ACLA WORKING PAPER SERIES

ADDRESSING MISSING DATA IN RANDOMIZED CLINICAL TRIALS: A CAUSAL INFERENCE PERSPECTIVE

ILJA CORNELISZ
PIM CUIJPERS
TARA DONKER
CHRIS VAN KLAVEREN

Working Paper 20201

http://www.acla.amsterdam/workingpapers-wp20201

AMSTERDAM CENTER FOR LEARNING ANALYTICS
van der Boechorststraat 7
1081 BT, Amsterdam
April, 2020

Addressing Missing Data in Randomized Clinical Trials: A Causal Inference Perspective

Ilja Cornelisz[1] Pim Cuijpers[2,3] Tara Donker[2,3] Chris van Klaveren[1]

[1]Amsterdam Center for Learning Analytics, Vrije Universiteit Amsterdam
[2] Department of Clinical, Neuro- and Developmental Psychology, Section Clinical Psychology, Vrije Universiteit Amsterdam
[3] Amsterdam Public Health Research Institute Amsterdam, Amsterdam, the Netherlands

Email: i.cornelisz@vu.nl, p.cuijpers@vu.nl, t.donker@vu.nl, c.p.b.j.van.klaveren@vu.nl (Corresponding Author)

Abstract

Background The importance of randomization in clinical trials has long been acknowledged for avoiding selection bias. Yet, bias concerns re-emerge with selective attrition. This study takes a causal inference perspective in addressing distinct scenarios of missing outcome data (MCAR, MAR and MNAR).

Methods This study adopts a causal inference perspective in providing an overview of empirical strategies to estimate the average treatment effect, improve precision of the estimator, and to test whether the underlying identifying assumptions hold. We propose to use Random Forest Lee Bounds (RFLB) to address selective attrition and to obtain more precise average treatment effect intervals.

Results When assuming MCAR or MAR, the often untenable identifying assumptions with respect to causal inference can hardly be verified empirically. Instead, missing outcome data in clinical trials should be considered as potentially non-random unobserved events (i.e. MNAR). Using simulated attrition data, we show how average treatment effect intervals can be tightened considerably using RFLB, by exploiting both continuous and discrete attrition predictor variables.

Conclusions Bounding approaches should be used to acknowledge selective attrition in randomized clinical trials in acknowledging the resulting uncertainty with respect to causal inference. As such, Random Forest Lee Bounds estimates are more informative than point estimates obtained assuming MCAR or MAR.

Keywords: Randomized Clinical Trials, Causal Inference, Attrition, Random Forest Lee Bounds

BACKGROUND

Causal inference and missing data

Empirical research is considered most valuable when it aims to answer a specific causal question (Angrist & Pischke, 2008), as is the case in a randomized clinical trial. Randomization avoids potential bias in the estimation of causal treatment effects by ensuring that a participant's treatment status is independent from both observed and unobserved characteristics. In other words, randomization enables internal validity (Rubin, 1974), such that differences between participants in a clinical trial assigned to different treatment arms can, apart from random error, be attributed to the treatment under investigation (Jüni, Altman & Egger, 2001).

For causal inference, it is important to empirically acknowledge the potential importance of missing data problems. These issues can introduce bias in subsequent data analyses, thereby undermining initial randomization. Several dimensions of missing data exist, which differ considerably in terms of potential causal inference concerns. We consider missing data on background characteristics, treatment status and outcomes.

Missing data on background characteristics can be addressed relatively easy, either by means of a missing indicator method or imputation, thereby rendering the sample of randomized participants complete. Given that background characteristics are independent from treatment status by virtue of randomization, the resulting causal inferences with respect to the average treatment effect are internally valid.

Missing data on treatment status occurs when participants discontinue, for whatever reason, the assigned treatment. When gathering data on outcomes for these participants is successful, the possibility to estimate an internally valid intent-to-treat effect is retained. This measure –in itself- is informative regarding treatment policy effects and provides a lower-bound estimate of the average treatment effect.

Missing data on outcomes is widespread in clinical trials (Hollis & Campbell, 1999; Pedersen et al., 2017). If such attrition co-varies with (un)observed characteristics and treatment assignment, this compromise initial randomization and causal inferences drawn from clinical trials (O'Neill & Temple, 2012). While attrition bias may thus render the results from a study non-informative, the problem of missing outcome data has received relatively little attention in the clinical trial literature (Sterne et al, 2009). Missing data is often inadequately addressed by means of simple fixes (Little et al., 2012), but also more advanced methods such as multiple imputation (MI) can yield biases as big as, or even bigger than, the bias in simple complete cases analysis results (Sterne et al., 2009).

Missing outcome data categories

Based on a widely established categorization of missing outcome data scenarios (Rubin, 1976), this study proposes empirical strategies for each scenario in the context of a randomized clinical trial and by adopting a causal inference perspective. An empirical solution for obtaining an internally valid treatment effect estimate in the presence of missing data is provided for each scenario. In addition, the underlying identifying assumptions that apply are made explicit, together with the extent to which –and how- they can be verified empirically.

Three missing data scenarios are distinguished, and each imposes different assumptions with respect to characterizing the missing data generation process:

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR)
3. Missing Not At Random (MNAR)

MCAR assumes that the occurrence of a missing outcome is a random event, unrelated to treatment assignment and –observed and unobserved-background characteristics. MAR assumes that observed background characteristics can account for the non-random nature of missing data. Finally, MNAR assumes that the occurrence of a missing outcome is a non-random event that can be structurally related to both observed and –importantly- unobserved background characteristics. From a causal inference perspective, each scenario entails different empirical strategies to test its underlying identifying assumptions and also to address missing data concerns.

Addressing selective attrition

This study argues that -from a causal inference perspective- it is most appropriate to assume that the missing outcome data generating process in a randomized clinical trial is MNAR. Both MCAR and MAR impose overly restrictive identifying assumptions that are crucial for claiming that the estimated effect is internally valid, yet often unrealistic and empirically unverifiable. In the case of MNAR, common empirical bounding approaches rely on exploring worst-case scenario bounds, often resulting in very wide, uninformative, treatment effect estimate intervals.

Therefore, this study proposes to use a new Random Forest Lee Bounds (RFLB) approach that can be used to estimate relatively tight average treatment effect intervals, when assuming MNAR. The potential of exploiting RFLB in the presence of selective attrition is demonstrate using data from a recent randomized clinical trial which established the treatment effect of virtual reality cognitive behavior therapy (VR CBT) in addressing acrophobia (Donker et al., 2019). We show how RFLB can considerably increase the statistical power of bias-corrected treatment effect estimates for randomized clinical trials facing non-random missing outcome data, by exploiting both continuous and discrete predictor variables for attrition.

METHODS

Assume a randomized clinical trial conducted with the objective to estimate the effect of treatment T on outcome y_i for individual i (with $i = 1, \dots, N$). The two potential outcomes for individual i can then be represented by:

$$y_i = \begin{cases} y_{1i} & \text{if } T_i = 1 \\ y_{0i} & \text{if } T_i = 0 \end{cases},$$

and the individual treatment effect for individual i as:

$$y_{0i} + (y_{1i} - y_{0i})T_i.$$

Individuals are randomly assigned to one of both treatment conditions, such that always only one of both potential outcomes y_{0i} and y_{1i} is observed and thus that individual treatment effects cannot be assessed (Rubin, 1974). This observation is referred to as the "fundamental problem of causal inference" (Holland, 1986). Yet, the average treatment effect for these individuals can be estimated empirically using ordinary least squares regression:

$$y_i = \alpha + \beta \cdot T_i + \varepsilon_i, \qquad (1)$$

where α and β are the parameters to be estimated and where ε represents the classical error term, assumed to be identically and independently distributed (i.i.d) with mean zero ($E(\varepsilon_i) = 0$). The estimated expected outcome for treated and non-treated individuals is $E(y_i|T = 1) = \alpha + \beta$ and $E(y_i|T = 0) = \alpha$, respectively, yielding the difference between these expected outcomes, β, to represent an unbiased estimated average treatment effect. Performing an independent sample t-test would yield the same average treatment effect.

A more precise estimate of this treatment effect can be obtained empirically by including observed background characteristics that are related to y_i in Equation 1:

$$y_i = \alpha + \beta \cdot T_i + \boldsymbol{X_i}'\boldsymbol{\delta} + \varepsilon_i. \qquad (2)$$

Vector **X** represents the vector of observed background characteristics and parameter vector $\boldsymbol{\delta}$ are the associated model parameters to be estimated. The estimated treatment effect is unbiased in both Equations 1 and 2, but the inclusion of background characteristics that are significantly related with the outcome variable reduces the standard error of β and thus increases the statistical power of empirically detecting this treatment effect.

RESULTS

Missing Completely At Random (MCAR)

MCAR thus assumes that missing outcomes are unrelated with treatment status and *all* background characteristics, whether observed or unobserved. Let indicator variable *M*(issing) denote 1 when outcome y_i is missing, and 0 otherwise. This *conditional mean zero* (CMZ) assumption of MCAR then implies $E(\varepsilon_{\text{i}}|T_i, \boldsymbol{X_i}) = 0$, meaning that both observed treatment T_i and background characteristics **X** are uncorrelated with the error term ε_i.

Let N_{fs} and N_{os} be the number of observations for, respectively, the full sample and the sample of individuals for which an outcome is observed and let $N_{os} \leq N_{fs}$. If the occurrence of missing outcomes is truly random, then *P(T|M)* = *P(T)* and an unbiased estimator of the treatment effect can be obtained empirically by estimating either Equation 1 or –for precision reasons- Equation 2.

The analysis can then be performed on the 'complete case' sample, also referred to as listwise deletion. A drawback of this procedure is that the use of the smaller sample N_{os} can severely reduce statistical power. More precisely, if the goal was to have 80% statistical power to empirically detect a treatment effect, this power reduction by using the smaller sample N_{os} is reflected by θ, for which holds:[1]

$$\theta = 2.8 \cdot \sqrt{\frac{N_{os}}{N_{fs}}} - 1.96. \qquad (3)$$

A power of 80% is associated with a θ-value of 0.84 (i.e. $N_{os} = N_{fs}$). Equation 3 illustrates that the statistical power is reduced by attrition as $N_{os} < N_{fs}$. For example, when $\sqrt{\frac{N_{os}}{N_{fs}}} = 0.75$, this implies a value for θ of 0.14, yielding a power of 56% and –thus- a power reduction of 24 percentage points (see Appendix A).

Therefore, a-priori knowledge about the extent to which missing outcomes will occur should be used to make an educated guess on the full-sample size needed (N_{fs}), as to retain the ability to perform a complete case analysis on the observed sample (N_{os}) with 80% statistical power.

A more fundamental problem with *MCAR* is that the randomness assumption of missing outcomes (CMZ) cannot be demonstrated, but merely be assumed. Empirically, this assumption can only be partially validated by two different estimation strategies. The first strategy is to estimate a (logistic/probalistic) regression model and show that treatment status T_i cannot be predicted empirically by the set of observed background characteristics **X** :

$$T_i = \alpha + \boldsymbol{X_i}'\boldsymbol{\delta} + \varepsilon_i. \qquad (4)$$

Generally a logistic or probability model is estimated, given the dichotomous nature of the dependent variable. Yet, for the validation purpose of causal inference considered here a simple linear regression model is sufficient to estimate, as we are not interested in the interpretation of the estimation parameters. The observed background characteristics should not explain treatment status if the identifying assumption of *MCAR* holds and the estimated coefficients of vector $\boldsymbol{\delta}$ should all be statistically insignificant and close to zero. A second estimation strategy is to estimate –and compare- Equations 5 and 6:

$$y_i = \alpha + \beta \cdot T_i + \varepsilon_i. \qquad (5)$$
$$y_i = \alpha + \beta \cdot T_i + \boldsymbol{X_i}'\boldsymbol{\delta} + \varepsilon_i. \qquad (6)$$

The estimated parameter β in the Equation 5 provides the t-test or Wald-test estimator of the treatment effect. The inclusion of observed background characteristics in Equation 6 should not statistically significantly change the estimated treatment effect β if the MCAR assumption of CMZ holds. The logic is similar to that of the first estimation strategy: if it is assumed that *P(T|M)* = *P(T),* then *P(T|M,* ***X****)* = *P(T)* must also hold. However, if missing outcomes are structurally related to relevant background characteristics, then the estimated treatment effect -or treatment assignment- will depend on whether these background characteristics are included or not.

A key insight from both strategies is that showing that T_i is independent from observable characteristics is sufficient to argue this will also hold for all unobserved characteristics. The estimated treatment effect β can properly be inferred causally if this identifying assumption holds, but this can thus merely be assumed and not demonstrated unequivocally.

Based on the above, it could be argued that the term missing completely at random is somewhat confusing. It suggests that missing outcomes are not problematic due to its random nature, yet it is precisely this random nature that

[1] See Appendix A for a full derivation of θ

cannot be proven. At best, one can show empirically that the estimated treatment effect is independent from observed characteristics and –then- to assume that this will also be the case for all unobserved characteristics.

Missing At Random (MAR)

Different from MCAR, MAR assumes that missing outcomes can actually be non-random in nature, such that the CMZ not hold and Equation 5 will yield a biased treatment effect estimate. Yet, with MAR it is assumed that the selective nature of missing outcomes can fully be accounted for by a set of observed background characteristics **X**. The term 'missing at random' therefore not defines missing outcomes as random events. Instead, the occurrence of missing outcomes are considered to be perfectly explained by certain observed background characteristics. This implies that P$(T|M, \boldsymbol{X}) = P(T|\boldsymbol{X})$, such that treatment status T_i -once controlled for observed background characteristics- is unrelated to missing observation status. This identifying assumption is referred to as *conditional mean independence* (CMI) and implies $E(\varepsilon_i|T_i, \boldsymbol{X}) = E(\varepsilon_i|\ \boldsymbol{X}) \neq 0$. This is unproblematic as long as we are only interested in inferring the causal effect of treatment T_i and not in the causal effect of background characteristics **X.** Conditional on X, T_i is as if randomly assigned, such that treatment T_i is uncorrelated with ε_i, but **X** can be correlated with ε_i. The CMI assumption then implies that $E(\varepsilon_i|T_i = 1, \boldsymbol{X}) = E(\varepsilon_i|T_i = 0, \boldsymbol{X})$, yielding $E(y_i|T = 1, \mathbf{X}) - E(y_i|T = 0, \mathbf{X}) = \beta$ and thus that the treatment effect estimate from Equation 6 is unbiased.

Whereas *MAR* is different from *MCAR* in assuming *conditional* randomness with respect to the missing outcomes, the underlying identifying assumption necessary to argue that the estimated treatment effect is causal is similar for both scenarios. As with *MCAR*, the validity of the identifying assumption for MAR –conditional mean independence- cannot be demonstrated. The two estimation strategies outlined before with *MCAR* can again be pursued, but are this time performed to show (1) how the probability of treatment depends on observed background characteristics, and (2) how the inclusion of background variables alters the estimated treatment effect.

The pivotal assumption underlying *MAR* is that the inclusion of the set of observed background characteristics **X** in Equation 6 is sufficient to also account for the potential effects of non-observed background characteristics on observed treatment status. This assumption is similar to situation described in *MCAR*, and –again- it is not possible to empirically validate this conditional mean independence assumption on non-observed background characteristics. In practical terms, this implies that *MCAR* and *MAR* are similar in the identifying assumptions made, with the notable exception that the inclusion of observed background characteristics in Equation 5 can alter the estimated treatment effect for *MAR*, while this is not allowed to be possible with *MCAR* (and are only included for precision).

The aforementioned observation that the estimated treatment effect is allowed to change with *MAR* due to the inclusion of ***X*** is what makes it distinctively different from *MCAR*. Whereas the estimation strategy in Equation 5 with *MCAR* serves to falsify its identifying assumption of missing outcomes as purely random events, it is used with *MAR* to empirically control for the non-random nature of missing outcomes. The identifying assumption for *MAR* that treatment status is then CMI on unobservables -after having included the observed characteristics **X-** is a strong assumption to make and empirically untractable.

Missing Not At Random (MNAR)

Given that randomness of missing outcomes cannot be demonstrated, a more appropriate point of departure is MNAR and to assume that missing outcomes *can* result from non-random events (Graham, 2009), such that the probability of observing a missing outcome can be structurally related to both observed and –importantly-unobserved characteristics. In contrast to MAR, MNAR thus implies that the inclusion of observed background characteristics **X** cannot control for the selective nature of missing values, rendering results thus obtained to be biased.

MNAR thus not assume CMZ or CMI, such that no unbiased point estimate for β can be generated using Equations 5 or 6. As such, additional assumptions regarding the potential selective nature of missing outcomes are required to estimate β. Empirically, this often implies performing a bounding procedure in which two contrasting scenarios regarding attrition (e.g. “worst”- and “best”-case) are considered; thereby yielding an interval estimate for β instead.

Bounding Procedures

Rather than correcting the point estimate of treatment effect β for potential bias, bounding procedures yield treatment effect interval estimates instead. In their seminal work on dealing with missing data in randomized experiments, Horowitz and Manski (2000) first provide a general assumption-free framework, with the objective to deal with non-random missing outcomes in an experimental setting. Their approach places very conservative bounds around treatment effect estimates. Essentially, this approach implies determining the worst- and best-case scenario for the missing outcomes, based on the observed data. By replacing the missing outcomes with the worst- and best-case outcomes, a lower and upper bound is constructed for the estimated treatment effect β. While this approach provides an intuitive benchmark, a major disadvantage is that these conservative bounds in clinical trials suffering from sizable attrition can often turn out to be largely uninformative, as they tend to be very wide and (thus) often include zero. Yet, this bounds analysis approach is valuable in being among the first to (1) acknowledge the potential severity of the imposed bias on the point estimate due to missing outcome data, and (2) provides an intuitive solution for dealing with this uncertainty by focusing on bound estimates for β, rather than on a single point estimate.

By imposing a monotonicity assumption on the selection mechanism of missing outcomes, Lee (2009) provides a -generally- more useful bounds procedure which generates relatively tighter bounds [10]. This assumption entails that assignment to treatment can only affect the likelihood of attrition in one direction, such that there is no heterogeneous effect –in terms of sign- of treatment assignment on attrition.

Let S_i be a binary indicator, with $S_i = 0$ indicating that outcome y_i is not observed due to attrition. Furthermore, let $1(\cdot)$ be the indicator function which counts the number of observations satisfying status indicated between parenthesis. The proportion of outcomes observed for treatment (T=1) and control (T=0) group can then be written as:

$$\begin{aligned} q_{T=1} &= \frac{\sum_i 1\cdot(T_i=1,S_i=1)}{\sum_i 1\cdot(T_i=1)} \\ q_{T=0} &= \frac{\sum_i 1\cdot(T_i=0,S_i=1)}{\sum_i 1\cdot(T_i=0)}. \end{aligned} \qquad (7)$$

The bounds procedure first determines which group suffers less from attrition. In accordance with the empirical example presented later in this section, and without loss of generalization, we assume that $q_{T=0} > q_{T=1}$. The Lee bounds procedure then trims the outcome distribution of the control group by removing $q = \frac{q_{T=0}-q_{T=1}}{q_{T=0}}$ observations from the lower (upper) end of the distribution such that an upper (lower) bound for the treatment effect β is estimated. Intuitively, the q-value indicates the proportion of control observations that ought to be removed in order to achieve a similar attrition rate in both the control and treatment group.

If the monotonicity assumption of sample selection holds, the treatment effect bounds estimated apply to so-called *never-attriters*. These are participants for whom the outcome will be observed ($S_i = 1$), irrespective of being assigned to the treatment (T=1) or control (T=0) group. By virtue of randomization, the treatment control difference in means will then provide bias-corrected bounds estimates of the average treatment effect for this particular group of participants. This subpopulation can be characterized by the distribution of **X** for participants with non-missing outcome data in the assignment group that remains untrimmed (here: treatment). Also, when the trimming proportion is zero, there is a limited test of whether the monotonicity assumption holds for observables (Lee, 2009) by estimating Equation 4 on the subsample of participants selected in the bounding procedure and by verifying that treatment status T_i cannot be predicted by **X.**

Empirical Example

To showcase the Lee bounds procedure, the randomized clinical trial data from Donker et al. (2019) is utilized. The objective of this study was to estimate the treatment effect of ZeroPhobia - a virtual reality cognitive behavior therapy (VR CBT) – on acrophobia symptoms. In a single-blind randomized clinical trial, 193 participants were randomly assigned to the treatment (n= 96) and a wait-list control group (n = 97). To measure acrophobia the Acrophobia Questionnaire was used (Cohen, 1977) and the mean (SD) AQ-Total outcome scores for the treatment and control group were 48.46 (24.33) and 74.68 (21.55), respectively.

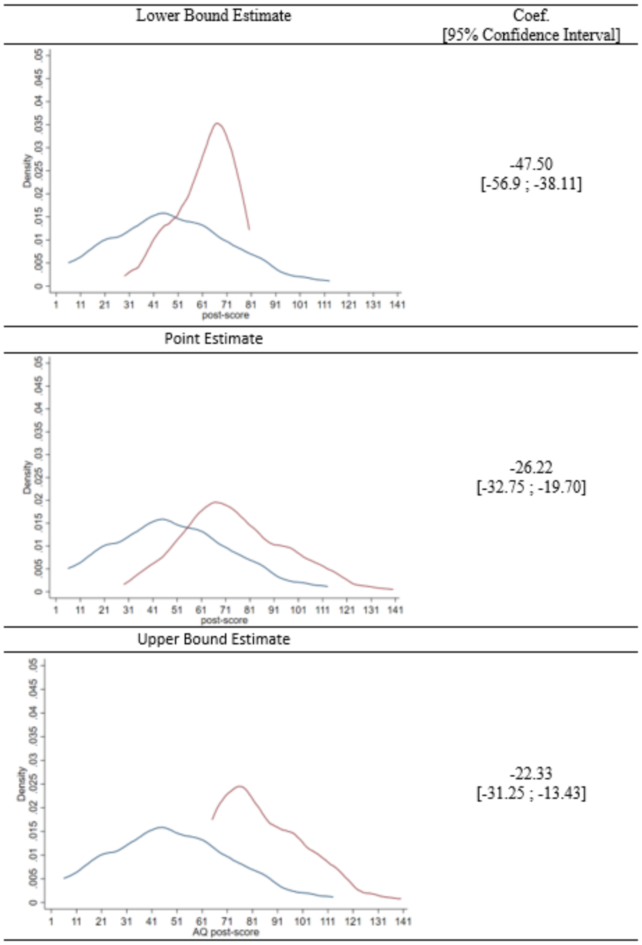
For 11 (40) participants of the control (treatment) group, no outcome was observed. The attrition proportion is therefore largest for the wait-list control group, such that the $q = \frac{q_{T=0} - q_{T=1}}{q_{T=0}} = \frac{(86/97) - (56/96)}{(86/97)} = 0.34$ (or 34 percent). To generate a lower and upper bound estimate for treatment effect β, the Lee bounds procedure then trims the outcome distribution of the wait-list control group by removing consecutively the 34 percent highest and lowest outcome values observed.

Figure 1 visualizes the Lee bounds, when applied to the data of Donker et al. (2019). Figure 1 shows three panels in which the blue line represents the untrimmed intervention distribution and the red line the distribution of the control group, which is either trimmed (i.e. the upper and lower panel) or untrimmed (i.e. the middle panel).

The middle panel shows the treatment effect point estimate as a reduction in acrophobia symptoms with 26.22 points (i.e. 74.68 - 48.46), when potential attrition bias as a result of MNAR is not taken into account. The distributions associated with the lower bounds estimate display that the upper 34 percent of the control distribution is trimmed (removed), yielding a lower bound estimate of -22.33. Moreover, the confidence interval of this lower bound estimate indicates that the mean difference of the AQ post-score is significantly smaller than zero. This is an important empirical result, as it indicates that even the most conservative Lee bound estimate indicates that ZeroPhobia significantly reduced acrophobia symptoms. The upper bound estimate of -47.50 points is obtained by trimming the lower 34 percent of the control distribution. The Lee bound results thus indicate that the estimated treatment effect is arguably somewhere between this interval of -22.33 and -47.50 points. When the confidence intervals for the point estimates of both bounds are also taken into consideration, the corresponding interval estimate for the average treatment effect is [-13.43 ; -56.90].

The Lee bound procedure can also be performed conditionally on discrete background characteristic variables, with the objective to further tighten the bounds thus obtained. This conditioned approach essentially implies that the trimming is performed separately for each category as defined by the discrete variable(s). In order to illustrate this ‘tightening by conditioning’- effect, suppose that only gender is used as the conditioning variable and that attrition occurs solely among female participants. In the unconditional bounding procedure, all extreme values of the control group outcome distribution are trimmed, regardless gender. In the conditional version, however, extreme observations of male participants in the control group are not trimmed (i.e. only extreme observations of women are). If we would trim only the 34 percent for women in the control group, then the conditional and unconditional Lee bounds are similar only if these trimmed female observations are all located at the extremes of the distribution. If this is not the case, the bounds will be relatively tight, as the resulting lower and the upper bound estimates will be closer around the initial point estimate of β.

Figure 1: Lee Bounds Trimming Procedure



Random Forest Lee Bounds

When assuming MNAR and estimating corresponding bounds for the average treatment effect, a limitation of the aforementioned conditioning procedure with Lee bounds is that it requires discrete groups to perform the trimming procedure. As such, the set of background characteristics that can be used to tighten the bounds is restricted to include only non-continuous variables. Thus, if attrition can to a large extend be explained by continuous variables, this information can unfortunately not be optimally exploited by the Lee bounds procedure. Therefore, we propose a Random Forest Lee Bounds (RFLB) approach (see Cornelisz, Cuijpers and Van Klaveren, 2019), such that also continuous variables can be used for obtaining tighter bounds around the estimate for treatment effect β.

The RFLB procedure first classifies data points efficiently into the attrition class they belong to. For this, a decision tree is used together with an entropy (E) function. The entropy function measures the purity of the data and -in the context of attrition- this indicates the proportion of attrition. The decision tree is used to partition data recursively into two groups such as to maximize data purity. This recursive data-splitting process can be explained using Figure 2, under the simplifying assumption that the observed input vector **X** consists of only two background characteristics, x_1 and x_2.[2] The red dots in the figure represent $\{x_1, x_2\}$ observations for missing outcomes, while the green dots represent non-missing outcome observations. In this example, the first algorithmic split occurred at $x_2 \geq a_2$, implying that the two data samples obtained by this split are more pure than the initial data sample. The second, third and fourth split occurred at $x_1 \geq a_4$, $x_1 \geq a_1$, and $x_2 \geq a_{3,}$ respectively.

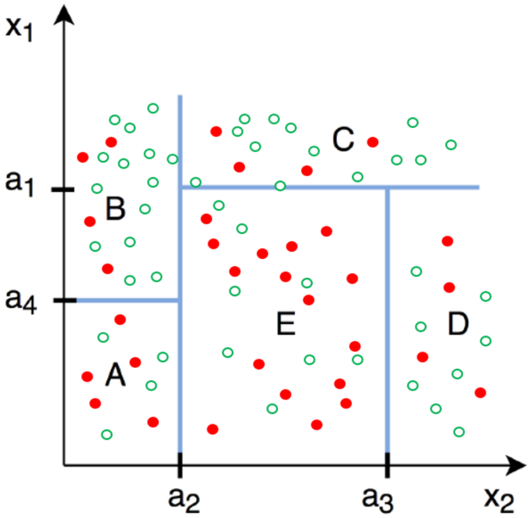
The algorithm stops when splitting the data does not yield more purified data samples and avoids further reduction of degrees of freedom for the adjusted residual sum of squares. We note that the random forest approach can be applied using multiple conditioning variables and is suitable for both categorical classification problems (such as shown in Figure 2), and continuous prediction problems. The example explained above can also be represented as a continuous outcome problem for which the random forest approach builds regression trees instead of classification trees and indicates – in our case - which characteristics best explain the observed variation in attrition.

The random forest approach avoids over-fitting by applying k-fold cross-validation and takes into account that small changes in the data can yield large changes in the final tree obtained, by selecting the splitting variable at each step from

[2] This example is taken from Plak et al. (2019)

m out of *x* randomly drawn input variables (Amit & Geman, 1997).

Figure 2: Decision Tree Data Splitting

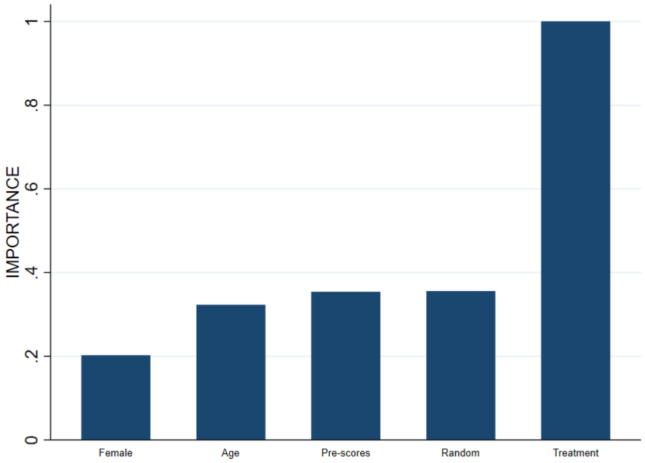


Note: this is a revised figure, taken from Zou & Schonlau (2018)

The Random Forest Lee Bounds (RFLB) procedure presented here can be summarized in four steps. First, a random forest is estimated using a set of background characteristics. Second, an importance graph is created which indicates how effective each characteristic reduces the variance (i.e. increases the prediction performance of the model). A random benchmark variable is included as covariate in the model, such that it can be evaluated if a covariate is more important than a random benchmark variable and –thus- whether it should be used to tighten the bounds. Third, a decision tree is estimated with only those variables included that were marked as important by the importance graph in the previous step. Fourth, discrete group indicator variables are generated using the decision tree from the previous step and used as conditioning variables in the Lee bounds procedure.

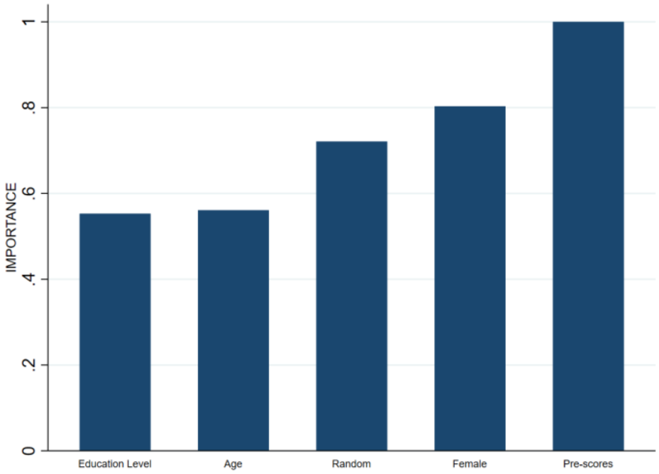
To reiterate, the rationale for using the random forest approach is to ensure that also continuous variables can be optimally exploited in the bounding procedure, and not to optimize the out-of-sample predictive power of the model. The obtained bounds will be tighter if these continuous variables statistically explain attrition. To see whether this is the case, random forest lee bounds are estimated applying the aforementioned procedure and using the ZeroPhobia randomized clinical trial data. The importance graph presented by Figure 3 indicates that the treatment indicator variable explains most of the attrition variation. This variable therefore receives an importance value of 1 and all other variables receive a value relative to that of the most important variable. The random variable receives a value of .3 and – importantly- all other variables (i.e, gender, age and AQ pre-scores) are considered less important in explaining variance than this random variable. This implies that the association between attrition and the variables gender, age and AQ pre-scores is not statistically significant and explains why Donker et al. (2019) perform an unconditional Lee bounds procedure. Yet, if the covariation between attrition and treatment assignment is indeed associated with some background characteristics, then the bounds can be tightened by conditioning on these variables.

Figure 3: Importance Graph Attrition Variation



To illustrate that additional conditioning on a continuous variable can further tighten the bounds, attrition is simulated in the data from Donker et al. (2019), such that the selective nature of the simulated attrition is now associated jointly with treatment status, gender and pre-score, while leaving the average treatment effect estimate unaffected. The simulation method is outlined in more detail in Appendix B. Figure 4 shows the importance graph and illustrates that the attrition is selective and conditional on the simulated characteristics gender and pre-scores.

Figure 4: Importance Graph Simulated Attrition



Note: the relative importance of the intervention variable is excluded from the figure, as only background characteristics are to be considered for tightening Random Forest Lee Bounds.

That attrition is selective with respect to gender and pre-score is a necessary but not sufficient condition for the bounds to be tightened by RFLB. It is only sufficient when the covariance between attrition and treatment assignment is conditional on these background characteristics as well. Figure 5 shows that these interaction terms are indeed important, which is why the variables Female and AQ pre-scores are selected for conditioning.

Figure 5: Importance Graph Simulated Attrition Interaction Terms

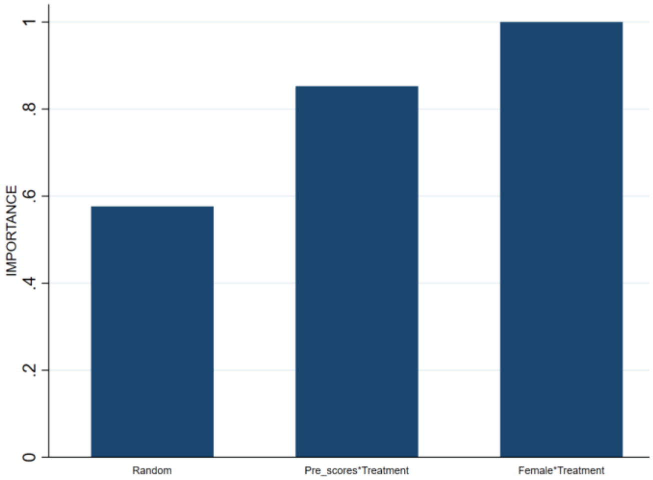


Table 1 summarizes the various bounding estimates obtained when performing unconditional Lee bounds, conditional Lee bounds and the RFLB introduced here. Columns 2 and 3 show the estimated lower and upper bound. Columns 4 and 5 show the 95% confidence intervals, which can be considered a more appropriate comparison between the different bounding approaches, as this also acknowledges precision loss resulting from the reduction of degrees of freedom when conditioning on background characteristics.

Table 1: Bounding Estimation Results

	Estimated Treatment Effect			
			95% conf. interval	
	Lower	**Upper**	**Lower**	**Upper**
Unconditional Lee- bound	-37.33	-17.89	-39.86	-15.44
	***Width**: 19.44*		***Width**: 24.42*	
Conditional Lee bound	-37.70	-19.85	-40.15	-17.49
(Female)	***Width**: 17.85*		***Width**: 22.66*	
Random forest Lee bound	-32.65	-20.39	-35.86	-17.94
(Female, Pre-scores)	***Width**:12.26*		***Width**:17.92*	

The estimation results of Table 1 show that the intervals are most tight when RFLB is performed. The 95% confidence intervals indicate that –relative to unconditional Lee bounds- the reduction in interval width is 7.2% with conditional Lee bounds, and 26.6% with RFLB. It follows that RFLB is the preferred bounding approach, as it provides the most precise bias-corrected interval estimate for the average treatment effect of ZeroPhobia.

More generally, it holds that RFLB will always produce interval widths that are smaller or similar compared to the other bounding approaches *if* the importance graph indicates that the interaction terms explain a significantly amount of the variation in attrition. This result is intuitive, since both conditional bounding approaches ensure that similar or less extreme values of the observed outcome distributions are trimmed, resulting tighter interval estimates. Moreover, RFLB allows for also conditioning on continuous variables, which is crucial for precision gains in randomized clinical trials as the pre-treatment scores is –often- the best predictor of the outcome observed at end point.

DISCUSSION

This study proposes to use a new Random Forest Lee Bounds (RFLB) approach that can be applied to estimate relatively tight average treatment effect intervals, while allowing for non-random missing outcome data (MNAR). An appealing feature of the RFLB procedure is that it allows for tightening intervals by exploiting the observed associations of attrition with both categorical *and* continuous background characteristics, as has been demonstrated empirically in this study.

In the case of a randomized clinical trial, the internal validity of RFLB entails an additional identifying monotonicity assumption. In the case of randomized clinical trials, this implies that if –for example- assignment to treatment induces some participants to drop out, it cannot simultaneously cause other participants to *no*t drop out. If this assumption holds, RFLB results are internally valid for the subpopulation of *non-attriters.*

Missing outcome data are unavoidable in randomized clinical trials and arguably depend on non-random unobserved data (Graham, 2009). Bounding approaches serve to accommodate such selective attrition processes and to acknowledge the resulting uncertainty with respect to causal inference. Yet, the potential of bounding procedures for generating informative –relatively precise- interval estimates for the average treatment effect depends crucially on the ability to statistically predict attrition using pre-treatment characteristics.

CONCLUSIONS

The importance of randomization in clinical trials has long been acknowledged for avoiding selection bias and enabling internal validity of the research results reported. Yet, many clinical trials suffer from attrition between randomization and follow-up data collection, such that attrition bias reintroduces concerns with respect to causal inference. This study takes a causal inference perspective when addressing distinct scenarios of missing outcome data (MCAR, MAR and MNAR) and provides an overview of empirical strategies that can be applied to estimate the average treatment effect, improve precision of the estimator, and to test whether the underlying identifying assumptions hold.

While many studies assume MCAR or MAR, the corresponding identifying assumptions are very restrictive, often untenable (Graham, 2009), and can hardly be verified empirically using the observed data. It is therefore appropriate to allow missing outcome data in clinical trials to be the result of non-random unobserved events (i.e. MNAR) and to address this issue accordingly. Yet, existing empirical approaches for dealing with MNAR rely on exploring worst-case scenario bounds and often result in very wide, uninformative, treatment effect estimate intervals.

By exploiting information from all observed background characteristics, the RFLB procedure proposed in this paper can considerably increase the statistical power of bias-corrected treatment effect estimates for randomized clinical trials facing non-random missing outcome data. As such, RFLB can hopefully facilitate researchers -whether involved in clinical trials or other disciplines- in allowing for post-assignment attrition to be non-random in nature and in addressing this accordingly.

LIST OF ABBREVIATIONS

MCAR: Missing Completely At Random; MAR: Missing At Random; MNAR: Missing Not At Random; RFLB: Random Forest Lee Bounds; MI: Multiple Imputation; Conditional Mean Zero: CMZ; Conditional Mean Independence: CMI; Acrophobia Questionnaire: AQ.

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

IC and CvK defined the baseline model, developed the method presented, performed the empirical analyses and were major contributors in writing the manuscript. PC provided input on the scope of the manuscript, and made contributions in writing the manuscript. TD provided the data used for demonstrating the methods presented in this paper and provided input on the simulated attrition data. All authors read and approved the final manuscript

Acknowledgements

Not applicable

REFERENCES

1. Angrist J. and Pischke J.S. (2008). Mostly Harmless Econometrics, Princeton University Press, NJ.

2. Cohen D.C. Comparison of self-report and overt-behavioral procedures for assessing acrophobia. *Behav Ther*. 1977;8(1):17-23. doi: 10.1016/S0005-7894(77)80116-0

3. Committee for Medical Products for Human Use. Guideline on missing data in confirmatory clinical trials. London: European Medicines Evaluation Agency, April 2009 (http://www.ema.europa.eu/pdfs/human/ewp/177699endraft.pdf).

4. Cornelisz, I., Cuijpers, P. & Van Klaveren, C. (2019) Treatment Effect Bounds with Missing Outcome Data: A Machine Learning Causal Inference Approach, *Mimeo*

5. Donker, T., Cornelisz, I., van Klaveren, C., et al. (2019). Effectiveness of Self-guided App-Based Virtual Reality Cognitive Behavior Therapy for Acrophobia: A Randomized Clinical Trial. JAMA Psychiatry. Published online March 20, 2019. doi:10.1001/jamapsychiatry.2019.0219

6. Gelman, A. & Hill, J. (2006) Data Analysis using Regression and Multlevel/Hierarchical Models, Cambridge University Press, New York.

7. Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual Review of psychology, 60, 549-576.

8. Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. Bmj, 319(7211), 670-674.

9. Horowitz, J. L., and Manski, C. F. (2000a) "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data" Journal of the American Statistical Association, 95, 77-84.

10. ICH harmonised tripartite guideline: choice of control group and related issues in clinical trials: International Conference on Harmonisation, Topic E10. London: European Medicines Agency, 2001 (http://www.ema.europa.eu/pdfs/human/ich/036496en.pdf).

11. Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. Bmj, 323(7303), 42-46.

12. Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Review of Economic Studies 76: 1071-1102

13. Little, R.J. et al. The Prevention and Treatment of Missing Data in Clinical Trials N ENG J MED 2012; 367;14

14. National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: National Academies Press, 2010. (http://www.nap.edu/catalog.php?record_id-12955).

15. O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther 2012;91:550-4.

16. Plak, S., Cornelisz, I., Meeter, M. & Van Klaveren, C. (2019) " Early Warning Systems for More Effective Student Counseling in Higher Education – Evidence from a Dutch Field Experiment". ACLA Working Paper Series 20191.

17. Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688–701.

18. Rubin, DB. Inference and missing data. Biometrika 1976;63: 581-92.

19. Statistical principles for clinical trials; step 5: Note for guidance on statistical principles for clinical trials: International Conference on Harmonisation, Topic E9. London: European Medicines Agency, 1998 (http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html).

20. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A.M., & Carpenter, J. R. (2009). Multiple imputation for

missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338, b2393.
21. Zou, R.Y. and Schonlau M. (2018) The Random Forest Algorithm for Statistical Learning with Applications in Stata.

APPENDICES

Appendix A

In Appendix A, we derive Equation 3. We define a full sample (*fs*) that includes all observed and missing outcome, and an observed sample (*os*). For expositional reasons, but without loss of generalization, assume that β can only be expected to be positive. Then, to find the required sample size such that exactly 80% of the average treatment effect estimates will be statistically significantly different from 0, the following should hold:

$$0 + 1.96 \cdot SE_{fs} = \beta - 0.84 \cdot SE_{fs}. \tag{7}$$

The left-hand side of this equation indicates that the treatment effect should be at least 1.96 standard errors away from 0 in order to attain statistical significance. The right-hand side reflects that the treatment effect β should be an additional 0.84 standard errors larger, as for the corresponding randomized clinical trial to have an associated statistical power of 80% (Gelman & Hill, 2006). Without loss of generality, we conveniently assume equal group sizes and that the variation is the same in both groups ($\sigma_{treatment} = \sigma_{control} = \sigma_{fs}$), such that $SE_{fs} = \frac{2\sigma_{fs}}{\sqrt{N_{fs}}}$ and we can solve for the required sample size conditional on the expected value for β (Gelman & Hill, 2006).

The following two equations have to be considered to link power reduction to missing outcome observations and a corresponding smaller observed sample size N_{os}:

$$\beta_{fs} = 1.96 \cdot SE_{fs} + 0.84 \cdot SE_{fs} = 2.8 \cdot SE_{fs} \tag{8}$$
$$\beta_{os} = 1.96 \cdot SE_{os} + \theta \cdot SE_{os}. \tag{9}$$

Equation 8 is the rearranged version of equation 7, such that β_{fs} is isolated on the left-hand side. Under the assumption that N_{fs} was chosen such that the power of the randomized clinical trial was set at exactly 80%, Equation 7 represents the implications for observing $N_{os} < N_{fs}$ in which θ reflects that the trial has less than 80% power due to the reduced sample size (i.e. $\theta < 0.84$).

An important feature is that both β_{fs} and β_{os} are unbiased estimators of the treatment effect if and only if missing outcome observations are truly random events. By acknowledging that $\beta_{fs} = \beta_{os}$ and $\sigma_{fs} = \sigma_{os}$, Equation 9 can be substituted in Equation 8, after which solving for θ leaves:

$$\begin{aligned} \theta &= 2.8 \cdot \frac{SE_{fs}}{SE_{os}} - 1.96 \\ &= 2.8 \cdot \sqrt{\frac{N_{os}}{N_{fs}}} - 1.96 \end{aligned} \tag{10}$$

Equation 10 shows the power reduction as a result of sample size reduction. A standard normal table, which present the values of the cumulative distribution function of the normal distribution, indicates that θ=0.84 is corresponding to a power of 80%. Yet, when $\sqrt{\frac{N_{os}}{N_{fs}}}$ is, for example, .75 generates a value for θ of 0.14, yielding a power of 56 percent and –thus- a power reduction of 24 percentage points.

Appendix B

This appendix provides a detailed outline of simulated data attrition for demonstrating the RFLB approach. First, the experimental sample was expanded by 10, such that the number of observations is larger (N=1930) and power issues are no longer a concern, while leaving the average treatment effect unaffected. The simulated attrition variable was constructed, such that it was jointly associated with treatment, pre-score and gender. It is important to recognized that conditioning on background characteristics can only produce tighter bounds if there exists a statistical dependency between attrition, treatment and, respectively, pre-scores and gender.

Attrition was simulated by conducting the following 3 steps. In step one, a random uniform (0,1) variate R was generated]. In step 2, pre-score were ranked -from low to high- and a categorical pre-score variable (*PreC)* was generated breaking down pre-scores into five evenly spaced categories. In step three, a binary attrition variable was simulated using the following 13 rules:

1. Attrition = 1 if $R \leq 0.05$ & $T = 0$
2. Attrition = 1 if $R \leq 0.15$ & $T = 1$
3. Attrition = 1 if $R \leq 0.10$ & $T = 0$ & $PreC = 2$
4. Attrition = 1 if $R \leq 0.15$ & $T = 0$ & $PreC = 3$
5. Attrition = 1 if $R \leq 0.25$ & $T = 0$ & $PreC = 4$
6. Attrition = 1 if $R \leq 0.45$ & $T = 0$ & $PreC = 5$
7. Attrition = 1 if $R \leq 0.10$ & $T = 1$ & $PreC = 1$
8. Attrition = 1 if $R \leq 0.20$ & $T = 1$ & $PreC = 2$
9. Attrition = 1 if $R \leq 0.40$ & $T = 1$ & $PreC = 3$
10. Attrition = 1 if $R \leq 0.60$ & $T = 1$ & $PreC = 4$
11. Attrition = 1 if $R \leq 0.80$ & $T = 1$ & $PreC = 5$
12. Attrition = 1 if $R \leq 0.25$ & $T = 0$ & $Female = 0$
13. Attrition = 1 if $R \leq 0.40$ & $T = 1$ & $Female = 0$

Rules 1 and 2 ensure that there is an attrition difference between the control and treatment group of 10 percent. Rules 3 up to 11 ensure (1) that attrition is an increasing function of pre- scores, and (2) that this relationship is stronger for participants in the treatment group. Rules 12 and 13 ensure that there is more attrition for men than for women and, -again- that attrition is even higher for men in the treatment group.

Figure 6 shows the increasing relationship between the simulated attrition and pre- scores. The mean simulated attrition rate was .36.

Figure 6: Simulated attrition by pre-scores

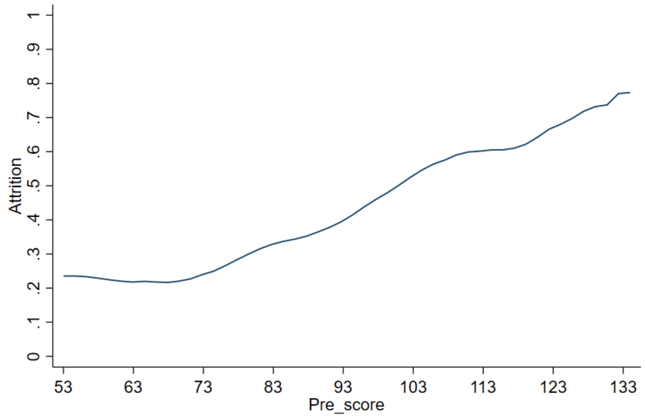
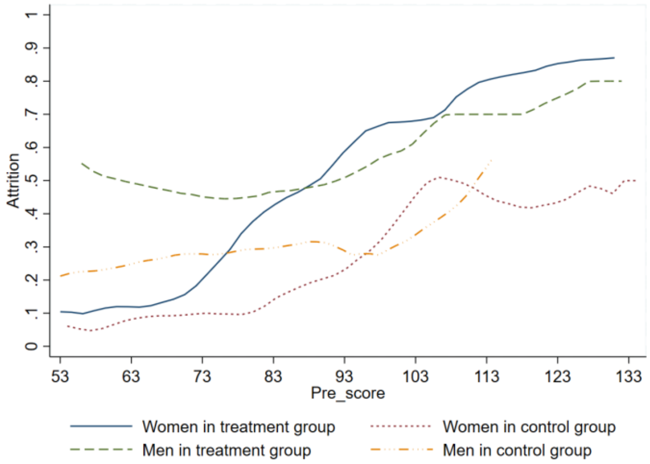


Figure 7 shows the resulting interdependency between treatment and, respectively, pre-scores and gender in relation to the simulated attrition variable after applying Rule 1-13.

Figure 7: Simulated attrition and pre-scores by gender and treatment status


Attrition
Pre_score
Women in treatment group
Women in control group
Men in treatment group
Men in control group