Panel studies: Missing binary data in a survey of Whitebark pine

Cynthia Hollimon

Department of Mathematical Sciences
Montana State University

May 8, 2013

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Cynthia Hollimon

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

May 15, 2013
Date

Dr. Kathryn M. Irvine
Writing Project Advisor

05/16/13
Date

Megan D. Higgs
Writing Projects Coordinator

Panel studies: Missing binary data in a survey of Whitebark pine

Cynthia Hollimon

Montana State University, Mathematical Sciences

Advisor: Dr. Kathryn M. Irvine, USGS

(Dated: May 8, 2013)

I. INTRODUCTION

A. Whitebark pine problem and study

White bark pine (*Pinus albicaulis*) has been identified as a "keystone species" in the subalpine zone of the northern Rocky Mountains, a large portion of which falls into the Greater Yellowstone Ecosystem (GYE). The trees contribute much to the ecosystem including assisting in snow accumulation and providing a major food source for the area's native grizzly bears, red squirrels, and Clark's nutcrackers [1]. In recent years, there has been an increase in infestation by the mountain pine beetle (*Cronartium ribicola*) which is hypothesized to be more deadly to trees already weakened by White pine blister rust (*Dendroctonus ponderosae*) infections [2].

An ongoing panel study visits randomly selected transects every two years to track mortality and pine beetle infestation in individual trees. This same study records blister rust infestation every other survey (that is, every four years). In order to evaluate the relationships between blister rust, pine beetle, and mortality, there must be a statistical decision made in how to best deal with the high level of missingness in the data (50%). Prior investigation indicated that the important blister rust states are likely presence/absence of bole cankers (individual infection sites), this is the state that we will be referencing throughout the paper [3]. Our application has several complicating factors. These include the longitudinal rotating panel design, the clustering of data due to transects, the binary nature of the covariate of interest, and the statistically rare occurrence of blister rust. For this analysis, we assume that the tree-level data can be treated as missing completely at random (MCAR). To be considered MCAR, the events that lead to the missingness are independent both of other observed variables and of anything unobserved that might tie them together [4, 5], the missingness in these data is random at the transect level but because trees cannot be random within transects the assumption is violated here.

B. Missing data methods

For our application, we explored methods used in working with clustered, longitudinal, binary data. A very commonly used method is complete case analysis, dropping incomplete observations from the data set. This is an easy approach that makes no assumptions regarding the missing data. However, if missingness is not at random, or if the rate of missingness is high, this can cause bias in estimates [6]. There are approaches to clustered data, taking into account the data structure when missingness is not at random, much as we see in our data [7]. However, these rely on some observed values within a cluster from which to take information for imputation, which we do not have since complete clusters are always missing.

Last observation carried forward (LOCF) uses the last known state for the missing value to fill the gap in data. This method performs well so long as the variable with missingness does not change value at a high rate [8]. Hot deck imputation is a common method for missing data. It takes the observed variables for a record with missingness and uses them to identify the most similar record without missingness in the data set; it then uses that information to fill the gap. There are various algorithms for identifying the most similar record [9]. Additionally, there are single and multiple imputation approaches with single relying on the one closest record, and multiple relying on identifying a pool of similar observations from which to randomly select [10]. Successful hot deck imputation relies on the ability to match observations based on important variables. The pattern of missingness in these data make that a problem. Since blister rust histories between panels are staggered, there are no complete records that can be matched on the most important variable of previous blister rust state at the same time periods.

Finding appropriate model-based approaches for binary data is an on-going effort. Traditionally, fitted probabilities have been to define binomial distributions for random draws [11, 12]. There are some very complex methods that are more statistically valid, but are challenging to implement with available software [13]. Recently, researchers have begun to use the normal approximation to the binomial so that more traditional model-based approaches can be used. Bernaards, et al. (2007) compare 3 methods of binary imputation[14]. The first is the "coin flip" method from defined binomial distributions, the second and third make draws from normal approximations based on fitted probabilities, one using .5 as a rounding threshold, and one using an adaptive threshold for rounding [15]. These approaches were compared using various sizes of data and rates of missingness. From this comparison, it was determined that for a larger data set with a higher rate of missingness, the best performing model-based approach to binary missingness is the normal approximation using adaptive rounding.

This paper will compare last observation carried forward, a coin flip approach using empirical cell proportions, and the Bernaards approach advocated for a large data set with high missingness. The predicted blister rust values for each of these methods will be compared to the true values from the data to compare performance. Additionally, these methods will be implemented in our actual data set and the estimates and associated errors will be compared with each other and a complete case analysis.

II. DATA

This study consists of 176 transects and 4770 trees. Transects can hold anywhere from 1 to 220 trees. The number of transects and trees in each panel can be seen in Table 1. Transects were established between 2004 and 2007, and before data collection in 2008 were randomly assigned to 1 of 4 panels.

Panel	Transects	Trees
1	43	1014
2	45	1088
3	44	1289
4	44	1379

TABLE I: Panel information

During the original establishment surveys, the trees were fully surveyed, meaning blister rust state is known. After the panels were established, 2 panels were visited each year with 2 different survey types implemented. There were full surveys in which mountain pine beetle status, blister rust status, and tree mortality were all recorded. There were also pine beetle only surveys in which only mountain pine beetle status and tree mortality were recorded. The pattern of panel visitation can be see in Figure 1, with X representing a full survey (known blister rust status) and O representing a partial survey (no record of current blister rust status). We can think of these as follows, with $\mathbf{X}_1$ indicating blister rust status:

$$\mathbf{X} = \{[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i], [\mathbf{Y}]\}$$

$$\mathbf{O} = \{[\mathbf{NA}, \mathbf{X}_2, \ldots, \mathbf{X}_i], [\mathbf{Y}]\}, \textit{where } i \textit{ is the total number of covariates}$$

2004-2007	Panel	2008	2009	2010	2011	2012
	1	X		O		X
observations	2		X		O	
prior to panel	3	O		X		O
establishment	4		O		X	

TABLE II: *Panel study design, X represents a full survey year for a given panel, O represents a "pine beetle only" survey year, that is, a year in which blister rust state was not recorded.*

As a result of this randomization, transects in panel 1 that underwent full surveys in 2008 could have been established anywhere from one to four years prior. This means that the last known blister rust status could be anywhere from one to four years out of date. Some transects in panel 4, receiving their first full survey since the implementation of the rotating panel design in 2011, last had blister rust status recorded in 2004, seven years prior. Our goal is to identify a method that best works with the data structure to take advantage of the information recorded, thereby minimizing wasted effort, and potentially decreasing error in model estimates. Although observations will be treated as missing completely at random (MCAR), in truth only the transects are MCAR and not the individual trees.

III. COMPARISON OF METHODS

This study aims to compare the predictive accuracy of last observation carried forward (LOCF), a binomial probability model using empirical proportions (BEP), and a normal approximation probability model with adaptive rounding which will be referred to as adaptive multiple imputation (AMI) from here. These will be compared based on measures of sensitivity and specificity, as well as Cohen's Kappa [16].

The data used for this section will be panel 1 data from 2012; we will predict the true 2012 blister rust values given known information. Our true interest is in predicting missing values 2 years after an actual observation, but this study will be predicting missing values 4 years after an observation. While the goal and approach are not completely aligned, if methods perform well over 4 years we would expect even better performance after only 2. The covariates used are last observed blister rust state, number of years since that status was first recorded, and the recorded diameter at breast height. Diameter was recorded during transect establishment, and never subsequently updated.

A. Imputation methods

Complete case analysis is not used in this section, but will come into play when comparing model estimates. The imputation methods compared using these three data sets are LOCF and model-based multiple imputation.

1. *Last observation carried forward*

Last observation carried forward (LOCF) takes advantage of the longitudinal nature of these data. It is a form of single imputation in which we assume no change in state from the most recent observation on the missing variable. In this case, we will fill our removed 2012 blister rust information using the observations taken for the same trees in 2008. LOCF can be an effective method for slow moving processes [8]. The literature states that blister rust is such a process, this is also supported empirically through our collected data. Using this process generates one complete, imputed data set.

2. *Empirical proportion imputation*

Using the 3 covariates of last observed blister rust state, number of years since that status was first recorded, and diameter at breast height (in 4 categories: (1) $DBH < 2.5$, (2) $2.5 < DBH < 10$, (3) $10 < DBH < 30$, (4) $30 < DBH$), we calculated the proportion of trees with blister rust and associated 95% confidence intervals. These empirical summaries can be found in Table III. Values marked with an asterisk were unobserved in any complete data, so were filled based on similar combinations of the variable levels.

DBH	Time persisting	Previous BR = 1 Proportion	Lower bound	Upper bound	Previous BR = 0 Proportion	Lower bound	Upper bound
1	1	1	1	1	0.33	0	1
	2	1	1	1	0.03	0	0.37
	3	0.81	0.05	1	0.03	0	0.37
	4	0.8	0.02	1	0.04	0	0.4
	5	1	1	1	0.03	0	0.39
	6	0.89	0.27	1	0.03	0	0.39
	7	0.90*			0.09	0	0.65
2	1	1	1	1	0.17	0	0.9
	2	0.94	0.48	1	0.05	0	0.47
	3	0.84	0.11	1	0.11	0	0.71
	4	0.91	0.35	1	0.1	0	0.68
	5	0.91	0.36	1	0.05	0	0.47
	6	0.91	0.34	1	0.14	0	0.83
	7	0.82	0.06	1	0.2	0	0.98
3	1	0.5	0	1	0	0	0
	2	1	1	1	0.08	0	0.62
	3	0.85	0.16	1	0.09	0	0.66
	4	0.88	0.23	1	0.14	0	0.82
	5	0.87	0.21	1	0.12	0	0.75
	6	0.92	0.38	1	0.29	0	1
	7	0.90*			0.07	0	0.58
4	1	0.5	0	1	0*		
	2	1	1	1	0	0	0
	3	0.5	0	1	0.11	0	0.73
	4	1	1	1	0.12	0	0.75
	5	1	1	1	0.17	0	0.9
	6	0	0	0	0.29	0	1
	7	0	0	0	0.25	0	1

TABLE III: Empirical proportions of trees with blister rust and associated 95% confidence intervals for the entire data set (observations from panel 1 in 2012). Values marked with (*) are unobserved, and inferred from similar combinations of covariates. The left columns are associated with trees that had blister rust bole cankers in the previous observation, and the right columns are associated with those that did not have blister rust previously.

For each tree, we use the proportion associated with its values on these three variables as the binomial probability for a single draw. We build 5 imputed data sets to account for the uncertainty of not knowing the truth. The literature on multiple imputation, even for binary processes, states that 5 is enough, and more than 5 does not improve efficiency of estimation [17].

3. Adaptive multiple imputation

Here we use a normal approximation to the binomial distribution with an adaptive threshold in order to impute our data. First, we model binomial probabilities using the same variables from which we calculated the empirical probabilities. Let p = previous blister rust state, t = time tree was known to be in that state, and d = diameter at breast height in cm, we will use these in a saturated model for prediction. This gives the following probability model for blister rust presence, using logistic regression:

$$\hat{Pr} = logit^{-1}(-2.843 + 4.815p + .075t + -.013d + -.034pt + -.016pd + .009td + -.006ptd + \epsilon_i)$$

For each observation missing a blister rust record, we input its values for previous state, known duration in that state, diameter at breast height, to generate a probability of blister rust in the bole for that particular tree. These fitted probabilities are then input to determine the mean and standard error for a normal approximation to the binomial distribution from which we draw an imputed probability, I_{Pr}.

$$I_{Pr}(BR = 1) \sim N(\hat{Pr}, \hat{Pr}(1 - \hat{Pr}))$$

Using these imputes, we then determine the binary value of the variable. A relatively unbiased method to accomplish this is to determine a threshold, T, based on an empirical mean.

$$T = \bar{\omega} - \Phi^{-1}(\bar{\omega})\sqrt{\bar{\omega}(1 - \bar{\omega})}$$

Using this model-based approach, we can again generate 5 impute data sets.

B. Comparison of methods

We compared performance of LOCF, BEP, and AMI. This included measures of sensitivity (true positive rate), specificity (true negative rate), and Cochran's Kappa (a relative measure of agreement between two sets of values). We present overall results, as well as those broken down by size class.

IV. RESULTS

Table IV gives a summary of the comparison of methods for all data, results by size class can be found in Tables V through VIII. We see that using all panel 1 observations, LOCF outperforms the other measures in both true negative rate and Kappa agreement. However, BEP has a generally higher true positive rate, and AMI performs closest to true when looking at the overall proportion instead of individual tree measures. Through all size classes, LOCF outperforms in specificity and Kappa agreement. This should be expected since most trees do not have blister rust to begin with, and most do not get it. Both of these measures, in this context, are essentially testing how well the model identifies a tree without blister rust, which is the most common state. Sensitivity and the overall proportion are more interesting measures for such a statistically

Data	Method	Specificity	Sensitivity	Kappa	Predicted proportion	True proportion
Full	LOCF	0.9622	0.5827	0.5981	0.1256	.1612
	AMI	0.9244	0.5512	0.4864	0.1523	
		0.8956	0.5591	0.4366	0.1777	
		0.9107	0.5591	0.4654	0.165	
		0.9047	0.5512	0.4473	0.1688	
		0.9032	0.5197	0.4189	0.165	
	BEP	0.9062	0.5906	0.4814	0.1739	
		0.8971	0.5906	0.4641	0.1815	
		0.9092	0.5669	0.4687	0.1675	
		0.9168	0.622	0.5271	0.1701	
		0.9047	0.5906	0.4785	0.1751	

TABLE IV: Performance comparison for all panel 1 data, 788 trees.

uncommon event. The range of AMI predicted proportions is more consistent with the true proportion for all but the smallest size class. This has a low true proportion, and AMI overestimates. For size class 4, all predictions are quite poor, and we have the lowest values for all measures. This is likely due to the small sample size of 38 in that size class. When judging performance by sensitivity, BEP outperforms the others, again for all but the smallest size class where all measure around .5. This may be due to the very small proportion of trees in this size class with evidence of blister rust.

Data	Method	Specificity	Sensitivity	Kappa	Predicted proportion	True proportion
Size 1	LOCF	0.9722	0.5	0.5228	0.0696	.0886
	AMI	0.9236	0.4286	0.3211	0.1076	
		0.9167	0.4286	0.3058	0.1139	
		0.9306	0.5714	0.4447	0.1139	
		0.8958	0.5714	0.3622	0.1456	
		0.9236	0.5714	0.4263	0.1203	
	BEP	0.9306	0.5	0.3926	0.1076	
		0.9514	0.5	0.4514	0.0886	
		0.9444	0.5	0.4306	0.0949	
		0.9583	0.5	0.4736	0.0823	
		0.9375	0.5	0.411	0.1013	

TABLE V: Performance comparison for size class 1, 158 trees.

Data	Method	Specificity	Sensitivity	Kappa	Predicted proportion	True proportion
Size 2	LOCF	0.9877	0.5417	0.6293	0.0997	.1649
	AMI	0.9465	0.5208	0.51	0.1306	
		0.8889	0.5417	0.4133	0.1821	
		0.9177	0.5208	0.4498	0.1546	
		0.9136	0.5	0.4242	0.1546	
		0.9218	0.5	0.4402	0.1478	
	BEP	0.9095	0.5208	0.4339	0.1615	
		0.9095	0.5417	0.4511	0.1649	
		0.9465	0.4792	0.4731	0.1237	
		0.9465	0.5625	0.5455	0.1375	
		0.93	0.4792	0.4385	0.1375	

TABLE VI: Performance comparison for size class 2, 291 trees.

V. APPLICATION

The real motivation behind this methods comparison was the high rate of error in mortality estimates generated in a previous analysis of these data [3]. Occurrence of blister rust is statistically quite low, and so estimates associated with postive blister rust values had especially high variance. We were interested in seeing the effect of doubling our sample size on these estimates by predicting blister rust values for the pine beetle only surveys in a given year. Here, we fit a model determined by our previous work to data from 2012 using the original data for the model as well as datasets supplemented by each of the methods explored. This means that the panel 3 data used includes observed bliser rust values and the panel 1 data used includes

Data	Method	Specificity	Sensitivity	Kappa	Predicted proportion	True proportion
Size 3	LOCF	0.9498	0.6452	0.6327	0.1728	.2060
	AMI	0.9079	0.6129	0.5271	0.1993	
		0.8954	0.6129	0.5053	0.2093	
		0.9079	0.5968	0.5139	0.196	
		0.8996	0.5806	0.486	0.1993	
		0.887	0.5323	0.4244	0.1993	
	BEP	0.8996	0.6774	0.5635	0.2193	
		0.8619	0.6613	0.4882	0.2458	
		0.8619	0.6613	0.4882	0.2458	
		0.8745	0.7097	0.5451	0.2458	
		0.8703	0.7097	0.5382	0.2492	

TABLE VII: Performance comparison for size class 3, 301 trees.

Data	Method	Specificity	Sensitivity	Kappa	Predicted proportion	True proportion
Size 4	LOCF	0.8286	0.3333	0.1006	0.1842	.0789
	AMI	0.8857	0.3333	0.1679	0.1316	
		0.8571	0.3333	0.1307	0.1579	
		0.8	0.3333	0.0757	0.2105	
		0.9143	0.6667	0.4452	0.1316	
		0.8	0.3333	0.0757	0.2105	
	BEP	0.8286	0.3333	0.1006	0.1842	
		0.8286	0.3333	0.1006	0.1842	
		0.8286	0.3333	0.1006	0.1842	
		0.8286	0.3333	0.1006	0.1842	
		0.8286	0.3333	0.1006	0.1842	

TABLE VIII: Performance comparison for size class 4, 38 trees.

imputed values.

We fit a mixed effects model for tree mortality (1=dead, 0=live) using `glmer` in the `lme4` package in `R 2.15.2`. The fixed effects include transect level effects for east facing aspect, and a standardized value for slope; tree level effects include presence/absence of mountain pine beetle, diameter at breast height, blister rust status, and their 2 and 3 factor interactions. The model includes a random effect for transect.

$$\begin{aligned}
Pr(Death) =& logit^{-1}(\beta_0 + \beta_1 e + \beta_2 s + \beta_3 br + \beta_4 mpb + \beta_5 dbh + \\
& \beta_6 br * mpb + \beta_7 br * dbh + \beta_8 mpb * dbh + \beta_9 br * mpb * dbh + \tau_{j(i)} transect_j(i) + \epsilon_i) \\
& \tau_{j(i)} \sim iidN(0, \sigma_\tau) \\
& \epsilon_i \sim iidN(0, \sigma)
\end{aligned}$$

e represents a measure of east facing aspect, s is standardized slope value, br, mpb, dbh are blister rust, mountain pine beetle and diameter, respectively. Transect is the random transect effect associated with the modeled tree. The parameter estimates vary for the two data sets; the fitted models are following:

Complete case:

$$\begin{aligned}
Pr(Death) =& logit^{-1}(-4.052 + -.931e + .800s + 4.169br + 1.209mpb + -.030dbh + -1.870br * mpb + \\
& -.492br * dbh + .096mpb * dbh + .469br * mpb * dbh + \tau_{j(i)} transect_{j(i)} + \epsilon_i) \\
& \tau_{j(i)} \sim iidN(0, 1.5087) \\
& \epsilon_i \sim iidN(0, \sigma)
\end{aligned}$$

LOCF:

$$\begin{aligned}
Pr(Death) =& logit^{-1}(-3.644 + -.668e + .155s + 2.190br + .632mpb + -.038dbh + .269br * mpb + \\
& -.079br * dbh + .110mpb * dbh + .016br * mpb * dbh + \tau_{j(i)} transect_{j(i)} + \epsilon_i) \\
& \tau_{j(i)} \sim iidN(0, 1.263) \\
& \epsilon_i \sim iidN(0, \sigma)
\end{aligned}$$

AMI:

$$
\begin{aligned}
Pr(Death) =& logit^{-1}(-3.665 + -.715e + .141s + 1.880br + .822mpb + -.036dbh + -.566br * mpb + \\
& - .062br * dbh + .105mpb * dbh + .027br * mpb * dbh + \tau_{j(i)} transect_{j(i)} + \epsilon_i) \\
& \tau_{j(i)} \sim iidN(0, 1.2856) \\
& \epsilon_i \sim iidN(0, \sigma)
\end{aligned}
$$

BEP:

$$
\begin{aligned}
Pr(Death) =& logit^{-1}(-3.482 + -.820e + .175s + 1.835br + .588mpb + -.038dbh + -.232br * mpb + \\
& - .063br * dbh + .103mpb * dbh + .023br * mpb * dbh + \tau_{j(i)} transect_{j(i)} + \epsilon_i) \\
& \tau_{j(i)} \sim iidN(0, 1.1977) \\
& \epsilon_i \sim iidN(0, \sigma)
\end{aligned}
$$

Table IX shows estimated probabilities of mortality and associated standard errors (on the probability scale) for the various combinations of Mountain pine beetle and blister rust states at tree diameters of 6,20,40 centimeters. The probabilities vary quite a bit, especially between complete case analysis and all of the other methods, and especially in the presence of mountain pine beetle. As far as which estimates are better, this is hard to say without a more in depth analysis of bias for the methods. However, we do see a decrease in standard error from the complete case analysis with every imputation method, which is promising.

VI. DISCUSSION

A. Summary

The various methods explored here each perform better under different measures. Last observation carried forward could best predict healthy trees, and matched better over all than any of the others according to Cohen's Kappa, though "best" by that measure was still middling. In general, the binomial approximation

		DBH	6	20	40
Data	MPB	BR	Probability(SE)	Probability(SE)	Probability(SE)
C.C.	No	No	0.02(0.60)	0.01(0.65)	0.01(0.79)
		Yes	0.05(0.67)	0.00(0.96)	0.00(1.00)
	Yes	No	0.09(0.63)	0.20(0.60)	0.48(0.64)
		Yes	0.45(0.74)	0.60(0.67)	0.78(0.88)
LOCF	No	No	0.02(0.57)	0.01(0.60)	0.01(0.71)
		Yes	0.10(0.59)	0.02(0.66)	0.00(0.82)
	Yes	No	0.07(0.59)	0.16(0.57)	0.45(0.60)
		Yes	0.36(0.65)	0.39(0.62)	0.43(0.76)
AMI	No	No	0.02(0.63)	0.01(0.66)	0.01(0.73)
		Yes	0.08(0.64)	0.02(0.69)	0.00(0.78)
	Yes	No	0.08(0.65)	0.18(0.63)	0.46(0.66)
		Yes	0.20(0.69)	0.28(0.66)	0.43(0.65)
BEP	No	No	0.02(0.63)	0.01(0.66)	0.01(0.73)
		Yes	0.09(0.64)	0.02(0.69)	0.00(0.79)
	Yes	No	0.07(0.66)	0.16(0.64)	0.40(0.67)
		Yes	0.23(0.69)	0.29(0.68)	0.40(0.59)

TABLE IX: Estimates for probability of mortality and associated standard errors at 4 combinations of blister rust state and Mountain pine beetle state at 3 different tree diameters (in cm).

approach appeared to do better for predicting the overall proportion, but less well in predicting for individual trees. The binomial coin flip approach did best for predicting infected trees. All of these comparisons varied some according to the magnitude of the true proportion and the size of the tree and sample. When applying imputation methods to earlier work in order to increase sample size, we see that we are able to decrease the standard errors of our estimates.

B. Further work

The most egregious assumption made and violated in this study is that of independence. Only transects are missing at random and the tree values should not truly be treated as independent since they are clustered. Finding a way to account for that lack in imputation for these data would likely help any method to perform better predictively. Additionally, due to the differences in fitted model parameters between the complete case analysis and all of the imputed analyses, it would be interesting to restart the model fitting process with an imputed data set.

[1] D. Mattson, K. Kendall, and D. Reinhart, Whitebark pine communities: ecology and restoration. Island Press, Washington, DC , 121 (2001).

[2] L. Koteen, Wildlife responses to climate change: North American case studies , 343 (2002).

[3] K. M. Irvine, C. Hollimon, E. Shanahan, and K. Legg, "Multi-agent disturbances: evaluating synergistic effects of an introduced pathogen and native bark beetle on mortality of a foundation species," (2013).

[4] D. B. Rubin, Biometrika **63**, 581 (1976).

[5] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, Vol. 539 (Wiley New York, 1987).

[6] I. R. White and J. B. Carlin, Statistics in medicine **29**, 2920 (2010).

[7] J. Ma, N. Akhtar-Danesh, L. Dolovich, L. Thabane, *et al.*, BMC medical research methodology **11**, 18 (2011).

[8] G. Frank Liu and X. Zhan, Journal of Biopharmaceutical Statistics **21**, 371 (2011).

[9] R. R. Andridge and R. J. Little, International Statistical Review **78**, 40 (2010).

[10] N. T. Longford, *Missing data and small-area estimation: Modern analytical equipment for the survey statistician* (Springer Science+ Business Media, 2005).

[11] J. L. Schafer, Statistical methods in medical research **8**, 3 (1999).

[12] S. Van Buuren, H. C. Boshuizen, D. L. Knook, *et al.*, Statistics in medicine **18**, 681 (1999).

[13] K. Lu, L. Jiang, and A. A. Tsiatis, Biometrics **66**, 1202 (2010).

[14] C. A. Bernaards, T. R. Belin, and J. L. Schafer, Statistics in medicine **26**, 1368 (2007).

[15] N. J. Horton, S. R. Lipsitz, and M. Parzen, The American Statistician **57**, 229 (2003).

[16] J. Cohen *et al.*, Educational and psychological measurement **20**, 37 (1960).

[17] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, Vol. 307 (Wiley, 1987).

[18] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear mixed-effects models using S4 classes* (2012), r package version 0.999999-0.