Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2010 *Paper* 73

The use of multiple imputation in molecular epidemiologic studies assessing interaction effects

Manisha Desai* Denise Esserman†
Marilie Gammon‡ Mary Beth Terry**

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

The use of multiple imputation in molecular epidemiologic studies assessing interaction effects

Manisha Desai, Denise Esserman, Marilie Gammon, and Mary Beth Terry

Abstract

Background: In molecular epidemiologic studies biospecimen data are collected on only a proportion of subjects eligible for study. This leads to a missing data problem. Missing data methods, however, are not typically incorporated into analyses. Instead, complete-case (CC) analyses are performed, which result in biased and inefficient estimates.

Methods: Through simulations, we characterized the bias that results from CC methods when interaction effects are estimated, as this is a major aim of many molecular epidemiologic studies. We also investigated whether standard multiple imputation (MI) could improve estimation over CC methods when the data are not missing at random (NMAR) and auxiliary information may or may not exist.

Results: CC analyses were shown to result in considerable bias while MI reduced bias and increased efficiency over CC methods under specific conditions. It improved estimation even with minimal auxiliary information, except when extreme values of the covariate were more likely to be missing. In a real study, MI estimates of interaction effects were attenuated relative to those from a CC approach.

Conclusions: Our findings suggest the importance of incorporating missing data methods into the analysis. If the data are MAR, standard MI is a reasonable method. Under NMAR we recommend MI as a tool to improve performance 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.

The Use of Multiple Imputation in Molecular Epidemiologic Studies Assessing Interaction Effects

Manisha Desai[1], Denise A Esserman[2], Marilie D Gammon[3], Mary Beth Terry[4]

Keywords: Missing data, molecular epidemiology, multiple imputation, auxiliary data, gene-environment interaction

Running Title: Missing data in molecular epidemiologic studies

* Send requests for reprints to: Manisha Desai, 1070 Arastradero Road, Palo Alto, CA 94304, manishad@stanford.edu

[1]Department of Medicine, Division of General Internal Medicine, Stanford University, Palo Alto, CA, 94304

[2]Department of Medicine, Division of General Medicine and Epidemiology, and Department of Biostatistics, University of North Carolina School of Medicine, Chapel Hill, NC, 27599

[3]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, 27599

[4]Department of Epidemiology and Herbert Irving Comprehensive Cancer Center, Columbia University, Mailman School of Public Health, NY, NY 10032

ABSTRACT

Background: In molecular epidemiologic studies biospecimen data are collected on only a proportion of subjects eligible for study. This leads to a missing data problem. Missing data methods, however, are not typically incorporated into analyses. Instead, complete-case (CC) analyses are performed, which result in biased and inefficient estimates.

Methods: Through simulations, we characterized the bias that results from CC methods when interaction effects are estimated, as this is a major aim of many molecular epidemiologic studies. We also investigated whether standard multiple imputation (MI) could improve estimation over CC methods when the data are not missing at random (NMAR) and auxiliary information may or may not exist.

Results: CC analyses were shown to result in considerable bias while MI reduced bias and increased efficiency over CC methods under specific conditions. It improved estimation even with minimal auxiliary information, except when extreme values of the covariate were more likely to be missing. In a real study, MI estimates of interaction effects were attenuated relative to those from a CC approach.

Conclusions: Our findings suggest the importance of incorporating missing data methods into the analysis. If the data are MAR, standard MI is a reasonable method. Under NMAR we recommend MI as a tool to improve performance 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.

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 study biomarkers that may provide insight into the underlying pathogenesis of disease or that may be predictive of prognosis. Often these investigations assess synergistic effects of the biomarker and another feature. A recent assessment of molecular epidemiologic studies revealed that 30% of such studies evaluate a gene-environment interaction (1). Generally, however, 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 (2) discuss 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 (3) found that less than 2% of papers published in epidemiology journals demonstrate the use of even accessible missing data methods like multiple imputation (MI). Instead, a common approach is to perform a complete-case (CC) analysis (1-3). More specifically, a CC analysis excludes subjects missing data on at least one variable considered in the analysis. Desai et al. recently assessed the handling of missing data specifically in molecular epidemiology studies and found that while the majority of studies acknowledged having missing data, 95% of these utilized a CC analysis (1).

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 estimates and confidence intervals that achieve nominal coverage) only when data 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 (4,5). See Rubin for a more complete discussion on statistical validity (5). If missingness is related only to observed variables, the data are considered missing at random (MAR). If, however, the reason for missing data is related to the unobserved values, the data are not missing at random (NMAR). An example of the latter would be if those in the study who provide a blood sample to measure folate were more likely to consume large amounts of vegetables and, as a result, have higher folate levels than those with unmeasured folate values. CC analyses conducted on data that are not MCAR can lead to biased and inefficient estimates.

Often one can infer whether missingness is related to 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 have to rely on assumptions based on biological, clinical and epidemiological understandings.

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) (6-9). 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 (10-11) .

In molecular epidemiologic studies there is often good reason to suspect the data are NMAR. For example, suppose tumor size is measured less frequently on smaller tumors. Furthermore, because it is not straightforward to distinguish between NMAR and MAR situations, analysts may incorrectly assume the data are MAR. Such studies may also make auxiliary data available. In the above example, a potentially useful auxiliary variable might be tumor site if it correlates with tumor size. Finally, many molecular epidemiologic studies focus on interaction effects, such as gene-environment effects. The bias that results specifically from estimating these effects using a CC approach has yet to be characterized.

The goals of this paper are to characterize the bias that arises from performing a commonly applied CC analysis when interaction effects are being assessed, and to discuss MI methods as a possible practical solution. We specifically investigate the consequences of applying MI, which in its standard form relies on the MAR assumption,

and assess the extent that auxiliary data (additional epidemiologic data) can help in estimation performance when data from a covariate are NMAR (i.e., when the MAR assumption is violated) and the interest lies in estimating an interaction effect, involving the covariate. We examine situations when the covariate and therefore the modifying variable of interest are missing data and evaluate the impact of the strength of the auxiliary information under three conditions of missingness: large values of the covariate are more likely to be missing; extreme values of the covariate are more likely to be missing; and the relationship between missingness and the covariate also depends on the outcome. We compare MI to the commonly applied CC approach on simulated and real data from a molecular epidemiologic study.

MATERIALS AND METHODS

Multiple Imputation (MI)

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 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 (11). The joint modeling approach relies on specifying a joint density for the data to derive the posterior predictive distribution of the missing values(10). 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 (11). These methods are available in easily accessible software. SAS, for example, utilizes MI based on the joint modeling approach via the PROC MIANALYZE procedure. 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 (12-14) in Appendix A. Other software implementing MI can be found in Horton and Kleinman's comprehensive review (15).

MI for Interaction Effects

Estimating interaction effects with MI is slightly more complicated than estimating main effects (16). This has to do with the assumptions under which the data are imputed. More specifically, MI methods that rely on parametric assumptions such as a multivariate normal distribution may produce reasonable results for the estimation of linear relationships, but not for higher-order relationships. There are several approaches to imputing interaction terms. The two main approaches are to 1) impute the variables involved in the interaction first and then generate the product term for inclusion in the analytic model or 2) generate the product term prior to imputation and then impute this term like one would any other variable. These methods and others are discussed in detail by von Hippel (17). We show example STATA code in Appendix A that implements both methods.

Design of Simulation Studies

We assessed the performance of CC and MI methods for estimating an interaction effect between two predictors (X1, which in some cases is continuous and in others is dichotomous, and a dichotomous predictor X2) on a dichotomous outcome (Y). One of the predictors (X1), and therefore the interaction term, is NMAR for a proportion of the subjects. An auxiliary variable (Z), generated as a linear function of X1 and random noise, is also available.

Table 1 describes the eighteen scenarios examined in our simulation study. Simulations were conducted to evaluate the impact of the following factors on the results: (1) the percentage missing, (2) the nature of the missingness, and (3) the relationship between X1 and Z. We chose the level of missingness to be representative of the real molecular epidemiologic study described below. To evaluate the impact of the nature of missingness, three conditions were considered. Under condition 1, the log odds of the probability of missing X1 is a linear function of X1. Under condition 2, X1 is more likely to be missing extreme values, that is, the log odds of the probability of missing X1 is a quadratic function of X1. Finally, condition 3 defines the log odds of missingness as a linear function of X1 given Y. If Y represented case-control status, for example, this would allow cases and controls to differ with respect to missingness. In our study, cases are more likely to be missing large values of X1, and controls are more likely to be missing small values of X1. To assess the effect of the strength of the relationship between X1 and the auxiliary variable on the results, nonexistent, moderate and strong relationships were considered. Moderate strength was defined as a correlation between X1 and Z of 0.57 for X1 continuous and an increase of 1 unit in Z for X1=1 versus 0 for X1 dichotomous. For X1 continuous, a strong auxiliary variable was one that had a correlation of 0.97 with X1. For X1 binary, a strong relationship was defined as a 3 or 4 unit increase in the auxiliary variable when X1 was 1 versus 0. Each scenario is based on 1000 iterations, each with a sample size of 1000.

Three different models are presented: (1) the full model or the model fit on the complete data; (2) the CC model; and (3) the MI-based model, where the missing values of X1 were imputed as a function of X2, Y and the auxiliary variable, Z. To

produce optimal results, we set *m*=10 as opposed to the more typical *m*=5, although we found negligible differences when comparing the two. For each scenario, the data were analyzed using a logistic regression model with Y as the outcome and X1, X2 and their interaction as predictors. Average point estimates, average model-based standard errors (SE), average biases, mean squared errors (MSE), mean squared errors relative to CC (RelMSE), and percentage coverage with 95% confidence intervals were calculated for X1 given X2=0, X1 given X2=1, and their interaction. The comparison of MI to CC using the RelMSE statistic is critical, as a comparison of MI to the method currently used in practice (CC) is more relevant than its comparison to an optimal method. For an ideal reference, however, the full model is presented.

Example Data Set

As an illustration of these methods, we compared a previously published CC analysis of a gene-environment interaction (18) to one based on MI methods using data from a population-based case-control study of breast cancer, the Long Island Breast Cancer Study Project (LIBCSP)(19). This particular analysis was undertaken to address a possible interaction effect between alcohol consumption and *ADH3* genotype on breast cancer risk. Details of the overall study design are provided in prior publications (18-19). In-person interviews were completed for 1,508 cases (82.1% of eligible cases) and 1,556 controls (62.8% of eligible controls). Seventy-three percent of both cases and controls who completed an interview donated a blood sample. As the CC approach adjusted for potential confounders, it further excluded those who were missing at least one variable and resulted in a data set of 1,008 cases and 1,055

controls. As previously published (18), subjects were more likely to donate blood if they were white, non-smokers, ever consumed alcohol, ever used hormone replacement therapy, breast-fed for six months or more, or ever had a mammogram.

RESULTS

Simulation results are presented in Table 2 as a function of the percentage missing for the three conditions and when there is a strong auxiliary variable. For completeness, estimates are presented for X1 given each level of X2 (denoted going forward as X1|X2) and their interaction. Performance, however, is based solely on estimation of X1|X2=0 and the interaction as their sum yields the estimate of X1|X2=1. Under condition 1, where X1 is 7.4 times more likely to be missing for X1=1 than for X1=0, CC overestimated the interaction effect. As the percentage missing increased, so did the magnitude of the bias. In addition, CC resulted in large standard error estimates due to small cell counts that occur when X1 is binary. MI, on the other hand, provided less biased and more efficient estimates of both parameters. For example, the RelMSE statistic showed improvements in estimating the interaction that increased with percentage missing from 92% to 99%. Under condition 2, where X1 is continuous and more likely to be missing extreme values, MI yielded slightly more biased estimates than CC, particularly of X1|X2=0, but smaller standard error estimates. As a consequence, the RelMSE statistic showed an overall improvement in performance by MI that increased with percentage missing, ranging from 14%-41% for the interaction effect. Under condition 3, the coverage probability for CC became increasingly worse as the percentage missing increased. The RelMSE statistic showed an improvement of MI over CC of 92%, 97% and 99% for the effect of X1|X2=0 for the three scenarios.

The nature of missingness had an effect on the direction of the bias for CC. Unlike condition 1, where CC overstated effects and condition 2, where CC did not result in large biases, CC underestimated X1|X2=0 and overstated the interaction effect under condition 3. The bias was also reflected in low coverage probabilities not found in the other conditions.

The impact of the strength of the auxiliary variable for the three conditions, where approximately 20% of the data are missing, is shown in Table 3. For each condition, there are three scenarios corresponding to non-informative, moderate, and strong auxiliary variables. Under condition 1 even when there is no auxiliary information, MI outperformed CC for both parameters, where the RelMSE statistic for the interaction effect showed a 93% improvement in estimation. Under condition 2, however, MI needed a strong auxiliary variable to compete with CC. CC did not yield biased results, but suffered a loss in efficiency. With moderate to weak auxiliary information, MI tended to underestimate the interaction effect and overestimate the effect of X1|X2=0. Overall it performed worse than CC and had an MSE that was 1.3 times greater than that of CC for X1|X2=0 and an MSE for the interaction that was 1.1 times greater. Even when the auxiliary variable was moderate, the MI MSE was twice that of the CC MSE for X1|X2=0, although this was counterbalanced by some improvement in estimating the interaction. Like condition 1, MI always improved performance over CC under condition 3. When there is no auxiliary data, MI and CC both underestimated X1|X2=0, but while CC overestimated the interaction effect, MI underestimated it. MI had a superior MSE statistic for both X1|X2=0 and the interaction. Although its MSE was 2.7 times worse than that of CC for X1|X2=1, this is because CC overestimated the interaction effect

which helped improve its estimation of X1|X2=1. With strong auxiliary information, MI improved estimation of X1|X2=0 and the interaction effect by 93% and 36%, respectively.

The results from analysis of the LIBCSP evaluating an interaction effect between alcohol consumption and the *ADH3* genotype on breast cancer risk are presented in Table 4. Adjusted odds ratios (ORs) from the previously published CC analysis, an MI-based analysis, and the percentage change in the beta coefficients or log-ORs for the interaction effect are shown. Both analyses involved fitting a logistic regression model adjusting for potential confounders (age at diagnosis; education; race; caloric intake; smoking status; and BMI). To impute values for genotype, these confounders as well as any variables identified as a risk factor for missingness were used. These possible auxiliary variables included: having ever breastfed; having ever used hormone replacement therapy; having ever used oral contraceptives; ever having a mammogram; income level; and having benign breast disease. Terry et al. previously reported a two-fold association (OR = 2.3, 95% CI 1.3-4.0) for moderate alcohol consumption (15-30 g/day) for fast metabolizers using a CC approach. MI resulted in a 39% reduction in the coefficient (OR = 1.7, 95% 1.0-2.8) (18). MI yielded parameter estimates that were smaller and closer to the null than those obtained by CC, where the percentage change in the beta coefficients ranged from 17% to >100%, and the median percentage change was 31%.

DISCUSSION

Studies of molecular epidemiology often involve collecting data on biomarkers. Issues with missing data arise when data are not fully observed for all subjects included in a study. The most common approach to analyzing these data is CC analysis (1-3), which has the advantage of computational ease but can result in estimates that are biased and inefficient. Using missing data methods in analyses, therefore, needs 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 (20). In molecular epidemiology studies, however, one may suspect that the data are NMAR or one may incorrectly assume the data are MAR. In addition, many molecular epidemiology studies evaluate interaction effects, for which the bias of CC estimates has not been fully characterized. Our goal, therefore, was to characterize this bias under CC and to investigate the performance of standard MI, a method that is as easy to implement and as accessible as CC, specifically in the context of assessing interaction effects when one of the predictors is NMAR.

Characterization of Bias Resulting from a CC Approach

Biased and inefficient estimates from the CC approach were observed in our simulation studies, indicating a strong need for missing data methods. The extent to which they were observed, however, varied by the nature of missingness. When large values of X1 were more likely to be missing (condition 1), CC tended to overestimate effects and produced large standard error estimates, particularly when X1 is not continuous. When extreme values of X1 are more likely to be missing (condition 2), CC

suffered a loss in efficiency. Finally, under condition 3, where missinginess is a function of both X1 and Y, the bias from CC was the most dramatic where it underestimated the effect of X1|X2=0 and overestimated the interaction effect.

Comparison of MI and CC Approaches in the Simulation Study

Improvements resulting from MI over CC varied by both nature of missingness and strength of auxiliary information. Specifically, under missingness conditions 1 and 3, there was no harm in using MI over CC even when there was no auxiliary information. Furthermore, when the auxiliary information was moderate to strong, large improvements were observed. Although we only show results under specification of a positive relationship between a binary X1 and missingness under condition 1, negative and positive relationships for a continuous X1 were also examined. While there was no impact on the magnitude of bias for MI, CC yielded a more overstated interaction effect under a negative relationship, yielding larger improvements of MI over CC. Larger gains in both efficiency and bias of MI over CC were observed, however, when X1 is binary rather than continuous. Under condition 2, MI was more misleading than CC, except when auxiliary information was strong, in which case, it yielded improvements.

Application of MI and CC Approaches to LIBCSP

In the LIBCSP example, data were not MCAR as blood donation was related to a variety of observed factors (18). We suspected that the data were NMAR as having the genotype for fast metabolism was related to alcohol intake, and alcohol intake was associated with providing a blood sample. Specifically, data on metabolism status were more likely to be missing for fast metabolizers and therefore, likely to be NMAR. While inference was similar between the two analytic approaches (the overall interaction effect

was not statistically significant by either method) and consistent with previous findings (21-22), MI estimates were attenuated toward the null relative to CC. Based on the findings of our simulation study, if we believe that either conditions 1 or 3 describe missingness or if we have strong auxiliary information, we would be more likely to believe our MI results. Although unlikely in this example, if condition 2 applies and we have weak auxiliary data, the truth may lie somewhere in between the CC and MI estimates. It makes sense in cases when CC and MI results are discrepant to present both analyses.

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

The intuition behind why MI performed well when auxiliary data were strong in our study has to do with the MAR assumption. Assuming the data are MAR is equivalent to assuming that the information needed to impute the missing values can be found in the observed data. This is a more reasonable assumption when the data include auxiliary information that is strongly related to the unobserved data. Thus, 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.

Our simulation study is limited in that it does not provide a precise definition for the strength of the relationship necessary for one to assume MAR. In our simulations, for X1 continuous, a strong auxiliary variable was one that had a correlation of 0.97 with X1. For X1 binary, a strong relationship was defined as a 3 or 4 unit increase in the auxiliary variable when X1 was 1 versus 0. Although extreme and perhaps not likely outside of longitudinal studies, we felt it was important to study the extremes (no

association and strong association) in addition to a moderate association (in our study, a correlation of 0.57 for X1 continuous and an increase of 1 unit in Z for X1=1 versus 0 for X1 dichotomous). A study examining the strength of the relationship needed to assume MAR is challenging, as many factors would have to be considered making it difficult to generalize. For practical purposes, we recommend making thoughtful and reasonable assumptions before proceeding. Below we give specific guidelines.

Practical Considerations

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. (23), 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. This is consistent with our findings for conditions 1 and 3. For condition 2, however, one would need to make stronger assumptions about the auxiliary information before using MI in its standard implementation if the data were NMAR. Thus, if one can assume conditions 1 or 3 and if potential auxiliary data are available, we recommend applying MI in an inclusive manner. If one were to suspect condition 2 applies, we recommend using standard MI only if one assumes the presence of strong auxiliary information. Alternatively, one could apply MI after modeling the missing data mechanism (10). The latter would require making explicit assumptions about the nature of missingness.

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.

In summary, 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 and strongly encourage the incorporation of missing data methods into the analysis when it is warranted. More specifically, if these 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.

REFERENCES

1. Desai M, Kubo J, Esserman D, Terry MB. The Handling of Missing Data in Molecular Epidemiologic Studies. COBRA. 2010.

2. 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.

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

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

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

6. 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.

7. 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.

8. 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.

9. 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.

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

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

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

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

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

15. 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.

16. Allison, Missing data. Sage Series: Quantitative Applications in the Social Sciences. 2002.

17. von Hippel, P.T. How to impute interactions, squares, andn other transformed variables. Sociological Methodology. 2009.

18. Terry, M. B., Gammon, M. D., Zhang, F. F., Knight, J. A., Wang, Q., Britton, J. A., Teitelbaum, S. L., Neugut, A. I., and Santella, R. M. ADH3 genotype, alcohol intake and breast cancer risk. Carcinogenesis, 27: 840-7, 2006.

19. Gammon, M. D., Neugut, A. I., Santella, R. M., Teitelbaum, S. L., Britton, J. A., Terry, M. B., Eng, S. M., Wolff, M. S., Stellman, S. D., Kabat, G. C., Levin, B., Bradlow, H. L., Hatch, M., Beyea, J., Camann, D., Trent, M., Senie, R. T., Garbowski, G., Maffeo, C., Montalvan, P., Berkowitz, G. S., Kemeny, M., Citron, M., Schnabel, F., Schuss, A., Hajdu, S., Vinceguerra, V., Collman, G. W., and Obrams, G. I. The Long Island Breast Cancer Study Project: Description of a multi-institutional collaboration to identify environmental risk factors for breast cancer. Breast Cancer Research and Treatment, 74: 235-54, 2002.

20. 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.

21. Smith-Warner, S.A., Spiegelman, D., Yaun SS, van den Brandt, P.A., Folsom, A.R., Goldbohm, R.A., Graham, S., Holmberg, L., Howe, G.R., Marshall, J.R., Miller, A.B., Potter, J.D., Speizer, F.E., Willett, W.C., Wolk, A., Hunter, D.J. Alcohol and breast cancer in women: a pooled analysis of cohort studies. Journal of the American Medical Association, 279: 535-540, 1998.

22. Kuper, H. Alcohol and breast cancer risk: the alcoholism paradox. British Journal of Cancer, 83: 949-951, 2000.

23. 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.

Table 1. Description of Scenarios Used in Simulation Study.

Table	Scenario	Median % Missing X1	Nature of Missing	Auxiliary Relationship	Variable Type*
2a. Impact of % Missing Under Condition 1	A	**20%**	**Condition 1[a]**	Strong[1]	X1 binary
	B	**30%**	**Condition 1[a]**	Strong[1]	X1 binary
	C	**40%**	**Condition 1[a]**	Strong[1]	X1 binary
2b. Impact of % Missing Under Condition 2	D	**20%**	**Condition 2[b]**	Strong[2]	X1 continuous
	E	**30%**	**Condition 2[b]**	Strong[2]	X1 continuous
	F	**40%**	**Condition 2[b]**	Strong[2]	X1 continuous
2c. Impact of % Missing Under Condition 3	G	**20%**	**Condition 3[c]**	Strong[2]	X1 continuous
	H	**30%**	**Condition 3[c]**	Strong[2]	X1 continuous
	I	**40%**	**Condition 3[c]**	Strong[2]	X1 continuous
3a. Impact of Auxiliary Relationship Under Condition 1	J	22%	**Condition 1[d]**	**None[3]**	X1 binary
	K	22%	**Condition 1[d]**	**Moderate[4]**	X1 binary
	L	22%	**Condition 1[d]**	**Strong[5]**	X1 binary
3b. Impact of Auxiliary Relationship Under Condition 2	M	20%	**Condition 2[b]**	**None[3]**	X1 continuous
	N	20%	**Condition 2[b]**	**Moderate[6]**	X1 continuous
	O	20%	**Condition 2[b]**	**Strong[2]**	X1 continuous
3c. Impact of Auxiliary Relationship Under Condition 3	P	20%	**Condition 3[c]**	**None[3]**	X1 continuous
	Q	20%	**Condition 3[c]**	**Moderate[6]**	X1 continuous
	R	20%	**Condition 3[c]**	**Strong[2]**	X1 continuous

*Coefficient values are 1 for X1 given X2=0; 2.5 for X1 given X2=1; and 1.5 for the interaction
[a]Condition 1: X1 is 7.4 times more likely to be missing if X1=1
[b]Condition 2: Extreme values of X1 are more likely to be missing (probability of missing is a quadratic function of X1 or the log odds of missing X1= $\gamma_0 + \gamma_1 X1 + \gamma_2 X1^2$, where γ_1=-1 and γ_2=2.)
[c]Condition 3: A 1-unit increase in X1 corresponds to a 7.4 times decrease in the probability of missing for controls, but a 7.4 times increase for cases
[d]Condition 1: X1 is 12.2 times more likely to be missing if X1=1
[1]Strong: Those with X1=1 have Z values that are 3 units higher on average than those with X1=0
[2]Strong: Average correlation between X1 and Z is 0.97
[3]None: X1 and Z are independent variables
[4]Moderate: Those with X1=1 have Z values that are 1 unit higher on average than those with X1=0
[5]Strong: Those with X1=1 have Z values that are 4 units higher on average than those with X1=0
[6]Moderate: Average correlation between X1 and Z is 0.57

Table 2: Impact of Percentage Missing Under Conditions 1, 2, and 3. Results From Fitting Full, Complete-Case, and Multiple Imputation Models to 1000 Simulated Data Sets With a Sample Size of 1000 Where the Covariate of Interest and as a Result the Interaction Term Were Missing for Some Subjects and the Auxiliary Information Was Strong.

a. Condition 1								
Scenario	Variable*	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
A: 20% missing								
	X1 (X2=0)	Full	1.002	0.211	0.002	0.044	0.543	95.0 (93.6,96.4)
		CC	1.000	0.284	-0.000	0.081	1.000	95.8 (94.6,97.0)
		MI	1.062	0.230	0.062	0.052	0.644	95.9 (94.7,97.1)
	X1 (X2=1)	Full	2.549	0.378	0.049	0.166	0.068	95.4 (94.1,96.7)
		CC	2.713	4.822	0.213	2.436	1.000	97.0 (95.9,98.1)
		MI	2.543	0.426	0.043	0.178	0.072	96.8 (95.7,97.9)
	Interaction	Full	1.547	0.434	0.047	0.204	0.082	95.7 (94.4,97.0)
		CC	1.713	4.894	0.213	2.503	1.000	96.7 (95.6,97.8)
		MI	1.482	0.483	-0.018	0.205	0.082	96.1 (94.9,97.3)
B: 30% missing								
	X1 (X2=0)	Full	1.003	0.211	0.003	0.044	0.358	94.6 (93.2,96.0)
		CC	0.992	0.345	-0.008	0.122	1.000	95.0 (93.6,96.4)
		MI	1.095	0.245	0.095	0.065	0.531	94.9 (93.5,96.3)
	X1 (X2=1)	Full	2.558	0.379	0.058	0.146	0.014	97.2 (96.2,98.2)
		CC	3.288	25.207	0.788	10.689	1.000	98.5 (97.7,99.3)
		MI	2.557	0.462	0.057	0.178	0.017	98.3 (97.5,99.1)
	Interaction	Full	1.555	0.435	0.055	0.195	0.018	96.1 (94.9,97.3)
		CC	2.295	25.290	0.795	10.904	0.018	96.1 (94.9,97.3)
		MI	1.462	0.515	-0.038	0.219	0.020	97.8 (96.9,98.7)
C: 40% missing								
	X1 (X2=0)	Full	1.020	0.211	0.020	0.045	0.258	95.6 (94.3,96.8)
		CC	0.991	0.411	-0.009	0.175	1.000	95.5 (94.2,96.8)
		MI	1.137	0.263	0.137	0.083	0.474	94.1 (92.6,95.6)
	X1 (X2=1)	Full	2.540	0.377	0.040	0.140	0.006	96.6 (95.5,97.7)
		CC	4.121	64.794	1.621	23.920	1.000	96.5 (95.4,97.6)
		MI	2.536	0.489	0.036	0.177	0.007	97.5 (96.5,98.5)
	Interaction	Full	1.520	0.433	0.020	0.191	0.008	96.0 (94.8,97.2)
		CC	3.131	64.890	1.631	24.283	1.000	97.0 (95.9,98.1)

		MI	1.399	0.538	-0.101	0.220	0.009	97.2 (96.2,98.2)
b. Condition 2								
Scenario	Variable	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
D: 20% missing								
	X1 (X2=0)	Full	1.007	0.108	0.007	0.012	0.590	95.5 (94.2,96.8)
		CC	1.010	0.147	0.010	0.021	1.000	95.9 (94.7,97.1)
		MI	1.059	0.115	0.059	0.017	0.820	92.5 (90.9,94.1)
	X1 (X2=1)	Full	2.552	0.279	0.052	0.089	0.909	93.7 (92.2,95.2)
		CC	2.548	0.298	0.048	0.098	1.000	93.4 (91.9,94.9)
		MI	2.593	0.281	0.093	0.095	0.966	94.5 (93.1,95.9)
	Interaction	Full	1.544	0.300	0.044	0.102	0.872	93.6 (92.1,95.1)
		CC	1.538	0.332	0.038	0.117	1.000	94.0 (92.5,95.5)
		MI	1.534	0.304	0.034	0.101	0.863	94.0 (92.5,95.5)
E: 30% missing								
	X1 (X2=0)	Full	1.003	0.108	0.003	0.012	0.437	94.5 (93.1,95.9)
		CC	1.008	0.168	0.008	0.027	1.000	94.6 (93.2,96.0)
		MI	1.079	0.118	0.079	0.020	0.741	92.4 (90.8,94.0)
	X1 (X2=1)	Full	2.532	0.277	0.032	0.078	0.798	95.7 (94.4,97.0)
		CC	2.529	0.317	0.029	0.098	1.000	96.2 (95.0,97.4)
		MI	2.612	0.284	0.112	0.091	0.928	95.6 (94.3,96.9)
	Interaction	Full	1.529	0.297	0.029	0.091	0.731	95.4 (94.1,96.7)
		CC	1.521	0.359	0.021	0.125	1.000	94.8 (93.4,96.2)
		MI	1.533	0.308	0.033	0.093	0.743	95.4 (94.1,96.7)
F: 40% missing								
	X1 (X2=0)	Full	1.005	0.108	0.005	0.011	0.330	95.8 (94.6,97.0)
		CC	1.007	0.197	0.007	0.035	1.000	95.9 (94.7,97.1)
		MI	1.109	0.123	0.109	0.026	0.749	87.5 (85.5,89.5)
	X1 (X2=1)	Full	2.551	0.279	0.051	0.081	0.620	95.7 (94.4,97.0)
		CC	2.563	0.355	0.063	0.131	1.000	96.2 (95.0,97.4)
		MI	2.675	0.295	0.175	0.112	0.852	94.3 (92.9,95.7)
	Interaction	Full	1.547	0.300	0.047	0.094	0.563	95.1 (93.8,96.4)
		CC	1.557	0.406	0.057	0.168	1.000	95.3 (94.0,96.6)
		MI	1.565	0.319	0.065	0.098	0.588	96.9 (95.8,98.0)
c. Condition 3								
Scenario	Variable	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
G: 20% missing								
	X1 (X2=0)	Full	1.012	0.108	0.012	0.012	0.076	94.7 (93.3,96.1)
		CC	0.616	0.119	-0.384	0.162	1.000	11.9 (9.9,13.9)
		MI	0.999	0.109	-0.001	0.012	0.077	94.3 (92.9,95.7)
	X1 (X2=1)	Full	2.549	0.279	0.049	0.080	0.667	96.1 (94.9,97.3)
		CC	2.320	0.289	-0.180	0.120	1.000	87.1 (85.0,89.2)
		MI	2.551	0.280	0.051	0.080	0.666	96.2 (95.0,97.4)
	Interaction	Full	1.537	0.300	0.037	0.091	0.635	96.1 (94.9,97.3)
		CC	1.704	0.313	0.204	0.143	1.000	92.6 (91.0,94.2)
		MI	1.552	0.301	0.052	0.092	0.642	95.7 (94.4,97.0)
H: 30% missing								
	X1 (X2=0)	Full	1.006	0.108	0.006	0.012	0.027	95.0 (93.6,96.4)
		CC	0.331	0.127	-0.669	0.466	1.000	0.3 (0,0.6)
		MI	0.983	0.110	-0.017	0.013	0.028	95.0 (93.6,96.4)
	X1 (X2=1)	Full	2.528	0.277	0.028	0.075	0.179	95.8 (94.6,97.0)
		CC	1.927	0.299	-0.573	0.419	1.000	48.6 (45.5,51.7)

		MI	2.492	0.277	-0.008	0.072	0.171	95.6 (94.3,96.9)
	Interaction	Full	1.522	0.297	0.022	0.084	0.743	96.0 (94.8,97.2)
		CC	1.596	0.325	0.096	0.113	1.000	95.7 (94.4,97.0)
		MI	1.509	0.298	0.009	0.081	0.713	96.1 (94.9,97.3)
I: 40% missing								
	X1 (X2=0)	Full	1.007	0.108	0.007	0.011	0.012	95.8 (94.6,97.0)
		CC	0.027	0.139	-0.973	0.966	1.000	0.0 (0.0,0.0)
		MI	0.968	0.110	-0.032	0.013	0.013	94.1 (92.6,95.6)
	X1 (X2=1)	Full	2.537	0.277	0.037	0.079	0.092	95.6 (94.3,96.9)
		CC	1.627	0.312	-0.873	0.864	1.000	23.1 (20.5,25.7)
		MI	2.461	0.276	-0.039	0.073	0.085	95.4 (94.1,96.7)
	Interaction	Full	1.531	0.298	0.031	0.089	0.683	96.1 (94.9,97.3)
		CC	1.600	0.342	0.100	0.130	1.000	95.2 (93.9,96.5)
		MI	1.493	0.297	-0.007	0.081	0.627	96.9 (95.8,98.0)

*Coefficient values are 1 for X1 given X2=0; 2.5 for X1 given X2=1; and 1.5 for the interaction

Table 3: Impact of Auxiliary Relationship Under Conditions 1, 2, and 3. Results From Fitting Full, Complete-Case, and Multiple Imputation Models to 1000 Simulated Data Sets With a Sample Size of 1000 Where the Covariate of Interest and as a Result the Interaction Term Were Missing for Approximately 20% of Subjects.

a. Condition 1

Scenario	Variable*	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
J: No Auxiliary								
	X1 (X2=0)	Full	0.998	0.211	-0.002	0.043	0.403	95.2 (93.9,96.5)
		CC	0.999	0.323	-0.001	0.106	1.000	95.3 (94.0,96.6)
		MI	1.106	0.321	0.106	0.101	0.951	95.4 (94.1,96.7)
	X1 (X2=1)	Full	2.542	0.378	0.042	0.161	0.031	95.2 (94.0,96.6)
		CC	2.890	11.419	0.390	5.271	1.000	95.7 (94.4,97.0)
		MI	2.330	0.615	-0.170	0.317	0.060	92.8 (91.2,94.4)
	Interaction	Full	1.544	0.434	0.044	0.195	0.037	96.0 (94.8,97.2)
		CC	1.891	11.500	0.391	5.321	1.000	96.4 (95.2,97.6)
		MI	1.223	0.690	-0.277	0.382	0.072	94.6 (93.2,96.0)
K: Moderate Auxiliary								
	X1 (X2=0)	Full	0.986	0.211	-0.014	0.043	0.410	96.1 (94.9,97.3)
		CC	0.984	0.323	-0.016	0.105	1.000	95.7 (94.4,97.0)
		MI	1.157	0.312	0.157	0.110	1.039	94.9 (93.5,96.3)
	X1 (X2=1)	Full	2.53	0.376	0.03	0.149	0.066	95.9 (94.7,97.1)
		CC	2.698	4.867	0.198	2.28	1.000	96.1 (94.9,97.3)
		MI	2.438	0.600	-0.062	0.251	0.110	96.3 (95.1,97.5)
	Interaction	Full	1.544	0.432	0.044	0.195	0.080	95.7 (94.4,97.0)
		CC	1.714	4.949	0.214	2.437	1.000	96.8 (95.7,97.9)
		MI	1.281	0.671	-0.219	0.317	0.130	96.7 (95.6,97.8)
L: Strong Auxiliary								
	X1 (X2=0)	Full	0.997	0.211	-0.003	0.046	0.403	95.4 (94.1,96.7)

		CC	0.999	0.324	-0.001	0.113	1.000	94.4 (93.0,95.8)
		MI	1.027	0.223	0.027	0.051	0.447	95.3 (94.0,96.6)
	X1 (X2=1)	Full	2.546	0.378	0.046	0.158	0.027	96 (94.8,97.2)
		CC	2.968	12.65	0.468	5.941	1.000	96.9 (95.8,98.0)
		MI	2.555	0.407	0.055	0.176	0.030	96.5 (95.0,97.4)
	Interaction	Full	1.549	0.434	0.049	0.200	0.034	96.1 (94.9,97.3)
		CC	1.969	12.730	0.469	5.971	1.000	96.6 (95.5,97.7)
		MI	1.528	0.462	0.028	0.205	0.034	96.2 (95.0,97.4)

b. Condition 2

Scenario	Variable	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
M: No Auxiliary								
	X1 (X2=0)	Full	1.008	0.108	0.008	0.011	0.496	95.6 (94.3,96.9)
		CC	1.013	0.146	0.013	0.023	1.000	93.8 (92.3,95.3)
		MI	1.090	0.145	0.090	0.029	1.254	92.0 (90.3,93.7)
	X1 (X2=1)	Full	2.551	0.280	0.051	0.083	0.869	95.6 (94.3,96.9)
		CC	2.550	0.299	0.050	0.096	1.000	95.4 (94.1,96.7)
		MI	2.357	0.293	-0.143	0.085	0.888	93.1 (91.5,94.7)
	Interaction	Full	1.543	0.300	0.043	0.093	0.804	94.9 (93.5,96.3)
		CC	1.537	0.333	0.037	0.116	1.000	95.5 (94.2,96.8)
		MI	1.267	0.325	-0.233	0.126	1.084	90.2 (88.4,92.0)
N: Moderate Auxiliary								
	X1 (X2=0)	Full	1.011	0.108	0.010	0.012	0.533	95.4 (94.1,96.7)
		CC	1.009	0.146	0.009	0.023	1.000	94.5 (93.1,95.9)
		MI	1.163	0.142	0.163	0.046	2.039	81.4 (79.0,83.8)
	X1 (X2=1)	Full	2.558	0.280	0.058	0.090	0.911	94.2 (92.8,95.6)
		CC	2.555	0.298	0.055	0.099	1.000	94.5 (93.5,96.3)
		MI	2.571	0.296	0.071	0.085	0.855	96.7 (95.6,97.8)
	Interaction	Full	1.547	0.300	0.047	0.103	0.845	94.9 (93.5,96.3)
		CC	1.545	0.333	0.045	0.122	1.000	94.7 (93.3,96.1)
		MI	1.408	0.328	-0.092	0.102	0.832	95.1 (93.8,96.4)
O: Strong Auxiliary								
	X1 (X2=0)	Full	1.002	0.108	0.002	0.012	0.518	94.5 (93.1,95.9)
		CC	1.004	0.146	0.004	0.024	1.000	94.2 (92.8,95.6)
		MI	1.053	0.114	0.053	0.016	0.693	93.0 (91.4,94.6)
	X1 (X2=1)	Full	2.530	0.278	0.030	0.079	0.860	95.2 (93.9,96.5)
		CC	2.530	0.297	0.030	0.092	1.000	95.0 (93.6,96.4)
		MI	2.573	0.280	0.073	0.084	0.908	95.7 (93.8,96.4)
	Interaction	Full	1.527	0.298	0.027	0.092	0.788	95.0 (93.6,96.4)
		CC	1.526	0.331	0.026	0.117	1.000	94.9 (93.5,96.3)
		MI	1.520	0.303	0.020	0.092	0.791	95.1 (93.8,96.4)

c. Condition 3

Scenario	Variable	Method	Mean β	Mean SE	Mean Bias	MSE	RelMSE	Coverage
P: No Auxiliary								
	X1 (X2=0)	Full	1.006	0.108	0.006	0.011	0.066	96.2 (95.0,97.4)
		CC	0.608	0.119	-0.392	0.168	1.000	10.2 (8.3,12.1)
		MI	0.680	0.119	-0.320	0.116	0.691	23.9 (21.3,26.5)
	X1 (X2=1)	Full	2.549	0.279	0.049	0.085	0.693	94.6 (93.2,96.0)
		CC	2.322	0.288	-0.178	0.122	1.000	85.3 (83.1,87.5)
		MI	1.974	0.274	-0.526	0.324	2.654	49.9 (46.8,53.0)
	Interaction	Full	1.543	0.299	0.043	0.097	0.640	95.1 (93.8,96.4)
		CC	1.714	0.312	0.214	0.152	1.000	92.3 (90.6,94.0)

		MI	1.294	0.297	-0.206	0.093	0.609	93.5 (92.0,95.0)
Q: Moderate Auxiliary								
	X1 (X2=0)	Full	1.003	0.108	0.004	0.012	0.073	94.5 (93.1,95.9)
		CC	0.610	0.119	-0.390	0.167	1.000	11.9 (9.9,13.9)
		MI	0.793	0.116	-0.207	0.057	0.340	54.4 (51.3,57.5)
	X1 (X2=1)	Full	2.535	0.277	0.035	0.085	0.657	93.5 (92.0,95.0)
		CC	2.309	0.287	-0.191	0.129	1.000	85.1 (82.9,87.3)
		MI	2.222	0.279	-0.278	0.141	1.093	80.9 (78.5,83.3)
	Interaction	Full	1.531	0.298	0.031	0.096	0.658	94.3 (92.9,95.7)
		CC	1.699	0.310	0.198	0.146	1.000	91.4 (89.7,93.1)
		MI	1.428	0.301	-0.072	0.074	0.508	95.4 (94.1,96.7)
R: Strong Auxiliary								
	X1 (X2=0)	Full	0.999	0.108	-0.001	0.012	0.069	95.4 (94.1,96.7)
		CC	0.603	0.118	-0.397	0.172	1.000	11.1 (9.2,13.0)
		MI	0.986	0.109	-0.014	0.012	0.072	94.8 (93.4,96.2)
	X1 (X2=1)	Full	2.533	0.277	0.033	0.081	0.635	95.0 (93.6,96.4)
		CC	2.303	0.287	-0.197	0.128	1.000	84.2 (81.9,86.5)
		MI	2.533	0.278	0.033	0.081	0.628	95.2 (93.9,96.5)
	Interaction	Full	1.533	0.298	0.033	0.091	0.638	95.3 (94.0,96.6)
		CC	1.700	0.310	0.200	0.142	1.000	92.8 (91.2,94.4)
		MI	1.547	0.299	0.047	0.091	0.640	95.8 (94.6,97.0)

*Coefficient values are 1 for X1 given X2=0; 2.5 for X1 given X2=1; and 1.5 for the interaction

Table 4: Results From Fitting Complete-Case and Multiple Imputation Models to Data from the Long Island Breast Cancer Study Project (15) Assessing the Effect of a Gene-Environment Interaction where m=10

Genotype/Alcohol Status	OR^{a}_{CC} N=2,063 (95% CI)	OR^{a}_{MI} N=3,064;m=10 (95% CI)	% Change in β Coefficient
Slow-Intermediate/Non-alcohol consumer	1.00	1.00	
Fast/Non-alcohol consumer	1.18 (0.88, 1.58)	1.14 (0.88, 1.48)	22.11%
Slow-Intermediate/<15 grams	1.16 (0.89, 1.50)	1.11 (0.89, 1.39)	25.38%
Fast/ < 15 grams	0.92 (0.69, 1.23)	0.95 (0.75, 1.20)	30.89%
Slow-Intermediate/15-30 grams	1.49 (0.99, 2.25)	1.27 (0.89, 1.82)	40.23%
Fast/ 15-30 grams	2.32 (1.35, 4.01)	1.68 (1.03, 2.75)	38.68%
Slow-Intermediate/ 30+ grams	0.72 (0.43, 1.21)	0.77 (0.49, 1.19)	17.40%
Fast/ 30+ grams	0.98 (0.52, 1.87)	0.86 (0.47, 1.56)	> 100%

CC=Complete Case; CI=Confidence Interval; MI= Multiple Imputation;OR=Odds Ratio
[a]Estimates are adjusted for age at diagnosis, education, race, caloric intake, smoking status and body mass index

APPENDIX A: STATA Code for Implementing MI

```
/* Data are generated under condition 1 */
/* 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
```