Handling missing outcome data in the one-year follow-ups of the three CHOICE trials: revised statistical analysis plan

Chris Rose, Norwegian Institute of Public Health, Oslo, Norway
14 November 2023

Introduction

This document outlines a revised plan for handling missing data — primarily due to loss to follow-up — in the analyses of the one-year follow-ups of the CHOICE trials performed in Kenya (PACTR202204883917313), Rwanda (PACTR202203880375077), and Uganda (PACTR202204861458660). The approach is based in part on the advice given in Jakobsen 2017. As of this writing, exploratory work on the issue has been performed for the Ugandan trial (see below), but the method has not yet been applied to any of the trials. The Methods section presents the planned approach. The Discussion outlines the logic behind the planned approach, along with alternatives considered.

Exploratory work shows that one year follow-up outcome data are missing for 1420 of the 4853 Ugandan students (i.e., 29%). Data are unlikely to be missing completely at random (MCAR) because missingness appears to be associated with being randomized to the intervention; scoring lower on the test used to assess the primary and secondary outcomes immediately after treatment; being older; attending a private school; and not attending a high-performing school. It seems likely that plausible explanations for these findings could easily be suggested.

Methods

Outcomes at one year were measured on students and teachers. In the interest of clarity, the following largely uses students as an exemplar, but an equivalent approach will be used to analyze data for teachers. Following Jakobsen, we will perform complete case analysis if outcome data are missing for no more than 5% of participants. In other words, missing outcome data is only problematic if a sufficiently large percentage of outcome data are missing.

If outcome data are missing for more than 5% of participants, we will use cross-validated elastic net logistic regression (Zou 2005) to automatically and impartially select variables associated with non-missingness. The variables entered into this model will include treatment allocation, the variables reported in the tables of baseline characteristics of the original trials (as applicable for the students or teachers), and score on the original test. The model will account for the cluster randomization for students (teachers are not clustered). If no variables are selected, and missing data are approximately balanced across trial arms, and there is a plausible rationale for outcome data being MCAR, then we will decide that the data are most likely MCAR and perform and report complete-case analyses.

If at least one variable is selected by the elastic net, then we will decide that outcome data are most likely missing at random (MAR; i.e., missingness is predictable using the available data). In this case all analyses will be performed using inverse probability weighting (IPW), similar to that originally planned for the Kenyan and Rwandan trials. We will compute the probabilities to use in an IPW analysis using the model obtained from the elastic net regression. If we decide that outcome data are likely MAR we will not also report complete-case estimates (because the MCAR assumption would not be justifiable).

If we decide that the outcome data are unlikely MCAR and unlikely MAR, we will consider data to most likely be missing not at random (MNAR; e.g., missingness may be caused by the value of the outcome variable). In this case we will perform and report complete case analyses (i.e., no IPW) and will not draw strong conclusions about treatment effect.

Irrespective of the above decisions, we will also estimate and report sensitivity analyses based on Lee (for continuous outcomes) and Manski-type (for dichotomous outcomes) interval estimates of treatment effect under conditions

that maximally favor and disfavor the intervention (Lee 2009; Horowitz 2000), again as originally planned for the Kenyan and Rwandan trials.

Note that outcome data for students and teachers may be missing for different reasons. Following the above plan, it is possible that, for example, an MAR decision may be made for students and an MCAR decision made for teachers.

Discussion

The logic behind the IPW approach is illustrated by the following example. Assume the probability[1] that outcome is non-missing for a given student to be 0.5. This implies that, on average, for every two students like that student, outcome data is missing for one of the two students. In the IPW analysis, the student who is not missing would therefore be weighted by 1/0.5 = 2. In other words, that one student would count as two students, one of whom is missing. IPW requires being able to estimate these probabilies sufficiently well. This should be reasonably achieved using the approach described above (but note the limitation below).

The IPW analysis planned differs from that originally planned in the Kenyan and Rwandan trials. The protocols for those trials planned to give each student "a weight equal to the inverse of the proportion of students in the school that completed the CTH Test". This approach cannot be used if none of the students in a school completed the test (loss of cluster — e.g., because of school closure): taking the inverse of zero gives an infinitely large weight. The method originally planned also implicitly assumes that the students who are not missing from a given class can be used as models for students who are missing from that class. This implies that missing and non-missing students are exchangeable within class, which contradicts the MAR assumption: if we think that students are lost to follow-up for a reason, rather than completely at random, then why would they be sufficiently similar to students not lost to follow-up? Finally, the original method does not extend to teachers (there is only one teacher per school).

We considered using multiple imputation (MI) to address the missing data issue. This is probably preferable to the IPW approach because MI accounts for uncertainty on the imputation model, while IPW does not account for uncertainty on the model used to compute the probability weights. However, from a pragmatic perspective, it would be very time-consuming to create an imputation model for every outcome for each of the three trials. Bayesian methods could be used, but suffer the same practical issues as MI.

References

Horowitz, Joel L., and Charles F. Manski. "Nonparametric analysis of randomized experiments with missing covariate and outcome data." *Journal of the American statistical Association*95.449 (2000): 77-84.

Jakobsen, Janus Christian, et al. "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts." *BMC Medical Research Methodology* 17.1 (2017): 1-10.

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

Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005): 301-320.

[1] If you think it is strange to talk about probabilities of events that have already occurred, you are correct. We should be using the word "likelihood" — i.e., the probability we would assign to an event had it not already occurred. However, the word "probability" is used in the term IPW, so I have used that word throughout.