Hindawi Publishing Corporation
Epidemiology Research International
Volume 2013, Article ID 875234, 6 pages
http://dx.doi.org/10.1155/2013/875234

Research Article

First Use of Multiple Imputation with the National Tuberculosis Surveillance System

Christopher Vinnard,[1] E. Paul Wileyto,[2] Gregory P. Bisson,[3] and Carla A. Winston[4]

[1] *Division of Infectious Diseases & HIV Medicine, Drexel University College of Medicine, 245 N 15th Street MS 461, New College Building 6314, Philadelphia, PA 19102, USA*
[2] *Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[3] *Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[4] *Division of Tuberculosis Elimination, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA*

Correspondence should be addressed to Christopher Vinnard; christopher.vinnard@drexelmed.edu

Received 27 July 2012; Accepted 18 December 2012

Academic Editor: Huibert Burger

Aims. The purpose of this study was to compare methods for handling missing data in analysis of the National Tuberculosis Surveillance System of the Centers for Disease Control and Prevention. Because of the high rate of missing human immunodeficiency virus (HIV) infection status in this dataset, we used multiple imputation methods to minimize the bias that may result from less sophisticated methods. *Methods*. We compared analysis based on multiple imputation methods with analysis based on deleting subjects with missing covariate data from regression analysis (case exclusion), and determined whether the use of increasing numbers of imputed datasets would lead to changes in the estimated association between isoniazid resistance and death. *Results*. Following multiple imputation, the odds ratio for initial isoniazid resistance and death was 2.07 (95% CI 1.30, 3.29); with case exclusion, this odds ratio decreased to 1.53 (95% CI 0.83, 2.83). The use of more than 5 imputed datasets did not substantively change the results. *Conclusions*. Our experience with the National Tuberculosis Surveillance System dataset supports the use of multiple imputation methods in epidemiologic analysis, but also demonstrates that close attention should be paid to the potential impact of missing covariates at each step of the analysis.

1. Background

Missing data is a common problem in epidemiologic research. Analytic techniques used in multivariable analysis, such as regression models, rely on methods that exclude cases with missing covariate data from analysis. This missing data approach has important limitations. First, case exclusion will always lead to loss of statistical power. Second, case exclusion will introduce bias into the analysis if excluded subjects differ from included subjects in ways that are relevant for the parameter of interest [1]. The potential for bias using case exclusion depends on the mechanism for missingness. For missing-at-random (MAR) data, the missingness of a particular observation depends only on observed covariates, and for missing-not-at-random (MNAR) data, missingness may depend on both observed and unobserved covariates. For either MAR or MNAR data, case exclusion will introduce bias, as subjects excluded from analysis will differ from subjects included in analysis according to either the measured or unmeasured covariates. In contrast, when data is missing-completely-at-random (MCAR), missingness can be considered a random deletion of observations without respect to measured or unmeasured covariates, and case exclusion does not lead to the introduction of bias (only the loss of statistical power) [2].

Multiple imputation methods were developed by Rubin to account for nonresponse bias in surveys [3], and later modified by Schafer [4]. The goal of multiple imputation is to

Table 1: Predictor variables included in the imputation model, with percent of nonmissing observations and P for association with the outcome.

Variable	No. evaluated (%) $N = 1614$	P
Isoniazid resistance	1614 (100)	0.02
Age categories	1614 (100)	<0.01
Race	1614 (100)	<0.01
Occupation	1614 (100)	<0.01
Gender	1614 (100)	<0.01
HIV positive	849 (53)	<0.01
Injecting drug use within the previous year	1367 (85)	<0.01
Drug use (noninjecting) within the previous year	1351 (84)	<0.01
Alcohol use within the previous year	1340 (83)	<0.01
Homeless within the previous year	1500 (93)	0.18
Abnormal chest radiograph	1525 (94)	<0.01
Resident of a correctional facility at diagnosis	1604 (99)	0.03
Foreign birth	1600 (99)	<0.01
Positive tuberculin skin test	953 (59)	<0.01
Positive AFB smear from a nonsputum specimen	1419 (88)	0.12

create several plausible values for the missing covariates, and consequently several complete datasets, with the imputed values generated from observed relationships between variables. The investigator determines which variables will be used to create the imputed datasets, specifies the mathematical relationships between these variables (the imputation model), and chooses the number of imputed datasets that will be created. All predictors of missingness should be included in the imputation model in order to satisfy the MAR assumption [4]. Once these complete datasets are generated, analysis is performed on each dataset according to the hypothesis being tested. The parameter estimates that are obtained from each imputed dataset are combined into a single-point estimate, and its associated error reflects uncertainty not only within each imputed dataset, but also between the imputed datasets [5].

Recently, we used multiple imputation methods in our analysis of the association of initial isoniazid resistance with death during therapy among cases of tuberculous meningitis (TBM) in the USA between 1993 and 2005, analyzing data collected by the National Tuberculosis Surveillance System (NTSS) [6]. Among 1614 patients with positive cerebrospinal fluid cultures for *M. tuberculosis*, we observed a significant association between initial isoniazid resistance and death during therapy (OR 2.07, 95% CI 1.30, 3.29). We limited our analysis to patients with a known result from initial isoniazid susceptibility testing, and therefore isoniazid resistance or susceptibility was completely known for all patients in the study. However, other clinical and demographic factors that were evaluated as potential confounders of the relationship between isoniazid resistance and death during therapy had varying degrees of missingness, as shown in Table 1.

The human immunodeficiency virus (HIV) status of the case patient was unknown in 47% of observations. HIV status is included in the national Report of a Verified Case of Tuberculosis (RVCT), with options including "positive," "negative," "indeterminate," "test done, results unknown," "not offered," "refused," and "unknown"; cases may also be submitted with missing data reported (i.e., no HIV variable response option selected). For cases reported from California during the time period of the study, HIV status was reported as missing for all cases. Matching was then performed through 2004 between the state tuberculosis surveillance dataset and the state acquired immunodeficiency syndrome (AIDS) registry, which only includes HIV-positive patients with a clinical diagnosis of AIDS. Consequently, this matching procedure did not identify patients who tested negative for HIV or HIV-positive patients without a clinical diagnosis of AIDS.

We sought to further explore the methodology of the multiple imputation in our analysis of the NTSS data. Our goal was to compare multiple imputation with case exclusion and to determine whether the use of increasing numbers of imputed datasets would have changed the inference regarding the association between initial isoniazid resistance and death during antituberculosis therapy.

2. Methods

2.1. Setting. The NTSS has collected aggregate tuberculosis incidence data in the USA since 1953 and individual-level data (including antituberculosis drug susceptibilities) since 1993 [7]. In order to be included in the national count, a case of tuberculosis must satisfy a standardized case definition. We examined all tuberculosis cases reported from January 1, 1993, through December 31, 2005, during which time drug-susceptibility and risk factor data were available for reported cases. Institutional review board approval was obtained from the University of Pennsylvania prior to the beginning of the study.

2.2. Subjects. Cases of tuberculosis are reported to the Centers for Disease Control and Prevention by state and local health agencies, using the RVCT. This report includes clinical and demographic information about the patient, as well as the anatomic sites of any positive mycobacterial cultures or smears. We selected subjects for inclusion in the cohort if there was a report of a meningeal site of involvement, as either a primary or secondary site of infection, with a positive culture for *M. tuberculosis* from cerebrospinal fluid. We limited enrollment to patients who were alive at diagnosis and who initiated antituberculosis therapy.

2.3. Analysis. The goal of our analysis was to measure the association of initial isoniazid resistance with death during therapy in patients with tuberculous meningitis. We first calculated the unadjusted association of initial isoniazid resistance with the outcome of death during therapy. We used a multivariable logistic regression model to simultaneously adjust for multiple confounders.

There are two general mathematical approaches for model-based multiple imputation: either based on the multivariate normal distribution [4] or fully conditional specification [8]. In a simulation study, both approaches produced similar results and were less biased than complete exclusion [9]. We employed the multiple imputation procedure for Stata written by Royston (*ice*) [10], which uses the approach of fully conditional specification (also known as the chained equation or regression switching approach). The imputation model included the primary exposure (isoniazid resistance), the outcome variable (death during antituberculosis therapy), and all potential confounding variables identified based on univariable associations with the outcome (based on $P < 0.25$, see Table 1). Exclusion of the outcome variable in imputation models may lead to biased estimates of regression coefficients [11]. Because age was not normally distributed, a categorical age variable was used in analysis.

We chose to generate 5 imputed datasets based on simulation studies demonstrating little gain in statistical power for higher numbers of imputations [3]. The 5 imputed datasets were combined for analysis according to Rubin's method [3], using the *mim* procedure written by Royston et al. [12], generating single estimates for the parameters of interest. These parameter estimates had associated error that reflected the degree of missingness in the original dataset.

We evaluated variables in the logistic regression model based on their ability to change the observed association between initial isoniazid resistance and death during therapy. A confounder was defined as a variable that changed the association of interest by greater than 15% when included in the model. The final multivariable logistic regression model included the following terms: isoniazid resistance, age, race/ethnicity, gender, and HIV status. Age and race/ethnicity were confounders of the association between isoniazid resistance and death. HIV status and gender were not confounders but remained in the model based on our *a priori* clinical reasoning. Of the four covariates that remained in the final logistic regression model, only HIV status had missing observations that had been imputed in the datasets.

Table 2: Logistic regression models for death: multiple imputation versus case exclusion.

Term	Multiple imputation	Case exclusion
Isoniazid resistance	2.07 (1.30, 3.29)	1.53 (0.83, 2.83)
Male gender	1.20 (0.88, 1.63)	1.16 (0.81, 1.66)
Age categories		
Age ≤ 1	0.076 (0.008, 0.69)	[+]
1 < Age ≤ 4	0.23 (0.067, 0.82)	0.61 (0.13, 2.79)
4 < Age ≤ 14	0.38 (0.102, 1.41)	[+]
14 < Age ≤ 24	1.22 (0.64, 2.34)	1.44 (0.71, 2.93)
24 < Age ≤ 34	Reference	
34 < Age ≤ 44	1.30 (0.87, 1.92)	1.11 (0.69, 1.78)
44 < Age ≤ 54	1.97 (1.23, 3.15)	1.87 (1.12, 3.13)
54 < Age ≤ 64	1.83 (1.09, 3.09)	1.13 (0.57, 2.23)
64 < Age ≤ 74	4.36 (2.48, 7.67)	5.85 (2.80, 12.19)
Age > 74	6.90 (3.85, 12.38)	2.99 (1.25, 7.12)
Race categories		
White	Reference	
Black	1.44 (1.01, 2.06)	1.10 (0.69, 1.76)
Hispanic	1.21 (0.74, 1.99)	1.19 (0.71, 1.98)
Asian[*]	0.64 (0.30, 1.35)	0.62 (0.30, 1.28)
American Indian	9.07 (2.65, 31.02)	28.5 (3.12, 260.2)
Other[**]	0.85 (0.20, 3.52)	0.81 (0.15, 4.43)
HIV positive	3.57 (1.87, 6.82)	4.58 (3.13, 6.71)

[*] Includes Native Hawaiian.
[**] Includes multiple race and unknown categories.
[+] There were no deaths among patients with known HIV status in this age group.

In the analyses presented here, we compare the inference obtained for the association between isoniazid resistance and death using multiple imputation with inferences obtained based on alternate approaches: (1) case exclusion and (2) varying the number of imputed datasets between 2 and 10.

3. Results

3.1. Multiple Imputation Compared with Case Exclusion. After excluding 47% of patients with missing HIV status, we repeated the regression model (including isoniazid resistance, age, race/ethnicity, gender, and HIV status) with the subset of subjects with HIV status reported as either "positive" or "negative" (case exclusion). We compared estimates of association for initial isoniazid resistance and other variables obtained after case exclusion with the estimates obtained after multiple imputation. We obtained notably different estimates for the association of the older age categories with death during therapy, as well as the association between initial isoniazid resistance and death during therapy (Table 2). Using case-exclusion approach, the association between isoniazid resistance and death was 1.53 (0.83, 2.83), while the association yielded by the multiple imputation approach was 2.07 (1.30, 3.29).

To explore why the case exclusion analysis moved the estimate of the OR for isoniazid resistance and death during therapy towards the null hypothesis, we examined the age

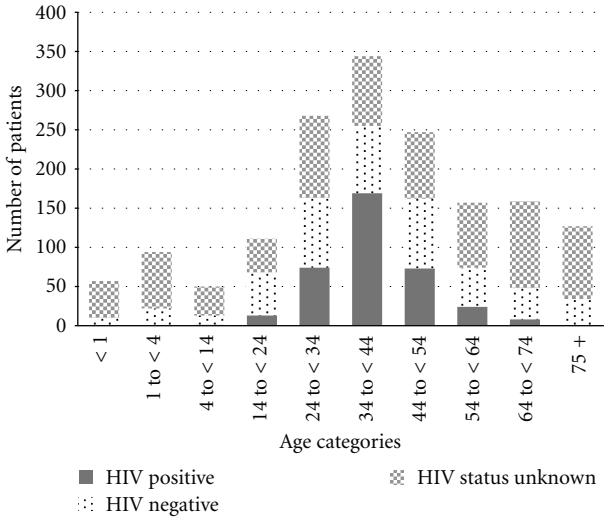

Figure 1: Missingness of HIV infection status with respect to age among TBM patients in the USA, from 1993 to 2005. "HIV status unknown" includes subjects with HIV status reported to be "indeterminate," "test done and results unknown," "not offered," "refused," and "unknown." This category also includes subjects with the HIV status response item left blank.

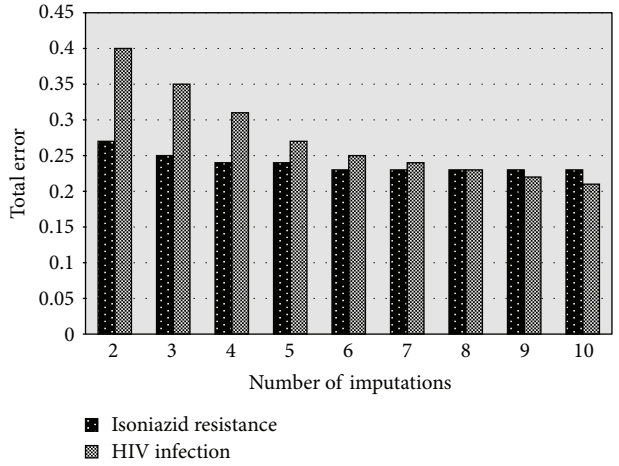

Figure 2: Total errors for odds ratios for isoniazid resistance and HIV in adjusted model (Adjusting for age, race/ethnicity, and gender), with increasing numbers of imputed datasets.

distributions of HIV positive subjects, HIV negative subjects, and subjects with missing HIV status (Figure 1). Missing HIV status was much more common in patients at the extremes of age. Exclusion of patients with missing HIV status preferentially excluded patients in the youngest and oldest age categories. Previously, we found that advancing age was a significant confounder of the association between initial isoniazid resistance and death during therapy. Older patients were more likely to die during therapy but less likely to be infected with an isoniazid resistant strain. As a result, the OR for initial isoniazid resistance and death was 1.61 (95% CI 1.08, 2.41) without adjusting for age and was 1.81 (95% CI 1.19, 2.75) after adjusting for age. Therefore, differential exclusion of patients at the extremes of age likely decreased the ability of the multivariable model to adjust for the confounding effect of age, masking the association between isoniazid resistance and death during therapy.

3.2. Varying the Number of Imputations between 2 and 10. Next, we examined whether the number of imputations influenced the observed association between initial isoniazid resistance and death during therapy. In the original analysis, we chose to use 5 imputed datasets based on simulation studies showing little gain of efficiency with additional rounds of imputation [3]. However, the optimal number of imputations may depend on properties of the particular dataset, as well as specific properties of the association of interest.

To further evaluate the optimal number of imputations to use in the study of isoniazid resistance and death, we repeated the analysis after varying the number of imputations from 2 to 10. We examined the error associated with the coefficient for initial isoniazid resistance in the logistic regression model, as well as the error associated with the coefficient for HIV infection in the same adjusted model, separating the total error into its within- and between-imputation components.

The results of this analysis are shown in Table 3. The error associated with the isoniazid-resistant coefficient did not require more than 5 imputed datasets to reach a plateau. As a result, the use of more than 5 imputations did not lead to an increase in precision for the estimate of the association of initial isoniazid resistance with death during therapy. In contrast, the error associated with the HIV coefficient continued to decrease with more than 5 imputed datasets, with more narrow confidence intervals as the number of imputed datasets approached 10 (Figure 2).

4. Discussion

Multiple imputation offers important advantages over other methods for handling missing data in epidemiologic studies, in particular with regards to its flexibility and wide applicability [1]. However, widespread use is limited by its theoretical complexity and lack of familiarity among audiences outside of statistical disciplines [13]. We have presented in detail our experiences in order to facilitate continued use of these techniques in the future.

We designed the imputation model in the context of a specific research question, with particular concern for the potential impact of missing HIV status. The matching mechanism to obtain HIV status of tuberculosis cases reported from California violates the "missing-at-random" status, since missingness will partly depend on unobserved covariates (the value of the HIV test itself). However, we previously demonstrated that imputing HIV status for all California cases, rather than relying on the HIV status from matching with the AIDS registry, did not influence the estimated association between isoniazid resistance and death [6].

Table 3: Errors in coefficient estimates for increasing numbers of imputed datasets.

	Number of imputed datasets								
	2	3	4	5	6	7	8	9	10
Coefficient for isoniazid resistance	0.71	0.72	0.72	0.73	0.72	0.73	0.73	0.74	0.74
Within-imputation error	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
Between-imputation error	0.11	0.08	0.07	0.06	0.05	0.05	0.05	0.05	0.05
Coefficient for HIV infection	1.44	1.33	1.29	1.27	1.25	1.24	1.23	1.24	1.23
Within-imputation error	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Between-imputation error	0.31	0.28	0.24	0.22	0.20	0.19	0.17	0.16	0.16

Because of the high proportion of tuberculosis cases with missing HIV status, we were concerned about the potential for bias with less sophisticated methods, such as the missing indicator method. Although frequently employed as a missing data approach, missing data categories are less optimal than imputation since they may be difficult to interpret as meaningful parameters, may produce large and unstable estimates, and in some situations, may introduce bias [14]. In our analysis, the use of a missing data category for HIV status did not significantly change the relationship between isoniazid resistance and death during therapy that was obtained by the multiple imputation approach (data not shown).

In contrast, we observed significant bias when we compared the results obtained from multiple imputation with the case-exclusion method that dropped subjects without known positive or negative HIV status from the analysis. Although HIV itself was not a confounder of the relationship between initial isoniazid resistance and death during therapy (unlike age and race/ethnicity), missing HIV status was predominantly seen in subjects at the extremes of age. Because age was a significant confounder of this relationship, exclusion of patients with missing HIV status led to an estimate of the OR for initial isoniazid resistance and death that was biased towards the null hypothesis.

The missing data problems that we addressed were primarily a result of our *a priori* decision to allow HIV status to remain in the model, regardless of its confounding effect on the relationship between isoniazid resistance and death. These challenges illustrate the importance of considering the effects of missing data when making *a priori* decisions for multivariable model selection. While we allowed HIV status to remain in the final multivariable model for clinical reasons, HIV status did not significantly influence the observed relationship between initial isoniazid resistance and death during therapy. The other variables in the final model were completely known and did not differ from one imputed data set to the next. For this reason, the use of increasing numbers of imputations did not lead to a decrease in the total error associated with the regression coefficient for isoniazid resistance, since most of the total error for this coefficient was within-imputation error, rather than between-imputation error. In contrast, the error associated with the coefficient for HIV status continued to decrease beyond 5 imputed datasets. If HIV status had been a significant confounder of the association between initial isoniazid resistance and death during therapy, we would have expected to see both errors continue to fall as we used increasing numbers of imputed datasets.

In summary, we found the approach of multiple imputation to be a useful method for dealing with the missing data problem in the NTSS of the USA, overcoming the bias inherent to case exclusion. The size of the dataset and completeness of reporting for variables such as age, gender, and race likely enhanced the robustness of our findings in these sensitivity analyses. While the use of more than 5 rounds of imputation did not lead to a more precise estimate for the association of initial isoniazid resistance and death during antituberculosis therapy, our findings would likely have been different if HIV status had exerted a stronger confounding effect on this relationship. While our analysis supports the use of multiple imputation methods in epidemiologic analysis, it also demonstrates that close attention should be paid to the potential impact of missing covariates at each step of the analysis.

Acknowledgments

The authors thank the local health departments who collected data for this analysis and Valerie Robison, Sandy Althomsons, and Carla Jeffries for their contribution to the analysis. This paper received no specific grant from any funding agency in the public, commercial, or nonprofit sectors. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

[1] J. W. Graham, "Missing data analysis: making it work in the real world," *Annual Review of Psychology*, vol. 60, pp. 549–576, 2009.

[2] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[3] D. B. Rubin, *Multiple Imputation For Nonresponse in Surveys*, John Wiley & Sons, New York, NY, USA, 1987.

[4] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York, NY, USA, 1997.

[5] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 3–15, 1999.

[6] C. Vinnard, C. A. Winston, E. P. Wileyto, R. R. MacGregor, and G. P. Bisson, "Isoniazid resistance and death in patients with tuberculous meningitis: retrospective cohort study," *British Medical Journal*, vol. 341, no. 7773, p. 596, 2010.

[7] "Trends in tuberculosis—United States, 2008," *Morbidity and Mortality Weekly Report (MMWR)*, vol. 58, pp. 249–253, 2009.

[8] S. van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, vol. 16, no. 3, pp. 219–242, 2007.

[9] K. J. Lee and J. B. Carlin, "Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation," *American Journal of Epidemiology*, vol. 171, no. 5, pp. 624–632, 2010.

[10] P. Royston, "Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables," *The Stata Journal*, vol. 9, no. 3, pp. 466–477, 2009.

[11] K. G. Moons, R. A. Donders, T. Stijnen, and F. E. Harrell Jr., "Using the outcome for imputation of missing predictor values was preferred," *Journal of Clinical Epidemiology*, vol. 59, pp. 1092–1101, 2006.

[12] P. Royston, J. B. Carlin, and I. R. White, "Multiple imputation of missing values: new features for mim," *The Stata Journal*, vol. 9, no. 2, pp. 252–264, 2009.

[13] M. A. Klebanoff and S. R. Cole, "Use of multiple imputation in the epidemiologic literature," *American Journal of Epidemiology*, vol. 168, no. 4, pp. 355–357, 2008.

[14] F. E. Harrell Jr., *Regression Modeling Strategies: With Applications To Linear Models, Logistic Regression, and Survival Analysis*, Springer, New York, NY, USA, 2001.