Published in final edited form as:

Am J Ophthalmol. 2009 December ; 148(6): 820–822. doi:10.1016/j.ajo.2009.07.027.

Missing Data: What a Little Can Do, and What Researchers Can Do in Response

Thomas R. Belin

Department of Biostatistics, UCLA School of Public Health Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine at UCLA 51-267 Center for Health Sciences Los Angeles, CA 90095-1772

“If only 5% of the data values are missing, is it o.k. to drop the cases with incomplete data and analyze only the cases with complete data?” Questions along these lines are frequently directed to statisticians by applied researchers. In considering the potential impact of missing data, we offer an example to illustrate why such an apparently simple question does not have a simple answer.

Suppose a two-arm randomized study seeks to compare two cataract surgery protocols (“standard” and “new”), with $n=20$ patients receiving each treatment and with the primary outcome, Y , being the number of lines of improvement between baseline and six months post-surgery on a Snellen visual acuity chart. Suppose further that baseline measurements are also obtained on a clinical characteristic, X (e.g., pre-operative macular thickness) that is highly predictive of outcomes in both arms of the study. A hypothetical set of data values for such a study is shown in Table, listing a patient identification number, macular thickness (X) in microns, and lines of improvement on visual acuity (Y) between baseline and 6 months. Note that the groups have identical distributions of X , so confounding by X is not a concern in comparing outcomes on Y . Note also that X and Y are highly correlated, with sample correlations of $r_{\text{std}} = 0.81$ and $r_{\text{new}} = 0.79$ based on the 20 patients in each group. If there were no missing data, one could calculate the average outcome $\bar{Y}_{\text{new}} = 2.30$ in the new-treatment group and $\bar{Y}_{\text{std}} = 1.60$ in the standard-treatment group; familiar inference procedures would yield a 95% confidence interval of (0.02, 1.38) for the difference between group means, corresponding to a p -value of $p = 0.044$ based on a two-sample t -test.

Suppose further, however, that one patient in each arm failed to appear for the 6-month evaluation, implying 5% missing data in each arm (i.e., $1/20 = 5\%$). In this example, one missing outcome is associated with a standard-treatment patient whose baseline macular thickness is tied for the smallest observed value, and the other is associated with a new-treatment patient whose baseline macular thickness is tied for the largest observed value. Analyzing the remaining 19 patients in each arm would yield average outcomes $\bar{Y}_{\text{new}}^* = 2.21$ and $\bar{Y}_{\text{std}}^* = 1.68$, with a 95% confidence interval for the mean difference of $(-0.14, 1.20)$, corresponding to a p -value of $\tilde{p} = 0.120$ from a two-sample t -test.

© 2009 Elsevier Inc. All rights reserved.

Address for correspondence: Thomas R. Belin, Ph.D. UCLA Department of Biostatistics 51-267 Center for Health Sciences Los Angeles, CA 90095-1772 Phone: 310-206-7361 Fax: 310-206-7361 tbelin@ucla.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

While it is worth reemphasizing that the widespread use of $\alpha = 0.05$ as a significance level is a scientific convention that should not be viewed as sacred, the distinction between $p < 0.05$ in the original scenario and $\tilde{p} > 0.10$ due to 5% missing data gives grounds for pause. Under these circumstances, it is reasonable to wonder whether the partially-observed cases with available baseline measurements on X can be used to enhance the precision of the outcome analysis, especially since baseline values of X are so strongly correlated with the 6-month outcomes Y.

One method for incorporating X into such an analysis is known as “multiple imputation”.¹⁻³ Multiple imputation is a well-established general strategy for handling missing data that makes use of available data (including covariate values) to fill in plausible values of missing items. To avoid exaggerating the precision of the inference, one produces several (e.g., 5) plausible values for each missing item and carries out a separate analysis for each “completed” data set. A key ingredient in the overall inference is to combine “between-imputation variance” (i.e., variation across imputed data sets in estimates of the target quantity, which reflects uncertainty due to values being missing) with “within-imputation variance” (i.e., the average of the squared standard errors from the separate analyses, which reflects uncertainty due to having one rather than another sample of size 20). To estimate the between-imputation variance, more than one imputation is needed; filling in only a single value (“single imputation”) would exaggerate precision by pretending that there is no uncertainty about the values of missing items.

To illustrate the method, we produced 5 multiple imputations for each of the missing items in the hypothetical example. Multiple imputation software is available in many statistical packages, although there are differences in programs and underlying modeling assumptions, and inferences can be sensitive to details of the imputation procedure. The implementation here used SAS software (specifically, the procedures PROC MI and PROC MIANALYZE), with missing values of visual-acuity improvement imputed based on a linear regression of visual-acuity improvement (Y) on macular thickness (X). (More generally, a modern statistical-computing strategy known as “Markov-chain Monte Carlo” can be used to produce imputations; this approach, which was used here, reduces to regression-based imputation when predictor variables are completely observed as in the present setting.) The five imputed values for the standard-treatment case (Y_{101}) were -1.47, 0.81, 0.48, 0.66, and -0.50, and the five imputed values for the new-treatment case (Y_{220}) were 3.39, 3.73, 3.59, 3.31, and 2.74. Using these values, the estimated mean difference between groups was estimated to be 0.67 with a 95% confidence interval of (0.01, 1.33), corresponding to an overall p-value of $p^* = 0.047$.

We also considered rounding the imputed values to the nearest whole number (i.e., -1, 1, 0, 1, 0 for Y_{101} and 3, 4, 4, 3, 3 for Y_{220}), which had only a slight impact on inferences (estimated mean difference: 0.66; 95% CI: 0.00, 1.31; $p^* = 0.048$). Either way, by making use of evidence in the data that X and Y are highly correlated, the procedure is able to recover information that was lost when values were omitted from the analysis.

Considering the possibility that chance variation might have played a role in producing a result that was just barely significant at the 0.05 level, we implemented the multiple-imputation procedure with an equally valid but slightly different user-supplied setting governing random number generation and obtained a p-value of $p^* = 0.055$. Five more analogous perturbations of the procedure yielded successively $p^* = 0.060, 0.050, 0.053, 0.059$, and 0.040. If greater precision were desired, the number of imputations could be increased, but although there is clearly a modest amount of sensitivity to user-supplied settings, it also seems clear that the finding of $\tilde{p} = 0.120$ based only on the cases with complete data understates the significance of the difference. The point of this example is not to suggest that multiple imputation will always produce significant results but rather that it can incorporate information from partially-

observed cases, which can mitigate bias and improve precision without exaggerating significance levels.

In the literature on incomplete-data analysis, distinctions are drawn among different types of mechanisms that might give rise to missing data.⁴ “Missing completely at random” (MCAR) refers to settings where missing values are like a random subsample of all values. When missingness is MCAR, complete-case analysis is valid (although it may not make full use of available information). “Missing at random” (MAR) has a similar sounding name but refers to the much broader set of scenarios where missingness on one variable is allowed to depend on an observed value of another variable. For example, if older individuals were less likely to return for a follow-up measurement, the missing values would not be MCAR but could be MAR. The missing-at-random assumption is not guaranteed to hold, but data sets typically do not contain information to contradict the missing-at-random assumption. General-purpose multiple-imputation software typically allows values to be missing at random.

In some settings, missingness may not be at random, as when dropping out of a study is related to an underlying, unmeasured characteristic. Multiple imputation may still be useful in such a scenario, but inferences will depend on assumptions that are not connected to available data. A classic example of inference assuming that missing values were not occurring at random⁵ was the decision to reinforce Allied planes in World War II based on an assumed “selection effect.” Certain surfaces on returning planes were seen to have more bullet holes than others, but rather than reinforcing surfaces of planes where returning planes were seen to have many bullet holes (which might make sense if it was thought that enemy planes were aiming at those areas), the decision was made to reinforce surfaces of planes where returning planes were seen to have few bullet holes (since it was thought that the reason fewer bullet holes were seen on certain surfaces was not because those areas were less frequently targeted but rather because planes that were hit on those surfaces were less likely to return). A good strategy for making the assumptions underlying general-purpose multiple-imputation software more plausible is to measure a wide array of characteristics and to incorporate those characteristics into imputation models.²

In the example presented here, the incomplete cases had very different values of macular thickness (X), which motivates the idea of regression-based imputation that controls for macular thickness. Inferences would not always be sensitive to having 5% missing data, but when observed outcomes are strongly related to a covariate and cases with missing outcomes have different covariate values across the groups being compared, there may be sensitivity in inferences, as shown here.

In summary, a reasonable answer to the question, “If only 5% of the data values are missing, is it o.k. to drop the cases with incomplete data and analyze only the cases with complete data?” would be, “It depends.” The good news is that recent advances such as multiple imputation and associated statistical-computing strategies provide statisticians and allied researchers with sophisticated techniques to address missing data.

Acknowledgments

a. The author has received government funding from several branches of NIH during the past two years including ongoing support from NIMH (R01 MH078853, P30 MH082760, P30 MH58017), NCI (R01 CA109650), and NIDA (R01 DA16850).

b. The author discloses multiple consulting roles in the past two years including participation on an external review of the Statistical Research Division of the U.S. Census Bureau, service on the data safety monitoring board of the USC Well Elderly Study, service as a dissertation reader for a Ph.D. student at RAND Graduate School on predictors of HIV testing, and participation in a study of the Center on Child Abuse and Neglect at the University of Oklahoma Health Sciences Center.

- c. The author was responsible for all aspects of the research summarized in this article.
- d. The development of this article conformed with the Declaration of Helsinki and all applicable federal and state laws of the United States of America.
- e. No other acknowledgements.

References

1. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; New York: 1987.
2. Rubin DB. Multiple imputation after 18+ years (with discussion). J Am Stat Assoc 1996;91:473–489. doi:10.2307/2291635.
3. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3–15. Medline. doi: 10.1191/096228099671525676. [PubMed: 10347857]
4. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2nd edition. John Wiley & Sons; New York: 2002.
5. Wainer H. Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. J Educ Stat 1989;14:121–140. doi:10.2307/1164602.

Table

Hypothetical data from two-arm randomized study: Macular Thickness (X) from baseline assessment, Visual Acuity Improvement (Y) based in part on 6-month assessment

Patient	Standard treatment		New treatment	
	Macular Thickness (X)	Visual Acuity Improvement (Y) *	Macular Thickness (X)	Visual Acuity Improvement (Y) *
101	180	0	180	0
102	180	0	180	1
103	195	1	195	1
104	195	1	195	2
105	210	0	210	1
106	210	2	210	2
107	225	1	225	2
108	225	2	225	3
109	240	1	240	2
110	240	2	240	3
111	255	1	255	1
112	255	2	255	3
113	270	1	270	2
114	270	2	270	3
115	285	2	285	2
116	285	3	285	3
117	300	2	300	3
118	300	3	300	4
119	315	3	315	4
120	315	3	315	4

* Values in boxes () treated as missing in some analyses