Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

Cancer Epidemiology, Biomarkers & Prevention

Minireview

See editorial, p. 1567

The Handling of Missing Data in Molecular Epidemiology Studies

Manisha Desai[1], Jessica Kubo[1], Denise Esserman[2], and Mary Beth Terry[3]

Abstract

Molecular epidemiology studies face a missing data problem, as biospecimen or imaging data are often collected on only a proportion of subjects eligible for study. We investigated all molecular epidemiology studies published as Research Articles, Short Communications, or Null Results in Brief in *Cancer Epidemiology, Biomarkers & Prevention* from January 1, 2009, to March 31, 2010, to characterize the extent that missing data were present and to elucidate how the issue was addressed. Of 278 molecular epidemiology studies assessed, most (95%) had missing data on a key variable (66%) and/or used availability of data (often, but not always the biomarker data) as inclusion criterion for study entry (45%). Despite this, only 10% compared subjects included in the analysis with those excluded from the analysis and 88% with missing data conducted a complete-case analysis, a method known to yield biased and inefficient estimates when the data are not missing completely at random. Our findings provide evidence that missing data methods are underutilized in molecular epidemiology studies, which may deleteriously affect the interpretation of results. We provide practical guidelines for the analysis and interpretation of molecular epidemiology studies with missing data. *Cancer Epidemiol Biomarkers Prev; 20(8); 1571–9. ©2011 AACR.*

Introduction

With the advent of new technology to measure biomarkers, studies in molecular epidemiology have become increasingly more common. As a result, many epidemiology studies now collect biospecimens such as blood, buccal, urine, or tissue samples to evaluate biomarkers that may provide insight into the underlying pathogenesis of disease or that may be predictive of prognosis. Imaging studies, such as mammography, positron emission tomography, and functional MRI, are also used to measure relevant biomarkers of disease.

Generally, biospecimens and image-based data are available only for a subset of the subjects in the study, posing a missing data problem. Occasionally, even when samples are available, measurements may be subject to censoring (i.e., partially missing) due to the detection limit of an assay. Missing data methods, however, are not typically being employed. In a 1995 study, Greenland and Finkle (1) discussed the underutilization of missing data methods in epidemiology studies due to their inaccessibility and complexity. Although missing data methods such as imputation are more readily available at present, a recent study by Klebanoff and Cole in 2008 (2) found that less than 2% of articles published in epidemiology journals make use of imputation-based methods. Instead, a common approach is to conduct a complete-case (CC) analysis (1, 2): exclusion of subjects missing data on at least one variable considered in the analysis. Our study characterizes the prevalence of missing data specifically in molecular epidemiology studies and provides an in-depth description of how the issue is addressed.

There are a variety of reasons biomarker data 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, that is, they provide unbiased point estimates and CIs that achieve nominal coverage (3), only when data are missing completely at random (MCAR), that is, 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 randomly selected samples for which measurements are not observed because of an instrumentation malfunction, as occurred in the study by Clendenen and colleagues (5); it is reasonable to assume that these data are MCAR. In this case, a CC analysis should not yield biased estimates, although the estimates may suffer from efficiency loss. 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 and colleagues (6), who

Authors' Affiliations: [1]Quantitative Sciences Unit, Department of Medicine, Stanford University, Palo Alto, California; [2]Department of Medicine, Division of General Medicine and Epidemiology, and Department of Biostatistics, University of North Carolina School of Medicine, Chapel Hill, North Carolina; and [3]Department of Epidemiology and Herbert Irving Comprehensive Cancer Center, Columbia University, Mailman School of Public Health, New York

Corresponding Author: Manisha Desai, Quantitative Sciences Unit, Department of Medicine, Stanford University, 1070 Arastradero Road, Palo Alto, CA 94304. Phone: 650-725-1946; Fax: 650-723-6450; E-mail: manishad@stanford.edu

doi: 10.1158/1055-9965.EPI-10-1311

©2011 American Association for Cancer Research.

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

examined the role of common single-nucleotide polymorphisms (SNP) in subtypes of breast cancer. These authors found that those eligible for study without samples for genotyping were more likely to have advanced stage breast cancer (III/IV). In this case, 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 and colleagues (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 observed variables, as Mavaddat and colleagues (6) conducted in their analysis comparing those included in the analysis with those excluded from the analysis, which may suggest MCAR is not a reasonable assumption for the variable in question. Distinguishing between NMAR and MAR patterns, however, is not feasible without making unjustifiable assumptions, as 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 epidemiologic understandings.

There are theoretically sound methods for analyzing data that are either MAR or NMAR. For MAR data, likelihood-based methods and standard multiple imputation (MI) are examples of statistically valid approaches. Furthermore, MI is particularly simple to implement and readily available (4). Analogous methods (likelihood-based and MI-based) exist for NMAR data, although they are not as easily accessible and are more complex to implement (4, 8–14). The increase in complexity is due to the need to model the missing data distribution (or missing data mechanism), whereas assuming the data are MAR generally allows one to ignore this aspect.

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

Materials and Methods

Missing data in molecular epidemiology studies

Cancer Epidemiology, Biomarkers & Prevention (*CEBP*) is a high-ranked journal that frequently reports on molecular epidemiology studies. We examined all molecular epidemiology studies published as Research Articles, Short Communications, or Null Results in Brief in this journal from January 1, 2009, through March 31, 2010. A molecular epidemiology study was defined as an observational study either using both epidemiologic data (such as demographic or clinical data) and molecular data obtained from a biospecimen such as tissue, saliva, or serum or using image-based data such as MRI or mammographic imaging data to address a research question. Studies that conducted meta-analyses were excluded for 2 reasons: (i) 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 (ii) these studies typically summarize results from individual studies. Pooled studies, on the other hand, were included, as these were viewed as single studies that applied consistent inclusion/exclusion criteria across subjects to combine cohorts to address a question.

Although any analytic approach including CC may be considered a method for handling missing data, we considered missing data methods to be tools applied for the purpose of including subjects in the analysis in the presence of both missing outcome and covariates. This includes likelihood-based methods, single and multiple imputations, and the use of missing data indicators. For longitudinal outcomes (such as time-to-event data or repeated-measures data), Cox proportional hazards models or mixed-effects models are examples of methods that accommodate missing outcome data; subjects can be included as long as they are measured for at least one time point, and validity relies on an MAR assumption for the outcome. If no additional attempt was made, however, to include subjects with missing covariates, the study was classified as not having employed a missing data method.

We characterized the most common types of study designs encountered in molecular epidemiology studies and calculated the percentage of studies that (i) had missing data, (ii) used availability of data as a criterion for inclusion into their study, (iii) used missing data methods when relevant, (iv) described differences between those included in and excluded from analysis, and (v) implemented a CC analysis.

Results

Molecular epidemiology studies included in our assessment

Of all 534 studies in the Research Articles, Short Communications, or Null Results in Brief sections of *CEBP*, there were 278 studies that satisfied our inclusion criteria. Of these, 38.1% were cross-sectional cohort studies, 28.1% were standard case–control studies (i.e., in which cases and controls were recruited by the authors for the purpose of the study), 17.3% were nested case–control studies (in which cases and/or controls were obtained from another observational or experimental study designed for a different purpose), 14% were longitudinal cohort studies, and 2.5% (7 total) were pooled studies (i.e., in which 4 were pooled case–control studies and 3 were pooled cohort studies (see Fig. 1 for a graphical display).

Characterization of missing data in molecular epidemiology studies

Figure 2 graphically describes the prevalence of missingness in the studies included for assessment and how the issue was addressed. Table 1 similarly tabulates

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

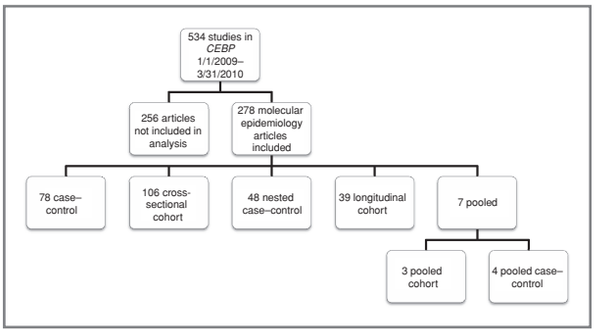

Figure 1. Articles considered for inclusion in assessment.

relevant frequencies. Of the 278 studies included in our assessment, 265 (95%) either had missing data or used availability of data as an inclusion criterion for study entry. More specifically, 66% (184) had missing data on at least one biomarker or key variable of interest. The percentage missing (on the biomarker or key variable) ranged from 0.1% to 98%, with a median percentage missing of 14% and a mean of 22% (for 10 of the studies with missing data, we could not determine the percentage of subjects with missing data). Of the 94 articles that did not have missing data, 81 articles (85%) used availability of data as an inclusion criterion for entry into the study. The remaining 13 articles neither had missing data nor used availability of data as inclusion criterion. Eleven of these articles were studies that appropriately defined the population of interest through use of a biospecimen, such as men with histologically confirmed prostate cancer. The remaining 2 studies did not claim to have any missing data, nor did they claim to use availability of data as an inclusion criterion, which was surprising (15, 16). For example, the study by Wang and colleagues (15), who investigated hepatocellular carcinoma, followed a cohort of 5,929 participants over an 8-year period. Although survival analytic techniques were applied to account for differences in lengths of follow-up (i.e., missing data on the outcome, time to development of hepatocellular carcinoma), 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 fasting blood glucose and nonfasting glucose levels were measured, indicating 2 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.

Of those studies with missing data, 85% acknowledged that they had missing data, although surprisingly, only 14% described differences in some aspect between those with and without available data. Nine of the 184 studies

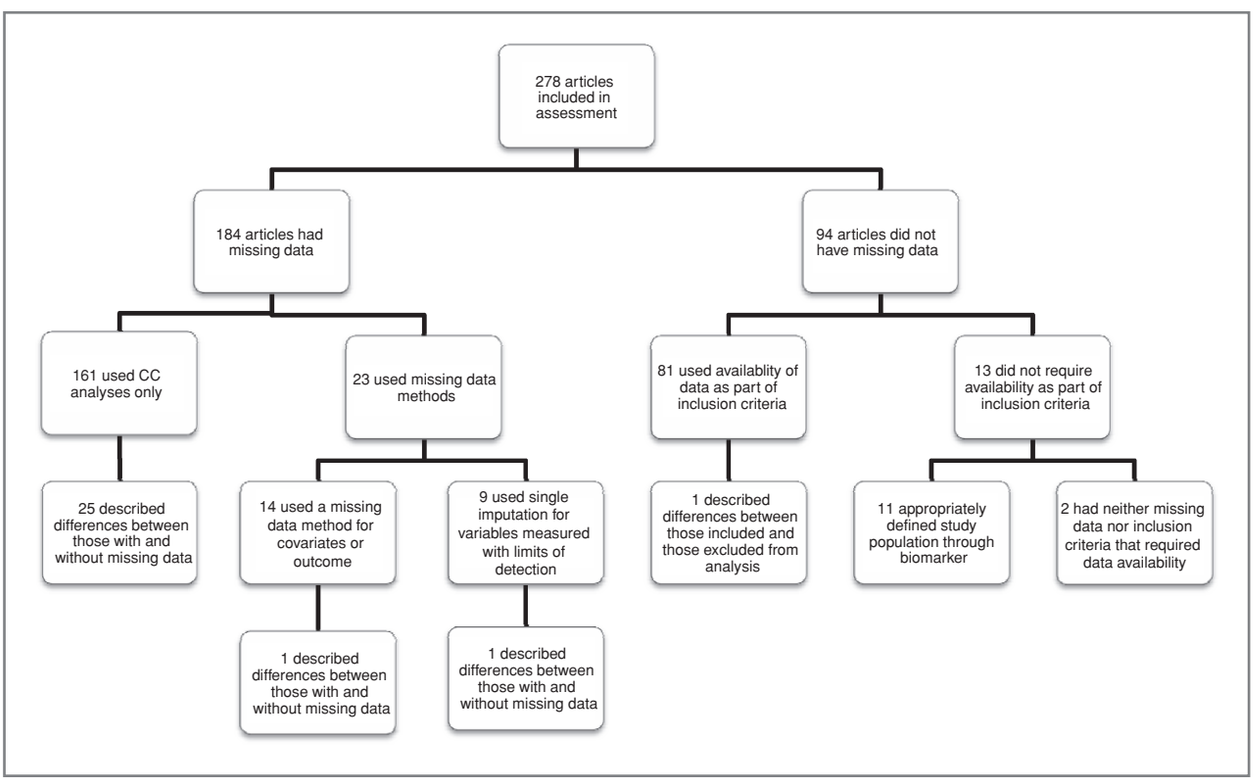

Figure 2. Presence of and techniques used to address missing data for studies assessed.

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

Table 1. Descriptive statistics relevant to missingness among studies assessed

Characteristic	Number of articles	%
Included in assessment among all *CEBP* Research Articles, Short Communications, and Null Results in Brief from January 1, 2009 to March 31, 2010 ($N = 534$)	278	52
Had missing data among articles included ($N = 278$)	184	66
Used data availability as inclusion criterion among articles included ($N = 278$)	126	45
Either had missing data or used availability of data as part of inclusion criteria among assessed ($N = 278$)	265	95
Acknowledged presence of missing data among those with missing data ($N = 184$)	157	85
Used CC analysis among those with missing data ($N = 184$)	161	88
Used missing data methods among those with missing data ($N = 184$)	23	13
Used single imputation to address limits of detection among those who used a method ($N = 23$)	9	39
Described differences between those included in and excluded from analysis among those with missing data ($N = 184$)	26	14
Described differences between those included in and excluded from analysis among those with either missing data or that used data availability as inclusion criterion ($N = 265$)	27	10

with missing data involved measures that were partially missing or censored because of assays with limits of detection. Only 23 studies with missing data used some type of missing data method (all 9 studies using assays with detection limits are included among these). All of these involved some form of single imputation (17 studies including all 9 with assay detection limits) and/or use of missing data indicators (7 studies). For example, although Platek and colleagues (17) 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 (88%) used a CC analysis.

Discussion

Missing data in molecular epidemiology studies

A large percentage of the studies we examined (66%) had missing data. Furthermore, a large percentage used availability of data (45%; often, but not always, the biospecimen or imaging-based 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, and the Surveillance Epidemiology and End Results (SEER) registry. Epidemiologic data (such as demographics and health behavior data) may be available on a large proportion in these cohorts, whereas data from biospecimens or images will typically only be available on a smaller proportion. If one defines the study population strictly by relevant patient characteristics such as age, gender, race, and particular disease features, one will inevitably face a missing data issue for research questions that involve data from a biospecimen or image. 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 the biomarker may exist, the potential for bias remains; excluding these individuals prior to study entry has the same impact as excluding them at the time of analysis.

Only 85% of the studies missing data made some mention of this in the article (e.g., 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 unaware that they were dealing with a missing data issue that could contribute to bias. This may partly explain the absence of comparisons between the participants and nonparticipants (only 10% of the studies described differences on some aspect) among those eligible or among those who would have been eligible and were having a biospecimen (or image) not part of the inclusion criteria. Another possible explanation for excluding this level of detail may be the word count restrictions of the journal.

Missing data methods used

Only a small percentage of the studies with missing data employed a missing data method (13%). Those that did used single imputation and/or missing data indicators. Advantages of the single imputation approach are as follows: standard complete data methods can be used; the approach has computational ease (there is only one set of imputations generated and no need for specialized software); and one can incorporate the investigator's knowledge into the imputation. The disadvantage, however, is that a singly imputed value reflects neither sampling variability about the actual value under a particular model for missingness nor variability corresponding to multiple models considered. This can result in an

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

overstatement of precision (14). The use of missing data indicators to retain a group of subjects is a popular approach that can be applied when the data are either categorical or continuous. Although it seems simple and intuitive, it is known to yield biased estimates even under an MCAR condition (14).

MI: A potential practical solution

Missing data methods that yield statistically valid results need to become more customary. Both MI and likelihood-based methods are statistically valid and rely on an assumption of missingness that is more flexible than that of CC and closer to what is expected from a typical molecular epidemiology study. MI is a simulation-based method for handling missing data, and similar to likelihood-based methods, it can simultaneously accommodate data that are missing on more than one variable. Unlike many likelihood-based methods, however, MI methods are readily available in mainstream software packages such as SAS, SPSS, STATA, and R. An additional advantage of MI over likelihood-based approaches is the additional ease in which auxiliary variables can be incorporated into MI, thereby enhancing the estimation. This is discussed in detail by Collins and colleagues (18).

There are 3 main steps involved in conducting an MI-based analysis. The first step consists of imputing plausible values for missing data from a specified distribution, the full complexity of which will be addressed in the following subsections. 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).

Limitations of MI. MI is theoretically sound under the MAR condition. In practice, however, its performance will not be robust if it is poorly applied or if the missing data mechanism is incorrectly specified. In the following text, we discuss 4 drawbacks to using MI: (i) the increase in burden on the user to assess whether the unverifiable assumption about the data mechanism (MAR) is appropriate; (ii) the varying performance of different imputation methods; (iii) the varying answers for each application of MI to the same data set; and (iv) the need for specialized software.

Limitation 1: Relying on unverifiable assumptions about the missing data mechanism for validity. An advantage of MI over CC is that it accommodates a more flexible assumption about missingness, whereas a disadvantage is that it increases the burden on the user by requiring careful specification of the imputation model in order for this assumption to hold. Although users rarely investigate whether the MCAR assumption holds, taking a CC approach does not avoid the need to assess assumptions about missingness, but the task may be easier: a simple check of whether those with and without data on the key variable differ can provide great insight.

Assuming the data are MAR, on the other hand, is equivalent to assuming that the information needed to impute the missing values can be found in the observed data. This requires both careful consideration of auxiliary information (variables that may enhance estimation, given the state of missingness) and one to rely on strong *a priori* knowledge of biological and clinical mechanisms. Incorporating auxiliary variables into missing data methods was studied extensively by Collins and colleagues (18). A useful auxiliary variable may be one that correlates with the variable for which the data are missing or one that correlates with missingness, or both. An example of the former may be tumor site, if it correlates with tumor size, where tumor size is missing. An example of an auxiliary variable that correlates with missingness may be advanced stage if those with more serious disease are more likely to provide biospecimens for genotyping.

Standard MI is not recommended when the data are suspected to be NMAR. For example, while Taylor and colleagues promote using MI to reduce nonresponse bias in epidemiology studies, they recommend doing so only when the MAR assumption is likely to hold (19). In molecular epidemiology studies, however, one may suspect that the data are NMAR based on *a priori* knowledge. Even so, the presence of strong auxiliary information may allow one to proceed with methods that assume MAR. It is difficult, however, to quantify the strength of the auxiliary variable(s) needed to assume MAR. This raises an important question, which is, were MAR not reasonable (e.g., if the data were truly NMAR and there were an absence of good auxiliary variables), would one be better off with a CC analysis than with a standard MI approach that assumes MAR? Desai and colleagues (20) compared the performances of CC and standard MI methods for this situation in a simulation study conducted in the context of missing exposure variables where interactions were being assessed. When extreme values of the covariate were more likely to be missing, MI yielded estimates that were more biased than those of CC. In the other situations examined (where the log odds of missing was a linear function of the variable with missing data), the bias from the 2 approaches was similar (although in all situations, the estimates achieved greater efficiency with MI than with CC). A follow-up question is: would a CC analysis be superior to a special application of an MI-based analysis appropriate for NMAR data? More specifically, if one were to suspect that the data are NMAR, one could use special models such as pattern mixture models or selection models that involve explicitly modeling the missing data mechanism in an MI-based analysis (14). Results, however, are known to be sensitive to misspecification of the model, for which the assumptions cannot be verified. It is, therefore, possible for such an approach to perform worse than CC.

Although making unverifiable assumptions about the missing data mechanism is a serious drawback to using

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

MI, a strength is its ability to incorporate the uncertainty of these assumptions into the results, where the assumptions may involve the missing data mechanism (NMAR versus MAR), different sets of auxiliary variables to include under the MAR condition, and various models under the NMAR condition. These results can also serve as a sensitivity analysis to get a sense of the robustness of the results.

Limitation 2: Varying performance of imputation methods. A second limitation is that different algorithms for producing MI vary in performance and thus the imputation method employed by the software can also have a bearing on the results. In general, the strategies for imputing fall into 1 of 2 classes: the joint modeling approach, which typically relies on a multivariate normality assumption and for which sound statistical properties have been established, or the fully conditional specification approach, which is more flexible in accommodating variables of different types but has less tractable statistical properties (13). SAS, for example, uses MI based on the joint modeling approach, whereas STATA uses the fully conditional approach. In his comparative study, van Buuren found the joint modeling approach to be more biased than the fully conditional specification approach. He recommends that the fully conditional specification approach be applied when no convenient and realistic joint distribution can be specified. For more details on the comparison of these approaches, see van Buuren (13).

Limitation 3: Varying answers for each application of MI. A third drawback is that because it is simulation based, MI produces a slightly different answer each time it is applied to the same data set. This may happen as the number of imputations varies, or even if the same number of imputations is used but the data sets are newly created. This undesirable property is not shared by CC- or likelihood-based methods. The implications of this limitation, however, are negligible in practice.

Limitation 4: Need for specialized software. Finally, the need for specialized software to apply MI can create a dependence on statistical support where it may not have existed when using the usual CC methods. Although this does add complexity over a CC analysis, MI methods are much more accessible than many of the likelihood-based methods (see Allison for a guide on accessing software for implementing likelihood-based methods; ref. 14) and are relatively easily implemented. To illustrate its ease of use, we provide example code implemented via the ICE and MICOMBINE procedures, developed by Patrick Royston for use in STATA (21–23) in Appendix A. Other software implementing MI can be found in the comprehensive review of Horton and Kleinman (24).

Censored measurements due to assays with limits of detection: a special case of missing data in molecular epidemiology studies

Much of our discussion has focused on data that are completely missing as opposed to partially missing, as when an assay does not detect levels below or above a certain point. For example, in HIV studies, the number of copies of HIV RNA per milliliter of plasma is an important marker of disease progression. The standard branched DNA assay for measuring HIV molecules, however, cannot detect levels below 50 HIV copies/mL (25). In this case, we may not know the exact number of HIV molecules in the sample, but we do know that the sample contains no more than 50 copies/mL.

Sixty-two percent of the studies we assessed made use of some sort of assay to measure a biomarker of interest, and 9 reported having issues with limit detection. In practice, assays have lower limits of detection and the resulting data are therefore left-censored. Methods specific for handling left-censored data should be considered in these situations and include both likelihood- and MI-based approaches. As the data are more likely to be censored for extreme values (outside of the range of the assay), the data are NMAR. Common approaches are to treat the data as missing, or to singly impute (as all 9 studies did in our assessment) as 0, the detection limit, or half the detection limit. These approaches are known to result in biased estimates, where the bias increases as the proportion of censored observations increases. Considerable work has been devoted to this area of study (e.g., refs. 26–28). For methods specific to data censored because of limits of detection, we refer the reader to the work of Hughes (27), in which performances between imputation- and likelihood-based methods are compared, a likelihood-based method is recommended, and software can be obtained by E-mailing the author.

Practical guidelines for handling missing data

In the following, we present a series of steps to incorporate into an analysis when faced with missing data.

Step 1: Describe targeted study population. A description of those eligible for study should be provided and unless relevant for defining subjects (e.g., those with histologically confirmed prostate cancer) should not include criteria relating to the availability of outcome or covariate data.

Step 2: Clearly describe derivation of analytic data set. Molecular epidemiology studies are often obtained from existing data sources, and, in these cases, the original sources should either be described or referenced. Critical to valid interpretation of the results, however, is the derivation of the analytic sample, that is, who was included and excluded. Ideally, this should match the description of the study population. In their study relating exposure of high-dose estrogen during adolescence to mammographic density in adulthood, Jordan and colleagues (29) provide an excellent example of a clear description of which patients were eligible to study and of which patients were ultimately analyzed by making use of a flowchart figure.

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

Step 3: Describe population characteristics of subjects in analytic data set including missing and/or censored data. Often description of the study population is provided. In cases wherein point estimates for describing relations are derived on the complete cases only (a subset of the study population), however, description of the analytic data set is crucial for interpretation, particularly if missingness is systematic.

Step 4: Describe differences between those with and without data on key variables with respect to population characteristics. In addition to aiding in interpretation of results, this also serves as an assessment of the MCAR assumption. Drawing from the earlier example, Jordan and colleagues observed that nonparticipants eligible for study (or those missing data on mammographic density) differed from participants in age at interview but not with respect to height, body mass index (BMI), and history of having had a breast biopsy (29).

Step 5: Investigate possible assumptions for missing data mechanism.

- Assume that the data are MCAR if there is no evidence of its violation (as determined in step 4) and there is no known mechanism for generating NMAR data.
- Assume that the data are MAR if the MCAR condition was violated, there is no known mechanism for generating NMAR data, and candidate auxiliary variables exist.
- Assume that the data are NMAR if *a priori* knowledge indicates that the missing values are related to the unobserved values. Common situations that give rise to NMAR data are when subjects are unable or unwilling to provide measurements because of the unobserved values (e.g., they are too sick to be evaluated and disease severity is measured), or the tool for measuring the variable cannot assess values outside its range (e.g., assays with limits of detection).

The latter 2 points need further qualification. For the first, if the MAR condition seems theoretically plausible, one must consider the presence of auxiliary variables that will make this assumption likely to hold in practice. For example, the analysis conducted by Jordan and colleagues (29) indicates that the data are not MCAR. Whether the data are MAR or whether it is possible that those with particular breast densities are more or less prone to participate is unknowable. It does not seem likely, however, that breast density would be related to participation and therefore it is reasonable to assume that the data are MAR. Auxiliary data are still required, however, for this condition to hold. Data that correlate with breast density (e.g., BMI) and/or missingness (age at interview) play an important role in satisfying this condition and need to be incorporated. For the second, even if the NMAR condition is suspected, auxiliary variables may still allow the MAR condition to hold. For example, if disease severity were being measured over several time points, and *a priori* knowledge indicates that those with missing values are more likely to be those who are too sick to come into the clinic for assessment, then the NMAR condition should be suspected. Assessments at other visits, however, may be correlated with missing scores at a given time point and could be used as auxiliary variables, making the MAR condition more plausible after conditioning on these variables. On the other hand, the MAR condition may not be appropriate in cases for which the study design requires only one disease severity assessment per patient.

Step 6: Conduct a CC analysis. If one can assume that the data are MCAR and a small proportion of the data are missing, this can serve as the sole and primary analysis. Otherwise, additional analyses should be conducted (see step 7).

Step 7: Choose an additional analytic approach if necessary. A standard MI analysis is a reasonable choice under MAR conditions (or MCAR in the presence of a large proportion of missing data for possible gains in efficiency). If NMAR is likely (even after considering auxiliary variables), special applications of MI-based methods can still be used; however, the missing data mechanism must be modeled and several NMAR models should be posed. These models are more complicated in both implementation and software accessibility and are discussed in greater detail by others (4, 14). If at least one of the missing variables is one that is subject to censoring due to limits of detection, a likelihood-based approach as described by Lyles and colleagues (26) and Hughes (27) is recommended.

Step 8: Implement the additional missing data methods. Auxiliary variables should be considered for optimal performance. More specifically, one needs to make sure that the variables specified in the imputation model include (i) all the variables in the scientific model including the dependent variable, (ii) any variables related to missingness of the variable(s), and (iii) any variable(s) correlated with those for which data are missing so that MAR can hold. When faced with the choice of which auxiliary variables to include in imputing the variable (s) of interest, however, Collins and colleagues (18) showed through simulation studies that being more inclusive even when doubtful of the usefulness of some auxiliary variables resulted in increased efficiency and reduced bias. In the study by Jordan and colleagues (29), this would involve including all variables in the scientific model (selected using forward-stepwise regression), the outcome (the mammographic measure), age at interview because it correlated with missingness, and any variables correlated with the mammographic measure (e.g., BMI). Although height and history of breast biopsy were not related to missingness, they should be included if they relate to the mammographic measure or they were included in the scientific model. The specific

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

form of the model will be dictated by the choice in software, as discussed in the limitations section. As suggested by van Buuren, this choice depends on whether the variables considered are of mixed type (i.e., some combination of continuous, categorical, ordinal, or binary variables) or whether a realistic joint distribution can be specified (13). If the latter is considered, software that uses the joint modeling approach (e.g., SAS) is recommended. Otherwise, software that employs the fully conditional specification approach is recommended (e.g., STATA or R).

Step 9: Conduct a sensitivity analysis. If missing data methods are employed, one should do a sensitivity analysis by fitting various models. If the proportion of data missing is not small and the data are MCAR, both CC and MI analyses can be done, wherein various MAR conditions (i.e., different sets of auxiliary variables) are considered. If NMAR is suspected, CC, MI approaches using different sets of auxiliary variables under MAR, and special application of MI approaches under several NMAR models should be conducted. The array of results can be presented to describe the robustness of the findings across the various assumptions. In addition, the uncertainty of these assumptions can be taken into account when presenting summarized findings. For other tips and tricks on implementing MI including how many imputations to be done, see Allison's excellent book on missing data, which provides more detailed practical guidelines and examples for applying MI in nontechnical language (14).

Step 10: Interpret results. If a consensus is easily reached across analyses, this makes interpretation (and presentation) straightforward. In situations where findings differ by assumptions, the authors will have to weigh in on what is most plausible, given their understandings of the biology, and present the array of results so the reader can understand how the assumptions affected the overall interpretation.

Conclusions

In summary, we have shown 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. Although 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 that the data are not MCAR, and MAR seems reasonable, we highly recommend the 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 presence of strong 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 it allows for incorporation of uncertainty of these factors into the results.

Appendix A: STATA Code for Implementing MI

```
/*Read in data set where interest is in estimating effects
  of 2 risk factors on case-control status*/
/* X1 and X2 are risk factors of interest and Z is a potential
  auxiliary variable */
    insheet using "~/scen1.csv",
    clear
/*Use ICE to fit imputation model creating 10 imputed
  data sets*/
/* All variables in scientific model are included in impu-
  tation model in addition to auxiliary variable */
    ice case x1 x2 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 scientific model (a
  logistic regression model that includes the risk factors
  of interest) and combine results across 10 data sets*/
    micombine logit case x1 x2
```

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Received December 15, 2010; revised March 8, 2011; accepted March 20, 2011; published OnlineFirst July 12, 2011.

References

1. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255–64.
2. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol 2008;168:355–7.
3. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473–89.
4. Little R, Rubin DB. Statistical analysis with missing data. New York: Wiley-Interscience; 1987.
5. Clendenen T, Koenig KL, Shore RE, Levitz M, Arslan AA, Zeleniuch-Jacquotte A. Postmenopausal levels of endogenous sex hormones and risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev 2009;18:275–81.
6. Mavaddat N, Dunning AM, Ponder B, Easton DF, Pharoah PD. Common genetic variation in candidate genes and susceptibility to subtypes of breast cancer. Cancer Epidemiol Biomarkers Prev 2009;18:255–9.
7. Gilcrease MZ, Kilpatrick SK, Woodward WA, Zhou X, Nicolas MM, Corley LJ, et al. Coexpression of α6β4 integrin and guanine nucleotide

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

exchange factor net1 identifies node-positive breast cancer patients at high risk for distant metastasis. Cancer Epidemiol Biomarkers Prev 2009;18:80–6.
8. Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics 1996;52:1071–8.
9. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. J R Stat Soc Ser B Stat Methodol 1999;61:173–90.
10. Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. Biometrika 2001;88:551–64.
11. Ibrahim JG, Lipsitz SR, Horton N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. Appl Stat 2001; 50:361–73.
12. Rubin DB. Multiple imputation for nonresponse surveys. New York: Wiley & Sons; 1987.
13. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16: 219–42.
14. Allison PD. Missing data. Sage series: quantitative applications in the social sciences. Thousand Oaks, CA: Sage Publications; 2002.
15. Wang CS, Yao WJ, Chang TT, Wang ST, Chou P. The impact of type 2 diabetes on the development of hepatocellular carcinoma in different viral hepatitis statuses. Cancer Epidemiol Biomarkers Prev 2009;18: 2054–60.
16. Salit IE, Tinmouth J, Chong S, Raboud J, Diong C, Su D, et al. Screening for HIV-associated anal cancer: correlation of HPV genotypes, p16, and E6 transcripts with anal pathology. Cancer Epidemiol Biomarkers Prev 2009;18:1986–92.
17. Platek ME, Shields PG, Marian C, McCann SE, Bonner MR, Nie J, et al. Alcohol consumption and genetic variation in methylenetetrahydrofolate reductase and 5-methyltetrahydrofolate-homocysteine methyltransferase in relation to breast cancer risk. Cancer Epidemiol Biomarkers Prev 2009;18:2453–9.
18. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6:330–51.
19. Taylor JMG, Cooper KL, Wei JT, Aruna VS, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey or urologic symptoms among African-American men. Am J Epidemiol 2002;56:774–82.
20. Desai M, Esserman D, Gammon M, Terry MB. Missing data in molecular epidemiologic studies assessing interaction effects. COBRA Preprint Ser. 2010. Article nr 73. Available from: http://biostats.bepress.com/cobra/ps/art73.
21. Royston P. Multiple imputation of missing values. Stata J 2004;4: 227–41.
22. Royston P. Multiple imputation of missing values. Stata J 2005;5: 118–201.
23. Royston P. Multiple imputation of missing values. Stata J 2005;5: 527–36.
24. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software used to fit incomplete data regression models. Am Stat 2007;61:79–90.
25. Anastassopoulou CG, Touloumi G, Katsoulidou A, Hatzitheodorou H, Pappa M, Paraskevis D, et al. Comparative evaluation of the QUANTIPLEX HIV-1 RNA 2.0 and 3.0 (bDNA) assays and the AMPLICOR HIV-1 MONITOR v1.5 test for the quantitation of human immunodeficiency virus type 1 RNA in plasma. J Virol Methods 2001;19: 67–74.
26. Lyles RH, Lyles CM, Taylor DJ. Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop outs. J R Stat Soc Ser C Appl Stat 2000;49:485–97.
27. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics 1999;55:625–9.
28. Paxton WB, Coombs RW, McElrath MJ, Keefer MC, Hughes J, Sinangil F, et al. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with $>=$ 400 CD4 lymphocytes: implications for applying measurements to individual patients. J Infect Dis 1997;175:247–54.
29. Jordan HL, Hopper JL, Thomson RJ, Kavanagh AM, Gertig DN, Stone J, et al. Influence of high-dose estrogen exposure during adolescence on mammographic density for age in adulthood. Cancer Epidemiol Biomarkers Prev 2010;19:121–9.

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.

Published OnlineFirst July 12, 2011; DOI: 10.1158/1055-9965.EPI-10-1311

Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

The Handling of Missing Data in Molecular Epidemiology Studies

Manisha Desai, Jessica Kubo, Denise Esserman, et al.

Cancer Epidemiol Biomarkers Prev 2011;20:1571-1579. Published OnlineFirst July 12, 2011.

Updated version	Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-10-1311

Cited articles	This article cites 25 articles, 7 of which you can access for free at: http://cebp.aacrjournals.org/content/20/8/1571.full#ref-list-1
Citing articles	This article has been cited by 3 HighWire-hosted articles. Access the articles at: http://cebp.aacrjournals.org/content/20/8/1571.full#related-urls

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions	To request permission to re-use all or part of this article, use this link http://cebp.aacrjournals.org/content/20/8/1571. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.

Downloaded from cebp.aacrjournals.org on December 8, 2020. © 2011 American Association for Cancer Research.