_1^{-1}\mathbf{X}_i(S_i(0) - p_i(0))$$

$$\mu_{0n} = \sum_{i=1}^{n_1} \left\{ \sum_{k=1}^{n_0} S_k(0)Y_k(0)p_k(1)(1-p_k(1))\mathbf{X}_k^T \right\} \times \mathbf{I}_0^{-1}\mathbf{X}_i(S_i(1) - p_i(1)) + \sum_{i=1}^{n_0} S_i(0)Y_i(0)p_i(1)$$

$$\mu_{0d} = \sum_{i=1}^{n_1} \left\{ \sum_{k=1}^{n_0} S_k(0)p_k(1)(1-p_k(1))\mathbf{X}_k^T \right\} \times \mathbf{I}_0^{-1}\mathbf{X}_i(S_i(1) - p_i(1)) + \sum_{i=1}^{n_0} S_i(0)p_i(1),$$

where $\mathbf{X}_k^T$ is the k-th row of the design matrix $\mathbf{X}^T$, and $\mathbf{I}_z^{-1}$ is the variance-covariance matrix of the coefficients of the logistic model fit for treatment z. Each term is a sum of independent and identically distributed random variables, the first being a summation over patients in treatment group 1, the second over patients in treatment group 0.

$p_i(z)$ is replaced by $\hat{p}_i(z)$ for $z = 0, 1$ and consequently expressions for ($\hat{\mu}_{1n}$, $\hat{\mu}_{1d}$, $\hat{\mu}_{0n}$, $\hat{\mu}_{0d}$) are obtained. Then the $n_1 \times 4$ matrix $\mathbf{C}_1$ is constructed, with the first summands in each of these expressions as columns. The $n_0 \times 4$ matrix $\mathbf{C}_0$ is constructed with the second summands of the expressions as columns. Let $\mathbf{E}_z$ be the n_z-column vector of ones and $\mathbf{1}_z$ the $n_z \times n_z$ identity matrix for $z = 0, 1$. Then the variance-covariance matrix of $(\mu_{1n}, \mu_{1d}, \mu_{0n}, \mu_{0d})$ can be estimated by

$$\mathbf{\Sigma} = \mathbf{C}_1^T(\mathbf{1}_1 - \mathbf{E}_1\mathbf{E}_1^T/n_1)\mathbf{C}_1 + \mathbf{C}_0^T(\mathbf{1}_0 - \mathbf{E}_0\mathbf{E}_0^T/n_0)\mathbf{C}_0. \tag{2.4}$$

The variance of (2.3) is then calculated from $\mathbf{\Sigma}$ and from the Jacobian matrix of $h(\mu_{1n}, \mu_{1d}, \mu_{0n}, \mu_{0d}) = \frac{\mu_{1n}}{\mu_{1d}} - \frac{\mu_{0n}}{\mu_{0d}}$ evaluated at ($\hat{\mu}_{1n}$, $\hat{\mu}_{1d}$, $\hat{\mu}_{0n}$, $\hat{\mu}_{0d}$), using the multivariate Delta method (Held and Sabanés Bové, 2014).

Alternatively, the variance of (2.3) can be calculated by bootstrap.

2.3 Implementation of Hayden's method in R

In this section, we describe the implementation of Hayden's method in the R function saceEstimator(). For completeness, we start by decribing the functions resWithCI() and flac() that are used in saceEstimator(). The corresponding R code can be found in Appendix A.2, while software information is available in Appendix A.1.

The function resWithCI() simply computes the desired confidence interval from the given estimate and standard error, and returns it with the other results as a list. The confidence interval is computed assuming normality of the estimate unless some degrees of freedom are specified. In that case, the quantile of the t-distribution is used. The significance level is assumed to be 0.05 unless otherwise specified. This function is used at the end of saceEstimator(), after that the SACE estimate, its standard error and the effective sample size analyzed are obtained.
The function flac() is provided in the supplementary material of Puhr *et al.* (2017) and computes Firth's logistic regression with added covariate (FLAC) from a given outcome vector and a matrix of covariate values. We added the last three lines of code to compute and extract the variance-covariance matrix for the estimation of the SACE variance. flac is needed in case of failure of the logistic regression models in saceEstimator() due to strata with few or no events. In fact, Firth's penalization (Firth, 1993) reduces the bias from the small-sample estimates in logistic regression and prevents infinite coefficients to occur (Heinze and Schemper, 2002). The "added covariate" correction removes the bias from the prediction probabilities created by Firth's logistic regression, introducing some bias in the effect estimates, which are however compensated by a decrease in the mean square error (Puhr *et al.*, 2017). The package logistf (Heinze and Ploner, 2018) is required to fit the logistic regression model with Firth's correction.
The saceEstimator() function takes as argument a data set dat, the name Z of the treatment variable in dat, the names X of the baselines covariates in dat, the name Y of the outcome variable in dat, the name S of the survival variable in dat, the significance level alpha (by default set to 0.05), the degrees of freedom (by default set to NA) and the logical variable flac_corr (by default set to FALSE), which allows to choose whether to use the FLAC correction or not. The data types and the missing values are first checked. In particular, the columns of dat named Z, Y, and S must be vectors. The columns named Z, X and S cannot have missing values and the column named Y can have a missing value only when the respective value in the column named S is 0. Then, the computation of the SACE estimate and of the SACE variance as proposed by Hayden *et al.* (2005) begins. dat is divided into two data sets: i, containing the patients assigned to treatment arm 1, and j, containing the patients assigned to treatment arm 0. Depending on whether flac_corr is set to TRUE or FALSE, a logistic regression model or a FLAC model is fitted for survival separately to i and j. The model fitted to i is then used to predict the survival of the patients in j (who actually received treatment 0) under treatment 1. Similarly, the model fitted to j is used to predict the survival of the patients in i (who actually received treatment 1) under treatment 0. The obtained survival probabilities pj.z1 and pi.z0 are used for the calculation of the SACE estimate mu_sace=mu1-mu0. The sum of the denominators of mu1 and mu0 is considered the effective sample size analyzed. For the computation of the SACE variance, the variance-covariance matrices of i and j are extracted from the logistic regression models or from the FLAC models, depending on the value assigned to flac_corr. The variance-covariance matrices are needed for the computation of the matrices C0 and C1, together with the design matrices of i and j, and the fitted values pi.z1 and pj.z0, as illustrated in Section 2.2. Once C0 and C1 are obtained, the variance-covariance matrix is easily estimated Sigma, as shown in (2.4). Finally, we compute the Jacobian matrix of the function h(x)=x[1]/x[2]-x[3]/x[4] at (mu1n, mu1d, mu0n, mu0d). This is done making use of the package numDeriv. The SACE variance is then obtained applying the multivariate Delta method.

2.4 Implementation of multiple imputation in R

To perform the multiple imputation analysis, we wrote the R function `multipleImputation()`, which can be found in Appendix A.2.4.

The implemented function takes as arguments a data set `dat`, the name `Y` of the outcome variable in `dat`, the name `Z` of the treatment variable in `dat`, the number of imputations `m` (by default set to 5), the name(s) `not_predictor` of the variable(s) to remove from the predictor matrix (by default set to `NA`), the confidence level `alpha` (by default set to 0.05) and the degrees of freedom `df` (by default set to `NA`). We used the package `mice` (Van Buuren and Groothuis-Oudshoorn, 2011) to generate multiple imputations, to analyze the imputed data and to pool the analysis results. For the variable imputation, we used the `pmm` method, which implements predictive mean matching (Little, 1988) and is the default method for numeric variables. The results are returned making use of the function `resWithCI()` described in the previous section.

2.5 Methods for the Epo trial

In this section, we present the setting and the methods for the Epo trial. These are based on the original publication (Natalucci *et al.*, 2016) and on the corresponding study protocol.

Figure 2.1: Infant born four months too early, at 23 weeks of gestation.

The Epo trial was a randomized, double-blind, placebo-controlled, multi-center trial on the effect of early prophylactic high-dose rhEPO in very preterm infants (Figure 2.1) on neurodevelopment at 2 years of age.
Preterm infants born between 26 weeks 0 days and 31 weeks 6 days gestation were eligible for enrollment within the first 3 hours after birth, when informed parental consent was obtained. Exclusion criteria were a genetically defined syndrome, a severe congenital malformation adversely affecting life expectancy or neurodevelopment, severe intraventricular hemorrhage before randomization, and a priori palliative care.

Patients were recruited from three university hospitals (Basel, Geneva and Zurich) and two district hospitals (Aarau and Chur) between 2005 and 2012, and the neurodevelopmental assessments at 2 years of age were completed in 2014.

The following baseline characteristics were measured in all patients before the treatment administration: gestational age (in days), weight at birth (in g), head circumference at birth (in cm), sex and Apgar score at 5 minutes (categorical, 11 categories from 0 to 10), which is a method to quickly summarize the health of a newborn 5 minutes after birth. For easier reading, in the following we will refer to the baseline characteristics simply as gestational age, weight, head circumference, sex and Apgar score, respectively.

Participants were randomly assigned to receive either rhEPO or placebo intravenously within 3 hours, at 12 to 18 hours, and at 36 to 42 hours after birth. A single dose of the active treatment consisted of 25 μg (3000 IU) rhEPO per kg of body weight dissolved in 1 mL distilled water. Similarly, the placebo dose consisted of 1 mL of isotonic saline (NaCl, 0.9%) per kg of body weight.

The primary outcome was cognitive development, assessed with the Mental Development Index (MDI, higher values indicate better function) of the Bayley Scales of Infant Development II (BSID II) at 2 years corrected age, i.e. 2 years after term (gestation week 40) equivalent age.

The flow of participants is illustrated in Figure 2.2 (for a more detailed diagram, see Natalucci *et al.*, 2016). 450 newborns were randomized to either rhEPO or placebo. 57 of these did not receive the randomized treatment but were analyzed as they were intended to be treated. Two infants were excluded after randomization as they were found to meet the exclusion criteria, 38 dropped out of the study and 20 were excluded because they underwent another developmental test. Of the patients remaining in the study (204 of the rhEPO group and 186 of the placebo group), 13 (6.4%) of the rhEPO group and 12 (6.5%) of the placebo group died before follow-up assessment. The outcome of interest was collected at 2 years of age from 191 patients in the rhEPO group and from 174 patients in the placebo group.

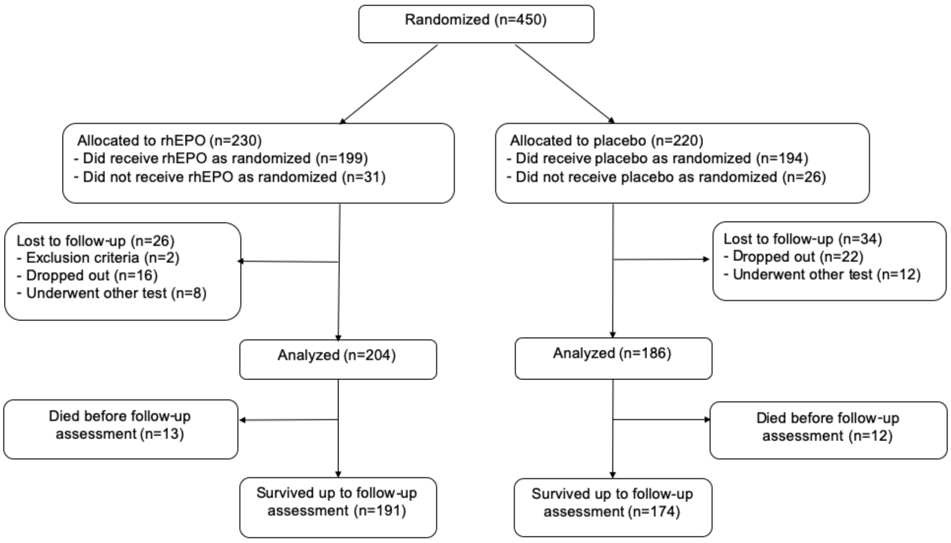
Randomized (n=450)
Allocated to rhEPO (n=230)
- Did receive rhEPO as randomized (n=199)
- Did not receive rhEPO as randomized (n=31)
Allocated to placebo (n=220)
- Did receive placebo as randomized (n=194)
- Did not receive placebo as randomized (n=26)
Lost to follow-up (n=26)
- Exclusion criteria (n=2)
- Dropped out (n=16)
- Underwent other test (n=8)
Lost to follow-up (n=34)
- Dropped out (n=22)
- Underwent other test (n=12)
Analyzed (n=204)
Analyzed (n=186)
Died before follow-up assessment (n=13)
Died before follow-up assessment (n=12)
Survived up to follow-up assessment (n=191)
Survived up to follow-up assessment (n=174)

Figure 2.2: Participant flow in the Epo trial.

In the original publication, losses to follow-up were ignored. The unadjusted treatment effect was determined using a linear regression model (equivalent to an unpaired *t*-test). A post hoc sensitivity analysis including the infants who died before follow-up was performed by imputing the worst observed outcome.

We reproduced the complete case analysis and the single imputation analysis reported in Natalucci *et al.* (2016). In addition, we analyzed the Epo trial using the implemented SACE estimator. We included all the baseline variables (gestational age, weight, head circumference,

Table 2.2: Overview of the investigated scenarios.

Scenario	Treatment effect on outcome	Treatment effect on survival
A	positive (MDI increased)	positive (survival probability higher)
B	positive (MDI increased)	negative (survival probability lower)
C	positive (MDI increased)	no effect
D	negative (MDI decreased)	positive (survival probability higher)
E	negative (MDI decreased)	negative (survival probability lower)
F	negative (MDI decreased)	no effect
G	no effect	positive (survival probability higher)
H	no effect	negative (survival probability lower)
I	no effect	no effect

Apgar score and sex) as arguments in `saceEstimator()` and thus as predictors of survival. We did not make use of the FLAC correction, since we did not expect failures with continuous covariates and we preferred to avoid the introduction of bias from FLAC.

2.6 Methods for the simulation study

We wrote a protocol of the simulation study in order to improve the quality of the simulation study itself. In fact, although often neglected by statisticians (Morris *et al.*, 2019), the protocol of a simulation study is as important as the protocol of a clinical trial, since it contributes to provide well-conducted and credible research. Writing a protocol forces one to think more deeply about the objectives of the study, to anticipate the possible obstacles that may be encountered, as well as to think in advance about possible solutions to the problems that may occur. In our case, before starting the actual simulation study, many questions arose about how the study should be conducted, analyzed and reported. Of course, it takes time to focus on all these questions without even seeing the code or the results, but this will be recovered in the simulation process, which will be more fluid and less prone to errors. In fact, a well-designed protocol prevents wasting time during the actual simulation, possible post-hoc changes and the repetition of the simulation several times due to bad planning. For our protocol, which is reported in Appendix A.4, we used the structure suggested by Burton *et al.* (2006).

The aim of our simulation study was to evaluate the performance of complete case analysis, Hayden's method and multiple imputation analysis in the estimation of a treatment effect from an RCT with a relevant proportion of outcomes truncated by death. The performance of the methods was evaluated under different scenarios in terms of bias, mean square error and coverage. For each scenario, we simulated 1300 small data sets (each containing 500 observations), using the Epo trial and the correlation structure of its variables as motivating example. For simplicity, only three baseline covariates were simulated (gestational age, head circumference and Apgar score). These were selected due to their importance for survival. All the data sets were analyzed by the three methods.

Since the methods estimate different quantities, the estimated treatment effects were compared with the "true values" of their respective estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died), which were obtained from the simulation of two large data sets (each containing 1'000'000 observations) for each scenario.

The investigated scenarios, in which we modeled different treatment effects on survival and on the outcome, are shown in Table 2.2. In particular, we modeled treatment effects on the outcome of -5, 0 or 5, and treatment effects on survival of $-log(2)$, 0 or $log(2)$ (corresponding to odds ratios of 0.5, 1 and 2).

As first step of the simulation study, we generated a matrix containing the seeds for all the simulations. Each column of the matrix contained the seeds for one scenario. The first two rows contained the seeds for the large data sets, the others contained seeds for the small data sets. As explained in the study protocol, the seeds for the small data sets were separated at least by their sample size, namely 500, and were separated from the seeds for the large data sets at least by the sample size of the large data sets, namely 1'000'000. The `R` code used to generate the matrix with the seeds is displayed in Appendix A.2.5.

The `R` function created for the simulation of the data sets is shown in Appendix A.2.6. The necessary variable values were extracted from the Epo trial data and were included as function's arguments.

Each data set was analyzed immediately after its simulation. For the analysis with Hayden's method, we used the `saceEstimator()` function described in Section 2.3, including gestational age, head circumference and Apgar score as baseline covariates. As for the analysis of the Epo trial, the FLAC correction was not needed. For the multiple imputation analysis, we used the `multipleImputation()` function described in Section 2.4, removing the survival variable from the predictor matrix since in our case it was perfectly associated with the missingness of the outcome and it was irrelevant as predictor of the outcome value. We used the default number of multiple imputations of `mice`, namely 5, since we performed many simulations and a larger number of imputed data sets was not necessary.

The seed, the mortality, the effective sample size analyzed by each method, and the results of the three analyses were stored after each simulation. In this way, the data sets could be reproduced and the storage of the whole data sets was not required.

2.6.1 Deviations from the protocol

The simulation study was not fully compliant with the protocol. Deviations from the protocol include the modeled treatment effects on survival, which we planned to be odds ratios of 0.9, 1 and 1.1 (depending on scenario), and which we substituted with odds ratios of 0.5, 1 and 2, respectively. In fact, the planned treatment effect was too weak to sufficiently differentiate between the performance of complete case analysis and Hayden's method.

An unexpected complication occurred in the estimation of the SACE variance from the large data sets ($n =$1'000'000). This was not achieved by the implemented `saceEstimator()`, since it involved the multiplication of 500'000×500'000-matrices and we could not obtain the amount of memory required. The SACE standard errors were estimated exploiting the "square-root law", which states that the accuracy of an estimator is inversely proportional to the square root of the sample size, and the following idea from Heyard and Held (2019): since $\text{SE}(n) = c \cdot \sqrt{1/n}$, we can estimate c using a weighted linear regression with outcome $\text{SE}(n)$, explanatory variable $\sqrt{1/n}$ and weight equal to n (`lm(se ~ sqrt(1/n), weight=n)` in `R`).

Moreover, in order to increase their precision, we estimated the true values of the estimands from two data sets of $n =$1'000'000 instead of from one. The two estimates were combined into one with the use of formulas from fixed effects meta-analysis.

Finally, the UZH math server for the simulation of the large data sets was not necessary and was not used.

Chapter 3

Results

3.1 Analysis of the Epo trial

In this section, the results of the Epo trial analysis are presented. After a descriptive analysis of the data, the treatment effect estimate provided by SACE approach is compared with those obtained by complete case analysis and by single imputation analysis.

Ideally, we would have analyzed all the 450 randomized patients, using multiple imputation for the “observable” missing values and then applying Hayden’s method because of the 25 outcomes truncated by death. However, in order to obtain a SACE estimate comparable with the estimates reported in Natalucci *et al.* (2016), the randomized patients who survived but were lost to follow-up were not analyzed. Consequently, the patients analyzed were 390: 191 survivors from the rhEPO group, 174 survivors from the placebo group, 13 non-survivors from the rhEPO group and 12 non-survivors from the placebo group.
The measure of head circumference was missing for 3 patients and that of Apgar score for 6 patients. For simplicity, these missing values were imputed using the mean of the observed measurements (in the case of Apgar score the mean was rounded to the closest integer).
Table 3.1 shows the baseline characteristics (after the imputation of the 9 missing values) for each treatment group.

	Placebo	rhEPO
n	186	204
Gestational age, days (mean (SD))	204.10 (11.83)	203.95 (11.54)
Weight at birth, g (mean (SD))	1197.77 (363.68)	1199.34 (319.67)
Head circumference at birth, cm (mean (SD))	26.80 (2.35)	26.84 (2.02)
Sex = male (%)	111 (59.7)	122 (59.8)
Apgar score at 5 minutes (median [IQR])	8.00 [6.00, 9.00]	8.00 [7.00, 9.00]

Table 3.1: Baseline characteristics in the treatment groups of the Epo trial.

The baseline characteristics were similar in the two groups. The distributions of gestational age, weight and head circumference were normally shaped, while Apgar score had a left-skewed distribution. The outcome of interest, MDI, was also normally distributed.
Figure 3.1 shows the association of survival up to 2 years of age with the treatment and the baseline variables. From the figure it can be seen that the proportion of survivors in the Epo trial was similar in the two treatment groups, as well as in male and female newborns. In contrast, gestational age, weight and head circumference had higher median values in the patients who survived. The distribution of Apgar score at 5 minutes also looked different in survivors and non-survivors, but too few data were present for non-survivors. This analysis suggested that gestational age, weight, head circumference and Apgar score could be predictive of survival up

to 2 years of age.

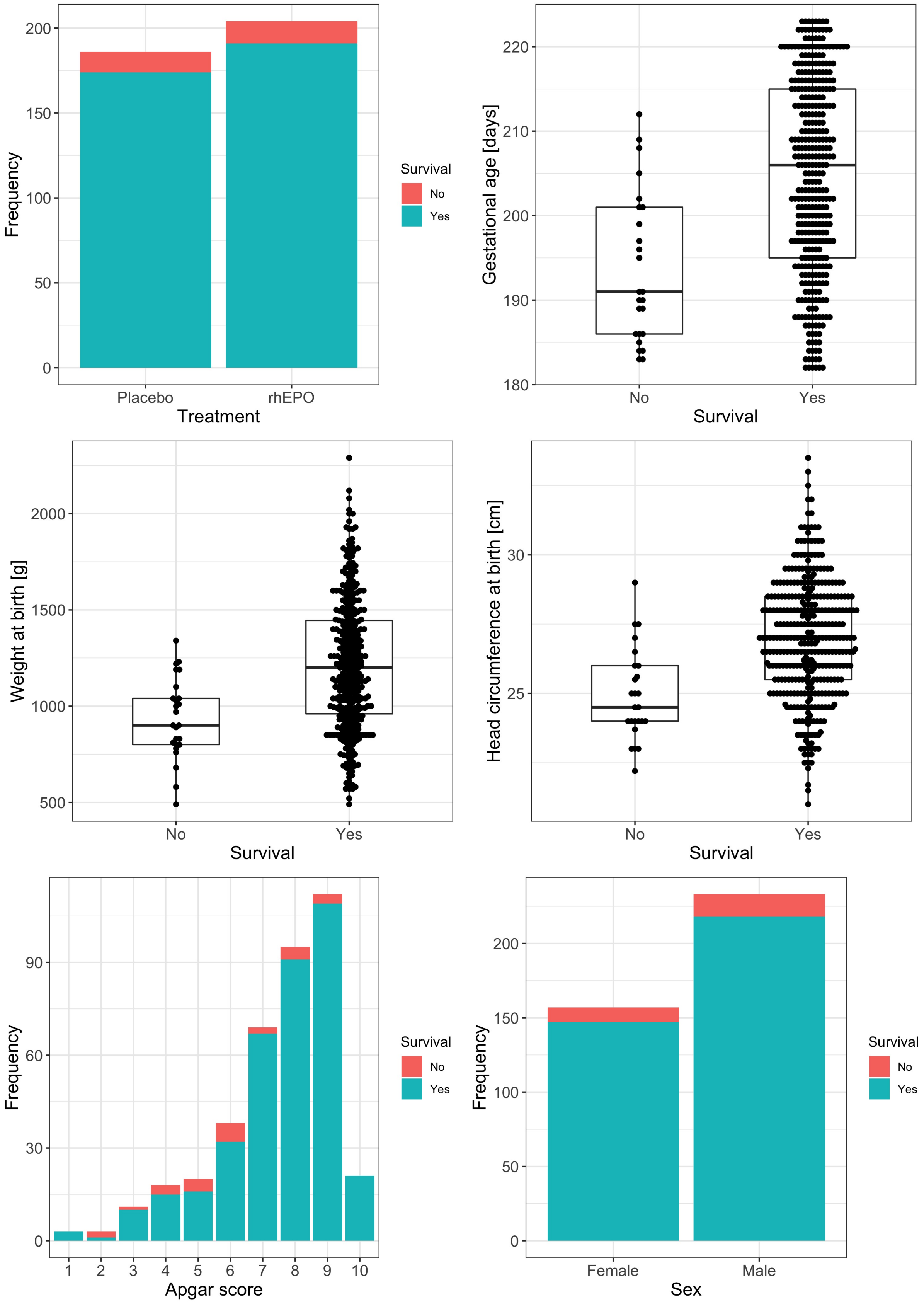

Figure 3.1: Association of survival up to 2 years of age with the assigned treatment and the baseline variables in the Epo trial.

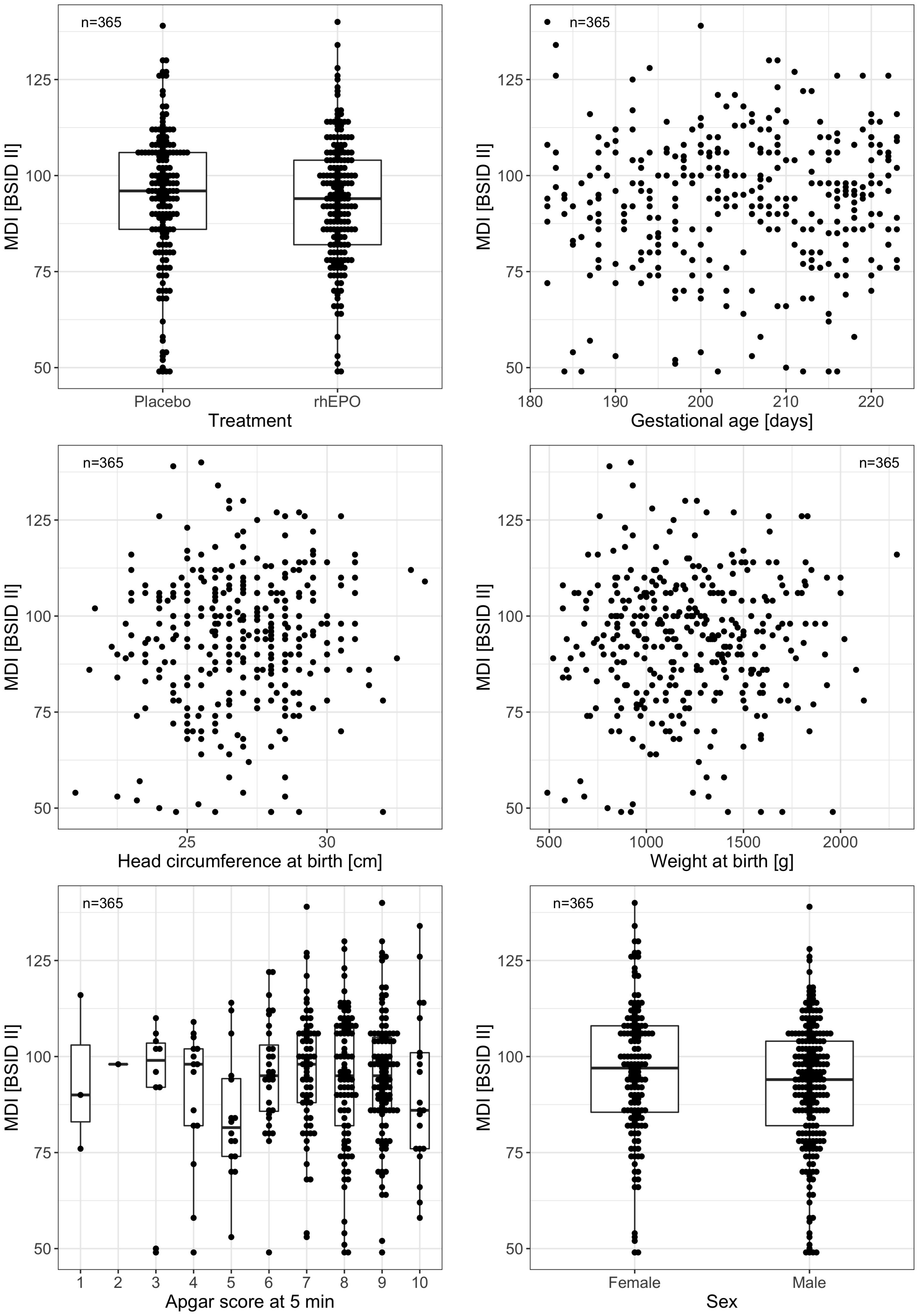

Figure 3.2: Association of the outcome of interest (MDI) with the assigned treatment and the baseline variables in the Epo trial. As the outcome was truncated by death for 25 patients, only 365 observations were available.

Figure 3.2 shows the association of the outcome of interest, MDI, with the treatment and the baseline variables. From the figure, no correlation between MDI and gestational age, head circumference, weight or Apgar score was visible. The distribution of MDI was similar in the two treatment groups, as well as in male and female newborns.

The treatment effect estimates resulting from the three analyses performed are displayed in Figure 3.3. With complete case analysis, a treatment effect on survivors of -1.02 (95% CI from -4.51 to 2.47) was obtained. With Hayden's estimator, a survivor average causal effect of -1.37 (95% CI from -4.83 to 2.09) was found. With the imputation of the worst observed outcome value to all the missing outcome measurements, a treatment effect of -0.92 (95% CI from -4.86 to 3.02) was estimated. The results derived by complete case analysis and by single imputation correspond to those reported in the original publication.

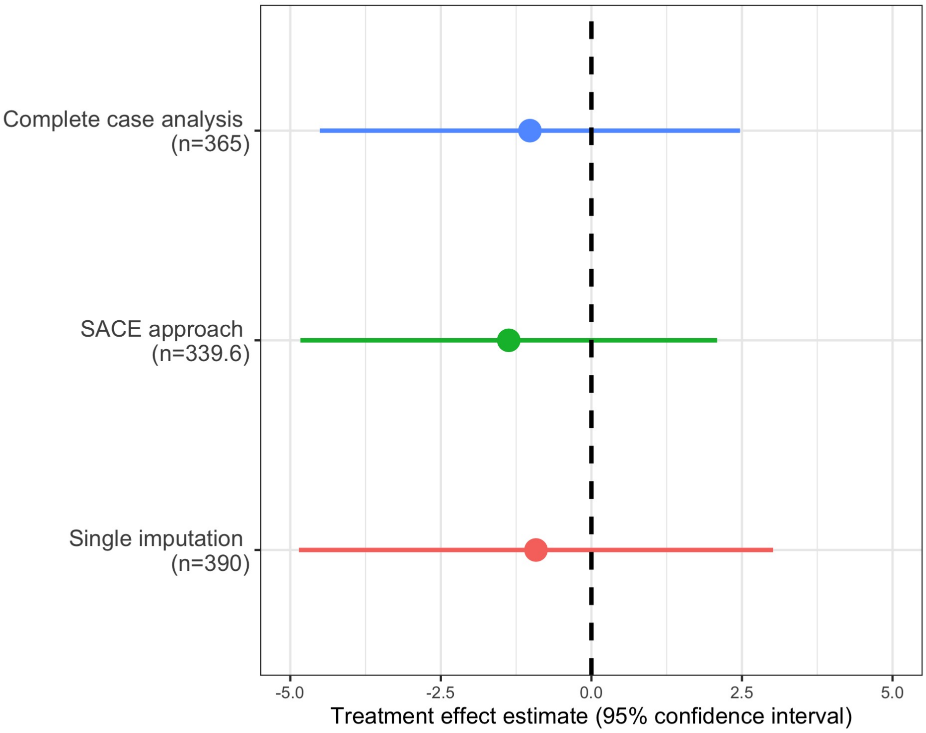

Figure 3.3: Estimates of the treatment effect on MDI at 2 years of age, derived by three different methods.

The effective sample size of patients analyzed was 365 for complete case analysis (the patients that survived up to 2 years of age), 339.6 for the analysis by Hayden's method (the sum of the survival probabilities under the treatment not assigned of patients who survived up to 2 years of age), and 390 for the single imputation analysis (the total number of patients).

From Figure 3.3 it can be seen that the estimates of the treatment effect on MDI provided by the three methods were similar. This was due to the fact that mortality in the Epo trial was low and similar in the treatment groups (6.4% in the rhEPO group, 6.5% in the placebo group; see also Figure 3.1, top left panel). As already discussed, when the treatment does not affect survival, complete case analysis and Hayden's method are expected to yield similar results. Moreover, when the overall mortality is low, there is a small number of outcomes truncated by death and the difference between the results provided by different methods to deal with them is obviously small.

It can also be noted that the variance of the single imputation estimate was larger than the others. This occurred because we imputed an extreme value to the missing data, which therefore increased the deviation from the mean outcome value.

3.2 Simulation study

3.2.1 Comparison between simulated data and Epo trial data

Figure 3.4 shows the distribution of the variables in a simulated data set (of sample size 500) and in the Epo trial data set (of sample size 390). The simulated data set was the first simulation of scenario I, where no treatment effects on the outcome and on survival were modeled (and thus seems the scenario most similar to the reality of the Epo trial). The distribution of gestational age in the Epo trial data was truncated due to the inclusion criteria. We relaxed the closeness to the Epo trial, allowing also slightly more extreme measures.

3.2.2 Check of simulation results

Once all simulations and analyses were performed, the stored results were explored. No missing values were present and no failures occurred. For each scenario and each method, the distributions of the treatment effect estimates and of their standard errors were visualized (Figures A.1 and A.2). No outliers or abnormalities were found.

3.2.3 True values of the estimands

Table 3.2 recalls the treatment effects modeled on the outcome ("T on O") and on survival ("T on S") for each investigated scenario, and shows the true values of the estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died) estimated from the large simulated data sets. Since the true values were estimated from two data sets of sample size 1'000'000, the standard errors are reported.
We expected the true value of θ_3 to be always the closest to the treatment effect modeled on the outcome, but it was not always the case. This may be due to the weak association of the outcome with the three covariates we used (see Figure 3.2). Ideally, for a proper multiple imputation analysis, we should have simulated other covariates than those important for survival, i.e. the variables associated with the outcome or potentially all the variables collected.
The true values of θ_2 and θ_3 were close in all scenarios, even though the estimands are different quantities, while the true value of θ_1 differed depending on the treatment effect modeled on survival. When treatment had no effect on survival (scenarios C, F and I), all the true values of the estimands were close.

Scenario	T on O	T on S	θ_1 (SE)	θ_2 (SE)	θ_3 (SE)
A	5	log2	4.901 (0.026)	5.011 (0.026)	5.024 (0.026)
B	5	-log2	5.124 (0.026)	5.008 (0.026)	5.014 (0.027)
C	5	0	5.008 (0.026)	5.008 (0.026)	5.015 (0.026)
D	-5	log2	-5.091 (0.026)	-4.976 (0.026)	-4.986 (0.028)
E	-5	-log2	-4.845 (0.026)	-4.957 (0.027)	-4.955 (0.026)
F	-5	0	-5.007 (0.026)	-5.009 (0.026)	-5.007 (0.026)
G	0	log2	-0.106 (0.026)	0.009 (0.026)	0.006 (0.025)
H	0	-log2	0.132 (0.026)	0.017 (0.026)	0.013 (0.026)
I	0	0	0.007 (0.026)	-0.002 (0.026)	0.013 (0.027)

Table 3.2: Treatment effect modeled on the outcome and on survival, and the true values of the estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died) estimated from the large simulated data sets.

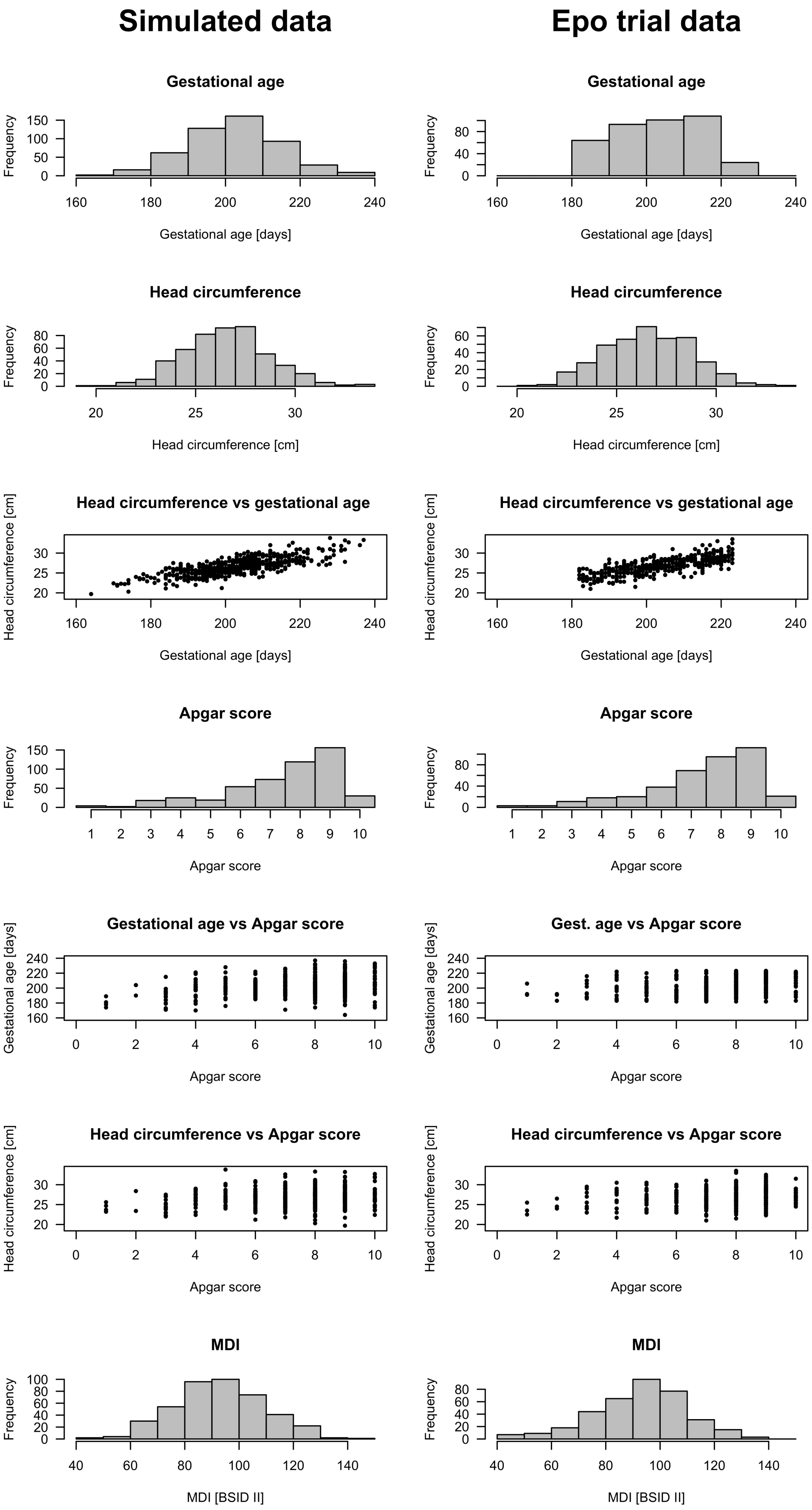

Figure 3.4: Distribution of the variables in a simulated data set from scenario I ($n = 500$, left column) and in the Epo trial data set ($n = 390$, right column).

3.2.4 Summary measures

Figure 3.5 shows the averages of the estimates provided by complete case analysis, Hayden's method and multiple imputation analysis, with the respective 95% confidence intervals, and the true values of the estimands θ_1, θ_2 and θ_3. It should be noted that, although the true values are represented as point estimates, some uncertainty about them is present (see Table 3.2).

In scenarios A, D and G, where treatment had a positive effect on survival, complete case analysis always resulted in more negative estimates than the SACE and multiple imputation estimates. More precisely, complete case analysis underestimated the positive treatment effect on the outcome (scenario A), overestimated the negative treatment effect (scenario D) and estimated a small negative effect when there was actually no effect (scenario G). This is coherent with the hypothesis that, since treatment increased survival, in the treatment group even less healthy patients survived, resulting in a placebo group with healthier patients than the treatment group. Consequently, complete case analysis estimated the treatment effect from non-comparable groups.

In contrast, in scenarios B, E and H, where treatment had a negative effect on survival, complete case analysis always resulted in more positive estimates than the SACE and multiple imputation estimates. More precisely, complete case analysis overestimated the positive treatment effect on the outcome (scenario B), underestimated the negative treatment effect (scenario E) and estimated a small positive effect when there was actually no effect (scenario H). This is coherent with the hypothesis that, since treatment decreased survival, in the treatment group the less healthy patients died, resulting in a treatment group with healthier patients than the placebo group. Also in this case, complete case analysis estimated the treatment effect from non-comparable groups.

As expected, in scenarios C, F and I, where treatment had no effect on survival, complete case analysis and Hayden's methods yielded similar results. In particular, in scenario I, where no treatment effects on outcome and on survival were present, the three methods estimated very similar treatment effects, even in presence of a significant proportion of outcomes truncated by death. This is coherent with the observation that, since treatment did not affect survival and the two groups had similar baseline characteristics due to randomization, the survivors also had similar characteristics and the groups remained comparable.

In summary, the complete case analysis estimates were very sensitive to changes in the treatment effect on survival, regardless of the treatment effect on the outcome, while the other methods were fairly stable.

The numerical summary measures (average of the estimates, empirical standard error, average of the standard errors), calculated for each scenario and each method, are displayed in Table A.2.

3.2.5 Performance measures

The performance of the methods was evaluated in terms of bias, MSE and coverage.

Bias

The bias of the three methods with respect to the three estimands θ_1, θ_2 and θ_3, with the 95% Monte Carlo confidence interval, is shown in Figure 3.6.

The methods were never found to be biased with respect to their targeted estimand, with the exception of multiple imputation in scenario A, which was biased with respect to θ_3. The reason may be, as already observed in Section 3.2.3, the weak association of the outcome with the three covariates we used, which did not allow multiple imputation to perform well.

As seen in the previous sections, in scenarios A, B, D, E, G and H, where a treatment effect on survival different from zero was modeled, the complete case analysis estimands and estimates substantially differed from those of Hayden's method and multiple imputation. In Figure 3.6, it is clearly visible that bias arose in these scenarios. More precisely, in scenarios A, B, D, G and

H, complete case analysis was biased with respect to θ_2 and θ_3, while in scenarios B, E, G and H, Hayden's method and multiple imputation were biased with respect to θ_1. Multiple imputation was also biased with respect to θ_1 in scenario D and with respect to θ_2 in scenario A.

In scenarios C, F and I, where treatment had no effect on survival, no method was found to be biased with respect to any estimand.

The numerical biases, calculated for each scenario and each method, are displayed in Table A.3.

MSE

Sometimes it is not possible to find an estimator that is both unbiased and has minimal variance. The MSE is a useful measure for comparative purposes as incorporates both measures. The MSE of the three methods with respect to the three estimands θ_1, θ_2 and θ_3 is shown in Figure 3.7, with the 95% Monte Carlo confidence interval.

From Figure 3.7, it can be seen that the MSE of the methods depended mainly on the scenario and less on the targeted estimand. This happened because, since the bias was small in magnitude, the variance of the estimate played a much larger role in the determination of the MSE. In fact, the MSEs were similar to the squared empirical standard errors shown in Table A.2.

Interestingly, in terms of MSE, all methods performed at worst in scenario A, where positive treatment effects on the outcome and on survival were modeled. In the same scenario, the treatment effect estimates provided by the three methods showed larger empirical standard errors (Table A.2).

The MSE of complete case analysis was always slightly smaller than that of Hayden's method. This may be caused by the fact that the effective sample size analyzed by complete case analysis was always greater than or equal to that analyzed by Hayden's method, and thus the variance of the complete case analysis estimates was always less than or equal to the variance of the Hayden's method estimates.

The numerical MSEs, calculated for each scenario and each method, are displayed in Table A.4.

Coverage

The coverage of the three methods with respect to the three estimands θ_1, θ_2 and θ_3 is shown in Figure 3.8, with the 95% Monte Carlo confidence interval.

The nominal coverage was achieved in most cases. However, multiple imputation showed overcoverage with respect to all estimands in scenario G, and complete case analysis showed overcoverage with respect to all estimands in scenario I. In scenario I, some tendency to overcoverage also by the other methods was observable, even though their Monte Carlo confidence intervals contained the nominal coverage of 95%. Overcoverage indicates that the results are too conservative, i.e. that more simulations would not find a significant result when this is present (Burton *et al.*, 2006).

The numerical coverages, calculated for each scenario and each method, are displayed in Table A.5.

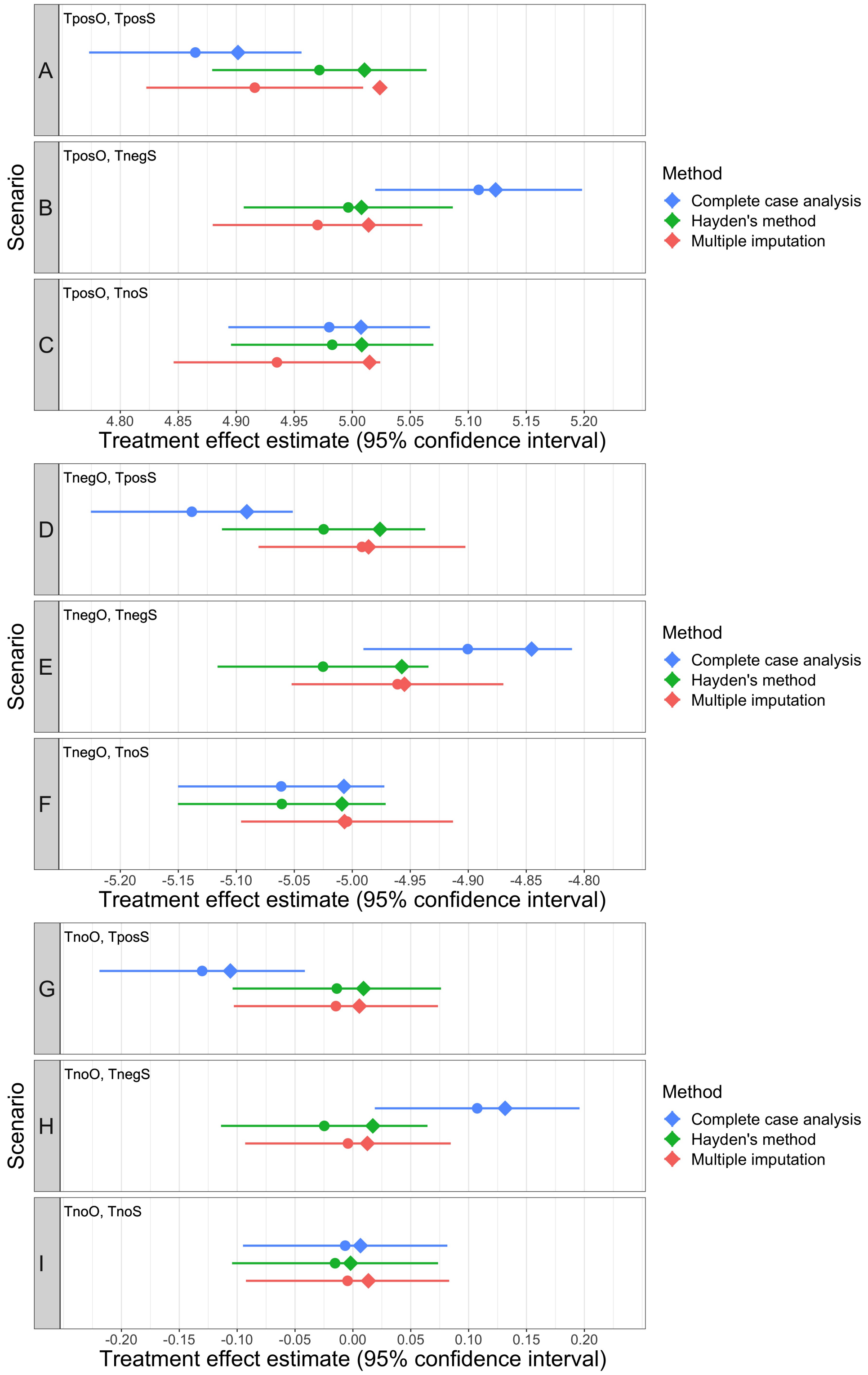

Figure 3.5: Averages of the treatment effect estimates (round points) provided by three methods, with 95% confidence intervals. Diamonds represent the true values of the respective estimands. The text “TposO”, “TnegO” or “TnoO” indicates whether the treatment effect modeled on the outcome in the given scenario was positive, negative or null, respectively. “TposS”, “TnegS” or “TnoS” indicates whether the treatment effect modeled on survival in the given scenario was positive, negative or null. Note the different x-axis scales for the three blocks of scenarios (A-C, D-F, G-I).

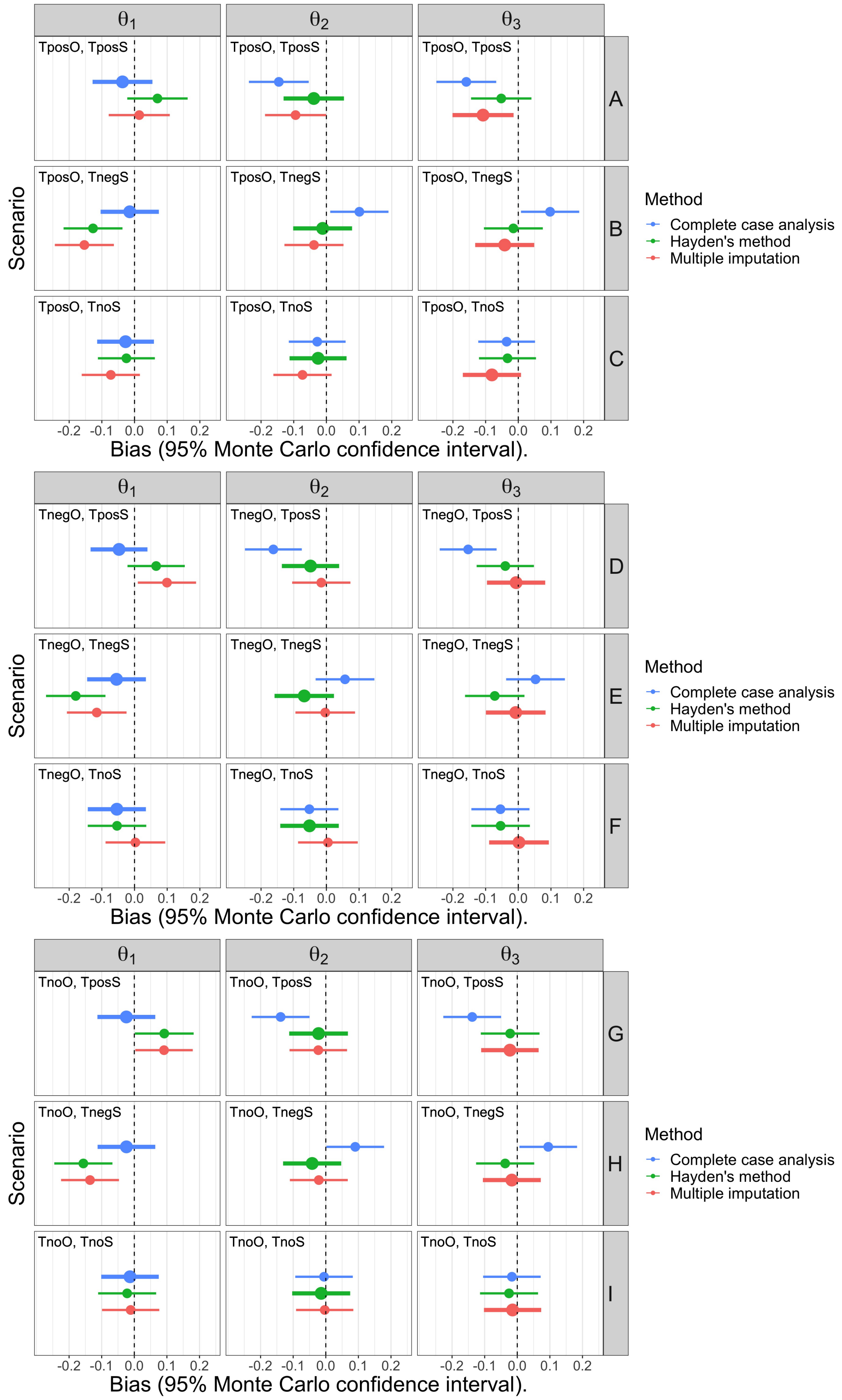

Figure 3.6: Bias of three methods with respect to θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients), with 95% Monte Carlo confidence interval. The dashed lines indicate no bias. In each subpanel, the estimate of the method that targets the column estimand is made thicker. "TposO", "TnegO" or "TnoO" indicates whether the treatment effect modeled on the outcome was positive, negative or null. "TposS", "TnegS" or "TnoS" indicates whether the treatment effect modeled on survival was positive, negative or null.

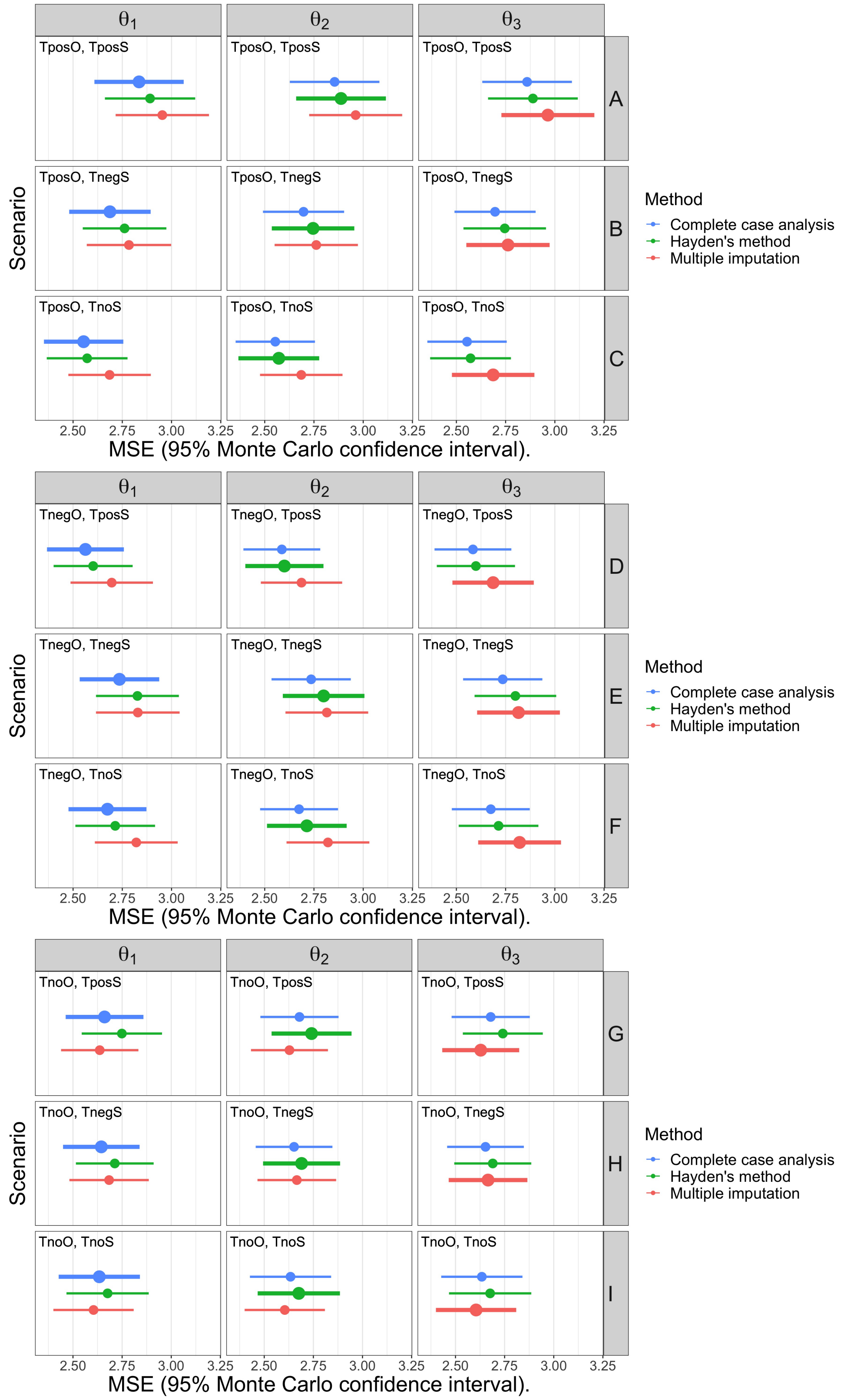

Figure 3.7: MSE of three methods with respect to θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients), with 95% Monte Carlo confidence interval. In each subpanel, the estimate of the method that targets the column estimand is made thicker. "TposO", "TnegO" or "TnoO" indicates whether the treatment effect modeled on the outcome was positive, negative or null. "TposS", "TnegS" or "TnoS" indicates whether the treatment effect modeled on survival was positive, negative or null.

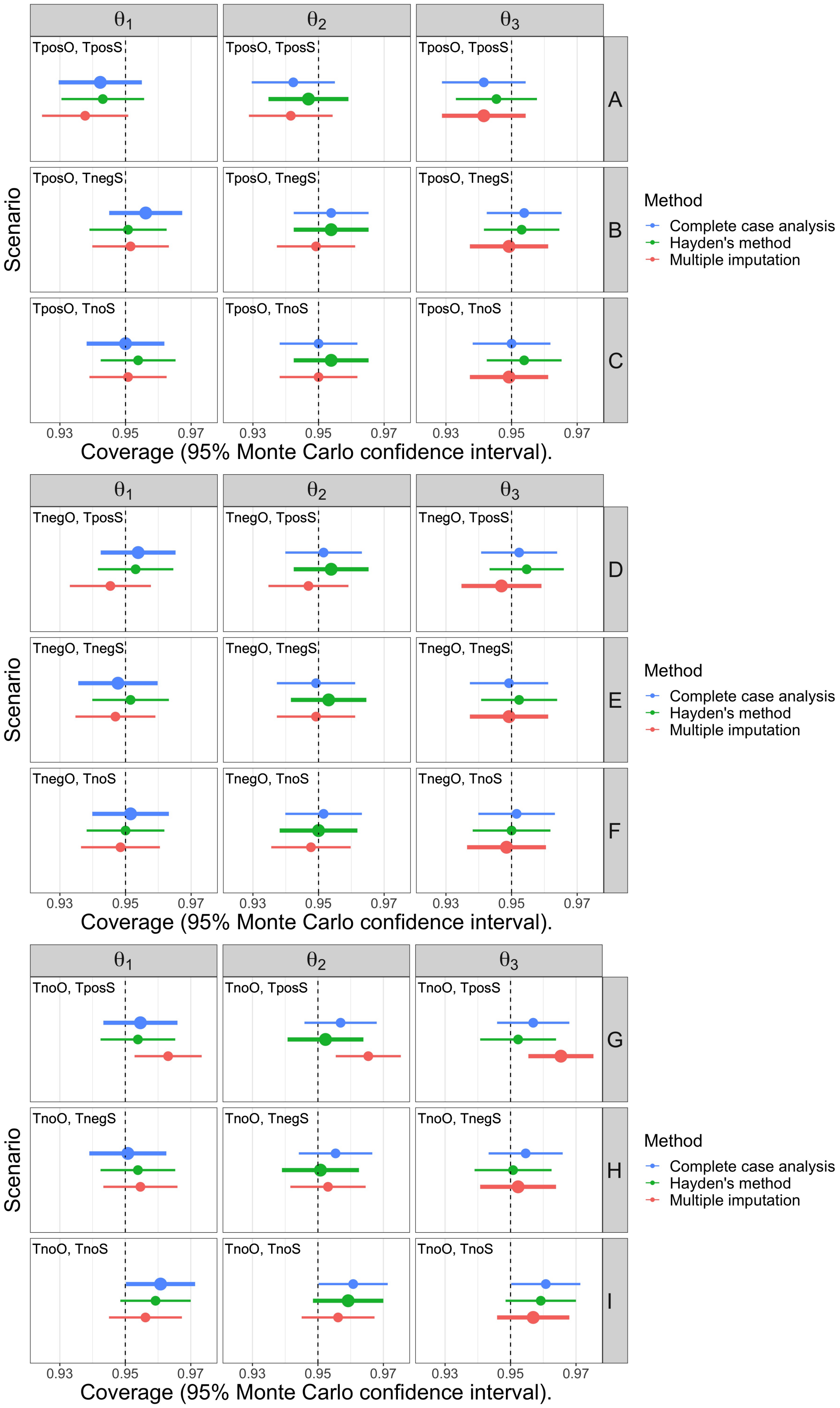

Figure 3.8: Coverage of three methods with respect to θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients), with 95% Monte Carlo confidence interval. The dashed lines indicate the nominal coverage. In each subpanel, the estimate of the method that targets the column estimand is made thicker. “TposO”, “TnegO” or “TnoO” indicates whether the treatment effect modeled on the outcome was positive, negative or null. “TposS”, “TnegS” or “TnoS” indicates whether the treatment effect modeled on survival was positive, negative or null.

Chapter 4

Discussion

This thesis aimed to implement Hayden's SACE estimator in R, to use it to reanalyze the Epo trial, and to compare it with complete case analysis and multiple imputation analysis under different scenarios through a simulation study.

Our findings from the Epo trial analysis confirmed those reported in Natalucci *et al.* (2016). Among the preterm infants of the Epo trial who received rhEPO, compared with those who received placebo, no evidence of a statistically significant difference on neurodevelopment at 2 years of age was found. However, some cognitive and physical differences may become evident later in life (Natalucci *et al.*, 2016) and will be monitored with a follow-up up to 5 years of age. The ongoing EpoRepair trial (Rüegger *et al.*, 2015) will give more insight about the effect of rhEPO on neurodevelopment at 5 years of age in very preterm infants suffering from intraventricular hemorrhage. Since the overall mortality in the Epo trial was low and similar in the treatment groups, the complete case analysis originally reported did not provide misleading results and yielded similar results to the SACE approach.

Concerning the findings from the simulation study, the treatment effects estimated by the compared methods varied between the scenarios and thus based on the modeled treatment effects. The unsuitability of complete case analysis to analyze studies with outcomes truncated by death was evident from the fact that the true value of its targeted estimand was very sensitive to changes in the treatment effect on survival. The true values of the estimands targeted by Hayden's method and multiple imputation were stable and similar to each other (even though the estimands are different quantities). Nevertheless, when survival was not affected by treatment, complete case analysis and Hayden's method always yielded similar treatment effect estimates. The same result was obtained by Colantuoni *et al.* (2018) when they compared complete case analysis with the SACE approach of Chiba and VanderWeele (2011), and is due to the fact that when survival is not affected by treatment the two methods estimate the same quantity. These considerations suggest that although complete case analysis is not generally advisable in the context of outcomes truncated by death, in some circumstances it may be able to estimate the SACE. On the contrary, although multiple imputation often provided similar results to the SACE approach, in the presence of truncation due to death it should be considered only if inference about an hypothetical population without deaths is desired. The methods were not biased with respect to their targeted estimands, except multiple imputation, which was biased with respect to θ_3 in scenario A, probably due to the weak association of the outcome with the simulated covariates, which made the setting not ideal for the application of this method. Bias with respect to the non-targeted estimands were present in scenarios where a treatment effect on survival different from zero was modeled. More precisely, the estimates derived by complete case analysis were biased with respect to the estimands targeted by Hayden's method and multiple imputation analysis, and/or vice versa. This result highlights the importance of choosing the statistical method to use based on the target estimand and on the expected scenario. In terms of MSE, the methods performed similarly across scenarios. The nominal coverage was achieved by all methods in all scenarios, except by multiple imputation in scenario G and complete case analysis in scenario I, where they showed overcoverage.

The approach used in our study has some limitations. The first is the not strong enough association of the Epo trial covariates with survival (see Figure 3.1). Good prediction of survival by covariates is necessary to make Hayden's assumptions to identify the SACE more plausible. Ideally, our analysis of the Epo trial should have been completed with a sensitivity analysis to investigate the robustness of our results to departures from those assumptions. We have also added a constant increase in mortality (to obtain 15% of outcomes truncated by death), which may have lead to a survival model with even lower discrimination. Thus, the performance of the SACE in both the analysis of the Epo trial and in the simulation study may have been decreased. If the use of Hayden's estimator was planned from the beginning of the study, this issue could have been reduced at the design stage, by collecting baseline variables that were likely to be predictive of survival. Moreover, the performance of multiple imputation has been limited due to the weak association of the Epo trial covariates with the outcome of interest (see Figure 3.2). As outlined by Sterne *et al.* (2009), for a proper multiple imputation analysis one should include a wide range of variables in the imputation model, even if they are not of interest in the substantive analysis, as failure to do so may mean that the missing at random assumption is not plausible and the results of the substantive analysis are biased. For simplicity, the covariates to be simulated in our study were selected by their importance for survival, and thus the created setting was not ideal for multiple imputation. This could have been improved by additionally simulating baseline variables that were likely to be predictive of the outcome and using them in the multiple imputation analysis. Lastly, due to time constraints, we estimated the true values of the estimands of interest from two data sets of sample size 1'000'000 per scenario. The variance of these estimates may not have been small enough to be considered as negligible.

In general, we believe that the SACE is a valuable approach to consider for many reasons. First, it provides a causal effect of the treatment. Second, the inference is made on a real population of subjects rather than on an hypothetical population, which may also be more relevant from the point of view of regulatory authorities. Moreover, as pointed out by Rubin (2006), the SACE approach is appropriate even when conclusions from a certain population are to be generalized to future healthier populations; a situation that can occur in real-world clinical trials, where experimental drugs are first tried with sicker patients (on which the approval is based), but then are used in broader and healthier populations. A point to consider is the fact that the SACE only makes sense when one is potentially able to manipulate the exposure (Holland, 1986). This is always the case in RCTs, but it would not be the case if one's aim is, for example, to estimate the SACE of age or sex on cognitive function, because it would not be ethical or possible to manipulate these exposures (Wen *et al.*, 2017).

We strongly recommend the SACE approach to analyze RCTs with outcomes truncated by death, even though there is no universally perfect choice of the method to use. As already mentioned, the choice should depend on the study research question and more specifically on the targeted estimand, as well as on the expected scenario. The importance of clearly defined estimands, in terms of targeted population, variable to measure, population level summary for the variable, possible post-randomization events and strategies to address them, is reported in the addendum on estimands and sensitivity analysis to the guideline on statistical principles for clinical trials (European Medicines Agency, 2017; US Food and Drug Administration, 2017), where principal stratification is one of the suggested strategies to deal with outcomes truncated by death. The motivation for the addendum derived from the not clear connectivity between study objectives, design, conduct, analysis and interpretation, and from the misalignment between missing data approaches and estimands of interest, as post-randomization events have been often approached as a missing data problematic (Lamarca, 2019). The final objective of a clinical trial must be to provide clear information on the effects of the treatments to patients and to clinicians.

We believe that further work could be done to make the different SACE estimators available and to give detailed guidance about them. Contributions already provided in this direction include the standalone application developed by Colantuoni *et al.* (2018), which implements the statistical methods discussed in their paper to compare non-mortality outcomes in RCTs with high mortality and can be found at `https://www.improvelto.com/stats-tools/`. The procedure proposed by Wang *et al.* (2017a) for comparing treatments based on the composite endpoint of both the functional outcome and survival is implemented in the `R` package `idem`. Their software is also demonstrated at `https://olssol.shinyapps.io/idem/`. The method of Wang *et al.* (2017b) to estimate the SACE with the use of a substitution variable is instead available in the `R` package `tbd`. Chiba and VanderWeele (2011) developed their simple technique with the aim of bringing the concepts of principal stratification to the epidemiology literature. We hope that our work will help increase familiarity with the SACE approach and availability of Hayden's method to deal with outcomes truncated by death.

Bibliography

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**, 4279–4292. 12, 22, 51, 52, 54

Chaix, B., Evans, D., Merlo, J., and Suzuki, E. (2012). Commentary: Weighing up the dead and missing - reflections on inverse probability weighting and principal stratification to address truncation by death. *Epidemiology*, **23**, 129–131. 1

Chiba, Y. (2012). The large sample bounds on the principal strata effect with application to a prostate cancer prevention trial. *The International Journal of Biostatistics*, **8**, 1. 3, 4

Chiba, Y. and VanderWeele, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology*, **173**, 745–751. 4, 27, 29

Colantuoni, E., Scharfstein, D. O., Wang, C., Hashem, M. D., Leroux, A., Needham, D. M., and Girard, T. D. (2018). Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*, **360**, 1756–1833. iii, 3, 5, 27, 29, 53

Egleston, B. L., Scharfstein, D. O., Freeman, E. E., and West, S. K. (2006). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, **8**, 526–545. 3, 4

European Medicines Agency (2017). ICH E9 (R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. iii, 28

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38. 9

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21–29. iii, 1, 2

Groenwold, R. H., Moons, K. G., and Vandenbroucke, J. P. (2014). Randomized trials with missing outcome data: how to analyze and what to report. *Canadian Medical Association Journal*, **186**, 1153–1157. 5

Hayden, D., Pauler, D. K., and Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, **61**, 305–310. iii, 1, 3, 4, 6, 7, 9, 44, 51

Heinze, G. and Ploner, M. (2018). *logistf: Firth's bias-reduced logistic regression*. R package version 1.23. 9

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409–2419. 9

Held, L. and Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 978–3. 8

Heyard, R. and Held, L. (2019). The quantile probability model. *Computational Statistics & Data Analysis*, **132**, 84–99. 13

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960. 28

Jemiai, Y., Rotnitzky, A., Shepherd, B. E., and Gilbert, P. B. (2007). Semiparametric estimation of treatment effects given baseline covariates on an outcome measured after a post-randomization event occurs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 879–901. 3, 4

Kurland, B. F., Johnson, L. L., Egleston, B. L., and Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical Science*, **24**, 211–222. 1, 6

Lamarca, R. (2019). *Estimands in clinical trials.* Astra Zeneca/SoCE. 28

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, **6**, 287–296. 10

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8**, 3–30. 52

Merchant, A. T., Sutherland, M. W., Liu, J., Pitiphat, W., and Dasanayake, A. (2018). Periodontal treatment among mothers with mild to moderate periodontal disease and preterm birth: reanalysis of OPT trial data accounting for selective survival. *International Journal of Epidemiology*, **47**, 1670–1678. 4

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102. 12, 55

Natalucci, G., Latal, B., Koller, B., Rüegger, C., Sick, B., Held, L., Bucher, H. U., and Fauchère, J.-C. (2016). Effect of early prophylactic high-dose recombinant human erythropoietin in very preterm infants on neurodevelopmental outcome at 2 years: a randomized clinical trial. *JAMA*, **315**, 2079–2085. iii, 1, 10, 11, 15, 27, 52, 54

Puhr, R., Heinze, G., Nold, M., Lusa, L., and Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, **36**, 2302–2317. 9, 51

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 1

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, **75**, 591–593. 4

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, **81**, 961–962. 4

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *John Wiley & Sons*, **81**, . 5

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Statistical Science*, **21**, 299–309. iii, 1, 2, 3, 28, 53

Rüegger, C. M., Hagmann, C. F., Bührer, C., Held, L., Bucher, H. U., Wellmann, S., and EpoRepair investigators (2015). Erythropoietin for the repair of cerebral injury in very preterm infants (EpoRepair). *Neonatology*, **108**, 198–204. 2, 27, 52

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*, **62**, 332–342. 3, 4

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**, b2393. 5, 28

US Food and Drug Administration (2017). ICH E9 (R1) Statistical principles for clinical trials addendum: estimands and sensitivity analysis in clinical trials. iii, 28

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, Articles*, **45**, 1–67. 10

Vickers, A. J. and Altman, D. G. (2013). Statistics notes: Missing outcomes in randomised trials. *BMJ*, **346**, f3438. 1

Wang, C., Scharfstein, D. O., Colantuoni, E., Girard, T. D., and Yan, Y. (2017a). Inference in randomized trials with death and missingness. *Biometrics*, **73**, 431–440. 29

Wang, L., Zhou, X.-H., and Richardson, T. S. (2017b). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika*, **104**, 597–612. 4, 29

Wen, L., Terrera, G. M., and Seaman, S. R. (2017). Methods for handling longitudinal outcome processes truncated by dropout and death. *Biostatistics*, **19**, 407–425. 28

Yang, F. and Small, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**, 299–318. 3, 4

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, **28**, 353–368. 1, 3, 4

Appendix A

Appendix

A.1 Software

All analyses were performed in the R programming language, R version 3.6.1 (2019-07-05), using base packages and the following analysis-specific packages: ggplot2 3.2.1, mice 3.6.0, lattice 0.20-38, numDeriv 2016.8-1.1, logistf 1.23, boot 1.3-22, truncnorm 1.0-8, MASS 7.3-51.4, fabricatr 0.10.0, stringr 1.4.0, tableone 0.10.0, dplyr 0.8.3, plyr 1.8.4, xtable 1.8-4, knitr 1.24. The computing environment on the author's personal computer had the following specifications: macOS Mojave, Version 10.14.6 (Operating system), 1.8 GHz Intel Core i5 (Processor) and 8 GB 1600 MHz DDR3 (Memory). This document was generated on January 7, 2020.

A.2 R code

A.2.1 Function resWithCI()

```
resWithCI <- function(est, se, m, alpha=0.05, df=NA) {
    if (is.na(df)){
        ci <- est+c(-1,1)*qnorm(1-alpha/2)*se
    } else {
        ci <- est+c(-1,1)*qt(p=1-alpha/2, df=df)*se
    }
    res <- list("estimate" = unname(est),
                "se" = unname(se),
                "ci" = ci,
                "m" = m)
    return(res)
}
```

A.2.2 Function flac()

```
library(logistf)

flac <- function(x,y) {
    # to calculate diagonal elements of hat matrix
    temp.fit1 <- logistf(y ~ x, pl=FALSE)
    # indicator variable of the augmented data set
    temp.pseudo <- c(rep(0,length(y)), rep(1, 2*length(y)))
    # weights of augmented data set
    temp.neww <- c(rep(1,length(y)), temp.fit1$hat/2, temp.fit1$hat/2)
    # ML estimation
    temp.fit2 <- logistf(c(y,y,1-y)~rbind(x,x,x)+temp.pseudo,
                         weights=temp.neww,
                         family=binomial(logit),
                         firth=FALSE,
                         pl=TRUE)
    # results
    res <- list()
    res$coefficients <- temp.fit2$coefficients[which(
        "temp.pseudo"!= names(temp.fit2$coefficients) )]
    res$fitted <- temp.fit2$predict[1:length(y)]
    res$linear.predictors <- temp.fit2$linear.predictors[1:length(y)]
    res$probabilities <- temp.fit2$probabilities[which(
        "temp.pseudo"!= names(temp.fit2$prob) )]
    res$ci.lower <- temp.fit2$ci.lower[which(
        "temp.pseudo"!= names(temp.fit2$ci.lower)) ]
    res$ci.upper <- temp.fit2$ci.upper[which(
        "temp.pseudo"!= names(temp.fit2$ci.upper)) ]
    vcov <- vcov(temp.fit2)
    res$vcov <- vcov[which("temp.pseudo"!= names(temp.fit2$coefficients) ),
                     which("temp.pseudo"!= names(temp.fit2$coefficients) )]
    return(res)
}
```

A.2.3 Function saceEstimator()

```
library(numDeriv)

saceEstimator <- function(Z, X, Y, S, alpha=0.05, flac_corr=FALSE,
                          df=NA, dat) {

    ####### 0) Check data type #######

    if(!is.character(Z)) stop("Z should be a character.")
    if(!is.character(X)) stop("X should be a character.")
    if(!is.character(Y)) stop("Y should be a character.")
    if(!is.character(S)) stop("S should be a character.")
    if(alpha>1 | alpha<0) stop("alpha should be between 0 and 1.")
    if(!(Z %in% colnames(dat))) stop("Z should be a column name of dat.")
    if(sum(X %in% colnames(dat))<length(X)) stop("X should be a vector of
                                                 column names of dat.")
    if(!(Y %in% colnames(dat))) stop("Y should be a column name of dat")
    if(!(S %in% colnames(dat))) stop("S should be a column name of dat")
    if(!is.vector(dat[ ,Z])) stop("dat[ ,Z] should be a vector.")
    if(!is.vector(dat[ ,Y])) stop("dat[ ,Y] should be a vector.")
    if(!is.vector(dat[ ,S])) stop("dat[ ,S] should be a vector.")
    if(!is.data.frame(dat)) stop("dat should be a data.frame.")
```

```
####### 1) Check missing values #######

if(sum(is.na(dat[ ,Z]))) stop("dat[ ,Z] should not have missing values.")
if(sum(is.na(dat[,X]))) stop("dat[ ,X] should not have missing values.")
if(sum(is.na(dat[ ,Y][dat[ ,S]==1]))) stop("dat[ ,Y] should not have
                                    missing values where dat[ ,S]==1.")
if(sum(is.na(dat[ ,S]))) stop("dat[ ,S] should not have missing values.")

####### 2) Compute sace estimate #######

# split patients according to assigned treatment
i <- dat[dat[ ,Z] == 1, ] # treatment arm z=1
j <- dat[dat[ ,Z] == 0, ] # treatment arm z=0
n1 <- nrow(i)
n0 <- nrow(j)

# fitting model for survival under treatment z=0
my.formula <- formula(paste(S, " ~ ", paste(X, collapse = "+")))
design.matrix.j <- model.matrix(my.formula, data = j)
if (flac_corr==FALSE) {
    mod.z0 <- glm(my.formula, data = j, family=binomial)
    betas.z0 <- coef(mod.z0)
} else {
    flac.z0 <- flac(design.matrix.j[,-1], j[,S])
    betas.z0 <- flac.z0$coefficients
}

# fitting model for survival under treatment z=1
design.matrix.i <- model.matrix(my.formula, data = i)
if (flac_corr==FALSE) {
    mod.z1 <- glm(my.formula, data = i, family=binomial)
    betas.z1 <- coef(mod.z1)
} else {
    flac.z1 <- flac(design.matrix.i[,-1], i[,S])
    betas.z1 <- flac.z1$coefficients
}

# calculation of survival probability under treatment z=0
# for the patients (i) who actually received treatment z=1
i$pi.z0 <- as.numeric(1 / (1 + exp(-design.matrix.i %*% betas.z0)))

# calculation of survival probability under treatment z=1,
# for the patients (j) who actually received treatment z=0
j$pj.z1 <- as.numeric(1 / (1 + exp(-design.matrix.j %*% betas.z1)))

# sace estimate mu = mu1 - mu0
mu1 <- sum(i[,Y] * i[,S] * i$pi.z0, na.rm = TRUE) / sum(i[,S] * i$pi.z0)
mu0 <- sum(j[,Y] * j[,S] * j$pj.z1, na.rm = TRUE) / sum(j[,S] * j$pj.z1)
mu_sace <- mu1 - mu0
```

```
# effective sample size analyzed
m_sace <- sum(i[,S] * i$pi.z0) + sum(j[,S] * j$pj.z1)

####### 3) Compute sace variance #######

# inverse information matrix for mod.z0
if (flac_corr==FALSE) {
    I0_inv <- vcov(mod.z0)
} else {
    I0_inv <- flac.z0$vcov
}

# inverse information matrix for mod.z1
if (flac_corr==FALSE) {
    I1_inv <- vcov(mod.z1)
} else {
    I1_inv <- flac.z1$vcov
}

# transposed design matrices
Xi <- t(design.matrix.i)
Xj <- t(design.matrix.j)

# prepare computation of long summands
# for mu1n
s1 <- i[,S] * i[,Y] * i$pi.z0 * (1 - i$pi.z0) * t(Xi)
ss1 <- t(apply(s1, 2, sum, na.rm = TRUE))
# for mu1d
s2 <- i[,S] * i$pi.z0 * (1 - i$pi.z0) * t(Xi)
ss2 <- t(apply(s2, 2, sum))
# for mu0n
s3 <- j[,S] * j[,Y] * j$pj.z1 * (1 - j$pj.z1) * t(Xj)
ss3 <- t(apply(s3, 2, sum, na.rm = TRUE))
# for mu0d
s4 <- j[,S] * j$pj.z1 * (1 - j$pj.z1) * t(Xj)
ss4 <- t(apply(s4, 2, sum))

# fitted survival probabilities pi1 and pj0
if (flac_corr==FALSE) {
    j$pj.z0 <- mod.z0$fitted.values
    i$pi.z1 <- mod.z1$fitted.values
} else {
    j$pj.z0 <- flac.z0$fitted
    i$pi.z1 <- flac.z1$fitted
}

# columns of C1
c11 <- i[,S] * i[,Y] * i$pi.z0
c11[which(is.na(c11))] <- 0 # S*Y=0 if S=0
```

```
    c12 <- i[,S] * i$pi.z0
    c13 <- t(ss3 %*% I0_inv %*% Xi) * (i[,S] - i$pi.z1)
    c14 <- t(ss4 %*% I0_inv %*% Xi) * (i[,S] - i$pi.z1)

    # C1
    C1 <- cbind(c11, c12, c13, c14)

    # columns of C0
    c01 <- t(ss1 %*% I1_inv %*% Xj) * (j[,S] - j$pj.z0)
    c02 <- t(ss2 %*% I1_inv %*% Xj) * (j[,S] - j$pj.z0)
    c03 <- j[,S] * j[,Y] * j$pj.z1
    c03[which(is.na(c03))] <- 0 # S*Y=0 if S=0
    c04 <- j[,S] * j$pj.z1

    # C0
    C0 <- cbind(c01, c02, c03, c04)

    # E0, E1, id1, id0
    E0 <- rep(1, n0)
    E1 <- rep(1, n1)
    id0 <- diag(n0)
    id1 <- diag(n1)

    # variance-covariance matrix of (mu1n, mu1d, mu0n, mu0d)
    Sigma <- t(C1) %*% (id1 - E1 %*% t(E1) / n1) %*% C1 +
             t(C0) %*% (id0 - E0 %*% t(E0) / n0) %*% C0

    # Jacobian matrix of h(mu1n, mu1d, mu0n, mu0d) = mu1n/mu1d - mu0n/mu0d
    mu1n <- sum(c11) + sum(c01)
    mu1d <- sum(c12) + sum(c02)
    mu0n <- sum(c13) + sum(c03)
    mu0d <- sum(c14) + sum(c04)
    x0 <- c(mu1n, mu1d, mu0n, mu0d)
    h <- function(x) {x[1] / x[2] - x[3] / x[4]}
    J <- jacobian(h, x0)

    # Delta method
    var_sace <- as.numeric(J %*% Sigma %*% t(J))

    ####### 4) Results #######
    res <- resWithCI(mu_sace, sqrt(var_sace), m_sace, alpha, df)
    return(res)
}
```

A.2.4 Function multipleImputation()

```
library(mice)

multipleImputation <- function(Y, Z, dat, m=5, not_predictor=NA,
                               alpha=0.05, df=NA){
    # preparation
    dat <- dplyr::rename(dat,"Y"=Y, "Z"=Z)
    ini <- mice(dat, m=m, print=FALSE, maxit=0)
    pred <- ini$predictorMatrix
    # remove specified variable(s) from predictor matrix
    if (!is.na(not_predictor)) {
        pred[, not_predictor] <- 0
    }
    # multiple imputations
    dat.mice <- mice(dat, m=m, pred=pred, print=TRUE)
    # analyze data sets
    fit.mice <- with(data=dat.mice, exp=lm(Y~Z))
    # pool estimates using Rubin's rules
    pool.mice <- pool(fit.mice)
    # results
    est_mice <- summary(pool.mice)[2,1]
    se_mice <- summary(pool.mice)[2,2]
    res <- resWithCI(est_mice, se_mice, m=nrow(dat), alpha=alpha, df=df)
    return(res)
}
```

A.2.5 Matrix with seeds for simulation

```
# Function to generate a vector of length n
# with integers separated at least by samplesize
generate_seeds <- function(n, samplesize) {
        # deterministic part
    det_part <- seq(1,
                    1 + (samplesize + 101) * (n - 1),
                    by = (samplesize + 101))
    # random part
    random_part <- round(runif(n, 0, 100))
    # seeds
    return(det_part + random_part)
}

nsim_plus <- 2000 # to have enough seeds in case of failures
n_small <- 500 # sample size of the small data sets
n_large <- 1000000 # sample size of the large data set
nscenarios <- 9 # number of columns of the matrix

# Generate matrix with seeds for the small data sets
set.seed(23092019)
m_seeds <- matrix(generate_seeds(n=nsim_plus*nscenarios, samplesize=n_small),
                                    nrow=nsim_plus, ncol=nscenarios)

# Seeds for the large data sets (first two rows)
m_seeds_large <- generate_seeds(nscenarios*2, samplesize=n_large)+
    n_large+max(m_seeds)
m_seeds[1,] <- m_seeds_large[1:9]
m_seeds[2,] <- m_seeds_large[10:18]
```

A.2.6 Function simulate_dataset()

```
# Function of which we want to find the root with respect to c
# c will be added in the simulation to the linear predictor of
# survival to decrease survival to 85%
target <- function(c, eta, target.prob=0.85){
    return(mean(inv.logit(eta+c)-target.prob))
}

library(fabricatr)
library(MASS)
library(truncnorm)
library(boot)

# Function to simulate one data set using the associations from the Epo trial
# data and specifying the direction of the treatment effects to model on the
# outcome and on survival

simulate_dataset <- function(n, # sample size of the data set to simulate
                        seed, # from the seeds matrix generated above
                        trt.eff.surv=c("pos", "neg", "no"),
                        trt.eff.out=c("pos", "neg", "no"),
# the values of the following variables are derived from the Epo trial data:
                        prop.apgar.epo, # distribution of apgar score
                        values.apgar.epo, # values of apgar score
                        mean.ga.apgarcat, # mean values of gestational
                        # age for each apgar category
                        mean.hc.apgarcat, # mean values of head circumference
                        # for each apgar category
                        Sigma.ga.hc, # covariance matrix of gest. age and head
                        # circumference
                        betas.out, # coefficients of covariates from the lm
                        # for the outcome
                        sigma.out, # residual standard deviation
                        # of the model for outcome
                        betas.surv){ # coefficients of covariates from the glm
                        # for survival

    ####### 0) Preparation #######

    # treatment effects on survival and on outcome
```

```r
trt.eff.surv <- ifelse(trt.eff.surv=="no", 0,
                       ifelse(trt.eff.surv=="pos", log(2), -log(2)))
trt.eff.out <- ifelse(trt.eff.out=="no", 0,
                      ifelse(trt.eff.out=="pos", 5, -5))

# initiate dataset with baseline covariates
dt <- data.frame(apgar5=rep(NA, n), gest.age.days=rep(NA, n),
                 hc.birth.cm=rep(NA, n))

# set the seed
set.seed(seed)

####### 1) Simulation of covariates #######

# simulate apgar score with distribution as in Epo trial data
dt$apgar5 <- draw_categorical(N=n, prob = prop.apgar.epo)

# simulate gest.age.days and hc.birth.cm as multivariate normal
for (i in values.apgar.epo){
    if(nrow(dt[dt$apgar5==i,])>0){
        multivars <- matrix(mvrnorm(
            n=nrow(dt[dt$apgar5==i,]),
            mu=c(mean.ga.apgarcat[i], mean.hc.apgarcat[i]),
            Sigma=Sigma.ga.hc), ncol=2)
        dt$gest.age.days[dt$apgar5==i] <- round(multivars[,1])
        dt$hc.birth.cm[dt$apgar5==i] <- round(multivars[,2],1)
    }
}

####### 2) Simulation of treatment #######

EPOrows <- sample.int(n, n/2, replace = FALSE)
dt$treatment <- 0
dt$treatment[EPOrows] <- 1

####### 3) Simulation of outcome #######

design.m <- model.matrix(~ gest.age.days + hc.birth.cm + apgar5 +
                             treatment, data=dt)
# add treatment effect
betas.out.trt <- matrix(c(betas.out, trt.eff.out), 5, 1)

# linear predictor
eta.out <- design.m %*% betas.out.trt

# simulate outcome
dt$mdi <- rnorm(n, eta.out, sigma.out)

####### 4) Simulation of survival #######
```

```
    # add treatment effect
    betas.surv.trt <- matrix(c(betas.surv, trt.eff.surv), 5, 1)

    # linear predictor
    xb <- design.m %*% betas.surv.trt

    # uniroot searches in the interval from lower to upper for a root
    # of the function specified, with respect to its first argument
    res <- uniroot(target, eta=xb, lower=-10, upper=10)
    c <- res$root # constant to add to the linear predictor
    # new linear predictor with decreased survival probability (0.85).
    eta.surv <- xb + c

    # survival probabilities given the covariates
    p <- inv.logit(eta.surv)

    # simulate survival
    dt$alive = rbinom(n = n, size = 1, prob = p)

    ####### 5) Simulation of outcomes truncated by death #######

    dt$mdi[dt$alive==0] <- NA

    ####### 6) Return simulated dataset #######

    return(dt)
}
```

A.3 Figures and tables

	Author	Title	Journal	Year	Type
1	Chaix et al.	Weighing up the dead and missing: reflections on inverse probability weighting and principal stratification to address truncation by death	Epidemiology	2012	Methodological
2	Checkley et al.	Inference for mutually exclusive competing events through a mixture of generalized gamma distributions	Epidemiology	2010	Methodological
3	Chiba and VanderWeele	A simple method for principal strata effects when the outcome has been truncated due to death	American Journal of Epidemiology	2011	Methodological
4	Chiba	The large sample bounds on the principal strata effect with application to a prostate cancer prevention trial	International Journal of Biostatistics	2012	Methodological
5	Chiba	Kaplan-Meier curves for survivor causal effects with time-to-event outcomes	Clinical Trials	2013	Methodological
6	Chiba	Sharp nonparametric bounds and randomization inference for treatment effects on an ordinal outcome	Statistics in Medicine	2017	Methodological
7	Chiba	Bayesian inference of causal effects for an ordinal outcome in randomized trials	Journal of Causal Inference	2018	Methodological
8	Choi et al.	Recurrent event frailty models reduced time-varying and other biases in evaluating transfusion protocols for traumatic hemorrhage	Journal of Clinical Epidemiology	2016	Methodological
9	Colantuoni et al.	Statistical methods to compare functional outcomes in randomized controlled trials with high mortality	British Medical Journal	2018	Methodological
10	Dawson and Lavori	Design and inference for the intent-to-treat principle using adaptive treatment.	Statistics in Medicine	2015	Methodological
11	Egleston et al.	Causal inference for non-mortality outcomes in the presence of death	Biostatistics	2007	Methodological
12	Egleston et al.	On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death	Biometrics	2009	Methodological
13	Egleston et al.	A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies	Clinical Trials	2010	Methodological
14	Egleston et al.	Latent class survival models linked by principal stratification to investigate heterogenous survival subgroups among individuals with early-stage kidney cancer	Journal of the American Statistical Association	2017	Methodological
15	Gilbert and Jin	Semiparametric estimation of the average causal effect of treatment on an outcome measured after a postrandomization event, with missing outcome data	Biostatistics	2010	Methodological
16	Huang et al.	Design and rationale of the reevaluation of systemic early neuromuscular blockade trial for acute respiratory distress syndrome	Annals of the American Thoracic Society	2017	Study protocol
17	Jemiai et al.	Serniparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs	Journal of the Royal Statistical Society Series B-statistics methodology	2007	Methodological
18	Kurland et al.	Longitudinal data with follow-up truncated by death: match the analysis method to research aims	Statistical Science	2009	Methodological
19	Lee and Daniels	Marginalized models for longitudinal ordinal data with application to quality of life studies	Statistics in Medicine	2008	Methodological
20	Lee et al.	Causal effects of treatments for informative missing data due to progression/death	Journal of the American Statistical Association	2010	Methodological
21	Leuchs et al.	Disentangling estimands and the intention-to-treat principle	Pharmaceutical Statistics	2017	Methodological
22	Liu and Ying	Joint analysis of longitudinal data with informative right censoring	Biometrics	2008	Methodological
23	Long and Hudgens	Comparing competing risk outcomes within principal strata, with application to studies of mother-to-child transmission of HIV	Statistics in Medicine	2012	Methodological
24	Long et al.	An investigation of selection bias in estimating racial disparity in stroke risk factors: the REGARDS study	American Journal of Epidemiology	2019	Application
25	Lou et al.	Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: noncompliance and missing data	Journal of Biopharmaceutical Statistics	2019	Methodological
26	Lu et al.	Rank-based principal stratum sensitivity analyses	Statistics in Medicine	2013	Methodological
27	MacKenzie et al.	The national study on costs and outcomes of trauma	Journal of Trauma-Injury Infection and Critical Care	2007	Application
28	MacKenzie et al.	The impact of trauma-center care on functional outcomes following major lower-limb trauma	Journal of Bone and Joint Surgery-American Volume	2008	Application
29	Mark et al.	Quality of life with defibrillator therapy or amiodarone in heart failure	New England Journal of Medicine	2008	Application
30	Mark et al.	Quality-of-life outcomes with coronary artery bypass graft surgery in ischemic left ventricular dysfunction: A randomized trial	Annals of Internal Medicine	2014	Application
31	Mark	Assessing quality-of-life outcomes in cardiovascular clinical research	Nature Reviews Cardiology	2016	Methodological
32	McGuinness et al.	Survival bias when assessing risk factors for age-related macular degeneration: A tutorial with application to the exposure of smoking	Ophtalmic Epidemiology	2017	Application
33	Mentz et al.	The palliative care in heart failure trial: rationale and design	American Hearth Journal	2014	Study protocol
34	Merchant et al.	Periodontal treatment among mothers with mild to moderate periodontal disease and preterm birth: reanalysis of OPT trial data accounting for selective survival	International Journal of Epidemiology	2018	Application
35	Needham et al.	Study protocol: the improving care of acute lung injury patients (ICAP) study	Critical Care	2006	Study protocol
36	Needham	Understanding and improving clinical trial outcome measures in acute respiratory failure	American Journal of Respiratory and Critical Care Medicine	2014	Methodological
37	Park et al.	Integrating tobacco treatment into cancer care: study protocol for a randomized controlled comparative effectiveness trial	Contemporary Clinical Trials	2016	Study protocol
38	Prada et al.	Level-I trauma center effects on return-to-work outcomes	Evaluation Review	2012	Application
39	Shardell et al.	Doubly robust estimation and causal inference in longitudinal studies with dropout and truncation by death	Biostatistics	2015	Methodological
40	Shepherd et al.	Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials	Biometrics	2006	Application
41	Shepherd et al.	Does finasteride affect the severity of prostate cancer? A causal sensitivity analysis	Journal of the American Statistical Association	2008	Application
42	Taguri and Chiba	A principal stratification approach for evaluating natural direct and indirect effects in the presence of treatment-induced intermediate confounding	Statistics in Medicine	2015	Methodological
43	Tchetgen Tchetgen	Identification and estimation of survivor average causal effects	Statistics in Medicine	2014	Methodological
44	Tchetgen Tchetgen et al.	A simple regression-based approach to account for survival bias in birth outcomes research	Epidemiology	2015	Methodological
45	Wang	Inference in randomized trials with death and missingness	Biometrics	2017	Methodological
46	Wang et al.	Identification and estimation of causal effects with outcomes truncated by death	Biometrika	2017	Methodological
47	Wen et al.	Methods for handling longitudinal outcome processes truncated by dropout and death	Biostatistics	2018	Methodological
48	Yang and Small	Using post-outcome measurement information in censoring-by-death problems	Journal of the Royal Statistical Society Series B-statistics methodology	2016	Methodological

Table A.1: Papers citing Hayden *et al.* (2005), collected from the Web of Science and PubMed.

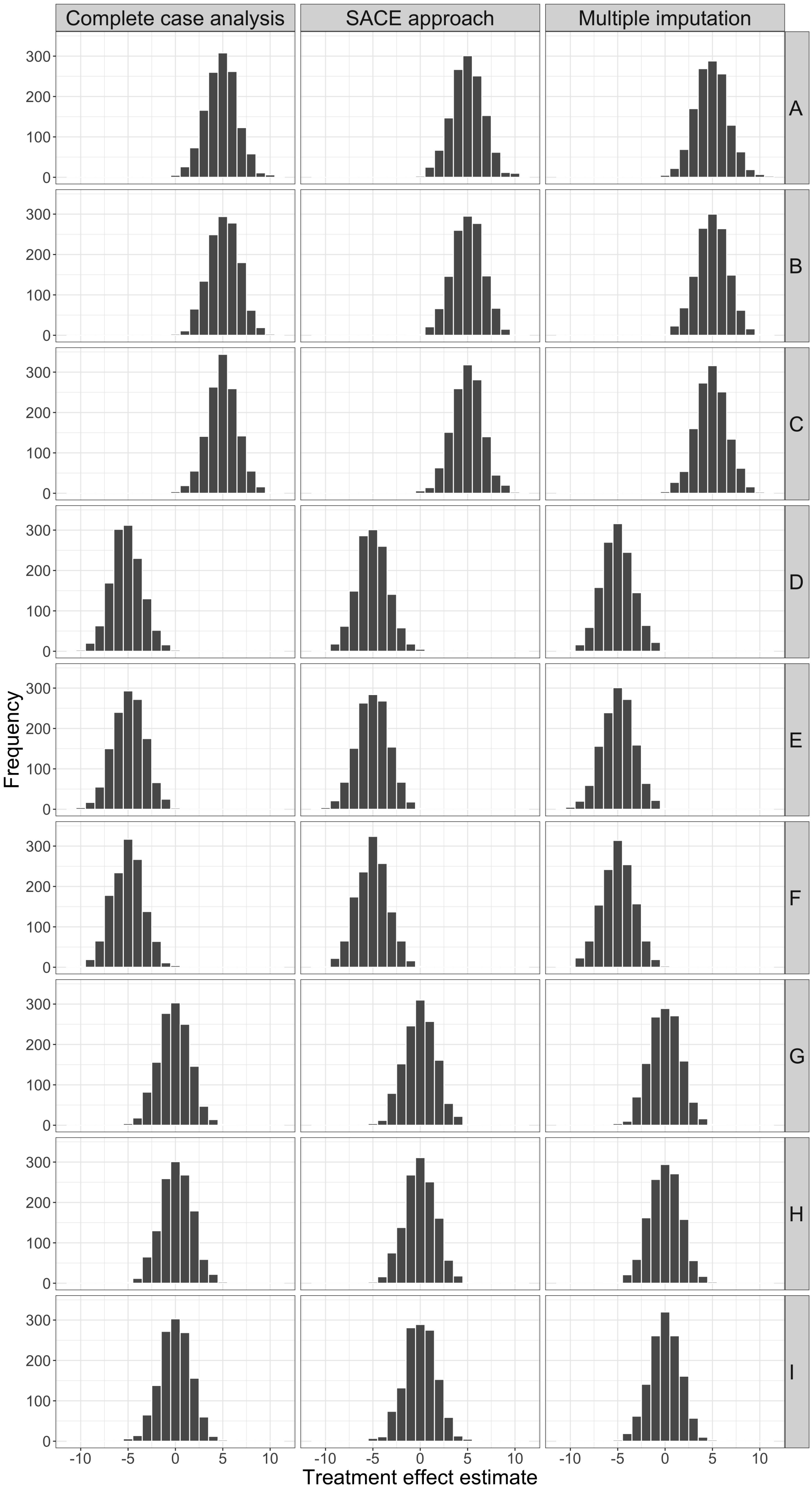

Figure A.1: Distribution of the 1300 treatment effect estimates provided by each method in each scenario of the simulation study.

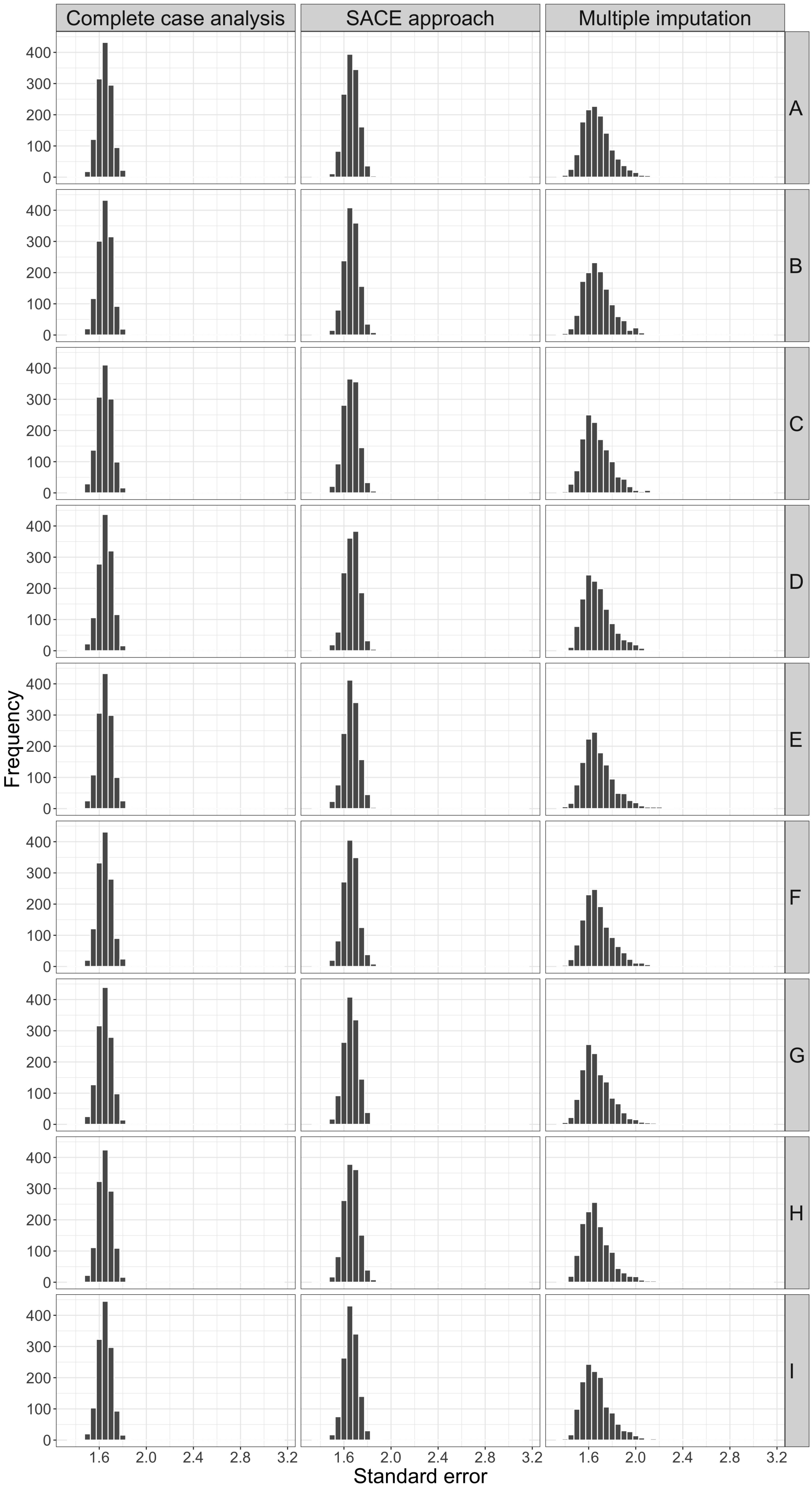

Figure A.2: Distribution of the 1300 standard errors provided by each method in each scenario of the simulation study.

Method	Average of estimates	Empirical SE	Average of SEs
Scenario A			
Complete case analysis	4.865	1.684	1.647
Hayden's method	4.972	1.699	1.662
Multiple imputation analysis	4.916	1.719	1.677
Scenario B			
Complete case analysis	5.109	1.640	1.649
Hayden's method	4.997	1.658	1.663
Multiple imputation analysis	4.970	1.662	1.682
Scenario C			
Complete case analysis	4.980	1.598	1.646
Hayden's method	4.983	1.604	1.659
Multiple imputation analysis	4.935	1.638	1.672
Scenario D			
Complete case analysis	-5.138	1.601	1.651
Hayden's method	-5.025	1.612	1.667
Multiple imputation analysis	-4.992	1.640	1.678
Scenario E			
Complete case analysis	-4.900	1.654	1.649
Hayden's method	-5.025	1.672	1.663
Multiple imputation analysis	-4.961	1.679	1.684
Scenario F			
Complete case analysis	-5.061	1.635	1.646
Hayden's method	-5.061	1.647	1.660
Multiple imputation analysis	-5.005	1.680	1.682
Scenario G			
Complete case analysis	-0.130	1.632	1.646
Hayden's method	-0.014	1.656	1.660
Multiple imputation analysis	-0.015	1.622	1.672
Scenario H			
Complete case analysis	0.107	1.627	1.648
Hayden's method	-0.025	1.640	1.663
Multiple imputation analysis	-0.004	1.633	1.670
Scenario I			
Complete case analysis	-0.007	1.624	1.647
Hayden's method	-0.015	1.636	1.661
Multiple imputation analysis	-0.005	1.614	1.666

Table A.2: Summary measures of the analyses of simulated data sets ($n_{\text{sim}} = 1300$ for each scenario) performed by three methods.

Method	θ_1 (MC SE)	θ_2 (MC SE)	θ_3 (MC SE)
Scenario A			
Complete case analysis	-0.037 (0.047)	-0.146 (0.047)	-0.159 (0.047)
Hayden's method	0.070 (0.047)	-0.039 (0.047)	-0.052 (0.047)
Multiple imputation analysis	0.014 (0.048)	-0.094 (0.048)	-0.108 (0.048)
Scenario B			
Complete case analysis	-0.015 (0.045)	0.101 (0.045)	0.097 (0.045)
Hayden's method	-0.127 (0.046)	-0.012 (0.046)	-0.015 (0.046)
Multiple imputation analysis	-0.154 (0.046)	-0.038 (0.046)	-0.042 (0.046)
Scenario C			
Complete case analysis	-0.027 (0.044)	-0.028 (0.044)	-0.036 (0.044)
Hayden's method	-0.025 (0.044)	-0.026 (0.044)	-0.033 (0.044)
Multiple imputation analysis	-0.073 (0.045)	-0.073 (0.045)	-0.081 (0.045)
Scenario D			
Complete case analysis	-0.047 (0.044)	-0.162 (0.044)	-0.153 (0.044)
Hayden's method	0.066 (0.045)	-0.049 (0.045)	-0.040 (0.045)
Multiple imputation analysis	0.099 (0.045)	-0.016 (0.045)	-0.007 (0.045)
Scenario E			
Complete case analysis	-0.055 (0.046)	0.057 (0.046)	0.053 (0.046)
Hayden's method	-0.180 (0.046)	-0.068 (0.046)	-0.072 (0.046)
Multiple imputation analysis	-0.116 (0.047)	-0.004 (0.047)	-0.008 (0.047)
Scenario F			
Complete case analysis	-0.054 (0.045)	-0.052 (0.045)	-0.055 (0.045)
Hayden's method	-0.054 (0.046)	-0.052 (0.046)	-0.054 (0.046)
Multiple imputation analysis	0.003 (0.047)	0.005 (0.047)	0.002 (0.047)
Scenario G			
Complete case analysis	-0.024 (0.045)	-0.139 (0.045)	-0.138 (0.045)
Hayden's method	0.092 (0.046)	-0.022 (0.046)	-0.022 (0.046)
Multiple imputation analysis	0.091 (0.045)	-0.023 (0.045)	-0.023 (0.045)
Scenario H			
Complete case analysis	-0.024 (0.045)	0.090 (0.045)	0.095 (0.045)
Hayden's method	-0.156 (0.045)	-0.042 (0.045)	-0.037 (0.045)
Multiple imputation analysis	-0.136 (0.045)	-0.021 (0.045)	-0.017 (0.045)
Scenario I			
Complete case analysis	-0.013 (0.045)	-0.005 (0.045)	-0.016 (0.045)
Hayden's method	-0.022 (0.045)	-0.014 (0.045)	-0.025 (0.045)
Multiple imputation analysis	-0.011 (0.045)	-0.003 (0.045)	-0.014 (0.045)

Table A.3: Bias of the three methods compared in the simulation study with respect to the three estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died), with the corresponding Monte Carlo standard error.

Method	θ_1 (MC SE)	θ_2 (MC SE)	θ_3 (MC SE)
Scenario A			
Complete case analysis	2.835 (0.116)	2.855 (0.116)	2.859 (0.116)
Hayden's method	2.891 (0.117)	2.887 (0.116)	2.889 (0.116)
Multiple imputation analysis	2.953 (0.121)	2.962 (0.120)	2.965 (0.120)
Scenario B			
Complete case analysis	2.687 (0.105)	2.697 (0.105)	2.696 (0.105)
Hayden's method	2.761 (0.108)	2.745 (0.107)	2.746 (0.107)
Multiple imputation analysis	2.784 (0.109)	2.762 (0.108)	2.762 (0.108)
Scenario C			
Complete case analysis	2.554 (0.103)	2.554 (0.103)	2.554 (0.103)
Hayden's method	2.571 (0.105)	2.571 (0.105)	2.572 (0.105)
Multiple imputation analysis	2.685 (0.107)	2.686 (0.107)	2.687 (0.107)
Scenario D			
Complete case analysis	2.563 (0.100)	2.587 (0.100)	2.584 (0.100)
Hayden's method	2.602 (0.102)	2.600 (0.101)	2.599 (0.101)
Multiple imputation analysis	2.696 (0.107)	2.687 (0.105)	2.687 (0.105)
Scenario E			
Complete case analysis	2.736 (0.103)	2.736 (0.103)	2.735 (0.103)
Hayden's method	2.827 (0.107)	2.799 (0.106)	2.800 (0.106)
Multiple imputation analysis	2.829 (0.108)	2.815 (0.107)	2.815 (0.107)
Scenario F			
Complete case analysis	2.675 (0.101)	2.674 (0.101)	2.675 (0.101)
Hayden's method	2.714 (0.103)	2.714 (0.103)	2.714 (0.103)
Multiple imputation analysis	2.821 (0.107)	2.821 (0.107)	2.821 (0.107)
Scenario G			
Complete case analysis	2.661 (0.101)	2.679 (0.101)	2.679 (0.101)
Hayden's method	2.749 (0.104)	2.741 (0.104)	2.741 (0.104)
Multiple imputation analysis	2.636 (0.100)	2.628 (0.100)	2.628 (0.100)
Scenario H			
Complete case analysis	2.644 (0.099)	2.652 (0.100)	2.653 (0.100)
Hayden's method	2.713 (0.101)	2.690 (0.100)	2.690 (0.100)
Multiple imputation analysis	2.684 (0.103)	2.666 (0.102)	2.666 (0.102)
Scenario I			
Complete case analysis	2.634 (0.105)	2.634 (0.105)	2.634 (0.105)
Hayden's method	2.676 (0.107)	2.676 (0.107)	2.676 (0.107)
Multiple imputation analysis	2.605 (0.104)	2.605 (0.104)	2.605 (0.104)

Table A.4: MSE of the three methods compared in the simulation study with respect to the three estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died), with the corresponding Monte Carlo standard error.

Method	θ_1 (MC SE)	θ_2 (MC SE)	θ_3 (MC SE)
Scenario A			
Complete case analysis	0.942 (0.006)	0.942 (0.006)	0.942 (0.007)
Hayden's method	0.943 (0.006)	0.947 (0.006)	0.945 (0.006)
Multiple imputation analysis	0.938 (0.007)	0.942 (0.007)	0.942 (0.007)
Scenario B			
Complete case analysis	0.956 (0.006)	0.954 (0.006)	0.954 (0.006)
Hayden's method	0.951 (0.006)	0.954 (0.006)	0.953 (0.006)
Multiple imputation analysis	0.952 (0.006)	0.949 (0.006)	0.949 (0.006)
Scenario C			
Complete case analysis	0.950 (0.006)	0.950 (0.006)	0.950 (0.006)
Hayden's method	0.954 (0.006)	0.954 (0.006)	0.954 (0.006)
Multiple imputation analysis	0.951 (0.006)	0.950 (0.006)	0.949 (0.006)
Scenario D			
Complete case analysis	0.954 (0.006)	0.952 (0.006)	0.952 (0.006)
Hayden's method	0.953 (0.006)	0.954 (0.006)	0.955 (0.006)
Multiple imputation analysis	0.945 (0.006)	0.947 (0.006)	0.947 (0.006)
Scenario E			
Complete case analysis	0.948 (0.006)	0.949 (0.006)	0.949 (0.006)
Hayden's method	0.952 (0.006)	0.953 (0.006)	0.952 (0.006)
Multiple imputation analysis	0.947 (0.006)	0.949 (0.006)	0.949 (0.006)
Scenario F			
Complete case analysis	0.952 (0.006)	0.952 (0.006)	0.952 (0.006)
Hayden's method	0.950 (0.006)	0.950 (0.006)	0.950 (0.006)
Multiple imputation analysis	0.948 (0.006)	0.948 (0.006)	0.948 (0.006)
Scenario G			
Complete case analysis	0.955 (0.006)	0.957 (0.006)	0.957 (0.006)
Hayden's method	0.954 (0.006)	0.952 (0.006)	0.952 (0.006)
Multiple imputation analysis	0.963 (0.005)	0.965 (0.005)	0.965 (0.005)
Scenario H			
Complete case analysis	0.951 (0.006)	0.955 (0.006)	0.955 (0.006)
Hayden's method	0.954 (0.006)	0.951 (0.006)	0.951 (0.006)
Multiple imputation analysis	0.955 (0.006)	0.953 (0.006)	0.952 (0.006)
Scenario I			
Complete case analysis	0.961 (0.005)	0.961 (0.005)	0.961 (0.005)
Hayden's method	0.959 (0.005)	0.959 (0.005)	0.959 (0.005)
Multiple imputation analysis	0.956 (0.006)	0.956 (0.006)	0.957 (0.006)

Table A.5: Coverage of the three methods compared in the simulation study with respect to the three estimands θ_1 (treatment effect on survivors), θ_2 (survivor average causal effect) and θ_3 (treatment effect on all patients, as if no one had died), with the corresponding Monte Carlo standard error.

A.4 Protocol of the simulation study

A.4.1 Aims and objectives

The aim of our simulation study is to evaluate the performance of three different methods in estimating the treatment effect from an RCT comparing an intervention with placebo, when the primary outcome is truncated by death for a relevant proportion of the patients randomized.
Estimates of the treatment effect will be derived by complete case analysis, by Hayden's method (Hayden *et al.*, 2005), and by an analysis involving multiple imputation of missing values. The estimates will be compared in terms of bias, mean square error and coverage. Since the compared methods estimate different quantities, we will discuss the results with reference to the topic of estimands. The estimands may be described as follows:

- Complete case analysis: treatment effect on the survivors, θ_1.
- Hayden's method: survivor average causal effect (SACE), which is the treatment effect on the subgroup of patients who would have survived under both treatments, θ_2.
- Analysis of multiply imputed data sets: treatment effect on all patients, as if no one had died (assuming all patients could have survived), θ_3.

As a subset of the randomized patients is analyzed, complete case analysis and Hayden's method do not conform with the ITT principle. In contrast, multiple imputation conforms with the ITT principle but creates data that could not be observed.
For each of the scenarios investigated, we will use one very large data set (of sample size 1'000'000) to obtain the "true" values of θ_1, θ_2, θ_3, and many smaller data sets, with realistic sample size, to compare the performance of the three methods.

A.4.2 Simulation procedures

Level of dependence between simulated datasets

For each scenario, we will simulate many data sets that will be analyzed by all three methods. We will employ "moderately independent" simulations (Burton *et al.*, 2006): we will use the same set of simulated independent data sets to compare the three statistical methods, but will generate a different set of independent data sets for each scenario investigated.

Allowance for failures

Failures may occur primarily for Hayden's method since involves fitting logistic regression models on survival in both treatment arms, which are used to predict survival probabilities of patients under the respective other treatment. Logistic regression models may fail to converge or yield very large standard errors in case of strata with few events. Samples with failures will be discarded and replaced. However, if we discard samples with failures, we may bias our results, as the simulated data sets will not cover the whole range we intend. Therefore, we implement Firth logistic regression with added covariate (FLAC, Puhr *et al.*, 2017) to reduce this problem. We currently do not expect failures for the other two methods. The number of failures, the reason and the method for which it occurred will be recorded.

Software to perform simulations

The simulation study will be performed in R version 3.6.1, with the exception of the simulation of the large data sets, which will be executed on the UZH math server, in R version 3.5.0. We will make use of the packages *boot*, *fabricatr*, *MASS*, *mice*, *truncnorm*.

Random number generator to use

We will use the default random number generator implemented in R, which is the "Mersenne-Twister" (Matsumoto and Nishimura, 1998).

Specification of starting seeds

For each simulation we will use starting seeds that are separated by at least the sample size of the simulated data sets. For example, if each simulated data set has a sample size of 500, then each simulation requires 500 random numbers, therefore the starting seed for each simulation should be separated by at least 500 (Burton *et al.*, 2006). We will generate all the necessary seeds in advance, we will store them in a matrix and then will choose them one after the other for each simulation. We will generate 2000 seeds for each scenario, more than n_{sim}, in order to have them available in case of failures.

A.4.3 Methods for generating the datasets

We will use the Epo trial (Natalucci *et al.*, 2016) as motivating example, but will modify some aspects of the data set to create different scenarios for our simulation study.

We will use a selection of the observed baseline covariates and the correlation structure among them as given, and will randomly allocate patients to treatment (rhEPO vs. Placebo). This first step will be the same in all scenarios.

We will then simulate the outcome, using the associations of the outcome with the covariates as observed in the Epo trial. In addition, we will model a treatment effect on the outcome which will vary between scenarios (positive effect, negative effect, no effect; note that the observed treatment effect in the Epo trial was neglectable).

Further, we will simulate survival up to 2 years follow-up, using the associations of death with the covariates as observed in the Epo trial and model a treatment effect on survival which will vary between scenarios (positive effect, negative effect, no effect). The proportion of children who died before 2 years of age will be increased for our simulation study, from an observed rate of around 6 % in the Epo trial to 15 %. The latter rate is expected in the EpoRepair trial, a similar, still ongoing RCT on long-term neurocognitive outcomes of very preterm infants suffering from intraventricular hemorrhage (Rüegger *et al.*, 2015).

Description of the Epo trial data

Baseline variables important for survival:

- Gestational age in days: normal (mean: 204.0, sd: 11.7)
- Head circumference at birth in cm: normal (mean: 26.8, sd: 2.2)
- Apgar score at 5 min after birth: categorical (categories from 0 to 10: 0% 0, 1% 1, 1% 2, 3% 3, 5% 4, 5% 5, 10% 6, 17% 7, 24% 8, 29% 9, 5% 10)

We will not simulate the variables sex and weight at birth: the first does not seem to be important for survival, while the second is strongly correlated with head circumference.
Treatment: binary (48% placebo, 52% rhEPO)
Outcome (MDI): normal (mean: 93.9, sd: 16.9)
Survival: binary (6% no, 94% yes)

Simulation of the data

For all scenarios we will simulate data sets with a sample size of 500, which is similar to the sample size of our real data set.

We will first simulate the Apgar score, with the same distribution as observed in the Epo trial. From the Epo trial data, we will also derive the mean gestational age and the mean head circumference for each Apgar score category, and we will use them to simulate gestational age and head circumference as multivariate normal variables in each Apgar score category. We will use the same variance (again derived from the Epo trial data) for all categories, to avoid a too data-driven simulation.

The patients will be randomly assigned to treatment. One half of the patients (250 patients) will receive rhEPO, one half will receive placebo.

The outcome will be simulated using the association with the covariates of the Epo trial and a treatment effect on the outcome of -5, 0 or 5, depending on the scenario considered.

Survival will be simulated using the association with the covariates of the Epo trial and a treatment effect on survival of -0.1, 0 or 0.1, depending on the scenario considered. The specified values are on the logit scale, and correspond to odds ratios of 0.9, 1 and 1.1, respectively. The overall survival probability will be 85%.

A.4.4 Scenarios to be investigated

Table A.6 shows the treatment effects on outcome and survival that we will assess in our simulation study. These effects will be examined in a fully factorial arrangement.

Table A.6: Overview of the planned simulation scenarios.

Scenario	Treatment effect on Outcome	Treatment effect on Survival
A	positive (MDI increased)	positive (survival probability higher)
B	positive (MDI increased)	negative (survival probability lower)
C	positive (MDI increased)	no effect
D	negative (MDI decreased)	positive (survival probability higher)
E	negative (MDI decreased)	negative (survival probability lower)
F	negative (MDI decreased)	no effect
G	no effect	positive (survival probability higher)
H	no effect	negative (survival probability lower)
I	no effect	no effect

For simplicity, we will assume no drop-outs due to withdrawal of informed consent or loss to follow-up for other reasons than death.

If treatment increases survival of less healthy individuals (possible in scenarios A, D, G), complete case analysis may underestimate the SACE (Colantuoni *et al.*, 2018). In fact, the survivors in the control group may be healthier than the survivors in the treatment group. If treatment decreases survival (scenarios B, E, H), the monotonicity assumption (survival under treatment is always at least as good as survival under control) does not hold. In reality this can happen, for example, if a treatment has unexpected negative side effects for a subgroup of people (Rubin, 2006). In this case, we expect the survivors in the treatment group to be healthier than the survivors in the placebo, and thus complete case analysis to overestimate the SACE.

If treatment has no effect on survival, complete case analysis provides an estimate of the causal effect of the treatment on the always survivors (Colantuoni *et al.*, 2018). Therefore, we expect complete case analysis and Hayden's method to yield similar results in scenarios C, F, I.

A.4.5 Statistical methods to be evaluated

1. "Naive" estimator of the treatment effect on survivors (complete cases analysis).

2. Hayden's estimator of the SACE (survivor average causal effect).

3. Estimator of the treatment effect on all patients with multiple imputation. 5 imputed data sets will be generated for each simulation and the estimates will be combined using Rubin's rule.

A.4.6 Estimates to be stored for each simulation and summary measures to be calculated over all simulations

For each simulation $i = \{1, \ldots, n_{\text{sim}}\}$ and each method $j = \{1, 2, 3\}$ we will store:

- the seed s_i,
- the proportion of deaths d_i
- the treatment effect estimate $\hat{\theta}_{ij}$,
- the standard error of the treatment effect estimate $\text{SE}(\hat{\theta}_{ij})$.
- the effective number of patients analyzed m_{ij}.

m_{ij} will be the number of patients alive for the complete case method, the total number of patients for multiple imputation and a non-integer number for Hayden's method (the sum of the survival probabilities under the not assigned treatment of the patients alive).

Once all simulations will be performed, we will compute for all combinations of scenario and method j:

- the average of the treatment effect estimates, $\bar{\hat{\theta}}_j = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_{ij}$,
- the empirical SE (standard deviation of the estimates) $\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_{ij} - \bar{\hat{\theta}}_j)^2}$,
- the average of the estimated within simulation SEs of the estimates, $\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \text{SE}(\hat{\theta}_{ij})$,

following the formulas of Burton *et al.* (2006).

A.4.7 Number of simulations to be performed

The number of simulations to perform for each scenario is based on the accuracy of the estimate of interest. The number of simulations n_{sim} is calculated using

$$n_{\text{sim}} = \left(\frac{Z_{1-\alpha/2} \cdot \sigma}{\delta} \right)^2$$

where δ is the specified level of accuracy of the estimate of interest we are willing to accept, i.e. the permissible difference from the true value, $Z_{1-\alpha/2}$ is the quantile of the standard normal distribution and σ^2 is the variance of the parameter of interest, which can be obtained from the real data (Burton *et al.*, 2006).

From the Epo trial we estimate a standard error of the SACE of $\sigma = 1.763$ (which may vary depending on the model used) and a standard error of the naive estimator of $\sigma = (2.5 + 4.5)/(2 * 1.96) = 1.786$ (confidence interval also reported in Natalucci *et al.*, 2016). We take the larger value of 1.8 to ensure an adequate number of simulations. Since we simulate data with a treatment effect of 5 (or 0 or -5) and we aim to have an accuracy of 2%, we have to perform at least 1245 simulations (if we would aim to have an accuracy of 1%, we would perform at least 4979 simulations). We decided to round it up and to perform $n_{\text{sim}} = 1300$ simulations.

Table A.7: Performance measures of method j with respect to estimand θ_k: estimates and Monte Carlo standard errors.

Performance measure	Estimate	Monte Carlo SE of estimate
Bias	$\frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}}\hat{\theta}_{ij}-\theta_k$	$\sqrt{\frac{1}{n_{\text{sim}}(n_{\text{sim}}-1)}\sum_{i=1}^{n_{\text{sim}}}(\hat{\theta}_{ij}-\bar{\hat{\theta}}_j)^2}$
Mean square error (MSE)	$\frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}}(\hat{\theta}_{ij}-\theta_k)^2$	$\sqrt{\frac{\sum_{i=1}^{n_{\text{sim}}}\left[(\hat{\theta}_{ij}-\theta_k)^2-\widehat{\text{MSE}}\right]^2}{n_{\text{sim}}(n_{\text{sim}}-1)}}$
Coverage	$\frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}}\mathbb{1}(\hat{\theta}_{low,ij}\leq\theta_k\leq\hat{\theta}_{upp,ij})$	$\sqrt{\frac{\widehat{\text{Coverage}}(1-\widehat{\text{Coverage}})}{n_{\text{sim}}}}$

A.4.8 Criteria to evaluate the performance of statistical methods for different scenarios

As mentioned above, the "true" value of each θ_k, $k=\{1,2,3\}$, will be computed using a data set of sample size 1'000'000. We expect the true value of θ_3 to be the closest to the treatment effect that we model on the outcome. The performance of each statistical method j with respect to the estimand θ_k on the smaller data sets will be evaluated using the criteria summarized in Table A.7, following the definitions given in Morris *et al.* (2019). In formulas of Table A.7, θ_k is intended as the true value of the estimand θ_k.

A.4.9 Presentation of the simulation results

For each scenario and each performance measure, we will summarize the results in tables combining the different methods with the different estimands. An example for the presentation of the results in terms of bias is given in Table A.8. The entries in italics display the bias magnitude that we expect and they will be substituted with the numeric values that will be found. The same kind of table will be created for the mean square error and the coverage.

Table A.8: Bias for scenario A, where a treatment effect on survival of 0.1 and a treatment effect on the outcome of 5 are modeled.

	θ_1	θ_2	θ_3
Complete case analysis	*small*	*larger*	*larger*
Hayden's method	*larger*	*small*	*larger*
Multiple imputation analysis	*larger*	*larger*	*small*