Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2010 *Paper* 72

The handling of missing data in molecular epidemiologic studies

Manisha Desai* Jessica Kubo†
Denise Esserman‡ Mary Beth Terry**

*Stanford University, manishad@stanford.edu
†Stanford University, jkubo@stanford.edu
‡University of North Carolina, denise.esserman@med.unc.edu
**Columbia University, mt146@columbia.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/cobra/art72

Copyright ©2010 by the authors.

The handling of missing data in molecular epidemiologic studies

Manisha Desai, Jessica Kubo, Denise Esserman, and Mary Beth Terry

Abstract

Background: Molecular epidemiologic studies face a missing data problem as biospecimen data are often collected on only a proportion of subjects eligible for study.

Methods: We investigated all molecular epidemiologic studies published in CEBP in 2009 to characterize the prevalence of missing data and to elucidate how the issue was addressed. We considered multiple imputation (MI), a missing data technique that is readily available and easy to implement, as a possible solution.

Results: While the majority of studies had missing data, only 16% compared subjects with and without missing data. Furthermore, 95% of the studies with missing data performed a complete-case (CC) analysis, a method known to yield biased and inefficient estimates.

Conclusions: Missing data methods are not customarily being incorporated into the analyses of molecular epidemiologic studies. Barriers may include a lack of awareness that missing data exists, particularly when availability of data is part of the inclusion criteria; the need for specialized software; and a perception that the CC approach is the gold standard. Standard MI is a reasonable solution that is valid when the data are missing at random (MAR). If the data are not missing at random (NMAR) we recommend MI over CC when strong auxiliary data are available. MI, with the missing data mechanism specified, is another alternative when the data are NMAR. In all cases, it is recommended to take advantage of MI's ability to account for the uncertainty of these assumptions.

Impact: Missing data methods are underutilized, which can deleteriously affect the interpretation of results.

INTRODUCTION

With the advent of new technology to measure biomarkers, studies in molecular epidemiology have become increasingly more common. As a result, many epidemiologic studies now collect biospecimens such as blood, buccal, urine or tissue samples in order to evaluate biomarkers that may provide insight into the underlying pathogenesis of disease or that may be predictive of prognosis. Generally, biospecimens are only available for a subset of the subjects in the study, posing a missing data problem. Missing data methods, however, are not typically being employed. In a 1995 study, Greenland and Finkle (1) discussed the underuse of missing data methods in epidemiologic studies due to their inaccessibility and complexity. Although missing data methods are more readily available at present, a recent study by Klebanoff and Cole in 2008 (2) found that less than 2% of papers published in epidemiology journals make use of accessible missing data methods like multiple imputation (MI). Instead, a common approach is to perform a complete-case (CC) analysis (1,2). More specifically, a CC analysis excludes subjects missing data on at least one variable considered in the analysis.

There are a variety of reasons data from biospecimens may be missing in molecular epidemiology studies, some of which may be related to the actual values of the biomarkers themselves and/or other variables; these underlying reasons matter. Specifically, CC approaches are statistically valid (i.e., they provide unbiased point estimates and confidence intervals that achieve nominal coverage (3)) only when data

COBRA
A BEPRESS REPOSITORY
Research Archive

Hosted by The Berkeley Electronic Press

are missing completely at random (MCAR); i.e., when missingness is unrelated to observed or unobserved data yielding a study sample that is representative of the larger cohort (3, 4). For example, consider a batch of samples randomly chosen for processing for which measurements are not observed due to an instrumentation malfunction, as occurred in the study by Clendenen et al. (5); these data are MCAR. If missingness is related only to observed variables, the data are considered missing at random (MAR). An example of this may be given by Mavaddat et al. (6) who examined the role of common SNPs in subtypes of breast cancer. The authors found that those eligible for study without samples for genotyping were more likely to have advanced stage breast cancer (III/IV). The data may be MAR if, conditional on stage, the probability of missing SNP information is not related to the unobserved SNP values. If, however, the reason for missing data is related to the unobserved values, the data are not missing at random (NMAR). For example, suppose tumor size is measured less frequently on smaller tumors as in the study described by Gilcrease et al. (7); these data would be considered NMAR. CC analyses conducted on data that are not MCAR (i.e., MAR or NMAR) can lead to biased and inefficient estimates.

Often one can infer whether missingness is related to the observed data, which may suggest that MCAR is not a reasonable assumption. Distinguishing between NMAR and MAR patterns, however, is not feasible without making unjustifiable assumptions since it is impossible to examine the nature of missingness for data that do not exist. Thus, one may rely on assumptions based on biological, clinical and epidemiological understandings.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art72

There are theoretically sound methods for analyzing data that are either MAR or NMAR. For MAR data, likelihood-based methods and standard MI are examples of statistically valid approaches that are simple to implement and readily available (4). Analogous methods exist for NMAR data although they are not as easily accessible and are more complex. The increase in complexity is due to the need to model the missing data distribution whereas assuming the data are MAR generally allows one to ignore this aspect. Valid likelihood-based methods for NMAR data include EM approaches to obtaining maximum-likelihood estimates and similar estimation strategies that exploit auxiliary data (defined as additional data that can be used to improve model performance given the missingness (8-11). While software has been developed for some cases under NMAR conditions, it has not been incorporated into mainstream statistical packages. Thus, access to specialized software presents a barrier to using these methods. MI, with the missing data distribution specified (such as pattern mixture models), is another alternative when the data are NMAR (12, 13) .

The goals of this paper are to characterize the prevalence of missing data in molecular epidemiology studies, to elucidate how the issue is being addressed, and to discuss MI as a possible practical solution.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

MATERIALS AND METHODS

Missing Data in Molecular Epidemiology Studies

Cancer Epidemiology Biomarkers and Prevention is a high-ranked journal that frequently reports on molecular epidemiology studies. We examined all molecular epidemiology studies in this journal from the year 2009. A molecular epidemiology study was defined as an observational study using both epidemiologic data (such as demographic or clinical data) and molecular data obtained from a biospecimen such as tissue, saliva, or serum to address a research question. Studies that performed meta-analyses were excluded for two reasons: 1) the state of missing data was difficult to assess because they involve multiple studies each of which has its own inclusion and exclusion criteria and 2) these studies typically summarize results from individual studies. Pooled studies, on the other hand, were included as these were viewed as single studies that combined cohorts of subjects to address a question. Of the 401 studies published in the Research Articles section, a total of 207 studies satisfied our inclusion criteria and were included in our assessment. We characterized the most common types of study designs encountered in molecular epidemiology studies and calculated the percentage of studies that 1) had missing data, 2) used the availability of data as a criterion for inclusion into their study, 3) utilized missing data methods when relevant, 4) described differences between those with and without available biomarker data, and 5) implemented a CC analysis.

Multiple Imputation

MI is a simulation-based method for handling missing data. There are three main steps involved in conducting an MI-based analysis. The first step consists of imputing

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

plausible values for missing data from a specified distribution. To incorporate the uncertainty of the imputed values, this is done *m* times to create *m* complete data sets, where *m* typically varies between 3 and 10. The data are analyzed separately for each of the *m* data sets in step 2, with the estimates appropriately combined to yield one summary result in step 3. The theoretical underpinnings of the method are described in Little and Rubin (4).

There are several approaches to specifying an appropriate distribution from which to draw the missing values required in the imputation step. In general, the strategies fall into one of two classes: the joint modeling approach or the fully conditional specification approach (13). The joint modeling approach relies on specifying a joint density for the data to derive the posterior predictive distribution of the missing values(12). The fully conditional specification approach, on the other hand, bypasses this step and imputes data on a variable-by-variable basis based on a specified conditional density. For more details on the comparison of these approaches see Van Buuren (13). These methods are available in easily accessible software. SAS, for example, utilizes MI based on the joint modeling approach via the PROC MIANALYZE procedures. We provide example code that uses the fully conditional approach implemented via the ICE and MICOMBINE procedures, developed by Patrick Royston for use in STATA (14-16) in Appendix A. Other software implementing MI can be found in Horton and Kleinman's comprehensive review (17).

Figure 1 graphically illustrates MI on a simulated data set. The data were generated such that a continuous covariate was missing data for 35% of the subjects. The covariate, generated under an NMAR condition, was 7.4 times more likely to be

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

missing data for every 1-unit increase in its value. A strong auxiliary variable, however, was used in the imputation model (where an increase in 1-unit of the auxiliary variable corresponded to a 4-unit increase in the covariate.) Figure 1a presents a histogram of the complete data with the observed data overlaid, where the latter were systemically missing the right tail of its complete distribution. Figure 1b overlays the histogram for the complete data with that of the imputed values (based on 10 imputed data sets). The comparability between the two complete and imputed distributions, shown clearly in Figure 1b, demonstrates the ability of MI to impute reasonable values under this condition.

RESULTS

Molecular Epidemiology Studies Included in Our Assessment

Of all 401 studies in the Research Articles section of CEPB, there were 207 studies that satisfied our inclusion criteria. Of these, 44% were cross-sectional cohort studies, 29% were standard case-control studies (that is, where cases and controls were recruited by the authors for the purpose of the study), 34 (16%) were nested case-control studies (where either cases and/or controls were obtained from another observational or experimental study designed for a different purpose), 19 (9%) were longitudinal cohort studies and 1% (3 total) were pooled studies, where 2 were pooled case-control studies and 1 was a pooled cohort study. Thirty percent of the studies were investigating a gene-environment interaction.

Characterization of Missing Data in Molecular Epidemiology Studies

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art72

Of the 207 studies included in our assessment, 200 (97%) either had missing data or used availability of data as an inclusion criterion for study entry. More specifically, 64% (132) had missing data on either the biomarker of interest (131 studies) or a key variable (1 study). The percentage missing ranged from 0.1% to 92% with a median percent missing of 14% and a mean of 23% (for three of the studies with missing data, we could not determine the percentage of subjects with missing data). Of the remaining 75 papers that did not have missing data, 68 papers (91%) used availability of data as an inclusion criterion into the study. There were 7 papers that neither had missing data nor used availability of data as inclusion criterion. Four of these were studies that appropriately defined the population of interest through use of a biospecimen, such as men with histologically confirmed prostate cancer. The remaining three studies did not claim to have any missing data, nor did they claim to use availability of data as an inclusion criterion, which was suspicious. For example, the study by Wang et al. (18), which investigated hepatocellular carcinoma, followed a cohort of 5,929 participants over an 8-year period. While survival analytic techniques were applied to account for differences in lengths of follow-up (i.e., missing data on follow-up), notably absent was any mention of missing baseline data. All 5,929 patients were successfully classified at baseline for hepatitis B and C infection as well as diabetes status using blood samples; for the latter both a fasting blood sugar level and a non-fasting glucose level were measured, indicating two blood draws. In addition, data on demographics and health behaviors were captured for the entire cohort. It is certainly possible that because hepatocellular carcinoma is relatively common in this population from southern Taiwan, the participants were highly motivated to comply. We

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

believe, however, that studies of this size without missing data are unusual. Of those studies with missing data, 84% acknowledged that they had missing data, although surprisingly, only 16% described differences in some aspect between those with and without available data. Most surprisingly, only 6 of 132 studies with missing data utilized some type of missing data method. All of these were some form of single imputation. For example, while Platek et al. (19) excluded subjects missing data on the biomarker, diet and alcohol consumption, they imputed the median value for the remaining continuous variables and created a missing category for categorical variables. All remaining studies with missing data (95%) utilized a CC analysis.

DISCUSSION

Missing Data in Molecular Epidemiology Studies

A large percentage of the studies we examined (63%) had missing data. Furthermore, a large percentage (47%) used availability of data as an inclusion criterion for study entry (and a small percentage fell into both categories). This was not surprising given the design of these studies. For example, nested case-control studies draw their subjects from other existing cohorts, such as the Women's Health Initiative (WHI), the Nurses Health Study, the Physician's Health Study or the Surveillance Epidemiology and End Results (SEER) registry. While epidemiologic data (demographics and health behavior data) may be available on a large proportion in these cohorts, biospecimens of interest will typically only be available on a smaller proportion. If one defines the study population strictly by patient characteristics such as age, gender, race, and particular disease features, one will inevitably face a missing

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

data issue for research questions that involve data from a biospecimen. Some investigators, however, alternatively used availability of data as part of the definition of the study population in the hope of avoiding a missing data problem. Unfortunately, as systematic differences between those with and without biospecimen data may exist, the potential for bias remains; excluding these individuals prior to study entry is not different than excluding them at the time of analysis.

Only 84% of the studies missing data made some mention of this in the manuscript (for example, by mentioning that not all study subjects contributed to the estimated point estimates, or that not all subjects who provided blood samples had corresponding genotype values due to an assay error.) Thus, it is likely that many investigators were not aware they were dealing with a missing data issue that could contribute to bias. This may explain the absence of comparisons between the participants and non-participants (only 16% of the studies described differences on some aspect) among those eligible for the study or among those who would have been eligible for study were having a biospecimen not part of the inclusion criteria.

Missing Data Methods Used

While a surprisingly small percentage of the studies with missing data employed a missing data method (6 studies total or 5%), those that did used single imputation (and in some cases where the missing data were categorical, a missing data indicator was used) to address the issue. An advantage of the single imputation approach is that standard complete-data methods can then be used. Additionally, with only one set of imputations being generated and without having to use specialized software, this approach has computational ease. Finally, one can incorporate the investigator's

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

knowledge into the imputation. The disadvantage, however, is that the singly imputed value reflects neither sampling variability about the actual value under a particular model for missingness nor variability corresponding to multiple models considered. Because the missing values are unknown and the methods used to analyze the filled-in data do not account for this, the extra variability due to the unknown missing values is not being taken into account. This can result in an overstatement of precision (20). Moreover, modern methods can implement methods like multiple imputation in a practical amount of computational time and accessibility to such software is increasing, making the computational ease argument for single imputation less compelling. Finally, the use of missing indicators to retain a group of subjects, while seeming simple and intuitive, is known to yield biased estimates (20).

Barriers to Using Missing Data Methods

One barrier to using missing data methods is the lack of awareness that there is a missing data problem. For example, if availability of data is used as an inclusion criterion, the investigators may not realize that there are any missing data. There is no difference, however, in the point estimates generated from excluding subjects with the appropriate characteristics prior to study entry or at analysis. A second barrier is the need to employ specialized methods to carry out an analysis that incorporates missing data methods. Finally, a third barrier may be the perceived notion that the CC approach is the gold standard and that journal reviewers may not be receptive to the use of methods like multiple imputation. Many may see MI as having cheated by filling in or making up data to favorably skew results or falsely inflate the sample size. Under the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

assumption of MAR (and MCAR), however, MI has been shown to be a statistically valid method.

Characterization of Bias Resulting from a CC Approach

The biases resulting from CC analyses have been characterized in several contexts (21-22). For example, a recent study on missing data in molecular epidemiologic studies demonstrated considerable bias in estimating interaction effects using CC methods and improvement with MI over CC particularly for situations when MAR is a reasonable assumption (22). As gene-environment interactions were being evaluated in 30% of the studies examined here, this is of particular relevance.

MI, the MAR Assumption, and Its Relationship to Auxiliary Variables

Because CC methods can result in biased and inefficient estimates if the data are not MCAR and because MCAR is often unlikely (and easily investigated), missing data methods need to become more customary. Standard MI is simple to implement and accessible but not recommended when the data are suspected to be NMAR. For example, while Taylor and colleagues promote using MI to reduce non-response bias in epidemiologic studies, they recommend doing so only when the MAR assumption is likely to hold (23). In molecular epidemiology studies, however, one may suspect that the data are NMAR or one may incorrectly assume the data are MAR. Even if one were to suspect the data are NMAR, the presence of strong auxiliary information may allow one to proceed with methods that assume MAR. It is difficult to quantify the strength of the relationship between auxiliary variable(s) and the variable with missing data that is needed in order to assume MAR. We recommend making thoughtful and reasonable assumptions before proceeding and give specific guidelines below.

COBRA
A BEPRESS REPOSITORY
Research Archive

Hosted by The Berkeley Electronic Press

Practical Considerations

When implementing MI in practice, one might be faced with the choice of which auxiliary variables to include in imputing the variable(s) of interest. In simulation studies described by Collins et al. (24), where this very issue was assessed for the MAR case, being more inclusive even when doubtful of the usefulness of some auxiliary variables resulted in increased efficiency and reduced bias. As mentioned above, MAR may be reasonable to assume in the presence of strong auxiliary variables when NMAR is suspected.

A nice feature of MI, however, is its ability to incorporate the uncertainty of these assumptions into the results, where the assumptions may involve the missing data mechanism (NMAR and MAR) as well as which auxiliary variables to include. One can also perform a sensitivity analysis of sorts that involves presenting results using different subsets of auxiliary variables in the MI analysis or in the case where MI is used after modeling the missing data mechanism, findings resulting from various assumptions of the missing data mechanism. This will give a sense of the robustness of the results. The CC analysis should be included among these.

We observed three important practices in our assessment that carry strong implications for interpreting results. One was that the distinction between the study population and the analytic data set was often nebulous. As the interpretation and generalizability of results is strongly linked to both the study population and those included in the analysis, it is crucial that investigators make clear the targeted study population, those ultimately included in the analysis, and how they arrived at the analytic data set. The second was how common it was to include availability of data in

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

the definition of the study population. Whenever possible, availability of data should be independent of characteristics for study eligibility as those without data may still be of interest to study and may differ from those with data. The third observation was that comparisons between participants and non-participants were rare. These comparisons are crucial, however, for 1) knowing how to proceed analytically, and 2) sufficiently describing the sample upon which findings are based. For example, they may provide some justification for doing a CC analysis. Interestingly, assumptions for validity of CC analyses (i.e., MCAR) or implications of violating this assumption were rarely mentioned.

In summary, we have demonstrated that molecular epidemiology studies face a particularly challenging missing data problem in that the majority of these studies will be missing data on the key variable of interest, the biomarker. While it seems sensible to study only those with the measured biomarker, we argue the importance of including those who would be eligible for study despite the missing biomarker. At the very least, we urge comparison of features between those with and without missing data. We strongly encourage the incorporation of missing data methods into the analysis when it is warranted. More specifically, if comparisons indicate the data are not MCAR, and MAR seems reasonable, we highly recommend use of standard MI. Even in cases where the data are MCAR, one can benefit from MI in efficiency. If it is likely that the data are NMAR and one can assume the strong presence of auxiliary information, standard MI may still be a reasonable estimation-enhancing tool. Otherwise, MI that models the missing data mechanism is a possibility. A useful feature of MI is that in either case it allows for incorporation of uncertainty of these factors into the results.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art72

REFERENCES

1. Greenland, S., and Finkle, W.D. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology, 142: 1255-1264, 1995.

2. Klebanoff, M.A., and Cole, S.R. Use of multiple imputation in the epidemiologic literature. American Journal of Epidemiology, 168 (4): 355-357, 2008.

3. Rubin, D. B. Multiple imputation after 18+ years. Journal of the American Statistical Association, 91: 473-489, 1996.

4. Little, R., and Rubin, D.B. Statistical analysis with missing data. Wiley-Interscience. 1987.

5. Clendenen T, Koenig KL, Shore RE, Levitz M, Arslan AA, and Zeleniuch-Jacquotte. Postmenopausal levels of endogenous sex hormones and risk of colorectal cancer. CEBP. 18(1): 275-281, 2009.

6. Mavaddat N, Dunning AM, Ponder B, Easton DF, and Pharoah PD. Common genetic variation in candidate genes and susceptibility to subtypes of breast cancer. CEBP. 18(1): 255-259, 2009.

7. Gilcrease MZ, Kilpatrick SK, Woodward WA, Zhou X, Nicolas MM, Corley LJ, Guller GN, Tucker SL, Diaz LK, Buchholz TA, and Forest JA. Coexpression of $\alpha 6\beta 4$ integrin and guanine nucleotide exchange factor net1 identifies node-positive breast cancer patients at high risk for distant metastasis. 18(1): 80-86, 2009.

8. Ibrahim, J. G., and Lipsitz, S. R. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics: 1071-1078, 1996.

9. Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. Journal of Royal Statistical Society, Series B: 173-190, 1999.

10. Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. Biometrika, 88: 551-564, 2001.

11. Ibrahim, J. G., Lipsitz, S. R., and Horton, N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. Applied Statistics, 50: 361-373, 2001.

12. Rubin, D. B. Multiple imputation for nonresponse surveys. 1987.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

13. Van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16: 219-242, 2007.

14. Royston, P. Multiple imputation of missing values. Stata Journal, 4: 227-241, 2004.

15. Royston, P. Multiple imputation of missing values. Stata Journal, 5: 118-201, 2005a.

16. Royston, P. Multiple imputation of missing values. Stata Journal, 5: 527-536, 2005b.

17. Horton, N. J., and Kleinman, K. P. Much ado about nothing: a comparison of missing data methods and software used to fit incomplete data regression models. The American Statistician, 61: 79-90, 2007.

18. Wang M, Long RE, Comunale MA, Junaidi O, Marrero J, Di Biscelglie AM, Block RM, and Mehta AS. Novel fucosylated biomarkers for the early detection of hepatocellular carcinoma. CEBP. 18(6):1914-1921, 2009.

19. Platek ME, Shields PG, Marian C, McCann SE, Bonner MR, Nie J, Ambrosone CB, Millen AE, Ochs-Balcom HM, Quick SK, Trevisan M, Russell M, Nochajski TH, Edge SB, Freudenheim JL. Alcohol consumption and genetic variation in methylenetetrahydrofolate reductase and 5-methyltetrahydrofolate-homocysteine methyltrasferase in relation to breast cancer risk. CEBP. 18(9): 2453-2459, 2009.

20. Allison PD. Missing data. Sage series: quantitative applications in the social sciences, 2002.

21. Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples A. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine. 22(4): 545-57, 2003.

22. Desai M, Esserman D, Gammon M, Terry MB. Missing data in molecular epidemiologic studies assessing interaction effects. COBRA, 2010.

23. Taylor, J. M. G., Cooper, K. L., Wei, J. T., Aruna, V. S., Raghunathan, T. E., and Heeringa, S. G. Use of multiple imputation to correct for nonresponse bias in a survey or urologic symptoms among African-American men. American Journal of Epidemiology, 56: 774-782, 2002.

24. Collins, L. M., Schafer, J. L., and Kam, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6: 330-351, 2001.

COBRA
Collection of Biostatistics
Research Archive

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

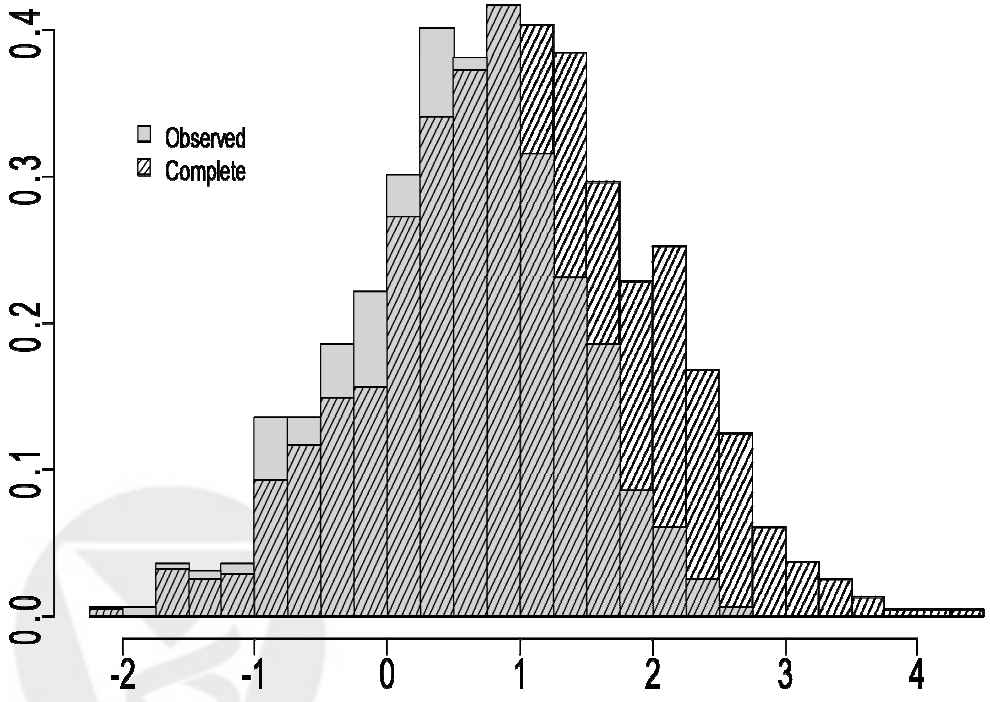

Figure 1a. Histograms for observed variable (gray) that is NMAR and generated such that 35% of the subjects missing data with a strong auxiliary variable, and complete

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics

variable (stripes).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

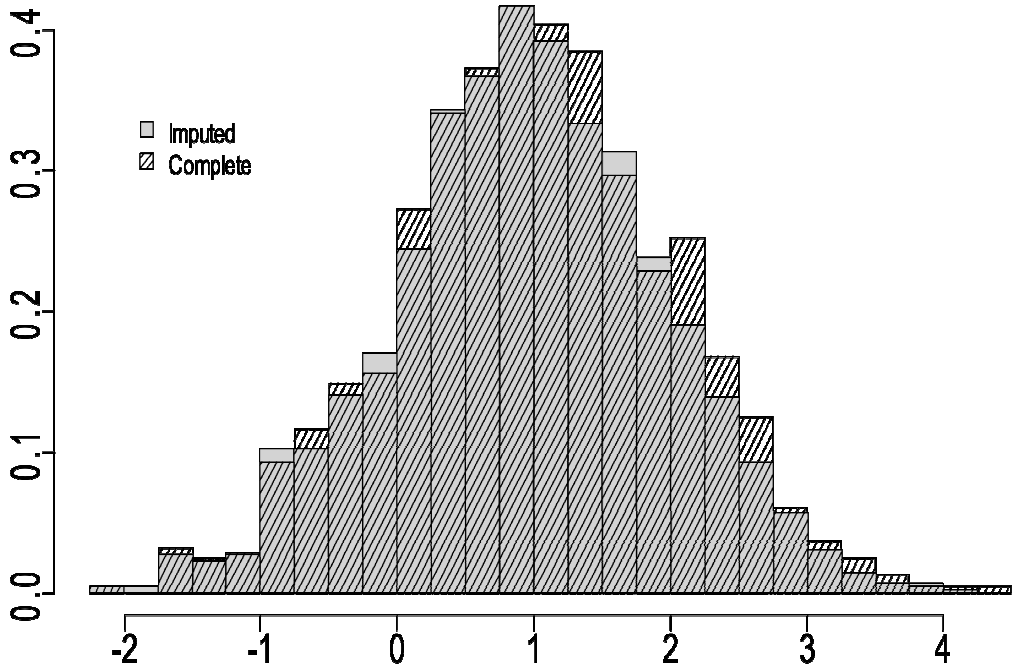

Figure 1b. Histograms for imputed (gray) and complete (stripes) variables. Data were imputed using a strong auxiliary variable and m=10. The histogram shows values resulting from all 10 imputations.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

APPENDIX A: STATA Code for Implementing MI

```
/* case is a binary indicator for case/control status */
/* x1 and x2 are binary variables, x1 is missing data on 20% of subjects and is NMAR */
/* z is a continuous auxiliary variable */

/*Read in data set where data were generated under condition 1*/

        insheet using "~/scen1.csv",
        clear

        /* Method 1 for Imputing Interaction Effects: Generate interaction term first and then
impute */

/*Create Interaction term*/
        gen theint=x1*x2

/*Use ICE to create 10 imputed data sets*/
                ice case x1 x2 theint z, saving(simimpute.dta) m(10) replace

/*Read in data set containing all 10 imputed data sets*/
        use simimpute.dta, clear

/*Use MICOMBINE to fit the desired model and combine results across 10 data sets*/
micombine logit case x1 x2 theint


        /* Method 2 for Imputing Interaction Effects: Impute first then create interaction term as
        is done in passive imputation */

/*Create Interaction term*/
        gen theint=x1*x2

/*Use ICE to create 10 imputed data sets*/
/* Using passive option to implement Method 2 for imputing interaction term */
        ice case x1 x2 theint z saving (simimpute.dta) m(10) passive (theint:x1*x2) replace

/*Read in data set containing all 10 imputed data sets*/
        use simimpute.dta, clear

/*Use MICOMBINE to fit the desired model and combine results across 10 data sets*/
micombine logit case x1 x2 theint
```

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press