HSR Health Services Research

DOI: 10.1111/1475-6773.12336
EDITORIAL

Editorial

Improving Disparity Research by Imputing Missing Data in Health Care Records

Several years ago, the Chief Scientist at Abt Associates asked my research assistant whether she was experienced at assembling analysis files. Her response was, "Yes, I make up data for Bill Rhodes all the time." After clarifying that she meant converting raw data into files suitable for statistical analysis, this episode became a humorous anecdote rather than cause for my dismissal. After all, for scientists, fabricating data is professional malfeasance.

In this issue of *Health Services Research*, Grundmeier et al. openly advocate fabricating data. Yet, instead of condemning the researchers, I concur with their recommendations and offer my congratulations on a valuable paper. This paradox requires explanation.

Often in scientific studies, the researcher works with a data file comprising N records (such as N patients, as in the Grundmeier et al. 2015, paper), but only $n < N$ of those records have complete data, and the remaining $N - n$ records have missing data for one or more variables (such as patient ethnicity/race, as in the Grundmeier paper). The widely adopted solution to this problem is to drop records with incomplete data, sometimes called list-wise deletion or complete case analysis, but this common approach has two disadvantages: (1) When omitted cases differ materially from included ones, list-wise deletion can bias statistics and (2) almost always using fewer cases (n rather than N) will lead to statistics with greater sampling variance.

An occasionally used alternative is to fabricate responses for the missing variables, or to use the justifiably less pejorative term, to *impute* responses for the missing items, and then use all N cases in the analysis. Justification for imputation comes from applying statistical principles that, under sanguinary

conditions, (1) cause the imputed responses to be true on average and (2) cause the estimated sampling variance of resulting statistics to correctly account for the fact that the imputed responses have measurement error—after all, they are only correct on average. The statistical principles justifying imputing data have formal (Schaefer 1997; Little and Rubin 2002) and more casual (Allison 2001; Enders 2010) justification and programs for performing imputations and for analyzing imputed data appear in statistical software (for example, StataCorp, 2009).

Ultimately, then, the reason why the authors can impute data rests on strong statistical principles. Establishing the foundation for these principles requires a seemingly arcane language (used by the authors) classifying missing data patterns as *missing completely at random*, *missing at random*, and *missing not at random*. Depending on the statistical model underlying the analysis, the principles require the conditions that missing completely at random or missing at random hold, but these conditions can never be confirmed, and some authorities argue that missing not at random is most likely to occur, vitiating the foundation for the principles.

Consequently, in most applications, and always when missing not at random occurs, the formal argument requires complementary assumptions that (1) the principles hold approximately and (2) adopting imputation procedures will yield statistics that are less biased and have lower sampling variance than statistics based on list-wise deletion. Put another way, from a bias reduction and mean squared error perspective, imputation is preferable to list-wise deletion. With justification, the Grundmeier paper asserts that imputation is also preferable to hot-deck sorting and mean-value imputation, both of which lack statistical justification.

Although the language is arcane, the notion of missing not at random is familiar to evaluators. From the potential outcome framework, evaluation research *always* poses a missing data problem (Rubin 2002; Imbens and Rubin 2015): Either an evaluator knows the outcome under the treatment state, or he knows the outcome under the untreated state, but she cannot know the outcome for any individual under both conditions, so the evaluator is forced to impute the outcomes for the unobserved state. When selection into treatment can be explained by variables observed in the data, so that no other variables explain both assignment and outcomes, evaluators sometimes say that *unconfoundedness* holds and a number of regression-

Address correspondence to William Rhodes, Ph.D., 261 Marlboro Road Sudbury, MA 01776; e-mail: WMRhodes0405@gmail.com.

based estimation procedures—such as propensity score analysis—are available. Unconfoundedness always holds for random design experiments, in which case we could say that data are *missing completely at random*. Unconfoundedness sometimes holds for observational studies, in which case we could say that data are *missing at random*. Otherwise, when unconfoundedness does not hold, we recognize selection bias, in which case we could say that data are *missing not at random*. These concepts, familiar to evaluators, reappear when performing data imputations in nonevaluation missing data settings.

Grundmeier et al. (2015) consider a simple imputation problem. Intended as a demonstration, they analyze medical records to determine health disparity based on ethnicity/race. The technical problem that they imagine is that race is sometimes missing, and their concern is that item nonresponse for race might bias univariate and multivariate statistics where race is a determinate of health disparities.

They *imagine* this problem because in fact their data are complete. No ethnicity/race variables have missing data until Grundmeier et al. (2015) induce a missing data problem by transforming some of the valid ethnicity/race variables from known to missing using a randomization process that is purposely missing not at random. Because they know the true values of the ethnicity/race variables whose missing responses are induced, they know the true values of the statistics derived from the complete dataset. This allows them to judge how well different approaches to imputation perform at recovering the "correct" statistics. They use a Monte Carlo simulation to form their judgment.

As far as imputation goes, their problem is very simple, because race/ethnicity is the only variable with missing values. We can think of their solution as, first, estimating a regression using the n records with known ethnicity/race as the dependent variable and, second, using the estimated regression parameters to impute ethnicity/race for the $N - n$ records with missing data. (In fact, their estimation procedure, based on the method of chained equations, is more complicated, but the regression analog provides a simple conceptualization.) The imputations are imprecise for two reasons: Even if the regression is correctly specified, (1) the regression parameters are estimates with sampling variance; and (2) predictions based on the regression have residual variance. The imputations may be biased for two reasons: (1) the regression may be misspecified or (2) data are missing not at random. As already noted, subsequent estimation takes the imprecision

into account, and the possible bias is ignored with the hope or expectation that the bias is small.

Grundmeier et al. (2015) are following an established procedure for performing imputations, and frankly, if their demonstration of imputation were the only contribution they made to the literature, I would find their paper to be no more than a research note. The intriguing part of their paper comes from variables that they have included on the regression's right-hand side.

When I have performed imputations, I have considered my data to be self-contained, so that right-hand-side variables come from within the data—the counterpart here is that the right-hand-side variables come from the medical records. In contrast, with expectation that prediction will be improved, the authors have introduced variables that are not part of the medical record. First, their imputations take advantage of the clinic's location, expecting that the distribution of ethnicity/race will vary across geography. Second, drawing on an algorithm developed by others (Fiscella and Freemont 2006; Elliot et al. 2008), they introduce surname as a predictor of ethnicity. Third, again drawing on an algorithm developed by others (Fiscella and Freemont 2006; Elliott et al. 2008), they introduce census data, linked with patient residence, to impute race/ethnicity. This is to say that they link the medical records with other sources—clinic location, a list of surname and ethnicity/race matching probabilities, and a list of residential addresses and ethnicity/race matching probabilities—to augment the regression's right-hand-side variables and to improve the likelihood that their imputed ethnicity/race variables will be correct on average.

As noted, the prediction based on the regression is not a precise measure of ethnicity/race. Instead, the regression provides a probability of being white, black, or other. After conditioning on variables other than ethnicity/race, if whites are 70 percent of the population, and blacks are 20 percent and other are 10 percent, the hope is that the regression will predict a probability near 0.7 for whites, 0.2 for blacks, and 0.1 for other. Any subsequent analysis of the data should take both the imprecision when estimating the regression (e.g., the regression parameters are estimated with error) and the imprecision in prediction (e.g., the 0.7, 0.2, and 0.1 distribution) into account. What the authors' paper shows is that statistics based on list-wise deletion are biased even to the point of altering the sign of the correlation between race/ethnicity and health disparity. (The authors built this perverse correlation into the problem by the way they simulated the missing items.) As they progressively add additional right-hand-side variables to the regression used for imputation, the perverse sign for the correlation disappears, and the bias diminishes. The bias does not

disappear, a consequence of their choice to induce missing data using a missing not at random selection mechanism. The authors show that their recommended technique does not overcome the problem with bias (because the data are missing not at random), but the improvement in the inferences are substantial.

The authors' paper provides helpful, practical advice for imputing ethnicity/race. I am impressed and have recommended the approach to some colleagues who are working on similar problems in a different setting.

However, the problem faced by the authors is narrow: Only a single variable is missing. My experience working with health care records (in the specific setting of nursing homes) is that multiple variables have missing items. When that is the case, the regression approach for imputing data still holds as an analogy, but the imputation modeling itself is more complicated. The authors did not want to tackle this problem in their paper because their main concern is with demonstrating the utility of going beyond medical records to introduce additional data to the imputation model. Their paper demonstrates the point, but I suspect that other researchers, after learning this lesson, will have to struggle with more complex imputation models.

As applied research, the paper does not demonstrate how to actually perform the statistical analysis using data derived from the imputation model. The Monte Carlo simulation identifies interquartile ranges by repeated application of a simulation. The paper does not actually explain how a researcher would derive confidence intervals for real-world problems. Fortunately, accessible sources (Allison 2001; Enders 2010) provide explanation and software (StataCorp, 2009, for example) provides application, so this omission need not be a stumbling block.

The focus of the Grundmeier et al.'s (2015) paper is narrow in another regard, namely, the study period is static, intended to identify health disparity at a specific point in time. This focus is worthwhile, of course, but concern is often with whether health disparity is changing over time. A complication is that ascribed and self-defined definitions of ethnicity/race vary over time, so constructing consistently defined time-series is challenging. Econometricians might reference this as a *measurement error problem.* Although I cannot be sure, it seems worthwhile to see if the techniques that the authors advocate for a missing values problem might inform techniques that econometricians advocate for measurement error.

Another observation is that the surname and Census data algorithms are calibrated for a specific population and there is no assurance that the association will hold for another population. Furthermore, street addresses may be

unknown, but other identifiers such as county of residence may be available. Some researchers might use the logic incorporated into the algorithm used by Grundmeier et al. (2015) without using the specific algorithm developed by coauthors at Rand.

Extending this idea, Grundmeier et al. (2015) note that their records do not report income, a variable that they and others see as important for disparity research. They use insurance as a proxy measure for income, but with the Affordable Care Act, I presume this becomes a decreasingly useful proxy. However, block-level Census data provide measures that are highly correlated with income, suggesting that the block-level measures might be used as proxy income measures.

The recommendations for improving imputations provided in this article are valuable, and moreover, I appreciate the implicit recommendation for matching health care records with other data—Census data specifically—for improving disparity research. I suggest that other researchers working on health disparity issues study the foundations of performing imputations and consider incorporating Grundmeier et al.'s (2015) approach into their applied research.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This paper is a commentary on a paper published in this issue of *HSR*. It is a methodological contribution written without financial support.

Disclosure: None.

Disclaimer: None.

William Rhodes

REFERENCES

Allison, P. 2001. *Missing Data.* Thousand Oaks, CA: Sage.

Elliott, M., P. Fremont, P. Morrison, and N. Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." *Health Services Research* 43: 1722–1736.

Enders, C. 2010. *Applied Missing Data Analysis.* New York: Guilford Press.

Fiscella, K., and A. Fremont. 2006. "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity." *Health Services Research* 41: 1482–1500.

Grundmeier, R., L. Song, M. J. Ramos, A. G. Fiks, M. N. Elliott, A. Fremont, W. Pace, R. C. Wasserman, and R. Localio. 2015. "Imputing Missing Race/Ethnicity in Pediatric Health Records: Reducing Bias with Use of US Census Location and Surname Data." *Health Services Research* 50 (4): 946–960.

Imbens, G., and D. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge, MA: Cambridge University Press.

Little, R., and D. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd Edition. Hoboken, NJ: Wiley.

Rosenbaum, P. 2002. *Observational Studies*, 2nd Edition. New York: Springer-Verlag.

Schaefer, J. 1997. *Analysis of Incomplete Missing Data.* Boca Raton, FL: Chapman & HALL/crc.

StataCorp. 2009. *Stata Multiple Imputation Reference Manual Release 11.* College Station, TX: StataCorp LP.