Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2011 August ; 20(8): 1567–1570. doi:
10.1158/1055-9965.EPI-11-0505.

Incomplete Data: What You Don't Know Might Hurt You

Daniel F. Heitjan[1]

[1]Department of Biostatistics & Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania

Abstract

Molecular epidemiology studies commonly exhibit missing observations. Methods for extracting correct and efficient analyses from incomplete data are well known in statistics, but relatively few such methods have diffused into applications. I review some areas of incomplete-data research that are relevant to molecular epidemiology, and appeal for greater efforts by statisticians to translate their methods into practice.

> [T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know.
>
> U. S. Secretary of Defense Donald H. Rumsfeld
>
> Department of Defense news briefing, February 12, 2002 (1)

Secretary Rumsfeld was referring to the problems of gathering and synthesizing accurate intelligence on terrorists and their plans. But the analogy to incomplete data in molecular epidemiology research is apt. There are the “known knowns” — the observed data that we analyze as best we can within the limits of sample size and available scientific information. Then there are the perilous “known unknowns” — the unobserved values of the missing data. These we can properly impute using the observed data and a few judiciously chosen assumptions. More dangerous still are the “unknown unknowns” — the data on subjects who were excluded from the study specifically because they had some missing items.

Desai et al. (2) review the statistical issues surrounding the analysis of incomplete data. They observe that a large fraction of studies published in this journal exhibit missing observations and that disclosure of the amount of missing data was inconsistent. Moreover only a handful of studies employed statistical methods tailored specifically for incomplete data. This is unfortunate, because the proper treatment of missing data has been a popular topic in the statistical literature for several decades. One can hardly lay the blame for this state of affairs at the feet of the scientists who publish in *CEBP*, however, as the statisticians who derived these methods have not always done their best to translate their findings into comprehensible prose and friendly software. Happily, Desai et al. (2) brims with practical advice for the analysis and reporting of incomplete data. I am hopeful that their work will have the intended effect. I offer here a few further observations intended to add some depth to the picture.

Daniel F. Heitjan, Department of Biostatistics & Epidemiology, University of Pennsylvania, 622 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104; voice: 215-573-7328; fax: 215-573-4865; dheitjan@upenn.edu.

Ignorability conditions

The pattern of missing data, like the observed data set itself, is a realization of a random process. Thus in principle one has to model and analyze the missing-data indicators just as one models other binary data. A thrust of missing-data research has been to identify *ignorability conditions*, or assumptions about the missing-data distribution that permit us to avoid modeling it. Ignorability can result in enormous simplification of the data analysis; rather than have separate models for the notional complete data and the missingness process, one simply treats the missing values as though there had never been any intention of collecting them.

Desai et al. (2) provide a concise summary of the standard ignorability conditions formally defined first in Rubin (3) and given their current form in Little and Rubin (4). The most restrictive is *missing completely at random* (MCAR), which we take to mean that the probability that a potential observation is missing is independent of its own value and of other data values, known and unknown. Slightly less restrictive is *missing at random* (MAR), defined to mean that the probability that a potential observation is missing, conditional on its value and the value of other data items, depends only on observed items. The negation of MAR is *missing not at random* (MNAR), which means that the probability of an observation being missing depends on the observation itself, even given all the other potential measured data.

It is well known that MCAR is sufficient to render correct a complete-case analysis — that is, an analysis that excludes all subjects who have missing items. Commonly we can test the null hypothesis of MCAR by comparing the distribution of a fully observed variable across groups defined by the presence or absence of some other variable. A significant test strongly suggests that the data are not MCAR.

The weaker condition MAR, together with the assumption that there are no *a priori* ties between the parameters of the data model and the missing-data model, implies that one can ignore the missing-data model in performing Bayesian or likelihood-based data analysis. Standard SAS analysis routines such as Procs Mixed and Glimmix assume MAR. To evaluate the MAR assumption, one can posit models that include MAR as a special case and test MAR as a null hypothesis. Unfortunately, such procedures are unreliable because they are exquisitely sensitive to unverifiable model assumptions. (4)

These oft-quoted results represent the most general versions of missing-data ignorability conditions, applicable in every situation. They are *sufficient* conditions, however, not *necessary*; thus their violation does not imply that ignorability does not hold. An example from molecular epidemiology is instructive. Suppose we have an outcome — disease incidence, survival time or some other phenotype — that is observed on all subjects in our study. We seek to relate this outcome to a panel of biomarkers via a regression model, where the biomarkers' effects will be evaluated in terms of functions of the regression coefficients — i.e., slopes, odds ratios or hazard ratios. The relevant fact is that a complete-case analysis of such data is perfectly valid for estimation of the regression model as long as the missing-data probability does not depend on the value of the outcome. That is, MCAR status of the biomarkers is not necessary for valid data analysis.

Why is MCAR not necessary here? The issue is what you seek to estimate. If you are only interested in the regression coefficients, then we obtain valid estimates however the subjects with missing items are chosen, as long as it does not depend on the value of the outcome itself. Even an NMAR mechanism — i.e., a mechanism where the probability that the biomarker is missing depends directly on the biomarker value — induces no bias.

The situation would be different if we were attempting to estimate a parameter of the marginal distribution of the outcome, such as its mean value in the population. If the outcome is associated with the biomarker, and the value of the biomarker determines the probability that an observation is missing, then the complete cases are a non-representative sample of the population, and consequently the mean of the outcome in the complete cases is biased. Thus for example in a cohort study in which one intends to relate a panel of SNPs to disease incidence or survival, we need not be concerned with the reasons that some SNP data are missing, as long as we can be certain that the missingness probability, given the SNP values and the outcome, does not depend on the outcome.

The need for imputation

This is not to say that the complete-case analysis is preferred, even when valid. In fact, missing data can have a profound effect on efficiency. To see this, consider a study relating an outcome to a panel of N biomarkers. If we assume that each biomarker is missing independently with probability q, then the probability that a subject has complete data is $(1-q)^N$. Table 1 shows the dependence of this probability on q and N. Note that even with a small proportion of SNPs missing, the fraction of complete cases in the data set is minute once the number of SNPs is substantial. For example, if only 2% of SNPs are missing, with 40 SNPs in the panel fewer than half of the subjects will have complete data. Complete independence gives a worst-case scenario and fortunately is not a realistic model. Under the more plausible assumption that the missing data will be concentrated within selected subjects, as would obtain if a fraction of subjects contributed insufficient material for evaluation of all biomarkers, the situation is less dire. Nevertheless, anyone who has attempted to conduct a stepwise regression on a data set with many missing predictor values has surely encountered this problem of the vanishing data.

Thus in this type of study the major concern is not nonignorability bias but loss of power and precision. Yet even with 10% missing SNPs, which would result in catastrophic data losses, on average subjects will have 90% of their SNP data, so presumably the fraction of information available on the SNP-outcome relationship far exceeds the fraction of complete cases. This is where imputation — the creation of substitute values for the missing observations — comes in. If we can impute data in a principled and robust way, we can hope to unlock that information and achieve the greatest possible efficiency.

Multiple imputation

Multiple imputation is the process of taking multiple draws from the predictive distribution of the missing observations given the complete observations under relevant model assumptions. (4) The idea is to fill in likely values for the missing data. We generate the imputations by a process of simulation that reflects our uncertainty about their true values. We create multiple data sets so as to avoid understating uncertainty about the true values of the missing items. One then analyzes each filled-in data set as a complete data set, finally combining the results across the imputations.

Imputation requires a model to describe the notional complete data, a model for the missing-data probability mechanism (typically assumed MAR), a numerical method for estimating the model, and a sampling algorithm to create the imputations. Some imputation procedures rely on implicit models; for example, predictive-mean matching selects imputations from subjects whose data are complete and that closely match the incomplete observations on a panel of fully observed predictors. (5) Such procedures can be valuable when the complete-data model is potentially complex. As a rule, the imputation model should be at least as rich as the analysis model. (6)

Extensions of models to coarse data

Desai et al. (2) hint that one can consider censored observations as a kind of partially missing data. That is, when a subject's survival is censored at, say, 5 years, we know only that his true survival time is some number larger than 5. Compare this to a completely missing observation, where all we know is that the survival time is something greater than 0. One can similarly describe other data types — data left-censored due to detection limits, or rounded, heaped or interval-censored data — in terms of inequalities on the true, unobserved data item. The recent statistical literature uses the term *coarsened data* to describe this more general form of incompleteness. (7) One can readily extend MAR and MCAR to the coarse-data model; the relevant generalizations are denoted *coarsened at random* (CAR) and *coarsened completely at random* (CCAR). (8) Contrary to the assertions of Desai et al. (2) and Little and Rubin (4), censored data should not be considered automatically NMAR; applying the CAR condition, censoring is nonignorable when the censoring limit and the true value are correlated. This would occur if subjects who enroll in the early stages of a clinical trial are more (or less) hardy than those who enroll later, or if subjects are preferentially lost to follow-up shortly before experiencing the event of interest.

Sensitivity analysis

As indicated above, MAR underlies many commonly used methods for analyzing and imputing incomplete data. When the missing data mechanism cannot reasonably be assumed to be MAR, one option is to fit models that explicitly assume dependence of the missingness probability on missing values. (9) This is both technically challenging and risky, however, as conclusions can be exquisitely sensitive to aspects of the assumed model that the data cannot robustly address.

A practical approach that has attracted interest recently is local sensitivity analysis. This involves assuming a provisional MNAR missing-data model that includes MAR as a special case, and evaluating the sensitivity of conclusions to small departures from MAR. The rationale is that if local sensitivity is modest — i.e., estimates of key parameters are unaffected by mild nonignorability — then we can trust the MAR assumption and avoid complex nonignorable modeling. Methods and workbench software exist for performing such an analysis in the generalized linear model with missing outcomes, the linear mixed model for longitudinal data with dropout, and censored data in observational studies and clinical trials. (10–13) As one would expect, sensitivity is modest if the fraction of incomplete data is small. Moreover, estimates of group-comparison parameters (hazard ratios, odds ratios, and differences in means) are insensitive to departures from MAR even if the fraction of incomplete data is large, as long as it is the same in the groups being compared.

Unknown unknowns: Missing data not disclosed

Desai et al. (2) found that 45% of the articles in their review used data availability as an inclusion criterion. This is in general a bad practice, as excluding data, either from the study data set or from a data analysis, invites bias in estimation of both summaries of marginal distributions (means, medians, proportions) and of relationships between outcomes and predictors (odds or hazard ratios, differences in means). If we know the fraction of subjects excluded, we can at least conduct a sensitivity analysis to evaluate whether nonignorability can affect conclusions. The problem with excluding subjects based on data availability is that the resulting database does not even allow us to count the excluded observations, and therefore we cannot perform even a rudimentary sensitivity analysis.

Conclusion

Desai et al. (2) have presented an excellent summary of the current status of analysis with missing data in molecular epidemiology. They have moreover proposed practical steps that can mitigate the potential biases and inefficiencies that arise with incomplete data. I applaud their work, and encourage my fellow biostatisticians to make greater efforts to translate their methods into this important area of research.

References

1. http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636
2. Desai M, Kubo J, Esserman D, Terry MB. The handling of missing data in molecular epidemiology studies. Cancer Epidemiology, Biomarkers & Prevention.
3. Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.
4. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2nd. New York: John Wiley & Sons; 2002.
5. Heitjan DF, Little RJA. Multiple imputation for the Fatal Accident Reporting System. Applied Statistics. 1991; 40:13–29.
6. Meng XL. Multiple-imputation inferences with uncongenial sources of input (with discussion). Statistical Science. 1994; 9:538–573.
7. Heitjan DF, Rubin DB. Ignorability and coarse data. Annals of Statistics. 1991; 19:2244–2253.
8. Heitjan DF. Ignorability in general incomplete-data models. Biometrika. 1994; 81:701–708.
9. Diggle PJ, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). Applied Statistics. 1994; 43:49–93.
10. Troxel A, Ma G, Heitjan DF. An index of local sensitivity to nonignorability. Statistica Sinica. 2004; 14:1221–1237.
11. Ma G, Troxel AB, Heitjan DF. An index of local sensitivity to nonignorable dropout in longitudinal modeling. Statistics in Medicine. 2005; 24:2129–2150. [PubMed: 15909292]
12. Zhang J, Heitjan DF. Nonignorable censoring in randomized clinical trials. Clinical Trials. 2005; 2:488–496. [PubMed: 16422309]
13. Zhang J, Heitjan DF. A simple sensitivity analysis tool for nonignorable coarsening: Application to dependent censoring. Biometrics. 2006; 62:1260–1268. [PubMed: 17156301]

Table 1

Percent of subjects having complete SNP data, as a function of the fraction of SNPs missing and the number of SNPs in the panel

Percent of SNPs missing (100×*q*)	Number of SNPs in the panel (*N*)						
	10	**20**	**30**	**40**	**50**	**100**	**500**
1	90	82	74	67	61	37	1
2	82	67	55	45	36	13	0
3	74	54	40	30	22	5	0
4	66	44	29	20	13	2	0
5	60	36	21	13	8	1	0
6	54	29	16	8	5	0	0
7	48	23	11	5	3	0	0
8	43	19	8	4	2	0	0
9	39	15	6	2	1	0	0
10	35	12	4	1	1	0	0