Bootstrap Inference When Using Multiple Imputation

M. Schomaker[*,a] and **C. Heumann**[b]

[a]University of Cape Town, Centre for Infectious Disease Epidemiology & Research, Falmouth Building, Observatory, 7925, South Africa

[b]Christian Heumann, Institut für Statistik, Ludwig-Maximilians Universität München, München, Germany

Abstract

Many modern estimators require bootstrapping to calculate confidence intervals because either no analytic standard error is available or the distribution of the parameter of interest is non-symmetric. It remains however unclear how to obtain valid bootstrap inference when dealing with multiple imputation to address missing data. We present four methods which are intuitively appealing, easy to implement, and combine bootstrap estimation with multiple imputation. We show that three of the four approaches yield valid inference, but that the performance of the methods varies with respect to the number of imputed data sets and the extent of missingness. Simulation studies reveal the behavior of our approaches in finite samples. A topical analysis from HIV treatment research, which determines the optimal timing of antiretroviral treatment initiation in young children, demonstrates the practical implications of the four methods in a sophisticated and realistic setting. This analysis suffers from missing data and uses the *g*-formula for inference, a method for which no standard errors are available.

Keywords

missing data; resampling; g-methods; causal inference; HIV

1. Introduction

Multiple imputation (MI) is a popular method to address missing data. Based on assumptions about the data distribution (and the mechanism which gives rise to the missing data) missing values can be imputed by means of draws from the posterior predictive distribution of the unobserved data given the observed data. This procedure is repeated to create M imputed data sets, the analysis is then conducted on each of these data sets and the M results (M point and M variance estimates) are combined by a set of simple rules [1].

During the last 30 years a lot of progress has been made to make MI useable for different settings: implementations are available in several software packages [2, 3, 4, 5], review articles provide guidance to deal with practical challenges [6, 7, 8], non-normal –possibly

[*]Correspondence to: Michael Schomaker, University of Cape Town, Centre for Infectious Disease Epidemiology & Research, Falmouth Building, Observatory, 7925, South Africa. michael.schomaker@uct.ac.za.

categorical–variables can often successfully be imputed [9, 3, 6], useful diagnostic tools have been suggested [3, 10], and first attempts to address longitudinal data and other complicated data structures have been made [11, 4].

While both opportunities and challenges of multiple imputation are discussed in the literature, we believe an important consideration regarding the inference after imputation has been neglected so far: if there is no analytic or no ideal solution to obtain standard errors for the parameters of the analysis model, and nonparametric bootstrap estimation is used to estimate them, it is unclear how to obtain valid inference – in particular how to obtain appropriate confidence intervals. Moreover, bootstrap estimation is also often used when a parameter's distribution is assumed to be non-normal and bootstrap inference with missing data is then not clear either. As we will explain below, many modern statistical concepts, often applied to inform policy guidelines or enhance practical developments, rely on bootstrap estimation. It is therefore necessary to have guidance for bootstrap estimation for multiply imputed data.

In general, one can distinguish between two approaches for bootstrap inference when using multiple imputation: with the first approach, M imputed datsets are created and bootstrap estimation is applied to each of them; or, alternatively, B bootstrap samples of the original data set (including missing values) are drawn and in each of these samples the data are multiply imputed. For the former approach one could use bootstrapping to estimate the standard error in each imputed data set and apply the standard MI combining rules; alternatively, the $B \times M$ estimates could be pooled and 95% confidence intervals could be calculated based on the 2.5^{th} and 97.5^{th} percentiles of the respective empirical distribution. For the latter approach either multiple imputation combining rules can be applied to the imputed data of each bootstrap sample to obtain B point estimates which in turn may be used to construct confidence intervals; or the $B \times M$ estimates of the pooled data are used for interval estimation.

To the best of our knowledge, the consequences of using the above approaches have not been studied in the literature before. The use of the bootstrap in the context of missing data has often been viewed as a frequentist alternative to multiple imputation [12], or an option to obtain confidence intervals after single imputation [13]. The bootstrap can also be used to create multiple imputations [14]. However, none of these studies have addressed the construction of bootstrap confidence intervals when data needs to be multiply imputed because of missing data. As emphasized above, this is however of particularly great importance when standard errors of the analysis model cannot be calculated easily, for example for causal inference estimators (e.g. the g-formula).

It is not surprising that the bootstrap has nevertheless been combined with multiple imputation for particular analyses. Multiple imputation of bootstrap samples has been implemented in [15, 16, 17, 18], whereas bootstrapping the imputed data sets was preferred by [19, 20, 21]. Other work doesn't offer all details of the implementation [22]. All these analyses give however little justification for the chosen method and for some analyses important details on how the confidence intervals were calculated are missing; it seems that

pragmatic reasons as well as computational efficiency typically guide the choice of the approach. None of the studies offer a statistical discussion of the chosen method.

The present article demonstrates the implications of different methods which combine bootstrap inference with multiple imputation. It is novel in that it introduces four different, intuitively appealing, bootstrap confidence intervals for data which require multiple imputation, illustrates their intrinsic features, and argues which of them is to be preferred.

Section 2 introduces our motivating analysis of causal inference in HIV research. The different methodological approaches are described in detail in Section 3 and are evaluated by means of both numerical investigations (Section 4) and theoretical considerations (Section 6). The implications of the different approaches are further emphasized in the data analysis of Section 5. We conclude in Section 7.

2. Motivation

During the last decade the World Health Organization (WHO) updated their recommendations on the use of antiretroviral drugs for treating and preventing HIV infection several times. In the past, antiretroviral treatment (ART) was only given to a child if his/her measurements of CD4 lymphocytes fell below a critical value or if a clinically severe event (such as tuberculosis or persistent diarrhoea) occurred. Based on both increased knowledge from trials and causal modeling studies, as well as pragmatic and programmatic considerations, these criteria have been gradually expanded to allow earlier treatment initiation in children: in 2013 it was suggested that all children who present under the age of 5 are treated immediately, while for older children CD4-based criteria still existed. By the end of 2015 WHO decided to recommend immediate treatment initiation in all children and adults. ART has shown to be effective and to reduce mortality in infants and adults [23, 24, 25], but concerns remain due to a potentially increased risk of toxicities, early development of drug resistance, and limited future options for people who fail treatment.

It remains therefore important to investigate the effect of different treatment initiation rules on mortality, morbidity and child development outcomes; however given the shift in ART guidelines towards earlier treatment initiation it is not ethically possible anymore to conduct a trial which answers this question in detail. Thus, observational data can be used to obtain the relevant estimates. Methods such as inverse probability weighting of marginal structural models, the g-computation formula, and targeted maximum likelihood estimation can be used to obtain estimates in complicated longitudinal settings where time-varying confounders affected by prior treatment are present — such as, for example, CD4 count which influences both the probability of ART initiation and outcome measures [26, 27].

In situations where treatment rules are dynamic, i.e. where they are based on a time-varying variable such as CD4 lymphocyte count, the g-computation formula [28] is *the* intuitive method to use. It is computationally intensive and allows the comparison of outcomes for different treatment options; confidence intervals are typically based on nonparametric bootstrap estimation. However, in resource limited settings data may be missing for administrative, logistic, and clerical reasons, as well as due to loss to follow-up and missed

clinic visits. Depending on the underlying assumptions about the reasons for missing data, this problem can either be addressed by the g-formula directly or by using multiple imputation. However, it is not immediately clear how to combine multiple imputation with bootstrap estimation to obtain valid confidence intervals.

3. Methodological Framework

Let $\mathscr{D}$ be a $n \times (p+1)$ data matrix consisting of an outcome variable $\mathbf{y} = (y_1, \ldots, y_n)'$ and covariates $\mathbf{X_j} = (X_{1j}, \ldots, X_{nj})'$, $j = 1, \ldots, p$. The $1 \times p$ vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ contains the ith observation of each of the p covariates and $\mathbf{X} = (\mathbf{x}_1{}', \ldots, \mathbf{x}_n{}')'$ is the matrix of all covariates. Suppose we are interested in estimating $\theta = (\theta_1, \ldots, \theta_k)'$, $k \geq 1$, which may be a regression coefficient, an odds ratio, a factor loading, or a counterfactual outcome. If some data are missing, making the data matrix to consist of both observed and missing values, $\mathscr{D} = \left\{ \mathscr{D}^{\text{obs}}, \mathscr{D}^{\text{mis}} \right\}$, and the missingness mechanism is ignorable, valid inference for θ can be obtained using multiple imputation. Following Rubin [1] we regard valid inference to mean that the point estimate $\hat{\theta}$ for θ is approximately unbiased and that interval estimates are randomization valid in the sense that actual interval coverage equals the nominal interval coverage.

Under multiple imputation M augmented sets of data are generated, and the imputations (which replace the missing values) are based on draws from the predictive posterior distribution of the missing data given the observed data $p(\mathscr{D}^{\text{mis}} | \mathscr{D}^{\text{obs}}) = \int p(\mathscr{D}^{\text{mis}} | \mathscr{D}^{\text{obs}}; \vartheta) p(\vartheta | \mathscr{D}^{\text{obs}}) d\vartheta$, or an approximation thereof. The point estimate for θ is

$$\hat{\bar{\theta}}_{\text{MI}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \quad (1)$$

where $\hat{\theta}_m$ refers to the estimate of θ in the mth imputed set of data $\mathscr{D}^{(m)}$, $m = 1, \ldots, M$. Variance estimates can be obtained using the between imputation covariance $\hat{V} = (M-1)^{-1} \sum_m (\hat{\theta}_m - \hat{\bar{\theta}}_{\text{MI}})(\hat{\theta}_m - \hat{\bar{\theta}}_{\text{MI}})'$ and the average within imputation covariance $\hat{W} = M^{-1} \sum_m \widehat{\text{Cov}}(\hat{\theta}_m)$:

$$\widehat{\text{Cov}}(\hat{\bar{\theta}}_{\text{MI}}) = \hat{W} + \frac{M+1}{M} \hat{V} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\text{Cov}}(\hat{\theta}_m) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\bar{\theta}}_{\text{MI}})(\hat{\theta}_m - \hat{\bar{\theta}}_{\text{MI}})'. \quad (2)$$

For the scalar case this equates to

$$\widehat{\text{Var}}(\hat{\bar{\theta}}_{\text{MI}}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\text{Var}}(\hat{\theta}_m) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\bar{\theta}}_{\text{MI}})^2.$$

To construct confidence intervals for $\hat{\bar{\theta}}_{\mathrm{MI}}$ in the scalar case, it may be assumed that

$\widehat{\mathrm{Var}}(\hat{\bar{\theta}}_{\mathrm{MI}})^{-\frac{1}{2}}(\hat{\bar{\theta}}_{\mathrm{MI}} - \theta)$ follows a t_R-distribution with approximately

$R = (M-1)[1 + \{M\widehat{W}/(M+1)\widehat{V}\}]^2$ degrees of freedom [29], though there are alternative approximations, especially for small samples [30]. Note that for reliable variance estimation M should not be too small; see White et al. [6] for some rules of thumb.

Consider the situation where there is no analytic or no ideal solution to estimate $\mathrm{Cov}(\hat{\theta}_m)$, for example when estimating the treatment effect in the presence of time-varying confounders affected by prior treatment using g-methods [31, 26]. If there are no missing data, bootstrap percentile confidence intervals may offer a solution: based on B bootstrap samples $\mathscr{D}_b^*$, $b = 1, \ldots, B$, we obtain B point estimates $\hat{\theta}_b^*$. Consider the ordered set of estimates $\Theta_B^* = \left\{\hat{\theta}_{(b)}^*; b = 1, \ldots, B\right\}$, where $\hat{\theta}_{(1)}^* < \hat{\theta}_{(2)}^* < \ldots < \hat{\theta}_{(B)}^*$; the bootstrap $1 - 2\alpha\%$ confidence interval for θ is then defined as

$$[\hat{\theta}_{\mathrm{lower}}; \hat{\theta}_{\mathrm{upper}}] = [\hat{\theta}^{*,\alpha}; \hat{\theta}^{*,1-\alpha}]$$

where $\hat{\theta}^{*,\alpha}$ denotes the α-percentile of the ordered bootstrap estimates Θ_B^*. However, in the presence of missing data the construction of confidence intervals is not immediately clear as $\hat{\theta}$ corresponds to M estimates $\hat{\theta}_1, \ldots, \hat{\theta}_M$, i.e. $\hat{\theta}_m$ is the point estimate calculated from the mth imputed data set. It seems intuitive to consider the following four approaches:

- **Method 1, MI Boot (pooled sample [PS]):** Multiple imputation is utilized for the data set $\mathscr{D} = \{\mathscr{D}^{\mathrm{obs}}, \mathscr{D}^{\mathrm{mis}}\}$. For each of the M imputed data sets $\mathscr{D}_m$, B bootstrap samples are drawn which yields $M \times B$ data sets $\mathscr{D}_{m,b}^*$; $b = 1, \ldots, B$; $m = 1, \ldots, M$. In each of these data sets the quantity of interest is estimated, that is $\hat{\theta}_{m,b}^*$. The pooled sample of ordered estimates $\Theta_{\mathrm{MIBP}}^* = \{\hat{\theta}_{(m,b)}^*; b = 1, \ldots, B; m = 1, \ldots, M\}$ is used to construct the $1 - 2\alpha\%$ confidence interval for θ:

$$[\hat{\theta}_{\mathrm{lower}}; \hat{\theta}_{\mathrm{upper}}]_{\mathrm{MIBP}} = [\hat{\theta}_{\mathrm{MIBP}}^{*,\alpha} \hat{\theta}_{\mathrm{MIBP}}^{*,1-\alpha}]; \quad (3)$$

  where $\hat{\theta}_{\mathrm{MIBP}}^{*,\alpha}$ is the α-percentile of the ordered bootstrap estimates Θ_{MIBP}^*.

- **Method 2, MI Boot:** Multiple imputation is utilized for the data set $\mathscr{D} = \{\mathscr{D}^{\mathrm{obs}}, \mathscr{D}^{\mathrm{mis}}\}$. For each of the M imputed data sets $\mathscr{D}_m$, B bootstrap samples are drawn which yields $M \times B$ data sets $\mathscr{D}_{m,b}^*$; $b = 1, \ldots, B$; $m = 1, \ldots, M$. The bootstrap samples are used to estimate the standard error of (each scalar

component of) $\hat{\theta}_m$ in each imputed data set respectively, i.e.

$\widehat{\mathrm{Var}}(\hat{\theta}_m) = (B-1)^{-1} \sum_b (\hat{\theta}_{m,b} - \bar{\hat{\theta}}_m)^2$ with $\bar{\hat{\theta}}_m = B^{-1} \sum_b \hat{\theta}_{m,b}$. This results in M point estimates (calculated from the imputed, but not yet bootstrapped data), and M standard errors (calculated from the respective bootstrap samples). One can thus calculate standard multiple imputation confidence intervals, possibly based on a t_R-distribution, as explained above.

- **Method 3, Boot MI (pooled sample [PS]):** B bootstrap samples $\mathscr{D}_b^*$ (including missing data) are drawn and multiple imputation is utilized in each bootstrap sample. Therefore, there are $B \times M$ imputed data sets $\mathscr{D}_{b,1}^*, \ldots, \mathscr{D}_{b,M}^*$ which can be used to obtain the corresponding point estimates $\hat{\theta}_{b,m}^*$. The set of the pooled ordered estimates $\Theta_{\mathrm{BMIP}}^* = \{\hat{\theta}_{(b,m)}^*; b = 1, \ldots, B; m = 1, \ldots, M\}$ can then be used to construct the $1 - 2\alpha\%$ confidence interval for θ:

$$[\hat{\theta}_{\mathrm{lower}}; \hat{\theta}_{\mathrm{upper}}]_{\mathrm{BMIP}} = [\hat{\theta}_{\mathrm{BMIP}}^{*,\alpha}; \hat{\theta}_{\mathrm{BMIP}}^{*,1-\alpha}] \quad (4)$$

where $\hat{\theta}_{\mathrm{BMIP}}^{*,\alpha}$ is the α-percentile of the ordered bootstrap estimates Θ_{BMIP}^*.

- **Method 4, Boot MI:** B bootstrap samples $\mathscr{D}_b^*$ (including missing data) are drawn, and each of them is imputed M times. Therefore, there are M imputed data sets, $\mathscr{D}_{b,1}^*, \ldots, \mathscr{D}_{b,M}^*$, which are associated with each bootstrap sample $\mathscr{D}_b^*$. They can be used to obtain the corresponding point estimates $\hat{\theta}_{b,m}^*$. Thus, applying (1) to the estimates of each bootstrap sample yields B point estimates $\bar{\hat{\theta}}_b^* = M^{-1} \sum_m \hat{\theta}_{b,m}^*$ for θ. The set of ordered estimates $\Theta_{\mathrm{BMI}}^* = \{\bar{\hat{\theta}}_{(b)}^*; b = 1, \ldots B\}$ can then be used to construct the $1 - 2P\alpha\%$ confidence interval for θ:

$$[\hat{\theta}_{\mathrm{lower}}; \hat{\theta}_{\mathrm{upper}}]_{\mathrm{BMI}} = [\hat{\theta}_{\mathrm{BMI}}^{*,\alpha}; \hat{\theta}_{\mathrm{BMI}}^{*,1-\alpha}] \quad (5)$$

where $\hat{\theta}_{\mathrm{BMI}}^{*,\alpha}$ is the α-percentile of the ordered bootstrap estimates Θ_{BMI}^*.

While all of the methods described above are straightforward to implement it is unclear if they yield valid inference, i.e. if the actual interval coverage level equals the nominal coverage level. Before we delve into some theoretical and practical considerations we expose some of the intrinsic features of the different interval estimates using Monte Carlo simulations.

4. Simulation Studies

To study the performance of the methods introduced above we consider four simulation settings: a simple one, to ensure that these comparisons are not complicated by the simulation setup; a more complicated one, to study the four methods under a more sophisticated variable dependence structure; a survival analysis setting to allow comparisons beyond a linear regression setup; and a complex longitudinal setting where time-dependent confounding (affected by prior treatment) is present, to allow comparisons to our data analysis in Section 5.

Setting 1

We simulate a normally distributed variable X_1 with mean 0 and variance 1. We then define $\mu_y = 0 + 0.4X_1$ and $\theta = \beta_{\text{true}} = (0, 0.4)'$. The outcome is generated from $N(\mu_y, 2)$ and the analysis model of interest is the linear model. Values of X_1 are defined to be missing with probability

$$\pi_{X_1}(y) = 1 - \frac{1}{(0.25y)^2 + 1}.$$

With this, about 16% of values of X_1 were missing (at random).

Setting 2

The observations for 6 variables are generated using the following normal and Bernoulli distributions: $\mathbf{X}_1 \sim N(0, 1)$, $\mathbf{X}_2 \sim N(0, 1)$, $\mathbf{X}_3 \sim N(0, 1)$, $\mathbf{X}_4 \sim B(0.5)$, $\mathbf{X}_5 \sim B(0.7)$, and $\mathbf{X}_6 \sim B(0.3)$. To model the dependency between the covariates we use a Clayton Copula [32] with a copula parameter of 1 which indicates moderate correlations among the covariates. We then define $\mu_y = 3 - 2X_1 + 3X_3 - 4X_5$ and $\theta = \beta_{\text{true}} = (3, -2, 0, 3, 0, -4, 0)'$. The outcome is generated from $N(\mu_y, 2)$ and the analysis model of interest is the linear model. Values of X_1 and X_3 are defined to be missing (at random) with probabilities

$$\pi_{X_1}(y) = 1 - \frac{1}{(ay)^2 + 1}, \quad \pi_{X_3}(X_4) = 1 - \frac{1}{bX_4^3 + 1.05}.$$

where a and b equate to 0.75 and 0.25 in a low missingness setting (and to 0.4 and 2.5 in a high missingness setting). This yields about 6% and 14% (45% and 38%) of missing values for X_1 and X_3 respectively.

Setting 3

This setting is inspired by the analysis and data in Schomaker et al. [33]. We simulate $\mathbf{X}_1 \sim \text{logN}(4.286, 1.086)$ and $\mathbf{X}_2 \sim \text{logN}(10.76, 1.8086)$. Again, the dependency of the variables is modeled with a Clayton copula with a copula parameter of 1. Survival times y are simulated from $-\log(U)/h_0\{\exp(X\beta)\}$ where U is drawn from a distribution that is uniform on the interval $[0, 1]$, $h_0 = 0.1$, and the linear predictor $X\beta$ is defined as $-0.3 \ln X_1 + 0.3 \log_{10} X_2$. Therefore, $\beta_{\text{true}} = (-0.3, 0.3)'$. Censoring times are simulated as $-\log(U)/0.2$. The observed

survival time T is thus $\min(y, C)$. Values of X_1 are defined to be missing based on the following function:

$$\pi_{X_1}(T) = 1 - \frac{1}{(0.075T)^2 + 1}.$$

This yields about 8% of missing values.

Setting 4

This setting is inspired by our data analysis from Section 5. We generate longitudinal data ($t = 0, 1, \ldots, 12$) for 3 time-dependent confounders ($\mathbf{L_t} = \{L_t^1, L_t^2, L_t^3\}$), an outcome ($Y_t$), an intervention ($A_t$), as well as baseline data for 7 variables, using structural equation models [34]. The data generating mechanism and the motivation thereof is described in Appendix B. In this simulation we are interested in a counterfactual outcome Y_t which would have been observed under 2 different intervention rules $\bar{d}^j$, $j = 1, 2$, which assign treatment (A_t) always or never. We denote these target quantities as ψ_1 and ψ_2 and their true values are −1.03 and −2.45 respectively. They can be estimated using the sequential g-formula, with bootstrap confidence intervals, see Appendix A for more details.

Values of L_t^1, L_t^2, L_t^3, Y_t are set to be missing based on a MAR process as described in Appendix B. This yields about 10%, 31%, 22% and 44% of missing baseline values, and 10%, 1%, 1%, and 2% of missing follow-up values.

In all 4 settings multiple imputation is utilized with `Amelia II` under a joint modeling approach, see Honaker et al. [3] and Section 6 for details. In settings 1-3 the probability of a missing observation depends on the outcome. One would therefore expect parameter estimates in a regression model of a complete case analysis to be biased, but estimates following multiple imputation to be approximately unbiased [35, 36].

We estimate the confidence intervals for the parameters of interest using the aforementioned four approaches, as well as using the analytic standard errors obtained from the linear model and the Cox proportional hazards model (method "no bootstrap") for the first three settings. The "no bootstrap" method serves therefore as a gold standard and reference for the other methods. We generate $n = 1000$ observations, $B = 200$ bootstrap samples, and $M = 10$ imputations. Based on $\mathcal{R} = 1000$ simulation runs we evaluate the coverage probability and median width of the respective confidence intervals.

Results

The computation time for Boot MI was always greater than for MI Boot, for example by a factor of 13 in the first simulation setting and by a factor of 1.3 in the fourth setting.

In all settings the point estimates for β were approximately unbiased.

Table 1 summarizes the main results of the simulations. Using no bootstrapping yields estimated coverage probabilities of about 95%, for all parameters and settings, as one would expect.

Bootstrapping the imputed data (MI Boot, MI Boot [PS]) yields estimated coverage probabilities of about 95% and confidence interval widths which are similar to each other, except for the high missingness setting of simulation 2. The standard errors for each component of β as simulated in the 1000 simulation runs were almost identical to the mean estimated standard errors under MI Boot, which suggests good standard error estimation of the latter approach. In the first simulation setting the coverage of MI Boot pooled is a bit too low for $M = 10$ (93%), but is closer to 95% if M is large ($M = 20$, Figure 1).

Imputing the bootstrapped data (Boot MI, Boot MI [PS]) led to overall good results with coverage probabilities close to the nominal level, except for the high missingness setting of simulation 2; however, using the pooled samples led to somewhat higher coverage probabilities and the interval widths were slightly different from the estimates obtained under no bootstrapping.

Figure 1 shows the coverage probability of the interval estimates for β_1 in the first simulation setting given the number of imputations.

As predicted by MI theory, using multiple imputation needs generally a reasonable amount of imputed data sets to perform well – no matter whether bootstrapping is used for standard error estimation or not (MI Boot, no bootstrap). Boot MI may perform well even for $M < 5$, but the pooled approach has a tendency towards coverage probabilities > 95%. For $M = 1$ the estimated coverage probability of Boot MI is too large in the above setting.

Figure 2 offers more insight into the behaviour of 'Boot MI (PS)' and 'MI Boot (PS)' by visualizing both the bootstrap distributions in each imputed data set (method MI Boot [PS]) as well as the distribution of the estimators in each bootstrap sample (method Boot MI [PS]): one can see the slightly wider spectrum of values in the distributions related to 'Boot MI (PS)' explaining the somewhat larger confidence interval in the first simulation setting.

More explanations and interpretations of the above results are given in Section 6.

5. Data Analysis

Consider the motivating question introduced in Section 2. We are interested in comparing mortality with respect to different antiretroviral treatment strategies in children between 1 and 5 years of age living with HIV. We use data from two big HIV treatment cohort collaborations (IeDEA-SA, [37]; IeDEA-WA, [38]) and evaluate mortality for 3 years of follow-up. Our analysis builds on a recently published analysis by Schomaker et al. [17].

For this analysis, we are particularly interested in the cumulative mortality difference between strategies (i) 'immediate ART initiation' and (ii) 'assign ART if CD4 count < 350 cells/mm^3 or CD4% < 15%', i.e. we are comparing current practices with those in place in 2006. We can estimate these quantities using the g-formula, see Appendix A for a

comprehensive summary of our implementation details and assumptions. The standard way to obtain 95% confidence intervals for this method is using bootstrapping. However, baseline data of CD4 count, CD4%, HAZ, and WAZ are missing: 18%, 28%, 40%, and 25% respectively. We use multiple imputation (using `Amelia II` [3]) to impute this data. We also impute follow-up data after nine months without any visit data, as from there on it is plausible that follow-up measurements that determine ART assignment (e.g. CD4 count) were taken (and are thus needed to adjust for time-dependent confounding) but were not electronically recorded, probably because of clerical and administrative errors. Under different assumptions imputation may not be needed. To combine the $M = 10$ imputed data sets with bootstrap estimation ($B = 200$) we use the four approaches introduced in Section 3: MI Boot, MI Boot (PS), Boot MI, and Boot MI (PS).

Three year mortality for immediate ART initiation was estimated as 6.08%, whereas mortality for strategy (ii) was estimated as 6.87%. This implies a mortality difference of 0.79%. The results of the respective confidence intervals are summarized in Figure 3: the estimated mortality differences are [−0.34%; 1.61%] for Boot MI (PS), [0.12%; 1.07%] for Boot MI, [−0.31%; 1.63%] for MI Boot (PS), and [−0.81%; 2.40%] for MI Boot.

Figure 3 shows that the confidence intervals vary with respect to the different approaches: the shortest interval is produced by the method Boot MI. Note that only for this method the 95% confidence interval does not contain the 0% when estimating the mortality difference, and therefore suggests a beneficial effect of immediate treatment initiation. The distributions of $\hat{\theta}^*_{b,m}$ for Boot MI (PS) and MI Boot (PS), as well as the distribution of $\hat{\tilde{\theta}}^*_b$ for Boot MI, are also visualized in the figure and are reasonably symmetric.

Figure 4 visualizes both the bootstrap distributions in each imputed data set (method MI Boot [PS]) as well as the distribution of the estimators in each bootstrap sample (method Boot MI [PS]). It is evident that the overall variation of the estimates is similar for these two approaches considered, which explains why their confidence intervals in Figure 3 are almost identical. Moreover, and of note, the top panel highlights the large variability of the point estimates used for the calculation of the MI Boot estimator. The graph indicates a large between imputation uncertainty of the point estimates, possibly due to the high missingness and complex imputation procedure. The large confidence interval of MI Boot in Figure 3, based on formula (2), reflects this uncertainty.

In summary, the above analyses suggest a beneficial effect of immediate ART initiation compared to delaying ART until CD4 count < 350 cells/mm^3 or CD4% < 15% when using method 3, Boot MI. The other methods produce larger confidence intervals and do not necessarily suggest a clear mortality difference.

6. Theoretical Considerations

For the purpose of inference we are interested in the observed data posterior distribution of $\theta \big| \mathcal{D}_{obs}$ which is

$$P(\theta|\mathcal{D}_{obs}) = \int P(\theta|\mathcal{D}_{obs}, \mathcal{D}_{mis})P(\mathcal{D}_{mis}|\mathcal{D}_{obs})d\mathcal{D}_{mis} \qquad (6)$$
$$= \int P(\theta|\mathcal{D}_{obs}, \mathcal{D}_{mis})\left\{\int P(\mathcal{D}_{mis}|\mathcal{D}_{obs}, \vartheta)P(\vartheta|\mathcal{D}_{obs})d\vartheta\right\}d\mathcal{D}_{mis}.$$

Please note that ϑ refers to the parameters of the imputation model whereas θ is the quantity of interest from the analysis model. With multiple imputation we effectively approximate the integral (6) by using the average

$$P(\theta|\mathcal{D}_{obs}) \approx \frac{1}{M}\sum_{m=1}^{M} P(\theta|\mathcal{D}_{mis}^{(m)}, \mathcal{D}_{obs}) \qquad (7)$$

where $\mathcal{D}_{mis}^{(m)}$ refers to draws (imputations) from the posterior predictive distribution $P(\mathcal{D}_{mis}|\mathcal{D}_{obs})$.

MI Boot and MI Boot (PS)

The MI Boot method essentially uses rules (1) and (2) for inference, where, for a given scalar, the respective variance in each imputed data set $\widehat{\mathrm{Var}}(\hat{\theta}_m)$ is not calculated analytically but using bootstrapping. This approach will work if the bootstrap variance for the imputed data set is close to the analytical variance. If there is no analytical variance, it all depends on various factors such as sample size, estimator of interest, proportion of missing data, and others. The data example highlights that in complex settings with a lot of missing data the between imputation variance can be large, yielding conservative interval estimates. As well-known from MI theory M should, in many settings, be much larger than 5 for good estimates of the variance [14]. Using bootstrapping to estimate the variance does not alter these conclusions. Using MI Boot should always be complemented with a reasonably large number of imputations. This consideration also applies to MI Boot (PS), which –as seen in the simulations–, can sometimes be even more sensitive to the choice of M.

Boot MI and Boot MI (PS)

Boot MI uses $\mathcal{D} = \left\{\mathcal{D}_{mis}, \mathcal{D}_{obs}\right\}$ for bootstrapping. Most importantly, we estimate θ, the quantity of interest, in each bootstrap sample using multiple imputation. We therefore approximate $P(\theta|D_{obs})$ through (6) by using multiple imputation to obtain $\hat{\theta}$ and bootstrapping to estimate its distribution – which is valid under the missing at random assumption.

However, if we simply pool the data and apply the method Boot MI (PS) we essentially pool all estimates $\hat{\theta}_{m,b}$: with this approach each of the $B \times M$ estimates $\hat{\theta}_{m,b}$ serves then as an estimator of θ (as we do not combine/average any of them). A possible interpretation of this observation is that each $\hat{\theta}_{m,b}$ estimates θ and since this is only a single draw from the

posterior predictive distribution $P(\mathcal{D}_{mis}|\mathcal{D}_{obs})$ we conduct multiple imputation with $M = 1$, i.e. we calculate $\hat{\bar{\theta}}_{MI} = \frac{1}{1}\sum_{m=1}^{1}\hat{\theta}_{m,b}$, $B \times M$ times. Such an estimator is statistically inefficient as we know from MI theory: the relative efficiency of an MI based estimator (compared to the true variance) is $(1 + \frac{\gamma}{M})^{-1}$ where γ describes the fraction of missingness (i.e. $V/(W + V)$) in the data. For example, if the fraction of missingness is 0.25, and $M = 5$, then the loss of efficiency is 5% [6]. The lower M, the lower the efficiency, and thus the higher the variance. This explains the results of the simulation studies: pooling the estimates is inefficient, does therefore overestimate the variance, and thus leads to confidence intervals with incorrect coverage.

It follows that one typically gets larger interval estimates when using Boot MI (PS) instead of Boot MI. Similarly, one can decide to use Boot MI with $M = 1$, which is not incorrect but often inefficient in terms of interval estimation.

Comparison

General comparisons between MI Boot and Boot MI are difficult because the within and between imputation uncertainty, as well as the within and between bootstrap sampling uncertainty, will determine the actual width of a confidence interval. If the between imputation uncertainty is large compared to between bootstrap sample uncertainty (as, for example, in the data example [Figure 4]) then MI Boot is large compared to Boot MI. However, if the between imputation uncertainty is small relative to the bootstrap sampling uncertainty, then Boot MI may give a similar confidence interval to MI Boot (as in the simulations [Figure 2]).

Another consideration is related to the application of the bootstrap. We have focused on the percentile method to create confidence intervals. However, it is also possible to create bootstrap intervals based on the t–distribution. Here, an estimator's variance is estimated with the sample variance from the B bootstrap estimates and symmetric confidence intervals are generated based on an appropriate t-distribution. In fact, MI Boot uses this approach because in each imputed dataset we estimate the bootstrap variance $\widehat{\text{Var}}(\hat{\theta}_m) = (B-1)^{-1}\sum_b (\hat{\theta}_{m,b} - \hat{\bar{\theta}}_{m,b})^2$, then calculate (2), followed by confidence intervals based on a t_R distribution, see Section 3. A similar approach would be possible when applying Boot MI. This method produces B point estimates $\hat{\bar{\theta}}_b^* = M^{-1}\sum_m \hat{\theta}_{b,m}^*$ for θ. One could estimate the variance as $(B-1)^{-1}\sum_b (\hat{\bar{\theta}}_b^* - \hat{\bar{\bar{\theta}}}^*)^2$, with $\hat{\bar{\bar{\theta}}}^* = B^{-1}\sum_b \hat{\bar{\theta}}_b^*$, and then create confidence intervals based on a t-distribution. This would however require that one assumes the estimator to be approximately normally distributed.

Bootstrapping as part of the imputation procedure

For each of the estimators introduced in Section 3, M *proper* multiply imputed data sets are needed. "Proper" means that the application of formulae (1) and (2) yield 1) approximately unbiased point estimates and 2) interval estimates which are randomization valid in the sense

that actual interval coverage equals the nominal interval coverage. Some imputation algorithms use bootstrapping to create proper imputations, and this may not be confused with the bootstrapping step *after* multiple imputation which we focus on in this paper.

To follow this argument in more detail it is important to understand that proper imputations are created by means of random draws from the posterior predictive distribution of the missing data given the observed data (or an approximation thereof). These draws can (i) either be generated by specifying a multivariate distribution of the data (joint modeling) and simulate the posterior predictive distribution with a suitable algorithm; or (ii) by specifying individual conditional distributions for each variable $\mathbf{X}_j$ given the other variables (fully conditional modeling) and iteratively drawing and updating imputed values from these distributions which will then (ideally) converge to draws of the theoretical joint distribution; or (iii) by the use of alternative algorithms.

An example for (i) is the EMB algorithm from the *R*-package `Amelia II` which assumes a multivariate normal distribution for the data, $\mathscr{D} \sim N(\mu, \sum)$ (possibly after suitable transformations beforehand). Then, *B* bootstrap samples of the data (including missing values) are drawn and in each bootstrap sample the EM algorithm [39] is applied to obtain estimates of μ and Σ which can then be used to generate proper multiple imputations by means of the sweep-operator [40, 11]. Of note, the algorithm can handle highly skewed variables by imposing transformations on variables (log, square root). Categorical variables are recoded into dummy variables based on the knowledge that for binary variables the multivariate normal assumption can yield good results [9].

An example for (ii) is imputation by chained equations (ICE, mice). Here, (a) one first specifies *individual* conditional distributions (i.e. regression models) $p(\mathbf{X}_j|\mathbf{X}_{-j}; \theta_j)$ for each variable. Then, (b) one iteratively fits all regression models and generates *random* draws of the coefficients, e.g. $\tilde{\beta} \sim N(\hat{\beta}, \widehat{\mathrm{Cov}}(\hat{\beta}))$. Values are (c) imputed as random draws from the distribution of the regression predictions. Then, (b) and (c) are repeated *k* times until convergence. The process of iteratively drawing and updating the imputed values from the conditional distributions can be viewed as a Gibbs sampler that converges to draws from the (theoretical) joint distribution. This method is among the most popular ones in practice and has been implemented in many software packages [4, 5]. However, there remain theoretical concerns as a joint distribution may not always exist for a given specifications of the conditional distributions [41]. A variation of (c) is a fully Bayesian approach where the posterior predictive distribution is used to draw imputations. Here, the bootstrap is used to model the imputation uncertainty and to draw the *M* imputations needed for the *M* imputed data sets. This variation yields approximate proper imputations and is implemented in the *R* library `Hmisc` [42].

An example for (iii) is the Approximate Bayesian Bootstrap [29]. Here, the (cross-sectional) data is stratified into several strata, possibly by means of the covariates of the analysis model. Then, within each stratum (a) one draws a bootstrap sample among the complete data (with respect to the variable to be imputed). Secondly, (b) one uses the original data set (with missing values) and imputes the missing data based on units from the data set created in (a),

with equal selection probability and with replacement. The multiply imputed data are obtained by repeating (a) and (b) M times.

It is evident from the above examples that many imputation methods use bootstrap methodology as part of the imputation model, that this does not replace the additional bootstrap step needed for the inference in the analysis model, and that – if they are combined – the resampling steps are nested.

7. Conclusion

The current statistical literature is not clear on how to combine bootstrap with multiple imputation inference. We have proposed that a number of approaches are intuitively appealing and three of them are correct: Boot MI, MI Boot, MI Boot (PS). Using Boot MI (PS) can lead to too large and invalid confidence intervals and is therefore not recommended.

Both Boot MI and MI Boot are probably the best options to calculate randomization valid confidence intervals when combining bootstrapping with multiple imputation. As a rule of thumb, our analyses suggest that the former may be preferred for small M or large imputation uncertainty and the latter for normal M and little/normal imputation uncertainty.

There are however other considerations when deciding between MI Boot and Boot MI. The latter is computationally much more intensive. This matters particularly when estimating the analysis model is simple in relation to creating the imputations. In fact, in our first simulation this affected the computation time by a factor of 13. However, MI Boot naturally provides symmetrical confidence intervals. These intervals may not be wanted if an estimator's distribution is suspected to be non-normal.

Acknowledgments

References

1. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996; 91(434):473–489.

2. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models. The American Statistician. 2007; 61:79–90. [PubMed: 17401454]

3. Honaker J, King G, Blackwell M. Amelia II: A program for missing data. Journal of Statistical Software. 2011; 45(7):1–47.

4. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45(3):1–67.

5. Royston P, White IR. Multiple imputation by chained equations (mice): Implementation in Stata. Journal of Statistical Software. 2011; 45(4):1–20.

6. White IR, Royston P, Wood AM. Multiple imputation using chained equations. Statistics in medicine. 2011; 30:377–399. [PubMed: 21225900]

7. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal. 2009; 339

8. Graham JW. Missing data analysis: Making it work in the real world. Annual Review of Psychology. 2009; 60:549–576.

9. Schafer J, Graham J. Missing data: our view of the state of the art. Psychological Methods. 2002; 7:147–177. [PubMed: 12090408]

10. Eddings W, Marchenko Y. Diagnostics for multiple imputation in Stata. Stata Journal. 2012; 12(3): 353–367.

11. Honaker J, King G. What to do about missing values in time-series cross-section data? American Journal of Political Science. 2010; 54:561–581.

12. Efron B. Missing data, imputation, and the bootstrap. Journal of the American Statistical Association. 1994; 89(426):463–475.

13. Shao J, Sitter RR. Bootstrap for imputed survey data. Journal of the American Statistical Association. 1996; 91(435):1278–1288.

14. Little, R., Rubin, D. Statistical analysis with missing data. Wiley; New York: 2002.

15. Briggs AH, Lozano-Ortega G, Spencer S, Bale G, Spencer MD, Burge PS. Estimating the cost-effectiveness of fluticasone propionate for treating chronic obstructive pulmonary disease in the presence of missing data. Value in Health. 2006; 9(4):227–235. [PubMed: 16903992]

16. Schomaker M, Heumann C. Model selection and model averaging after multiple imputation. Computational Statistics & Data Analysis. 2014; 71:758–770.

17. Schomaker M, Davies MA, Malateste K, Renner L, Sawry S, N'Gbeche S, Technau K, Eboua FT, Tanser F, Sygnate-Sy H, et al. Growth and mortality outcomes for different antiretroviral therapy initiation criteria in children aged 1–5 years: A causal modelling analysis from West and Southern Africa. Epidemiology. 2016; 27:237–246. [PubMed: 26479876]

18. Worthington H, King R, Buckland ST. Analysing mark-recapture-recovery data in the presence of missing covariate data via multiple imputation. Journal of Agricultural, Biological, and Environmental Statistics. 2015; 20:28.

19. Wu W, Jia F. A new procedure to test mediation with missing data through nonparametric bootstrapping and multiple imputation. Multivariate Behavioral Research. 2013; 48(5):663–691. [PubMed: 26741058]

20. Baneshi MR, Talei A. Assessment of internal validity of prognostic models through bootstrapping and multiple imputation of missing data. Iranian Journal of Public Health. 2012; 41(5):110–115. [PubMed: 23113185]

21. Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HCW. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Medical Research Methodology. 2007; 7

22. Chaffee BW, Feldens CA, Vitolo MR. Association of long-duration breastfeeding and dental caries estimated with marginal structural models. Annals of Epidemiology. 2014; 24(6):448–454. [PubMed: 24636616]

23. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, Gange SJ, Hernan MA. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Statistics in medicine. 2012; 31(18):2000–2009. [PubMed: 22495733]

24. Edmonds A, Yotebieng M, Lusiama J, Matumona Y, Kitetele F, Napravnik S, Cole SR, Van Rie A, Behets F. The effect of highly active antiretroviral therapy on the survival of HIV-infected children in a resource-deprived setting: a cohort study. PLoS Medicine. 2011; 8(6):e1001 044.

25. Violari A, Cotton MF, Gibb DM, Babiker AG, Steyn J, Madhi SA, Jean-Philippe P, McIntyre JA. Early antiretroviral therapy and mortality among HIV-infected infants. New England Journal of Medicine. 2008; 359(21):2233–2244. [PubMed: 19020325]

26. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. Statistics in Medicine. 2013; 32(9):1584–618. [PubMed: 23208861]

27. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. Journal of Causal Inference. 2014; 2:147–185. [PubMed: 25909047]

28. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7(9–12):1393–1512.

29. Rubin D, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association. 1986; 81:366–374.

30. Lipsitz S, Parzen M, Zhao L. A degrees-of-freedom approximation in multiple imputation. Journal of Statistical Computation and Simulation. 2002; 72:309–318.

31. Robins, J., Hernan, MA. Estimation of the causal effects of time-varying exposures. CRC Press; 2009. p. 553-599.

32. Yan J. Enjoy the joy of copulas: with package copula. Journal of Statistical Software. 2007; 21:1–21.

33. Schomaker M, Hogger S, Johnson LF, Hoffmann C, Brnighausen T, Heumann C. Simultaneous treatment of missing data and measurement error in HIV research using multiple overimputation. Epidemiology. 2015; 26:628–636. [PubMed: 26214336]

34. Sofrygin, O., van der Laan, MJ., Neugebauer, R. simcausal: Simulating Longitudinal Data with Causal Inference Applications. 2016. URL https://CRAN.R-project.org/package=simcausal, r package version 0.5.3

35. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression-coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89(427):846–866.

36. Little RJA. Regression with missing X's - a review. Journal of the American Statistical Association. 1992; 87(420):1227–1237.

37. Egger M, Ekouevi DK, Williams C, Lyamuya RE, Mukumbi H, Braitstein P, Hartwell T, Graber C, Chi BH, Boulle A, et al. Cohort profile: The international epidemiological databases to evaluate AIDS (IeDEA) in sub-Saharan Africa. International Journal of Epidemiology. 2012; 41(5):1256–1264. [PubMed: 21593078]

38. Ekouevi DK, Azondekon A, Dicko F, Malateste K, Toure P, Eboua FT, Kouadio K, Renner L, Peterson K, Dabis F, et al. 12-month mortality and loss-to-program in antiretroviral-treated children: The iedea pediatric west african database to evaluate aids (pwada), 2000–2008. Bmc Public Health. 2011; 11:519. [PubMed: 21718505]

39. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society B. 1977; 39:1–38.

40. Goodnight JH. Tutorial on the sweep operator. American Statistician. 1979; 33(3):149–158.

41. Drechsler, J., Rssler, S. Does convergence really matter?. Springer; 2008. p. 342-355.

42. Harrell, FE, Jr. Hmisc: Harrell Miscellaneous. 2016. with contributions from Charles Dupont, many othersURL https://CRAN.R-project.org/package=Hmisc, R package version 4.0-1

43. Schomaker M, Egger M, Ndirangu J, Phiri S, Moultrie H, Technau K, Cox V, Giddy J, Chimbetete C, Wood R, et al. When to start antiretroviral therapy in children aged 2–5 years: a collaborative causal modelling analysis of cohort studies from southern Africa. Plos Medicine. 2013; 10(11):e1001 555.

44. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernan MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Statistics in biosciences. 2011; 3(1):119–143. [PubMed: 24039638]

45. Petersen ML. Commentary: Applying a causal road map in settings with time-dependent confounding. Epidemiology. 2014; 25(6):898–901. [PubMed: 25265135]

A. Details of the G-formula Implementation

We consider n children studied at baseline ($t = 0$) and during discrete follow-up times ($t = 1$, …, T). The data consists of the outcome Y_t, an intervention variable A_t, q time-dependent covariates $\mathbf{L}_t = \{L_t^1, ..., L_t^q\}$, and a censoring indicator C_t. The covariates may also include baseline variables $\mathbf{V} = \{L_0^1, ..., L_0^{qV}\}$. The treatment and covariate history of an individual i up to and including time t is represented as $\bar{A}_{t,i} = (A_{0,i}, ..., A_{t,i})$ and $\bar{L}_{t,i}^s = (L_{0,i}^s, ..., L_{t,i}^s)$ respectively. C_t equals 1 if a subject gets censored in the interval $(t - 1, t]$, and 0 otherwise. Therefore, $\bar{C}_t = 0$ is the event that an individual remains uncensored until time t.

The counterfactual outcome $Y_{t,i}^{\bar{a}_t}$ refers to the hypothetical outcome that would have been observed at time t if subject i had received, possibly contrary to the fact, the treatment history $\bar{A}_{t,i} = \bar{a}_t$. Similarly, $\mathbf{L}_{t,i}^{\bar{a}_t}$ are the counterfactual covariates related to the intervention $\bar{A}_{t,i} = \bar{a}_t$. The above notation refers to *static* treatment rules; a treatment rule may, however, depend on covariates, and in this case it is called *dynamic*. A dynamic rule $d_t(\bar{\mathbf{L}}_t)$ assigns treatment $A_{t,i} \in \{0, 1\}$ as a function of the covariate history $\bar{\mathbf{L}}_{t,i}$. The vector of decisions d_t, $t = 0, ..., T$, is denoted as $\bar{d}_T = \bar{d}$. The notation $\bar{A}_t = \bar{d}$ refers to the treatment history up to time t according to the rule $\bar{d}$. The counterfactual outcome related to a dynamic rule $\bar{d}$ is $Y_{t,i}^{\bar{d}}$, and the counterfactual covariates are $\mathbf{L}_{t,i}^{\bar{d}}$.

In our setting we study $n = 5826$ children for $t = 0, 1, 3, 6, 9, ...$ where the follow-up time points refer to the intervals $(0, 1.5)$, $[1.5, 4.5)$, $[4.5, 7.5)$, …, $[28.5, 31.5)$, $[31.5, 36)$ months respectively. Follow-up measurements, if available, refer to measurements closest to the middle of the interval. In our data Y_t refers to death at time t (i.e. occurring during the interval $(t - 1, t]$). A_t refers to antiretroviral treatment (ART) taken at time t. $\mathbf{L}_t = (L_t^1, L_t^2, L_t^3, L_t^{1m}, L_t^{2m}, L_t^{3m})$ are CD4 count, CD4%, and weight for age z-score (WAZ, which serves as a proxy for WHO stage, see [43] for more details) as well as three indicator variables whether these variables have been measured at time t or not. $\mathbf{V} = \mathbf{L}_0^V$ refer to baseline values of CD4 count, CD4%, WAZ, height for age z-score (HAZ) as well as sex, age, and region. The two treatment rules of interest are:

$$\bar{d}_t^1 = \left\{ c_t = 0;\ l_t^{1m} = l_t^{2m} = l_t^{3m} = 1;\ a_t = 1\ \text{for}\ \forall t \right.$$

$$\bar{d}_t^2 = \begin{cases} c_t = 0;\ l_t^{1m} = l_t^{2m} = l_t^{3m} = 1;\ a_t = 1\ \text{if}\ \text{CD4 count}_t^{\bar{d}^2} < 350\ \text{or}\ \text{CD4\%}_t^{\bar{d}^2} < 15\% \\ c_t = 0;\ l_t^{1m} = l_t^{2m} = l_t^{3m} = 1;\ a_t = 0\ \text{otherwise} \end{cases}$$

The quantity of interest is cumulative mortality after $T = 36$ months, under (the intervention of) no censoring, regular 3 monthly follow-up and for treatment assignment according to $\bar{d}^j$, that is $\psi = \sum_{t=1}^{T} \mathbb{P}(Y_t^{\bar{d}^j} = 1)$.

Under the assumption of *consistency*, i.e. $Y^{\bar{d}_t} = Y$ if $\bar{A}_t = \bar{d}_t$ and $\overline{\mathbf{L}}_t^{\bar{d}_t} = \overline{\mathbf{L}}_t$ if $\bar{A}_{t-1} = \bar{d}_{t-1}$, *sequential conditional exchangeability* (or *no unmeasured confounding*), i.e. $Y^{\bar{d}_t} \amalg A_t | \overline{\mathbf{L}}_t, \bar{A}_{t-1}$ for $\forall \bar{A}_t = \bar{d}_t, \overline{\mathbf{L}}_t = \bar{\mathbf{l}}_t, t \in \{0, \ldots, T\}$ and *positivity*, i.e. $P(A_t = \bar{d}_t | \overline{\mathbf{L}}_t = \bar{\mathbf{l}}_t, \bar{A}_{t-1} = \bar{d}_{t-1}) > 0$ for $\forall t, \bar{d}_t, \bar{\mathbf{l}}_t$ with $P(\overline{\mathbf{L}}_t = \bar{l}_t, \bar{A}_{t-1} = \bar{d}_{t-1}) \neq 0$, the g-computation formula can estimate ψ as:

$$\psi = \sum_{t=1}^{T} \mathbb{P}(Y_t^{\bar{d}^j} = 1) = \sum_{t=1}^{T} \int_{\bar{\mathbf{l}} \in \overline{\mathbf{L}}_t} \left\{ \mathbb{P}(Y_t = 1 | \bar{A}_{t-1} = \bar{d}_{t-1}, \overline{\mathbf{L}}_t = \bar{\mathbf{l}}_t, Y_{t-1} = 0) \right. \quad (8)$$
$$\left. \times \prod_{t=1}^{T} f(l_t | \bar{A}_{t-1} = \bar{d}_{t-1}, \overline{\mathbf{L}}_{t-1} = \bar{\mathbf{l}}_{t-1}, Y_{t-1} = 0) \right\} d\bar{\mathbf{l}}$$

see [23] and [44] about more details and implications of the representation of the g-formula in this context. Note that the inner product of (8) can be written as

$$\prod_{t=1}^{T} \prod_{s=1}^{q} f(l_t^s | \bar{A}_{t-1} = \bar{d}_{t-1}, \overline{\mathbf{L}}_{t-1} = \bar{\mathbf{l}}_{t-1}, L_t^1 = l_t^1, \ldots, L_t^{s-1} = l_t^{s-1}, \bar{Y}_{t-1} = 0).$$

In the above representation of the g-formula we assume that the time ordering is $L_t^1 \rightarrow L_t^2 \rightarrow L_t^3 \rightarrow A \rightarrow C \rightarrow Y$.

There is no closed form solution to estimate (8), but ψ can be approximated by means of the following algorithm; Step 1: use additive linear and logistic regression models to estimate the conditional densities on the right hand side of (8), i.e. fit regression models for the outcome variables CD4 count, CD4%, WAZ, and death at $t = 1, 3, .., 36$ using the available covariate history and model selection. Step 2: use the models fitted in step 1 to stochastically generate $\mathbf{L}_t$ and Y_t under a specific treatment rule. For example, for rule (ii), draw $\mathbf{L}_1^1 = \sqrt{\text{CD4 count}_1}$ from a normal distribution related to the respective additive linear model from step 1 using the relevant covariate history data. Set $A_1 = 1$ if the generated CD4 count at time 1 is < 350 cells/mm^3 or CD4% < 15% (for rule $\bar{d}^2$). Use the simulated covariate data and treatment as assigned by the rule to generate the full simulated data set forward in time and evaluate cumulative mortality after 3 years of follow-up. We refer the reader to [17], [23], and [44] to learn more about the g-formula in this context.

Note that the so-called sequential g-formula, used in the simulation study, shares the idea of standardization in the sense that one sequentially marginalizes the distribution with respect

to $\mathbf{L}$ given the intervention rule of interest. It is just a re-expression of (8) where integration with respect $\mathbf{L}$ is not needed [45]:

$$\mathbb{E}(Y_T^{\bar{d}}) = \mathbb{E}(\mathbb{E}(\dots \mathbb{E}(\mathbb{E}(Y_T | \bar{A}_T = \bar{d}_T, \bar{\mathbf{L}}_T) | \bar{A}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_{T-1}) \dots | \bar{A}_0 = d_0, \mathbf{L}_0) | \mathbf{L}_0). \quad (9)$$

B. Data Generating Process in the Simulation Study

Both baseline data ($t = 0$) and follow-up data ($t = 1, \dots, 12$) were created using structural equations using the *R*-package `simcausal` [34]. The below listed distributions, listed in temporal order, describe the data-generating process. Baseline data refers to region, sex, age, CD4 count, CD4%, WAZ and HAZ respectively (V^1, V^2, V^3, L_0^1, L_0^2, L_0^3, Y_0). Follow-up data refers to CD4 count, CD4%, WAZ and HAZ (L_t^1, L_t^2, L_t^3, Y_t), as well as an antiretroviral treatment (A_t) and censoring (C_t) indicator. For simplicity, no deaths are assumed. In addition to Bernoulli (B), uniform (U) and normal (N) distributions, we also use truncated normal distributions which are denoted by $N_{[a,b]}$ where a and b are the truncation levels. Values which are smaller a are replaced by a random draw from a $U(a_1, a_2)$ distribution and values greater than b are drawn from a $U(b_1, b_2)$ distribution. Values for (a_1, a_2, b_1, b_2) are (0, 50, 5000, 10000) for L^1, (0.03, 0.09, 0.7, 0.8) for L^2, and (−10, 3, 3, 10) for both L^3 and Y. The notation $\overline{\mathscr{D}}$ means "conditional on the data that has already been measured (generated) according the the time ordering". The distributions are as follows:

For $t = 0$:

$$V^1 \sim B(p = 4392/5826)$$

$$V^2 | \overline{\mathscr{D}} \sim \begin{cases} B(p = 2222/4392) & \text{if } V^1 = 1 \\ B(p = 758/1434) & \text{if } V^1 = 0 \end{cases}$$

$$V^3 | \overline{\mathscr{D}} \sim U(1, 5)$$

$$L_0^1 | \overline{\mathscr{D}} \sim \begin{cases} N_{[0, 10000]}(650, 350) & \text{if } V^1 = 1 \\ N_{[0, 10000]}(720, 400)) & \text{if } V^1 = 0 \end{cases}$$

$$\tilde{L}_0^1 | \overline{\mathscr{D}} \sim N((L_0^1 - 671.7468) \big/ (10 \cdot 352.2788 + 1, 0)$$

$$L_0^2 | \overline{\mathcal{D}} \sim N_{[0.06, 0.8]}(0.16 + 0.05 \cdot (L_0^1 - 650 \Big/ 650, 0.07)$$

$$\widetilde{L}_0^2 | \overline{\mathcal{D}} \sim N((L_0^2 - 0.1648594) \Big/ (10 \cdot 0.06980332 + 1, 0)$$

$$L_0^3 | \overline{\mathcal{D}} \sim \begin{cases} N_{[-5, 5]}(-1.65 + 0.1 \cdot V^3 + 0.05 \cdot (L_0^1 - 650) \Big/ 650 + 0.05 \cdot (L_0^2 - 16) \Big/ 16, 1) & \text{if } V^1 = 1 \\ N_{[-5, 5]}(-2.05 + 0.1 \cdot V^3 + 0.05 \cdot (L_0^1 - 650) \Big/ 650 + 0.05 \cdot (L_0^2 - 16) \Big/ 16, 1)) & \text{if } V^1 = 0 \end{cases}$$

$$A_0 | \overline{\mathcal{D}} \sim B(p = 0)$$

$$C_0 | \overline{\mathcal{D}} \sim B(p = 0)$$

$$Y_0 | \overline{\mathcal{D}} \sim N_{[-5, 5]}(-2.6 + 0.1 \cdot I(V^3 > 2) + 0.3 \cdot I(V^1 = 0) + (L_0^3 + 1.45), 1.1)$$

For $t > 0$:

$$L_t^1 | \overline{\mathcal{D}} \sim$$

$$\begin{cases} N_{[0, 10000]}(13 \cdot \log(t \cdot (1034 - 662)/8) + L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{1, 2, 3, 4\} \\ N_{[0, 10000]}(4 \cdot \log(t \cdot (1034 - 662)/8) + L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{5, 6, 7, 8\} \\ N_{[0, 10000]}(L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{9, 10, 11, 12\} \end{cases}$$

$$L_t^2 | \overline{\mathcal{D}} \sim N_{[0.06, 0.8]}(L_{t-1}^2 + 0.0003 \cdot (L_{t-1}^1 - L_{t-1}^1) + 0.0005 \cdot (L_{t-1}^3) + 0.0005 \cdot A_{t-1} \cdot \widetilde{L}_0^1, 0.02)$$

$$L_t^3 | \overline{\mathcal{D}} \sim N_{-5, 5}(L_{t-1}^3 + 0.0017 \cdot (L_t^1 - L_{t-1}^1) + 0.2 \cdot (L_t^2 - L_{t-1}^2) + 0.005 \cdot A_{t-1} \cdot \widetilde{L}_0^2, 0.5)$$

$$A_t | \overline{\mathcal{D}} \sim \begin{cases} B(p = 1) & \text{if } A_{t-1} = 1 \\ B(p = 1/(1 + \exp(-[-2.4 + 0.015 \cdot (750 - L_t^1) + 5 \cdot (0.2 - L_t^2) - 0.8 \cdot L_t^3 + 0.8 \cdot t]))) & \text{if } A_{t-1} = 0 \end{cases}$$

$$C_t | \overline{\mathcal{D}} \sim B(p = 1/(1 + \exp(-[-6 + 0.01 \cdot (750 - L_t^1) + 1 \cdot (0.2 - L_t^2) - 0.65 \cdot L_t^3 - A_t])))$$

$$Y_t | \overline{\mathcal{D}} \sim N_{[-5,5]}(Y_{t-1} + 0.00005 \cdot (L_t^1 - L_{t-1}^1) - 0.000001 \cdot \left((L_t^1 - L_{t-1}^1) \cdot \sqrt{\widetilde{L}_0^1}\right)^2 + 0.01 \cdot (L_t^2 - L_t^2)$$
$$- 0.0001 \cdot \left((L_t^2 - L_{t-1}^2) \cdot \sqrt{\widetilde{L}_0^2}\right)^2 + 0.07 \cdot ((L_t^3 - L_{t-1}^3) \cdot (L_0^3 + 1.5135)) - 0.001$$
$$\cdot ((L_t^3 - L_{t-1}^3) \cdot (L_0^3 + 1.5135))^2 + 0.005 \cdot A_t + 0.075 \cdot A_{t-1} + 0.05 \cdot A[t] \cdot A[t-1], 0.01)$$

The data generating process leads to the following baseline values: region A = 75.5%; male sex = 51.2%; mean age = 3.0 years; mean CD4 count = 672.5; mean CD4% = 15.5%; mean WAZ = −1.5; mean HAZ = −2.5. At $t = 12$ the arithmetic mean of CD4 count, CD4%, WAZ and HAZ are 1092, 27.2%, −0.8, −1.5 respectively. The target quantities ψ_1 and ψ_2 are defined as the expected value of Y at time T, under no censoring, for a given treatment rule $\bar{d}^j$, where

$$\bar{d}_t^1 = \{c_t = 0; \ a_t = 1 \ \text{for} \ \forall t \ \text{and} \ \bar{d}_t^2 = \{c_t = 0; \ a_t = 0 \ \text{for} \ \forall t$$
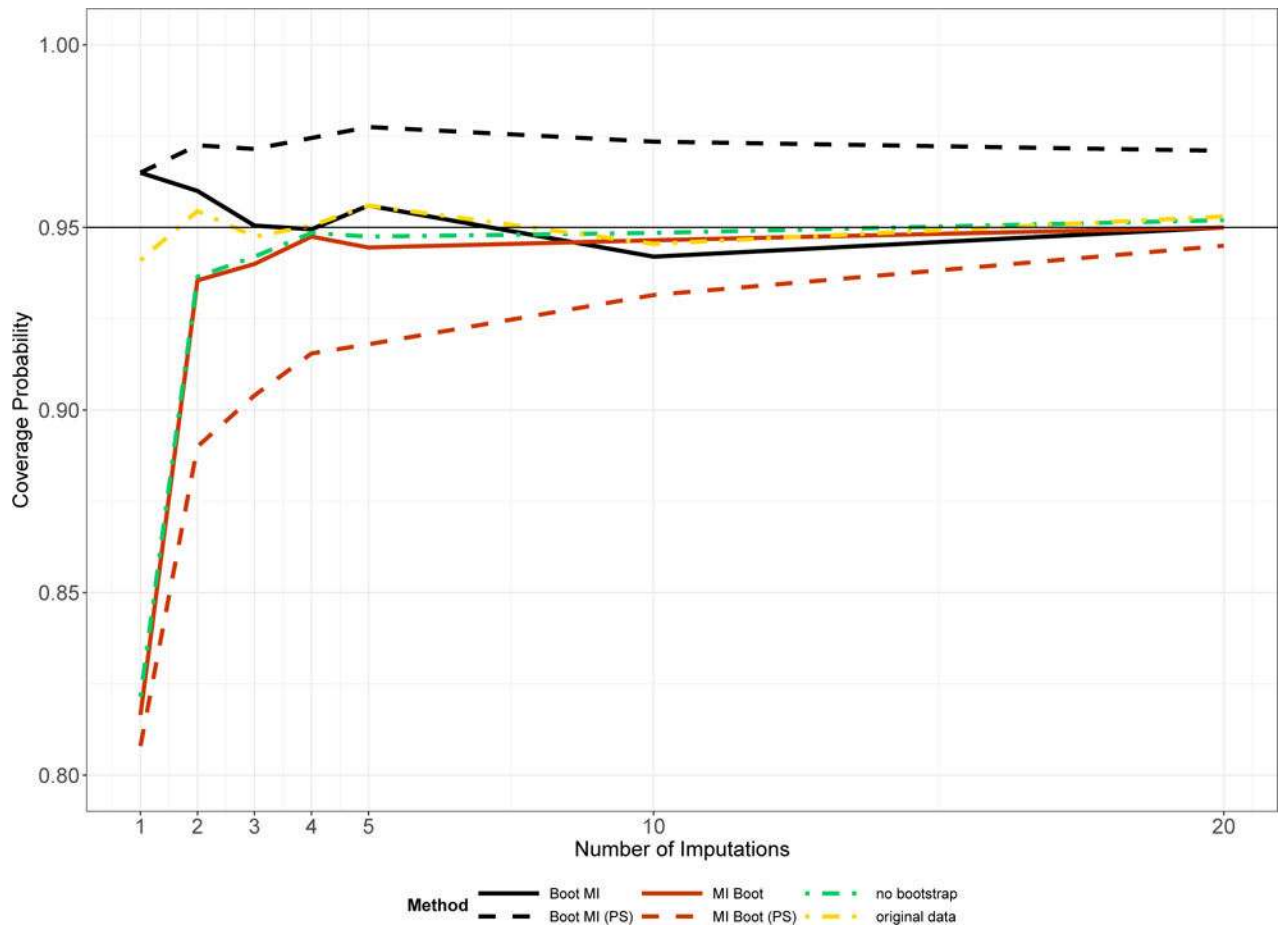
and are −1.03 and −2.45 respectively. Missing baseline and follow-up data were created based on the following functions:

$$\pi_{(L_t^1)} = 0.1;$$

$$\pi_{(L_0^2)}(L_0^1) = 1 - \frac{1}{(0.001 \cdot L_0^1)^2 + 1}; \quad \pi_{(L_t^2)}(t, L_t^1) = 1 - \frac{1}{(0.00005 \cdot t \cdot L_t^1)^2 + 1};$$

$$\pi_{(L_0^3)}(Y_0) = 1 - \frac{1}{(0.2 \cdot |Y_0|)^2 + 1}; \quad \pi_{(L_t^3)}(t, Y_t) = 1 - \frac{1}{(0.015 \cdot t \cdot |Y_t|)^2 + 1};$$

$$\pi_{(Y_0)}(L_0^3) = 1 - \frac{1}{(0.7 \cdot |L_0^3|)^2 + 1}; \quad \pi_{(Y_t)}(t, L_t^3) = 1 - \frac{1}{(0.015 \cdot t \cdot |L_t^3|)^2 + 1}.$$

Figure 1.
Coverage probability of the interval estimates for β_1 in the first simulation setting dependent on the number of imputations. Results related to the complete simulated data, i.e. before missing data are generated, are labelled "original data".
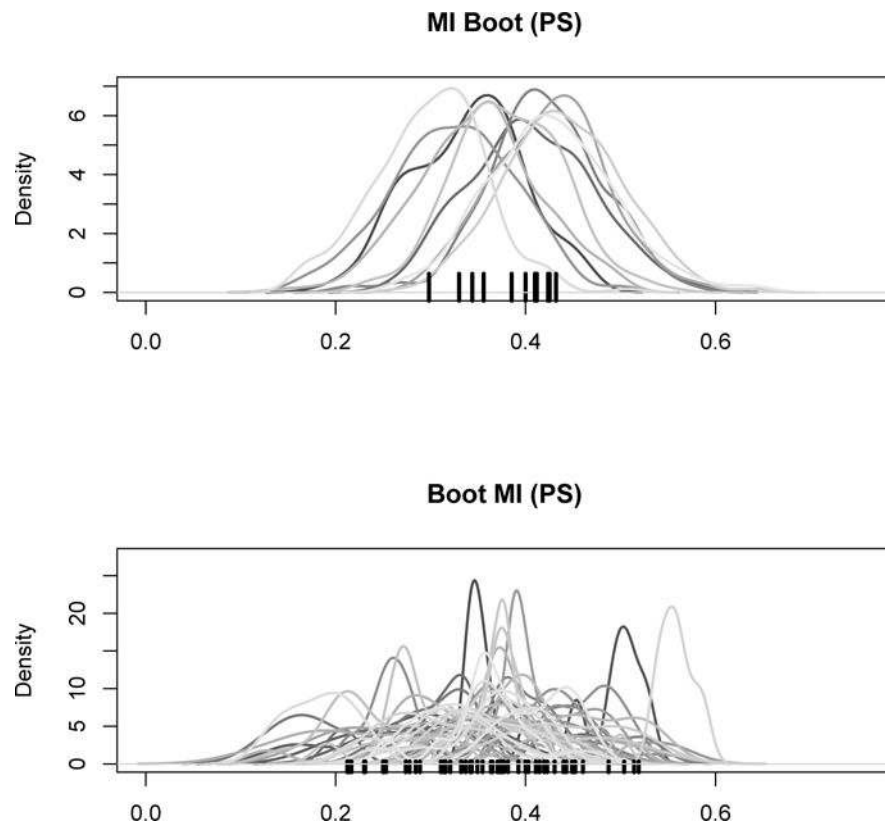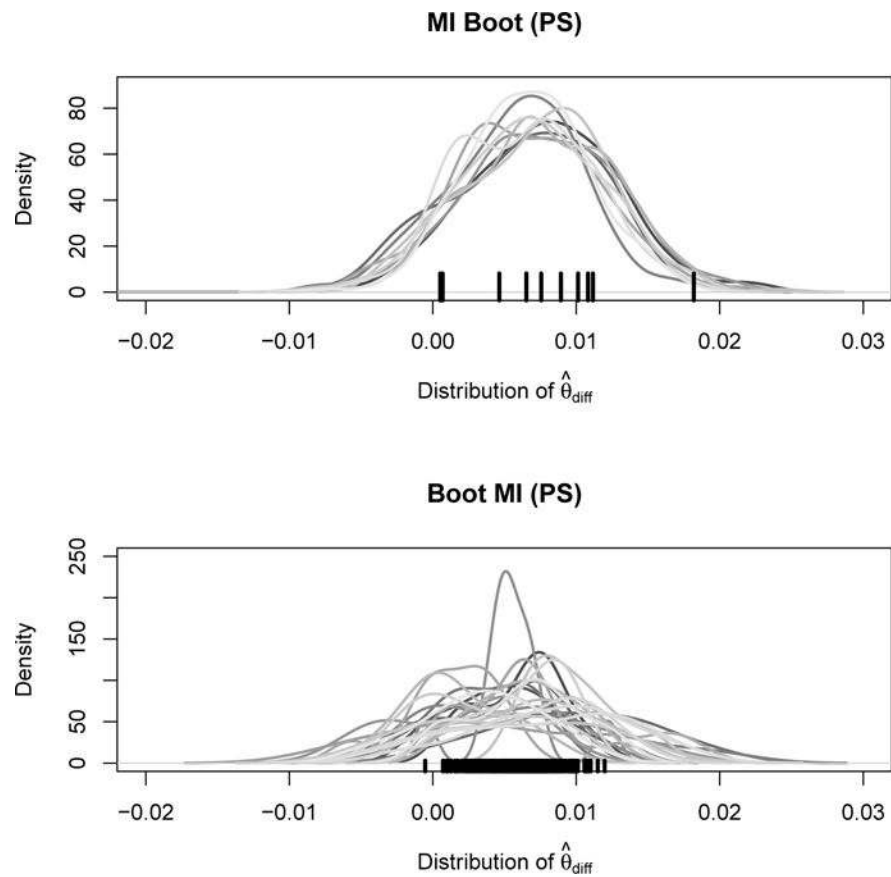
Figure 2.
Estimate of β_1 in the first simulation setting, for a random simulation run: distribution of 'MI Boot (pooled)' for each imputed dataset (top) and distribution of 'Boot MI (PS)' for 50 random bootstrap samples (PS). Point estimates are marked by the black tick marks on the x-axis.
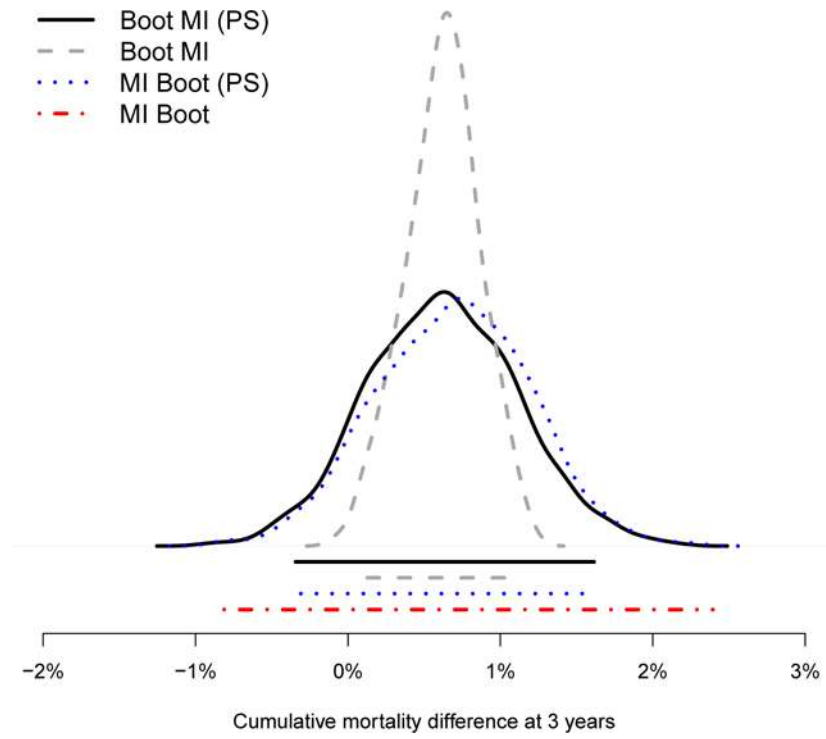
Figure 3.
Estimated cumulative mortality difference between the interventions 'immediate ART' and '350/15' at 3 years: distributions and confidence intervals of different estimators

Figure 4.
Estimated cumulative mortality difference: distribution of 'MI Boot (PS)' for each imputed dataset (top) and distribution of 'Boot MI (PS)' for 25 random bootstrap samples (bottom). Point estimates are marked by the black tick marks on the x-axis.

Table 1

Results of the simulation studies: estimated coverage probability (top), median confidence intervals width (middle), and standard errors for different methods (bottom). The bottom panel lists standard errors estimated from the 1000 point estimates of the simulation ("simulated") and the mean estimated standard error across the simulated data sets, for both the analytical standard error ("no bootstrap") and the bootstrap standard error ("MI Boot"). All results are based on 200 bootstrap samples and 10 imputations.

	Method	Setting 1	Setting 2 (low missingness)					Setting 3	
		β_1	β_2	β_3	β_4	β_5	β_6	β_1	β_2
Coverage Probability	1) MI Boot (PS)	93%	95%	94%	94%	95%	95%	95%	95%
	2) MI Boot	95%	95%	95%	94%	95%	95%	95%	95%
	3) Boot MI (PS)	97%	95%	96%	95%	96%	96%	96%	96%
	4) Boot MI	94%	94%	94%	94%	94%	94%	94%	94%
	5) no bootstrap	95%	95%	95%	95%	95%	96%	95%	95%
Median CI Width	1) MI Boot (PS)	0.30	0.33	0.33	0.60	0.68	0.62	0.25	0.31
	2) MI Boot	0.31	0.34	0.34	0.61	0.69	0.63	0.26	0.31
	3) Boot MI (PS)	0.35	0.35	0.35	0.64	0.72	0.66	0.26	0.31
	4) Boot MI	0.30	0.33	0.33	0.60	0.67	0.62	0.24	0.30
	5) no bootstrap	0.31	0.34	0.34	0.61	0.69	0.63	0.25	0.31
Std. Error	simulated	0.08	0.09	0.09	0.16	0.18	0.16	0.06	0.08
	no bootstrap	0.08	0.09	0.09	0.16	0.18	0.16	0.06	0.08
	MI Boot	0.08	0.09	0.09	0.16	0.18	0.16	0.07	0.08

	Method	Setting 2 (high missingness)						Setting 4	
		β_1	β_2	β_3	β_4	β_5	β_6	ψ_1	ψ_2
Coverage Probability	1) MI Boot (PS)	89%	91%	92%	91%	92%	92%	94%	94%
	2) MI Boot	91%	93%	94%	94%	94%	94%	94%	94%
	3) Boot MI (PS)	96%	97%	98%	98%	97%	98%	94%	94%
	4) Boot MI	90%	93%	93%	95%	94%	94%	94%	92%
	5) no bootstrap	91%	93%	94%	94%	94%	94%	–	–
Median CI Width	1) MI Boot (PS)	0.44	0.40	0.44	0.79	0.87	0.78	0.20	0.21

Method	Setting 1	Setting 2 (low missingness)						Setting 3	
	β_1	β_1	β_2	β_3	β_4	β_5	β_6	β_1	β_2
2) MI Boot	0.48	0.44	0.49	0.87	0.95	0.86	0.20	0.22
3) Boot MI (PS)	0.58	0.51	0.59	1.03	1.12	1.01	0.21	0.23
4) Boot MI	0.47	0.43	0.47	0.84	0.92	0.82	0.20	0.22
5) no bootstrap	0.48	0.44	0.49	0.87	0.95	0.86	–	–
Std. Error									
simulated	0.12	0.11	0.12	0.22	0.24	0.21	–	–
no bootstrap	0.12	0.12	0.12	0.22	0.24	0.22	–	–
MI Boot	0.12	0.11	0.12	0.22	0.24	0.21	–	–