

Addressing Missing Data in Clinical Studies of Kidney Diseases

Maria E. Montez-Rath,* Wolfgang C. Winkelmayr,* and Manisha Desai†

Abstract

Missing data constitute a problem present in all studies of medical research. The most common approach to handling missing data—complete case analysis—relies on assumptions about missing data that rarely hold in practice. The implications of this approach are biased and inefficient descriptions of relationships of interest. Here, various approaches for handling missing data in clinical studies are described. In particular, this work promotes the use of multiple imputation methods that rely on assumptions about missingness that are more flexible than those assumptions relied on by the most common method in use. Furthermore, multiple imputation methods are becoming increasingly more accessible in mainstream statistical software packages, making them both a sound and practical choice. The use of multiple imputation methods is illustrated with examples pertinent to kidney research, and concrete guidance on their use is provided.

Clin J Am Soc Nephrol 9: 1328–1335, 2014. doi: 10.2215/CJN.10141013

Introduction

Missing data are a common problem in medical research. For example, in clinical trials where patients are followed over time, patients may drop out of the study or not attend every scheduled clinic visit, giving rise to missing data on relevant measurements (1). Missing data are common in observational studies as well. For example, disease registries, such as the US Renal Data System (USRDS) (2) or the Scientific Registry for Transplant Recipients (3), or electronic health records of dialysis providers face similar challenges with missing data, although they may arise for different reasons.

There are numerous ways to analyze data in the presence of missing outcomes or covariates. The approach can affect descriptions of relationships of interest. Consider the following hypothetical example. Suppose investigators are interested in characterizing the natural trajectory of eGFR among kidney transplant patients with some recurrent disease. To study this question, subjects are recruited into a study and followed over time, and eGFR is measured at regularly scheduled visits. As with any longitudinal study, some subjects are lost to follow-up. Suppose also that patients whose disease progresses are less likely to come into the clinic for their scheduled follow-up than patients whose eGFR is more stable. Figure 1 shows three eGFR trajectories: the true trajectory, the trajectory estimated based on observed measurements, and a trajectory estimated on both observed measurements and measurements assuming the last eGFR value for those patients who did not return to the clinic. The latter is a method commonly used to retain observations, and it makes a seemingly innocuous assumption that, for subjects with missing data, their status is unchanged from the last visit. The observed

trajectory shows a more modest decline in eGFR over the study period than what is actually true. This is because those patients with complete data fared better than those patients with incomplete data. The trajectory that incorporates the last value carried forward shows only mild progression of disease and may lead investigators to underestimate the true extent of eGFR decline. This simplistic example illustrates the implications of different ways of handling missing data, the idea of which is not limited to when the outcome is missing but also extends to missing covariates.

Complete Case Analyses

The default approach for handling missing data in mainstream statistical software is to perform what is called a complete case analysis. More specifically, if a subject does not have all relevant information considered in the model, the subject is excluded from analysis. Consider an example where investigators are interested in which factors among eGFR, body mass index (BMI), and BP at diagnosis of kidney disease are predictive of initiating dialysis. Suppose a subject has her eGFR and BP observed but not BMI. This subject would be excluded from the analysis. Complete case analyses lead to valid inferences about population quantities (*i.e.*, unbiased estimates of relevant relationships) when the subjects included in the analysis are no different than the subjects excluded with respect to features that influence relationships of interest. Even when this case is true, excluding subjects with only partial information results in a loss of efficiency. Because no special software is required for its use, complete case analysis is the most common approach to handling missing data (4,5).

*Division of Nephrology and
†Quantitative Sciences Unit, Department of Medicine, Stanford University School of Medicine, Palo Alto, California

Correspondence:
Dr. Maria E. Montez-Rath, Division of Nephrology, Stanford University School of Medicine, 1070 Arastradero Road, Suite 3C11C, Palo Alto, CA 94304. Email: mmrath@stanford.edu

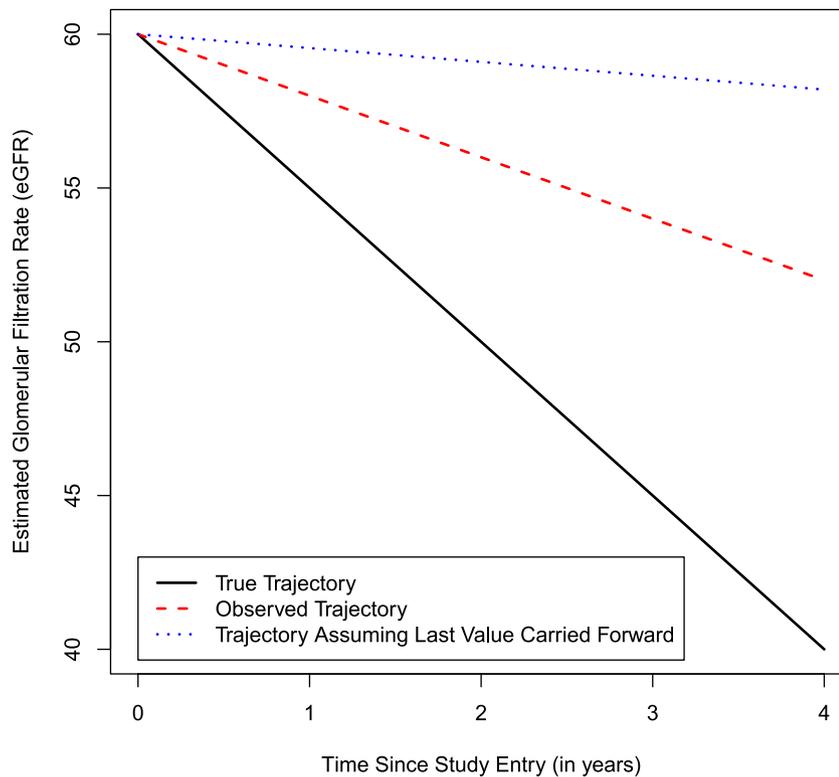


Figure 1. | Trajectories of eGFR measurements for hypothetical study of kidney transplant patients followed longitudinally. Three eGFR trajectories are shown: the true trajectory, the trajectory estimated based on observed measurements, and the trajectory estimated on both observed measurements and measurements assuming the last eGFR value for those patients who did not return to the clinic.

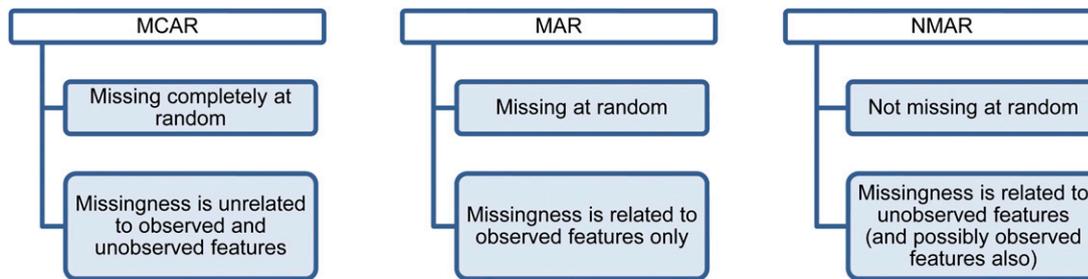


Figure 2. | Missing data mechanisms. MAR, missing at random; MCAR, missing completely at random; NMAR, not missing at random.

Assumptions Regarding Missing Data

All statistical methods rely on assumptions about missing data for statistical validity of findings. To choose an appropriate method, one must consider why data are missing. There are three main categories that characterize reasons for missingness or missing data mechanisms, and they are described in greater detail in the seminal book by Little and Rubin (6) on missing data and depicted in Figure 2. They are that data are (1) missing completely at random (MCAR) or that those patients included in the analysis are no different than the patients excluded from the analysis; (2) missing at random (MAR), a more flexible assumption that allows missingness to be related to observed features only; and (3) not missing at random (NMAR), the most flexible of the three assumptions, which allows missingness to be related to observed and unobserved features.

Consider the hypothetical study described above, where eGFR is measured over time among transplant patients. An example of the eGFR data being MCAR may be one where the study coordinator neglected to schedule subjects whose last names began with the letter A for a follow-up visit. Because missingness is not related to eGFR itself or other relevant demographics, the data can be considered MCAR. If, however, subjects with missing eGFR tend to be younger patients (*i.e.*, missingness is related to age) and additionally, among younger subjects, missingness does not depend on eGFR values and among older patients, missingness does not depend on eGFR, we could consider the data to be MAR. The data would be considered NMAR if subjects with more favorable eGFR measures are less likely to return to the clinic than those subjects with unfavorable eGFR measures.

We are limited in our ability to distinguish among these conditions. The MCAR assumption can be disproven by checking whether missingness is related to any of the observed variables. Note that MCAR cannot, however, be proven; if missingness is not related to observed variables, it may still be related to unobserved features. Similarly, it is not possible to distinguish between MAR and NMAR conditions without evaluating the missing data itself. Instead, one must rely on *a priori* knowledge of the subject.

Other Approaches to Missing Data

Single Imputation Approaches

Single imputation methods involve filling in the missing values with a single value. These methods include missing indicators, mean imputation, and last value carried forward for longitudinal studies.

Missing indicators involve creating a separate category for those subjects with missing data for categorical variables, creating an indicator for missingness, and imputing a value (say the mean) for continuous variables. Although it seems an intuitive approach to retain all the observations, it has been shown to provide biased estimates and incorrect SEs (7–10), even under the MCAR condition (8).

The last value carried forward method is commonly applied in longitudinal studies (11,12) but known to provide biased estimates (13–16). The approach involves filling in missing values with the last recorded observation. It assumes that, for those patients with missing values, there is no change over time in the relevant variable, a strong assumption unlikely to hold in most settings.

In general, single imputation methods are not recommended. Although these methods are easy to implement, they provide estimates that are biased, even when the data are MCAR, and they yield incorrect SEs (7).

Likelihood-Based, Weighting, and Multiple Imputation–Based Methods

Theoretically sound approaches to handling missing data include maximum likelihood (ML)–based (17) approaches performed on the complete data likelihood (15,18–21), weighting methods (22–26), and multiple imputation (MI)–based approaches (6,22,27–31). ML-based approaches are considered ideal, because they are efficient and provide valid results under the MAR condition. Although software exists for a subset of ML-based approaches, it has not been incorporated into mainstream software for many cases (32–34). Similarly, software for many weighting procedures has not been incorporated into mainstream statistical packages, which poses a barrier to their use. However, MI, which results in estimators with properties similar to ML-based estimators under MAR, is becoming increasingly more accessible. Mainstream packages like SAS, Stata, and R have user-friendly MI procedures. Horton and Kleiman (35) give a comprehensive overview on available MI software. When the data are NMAR, both ML- and MI-based methods are reasonable choices, although their use increases in complexity, because the nature of missingness needs to be explicitly modeled (7). In this paper, we focus our attention on standard MI under MAR because of its accessibility and reasonable flexibility of assumptions.

Multiple Imputation

MI is a simulation-based technique for handling missing data. Figure 3 illustrates the steps required when linear regression is of interest. Briefly, there are three steps: (1) the imputation step, (2) the model fitting step, and (3) the summarization step. In the first step, one samples values from a plausible distribution to fill in the missing observations yielding m full datasets, where $m=5$ is typical. In the second step, the analytic model is fitted to each of the m datasets. Finally, in the third step, the results are combined to summarize the findings, where the overall estimate is the average of individual estimates and the overall SE takes into account the uncertainty of both the sampled values and the imputation process.

Important Considerations When Applying MI

Although MI has desirable statistical properties in theory, it can be challenging to use in practice. Several choices need to be made to implement MI appropriately. For example, there are two main approaches to doing imputation (a joint modeling approach and a fully conditional specification approach). This choice and others affect the results. Below, we highlight aspects key to any imputation-based analysis, which are also discussed in the excellent book by van Buuren (36).

Assessing the Missing Data Mechanism

The first step to any analysis where missing data are present is to consider the missing data mechanism. If MAR seems plausible, one needs to consider on which variables to condition, such that the MAR assumption will hold in practice. Essentially, the MAR assumption implies that information about missingness can be gleaned from the observed data. Under NMAR, MI can still be applied. However, the analysis is more complicated, because the missing data mechanism must be specified explicitly. We refer the reader to several sources that describe analyses under this condition (6,7,27,32–34,37).

Consistency between the Imputation Model and the Analytic Model

The imputation model (*i.e.*, the model specified in the imputation step on which to base imputations) should include all variables in the analytic model, including the outcome (38). The idea is that important inter-relationships among the variables need to be preserved to produce statistically valid results regarding the relationship of interest. For example, suppose it is of interest to describe the association between BMI and progression of CKD after adjusting for eGFR in a model regressing progression on BMI and eGFR. If, for example, CKD progression is not included in the imputation model, bias could be introduced, because BMI and eGFR would have been imputed under the assumption that the correlation between each variable and CKD progression was zero.

Inclusion of Auxiliary Variables in the Imputation Model

The imputation model should additionally include auxiliary variables—variables that may help with the imputation—to ensure that the assumption of MAR holds. Auxiliary variables are those variables that are related to missingness and/or the variable(s) with missing data (39). Which

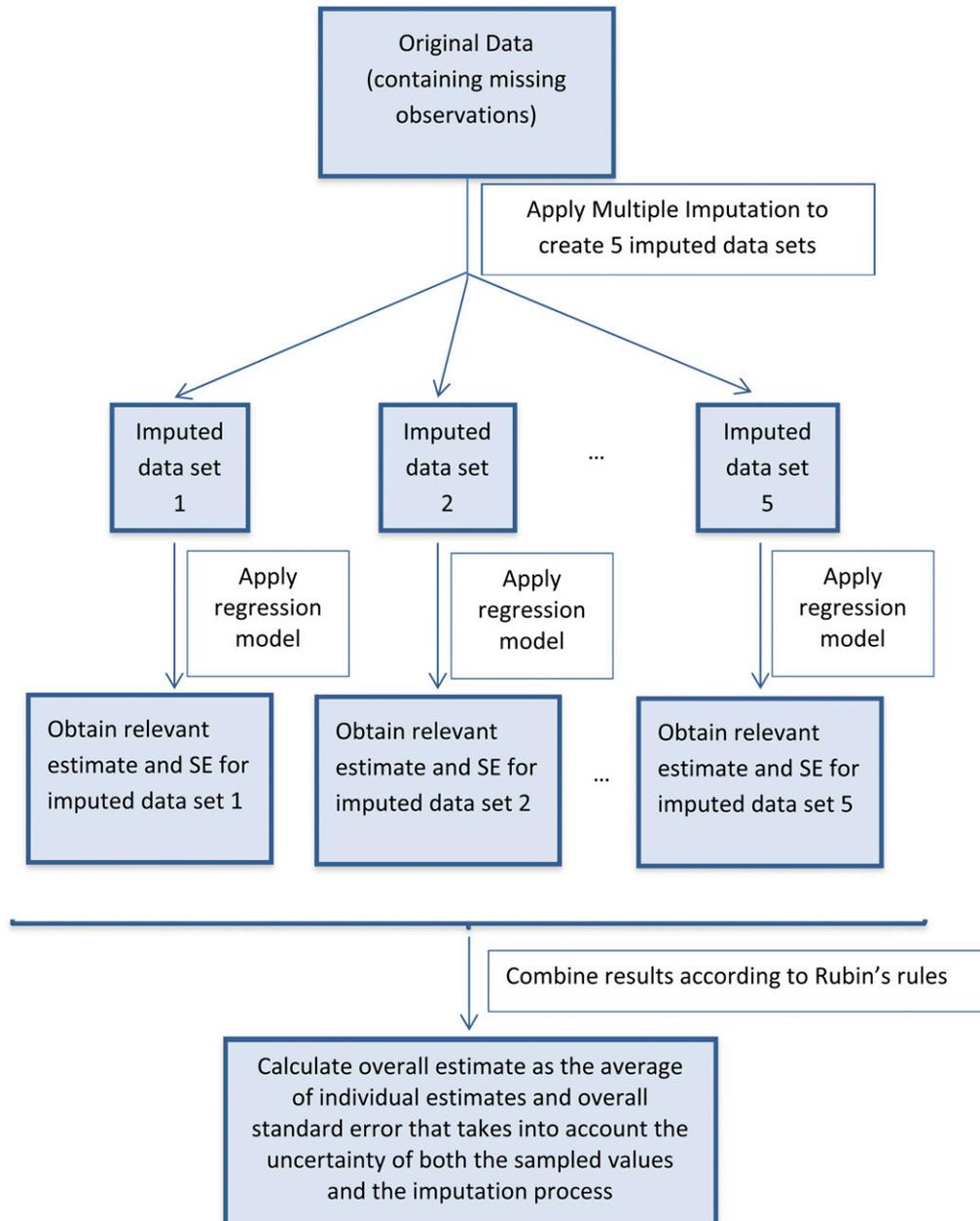


Figure 3. | Illustrative flow chart of implementation of multiple imputation when linear regression is of interest and five imputations are performed.

auxiliary variables to include has been an ongoing topic of study (38,40–42). Collins *et al.* (39) showed that being more inclusive when doubtful of the usefulness of some variables results in decreased bias and increased efficiency. In fact, even when the data are NMAR, the inclusion of auxiliary terms may make the MAR assumption more reasonable, allowing the application of standard MI.

The Number of Imputations to Perform

The general rule of thumb has been that 5–10 imputed datasets are sufficient to obtain stable estimates based on relative efficiency (38). Recent research, however, suggests that the appropriate number of imputations may be larger

(39–41). Both Bodner (43) and White *et al.* (44) suggest a rule based on precision that the number of imputations should roughly equal the percentage of subjects with incomplete data. We suggest using this guideline when computationally feasible as well as sensitivity analyses that vary the number of imputations to guarantee that stability of point estimates and SEs is achieved.

The Role of Sensitivity Analyses and Diagnostics

Robustness of findings to various conditions should be assessed in any MI-based analysis. Assuming MCAR and performing a complete case analysis is a good starting point. If MAR or NMAR is plausible, various MI models

under these missing data mechanisms would be important to explore. Variation under different MAR models (*i.e.*, varying which auxiliary variables to include and/or the functional form of the auxiliary variables in the imputation model) can provide insight. Similar approaches should be taken under NMAR if NMAR is plausible. An appealing property of MI is the ability to account for uncertainty of assumptions, including the missing data mechanism. The variation in findings can provide insight to the reader into how the assumptions impact the findings (if at all).

In addition to sensitivity analyses, there are diagnostic procedures that one can pursue that involve the plausibility of the imputations. A more detailed discussion is presented in the work by van Buuren (36).

To sum, these considerations suggest some of the complexities involved in performing an MI-based analysis and show that such analyses should not be automated. Although MI software programs may set parameters to default values, such as the number of imputations and the imputation method, we recommend that the analyst be proactive and aware of the choices involved.

Illustrative Example

By way of illustration, we use a cohort study of patients available from the USRDS registry. This cohort was also evaluated in a study by Lenihan *et al.* (45), where subjects with at least 1 year of uninterrupted Medicare (parts A and B) coverage before receiving their first kidney transplant between the years of 1997 and 2009 were studied. One objective was to evaluate the association between atrial fibrillation before transplantation and stroke. To that end, a Cox proportional hazards model was used to fit the hazard of stroke as a function of atrial fibrillation after adjusting for potential confounders. Of 62,706 patients, approximately 42% had at least one observation missing, where the proportion of missing data per variable ranged from less than 1% (race) to about 20% (BMI). Table 1 shows a hypothetical example of some of these relevant measurements for seven patients.

Conducting a Complete Case Analysis

Suppose that, when fitting the analytic model, all the variables listed in Table 1 are of interest, with the exception

of BMI at time of listing for the transplant. Because observations 2–5 are missing either BMI at transplant or panel reactive antibody (Table 2), they are excluded from a complete case analysis, whereas observations 1, 6, and 7 remain in the analysis. Logistic regression models of missingness of BMI at transplant on each variable in the analytic model (including the outcome) revealed that missingness was related to several observed features, which suggested that the MCAR assumption did not hold and furthermore, an alternative approach to a complete case analysis should be considered.

Applying MI

Using the MI procedure in SAS 9.2, the imputation model consisted of the variables in the analytic model, including the outcome (both the censoring indicator and the log-transformed length-of-time to stroke) and additional auxiliary variables. Although BMI at transplant was of interest, it was missing for about 20% of the patients. BMI at the time of listing for transplant was available, however, for most patients. Because we expect these variables to be highly correlated, BMI at the time of listing may be an excellent auxiliary variable. Another potential auxiliary variable is the lag time from listing to time of transplant, because BMI at transplant is a function of both BMI at the time of listing and the time lag. Table 3 shows a hypothetical example of the data after five imputations for three patients. Each patient has five rows of data, where each row represents data from one of five imputed datasets. Patient 1 had no missing data, and therefore, the corresponding rows simply contain the observed data. Patient 2 was missing BMI and has an imputed value filled in for each of five imputed datasets. For each imputed dataset, a Cox proportional hazards model was fit. Results were combined (using the MIANALYZE procedure in SAS 9.2) by applying Rubin's rules to summarize findings.

Sensitivity Analyses

We performed sensitivity analyses, which included a complete case analysis, and MI-based analyses under several MAR models. Specifically, we used an imputation model that included only the variables in the analytic model and one model that additionally included all possible

Table 1. Hypothetical example of data with incomplete information

ID	Age (yr)	Sex	BMI at Transplant (kg/m ²)			BMI at Listing (kg/m ²)			PPRA (%)		Stroke	Years Followed
			15–18.5	25–30	30–45	15–18.5	25–30	30–45	11–80	80–100		
1	39	M	0	1	0	0	0	0	0	1	No	8.4
2	44	W	—	—	—	0	0	0	0	1	Yes	10.9
3	29	W	—	—	—	—	—	—	0	0	No	10.0
4	67	M	0	0	0	0	0	0	—	—	No	11.6
5	44	M	—	—	—	0	0	0	0	0	Yes	3.9
6	43	W	0	0	1	0	0	1	1	0	Yes	1.5
7	68	M	0	0	1	—	—	—	1	0	No	4.3

BMI=18.5–25 (normal) and PPRA=0–10 (normal) are used as reference groups and represented by a zero in all dummy variables pertaining to each variable. ID, identification; BMI, body mass index; PPRA, panel reactive antibody; M, man; W, woman.

Table 2. Hypothetical example of data used in a complete case analysis

ID	Age (yr)	Sex	BMI at Transplant (kg/m ²)			PPRA (%)		Stroke	Years Followed
			15–18.5	25–30	30–45	11–80	80–100		
1	39	M	0	1	0	0	1	No	8.4
6	43	W	0	0	1	1	0	Yes	1.5
7	68	M	0	0	1	1	0	No	4.3

BMI=18.5–25 (normal) and PPRA=0–10 (normal) are used as reference groups and represented by a zero in all dummy variables pertaining to each variable. ID, identification; PPRA, panel reactive antibody; M, man; W, woman.

Table 3. Hypothetical example of data with five imputed datasets

Imputed Dataset	ID	Age (yr)	Woman	BMI (kg/m ²)			PPRA (%)		Stroke	Years Followed
				15–18.5	25–30	30–45	11–80	80–100		
1	1	39	No	0	1	0	0	1	No	8.4
2	1	39	No	0	1	0	0	1	No	8.4
3	1	39	No	0	1	0	0	1	No	8.4
4	1	39	No	0	1	0	0	1	No	8.4
5	1	39	No	0	1	0	0	1	No	8.4
1	2	44	Yes	0.35 ^a	0.34 ^a	−0.21 ^a	0	1	Yes	10.9
2	2	44	Yes	0.12 ^a	0.45 ^a	0.03 ^a	0	1	Yes	10.9
3	2	44	Yes	0.21 ^a	0.27 ^a	−0.47 ^a	0	1	Yes	10.9
4	2	44	Yes	−0.01 ^a	0.97 ^a	−0.44 ^a	0	1	Yes	10.9
5	2	44	Yes	0.38 ^a	0.80 ^a	0.64 ^a	0	1	Yes	10.9
1	4	67	No	0	0	0	−0.21 ^a	0.08 ^a	No	11.6
2	4	67	No	0	0	0	0.04 ^a	−0.33 ^a	No	11.6
3	4	67	No	0	0	0	0.25 ^a	0.21 ^a	No	11.6
4	4	67	No	0	0	0	0.31 ^a	−0.04 ^a	No	11.6
5	4	67	No	0	0	0	0.69 ^a	0.07 ^a	No	11.6

BMI=18.5–25 (normal) and PPRA=0–10 (normal) are used as reference groups and represented by a zero in all dummy variables pertaining to each variable. ID, identification; PPRA, panel reactive antibody.
^aMultiple imputation may lead to data that are not consistent with the original format; in this case, values imputed for missing observations of categorical (binary) data are continuous. Furthermore, although original categories of a variable may be mutually exclusive, imputed data may not be mutually exclusive, which is appropriate, because the imputed values, *per se*, do not have any meaning.

auxiliary variables. We also varied the number of imputations (*m*) to be 8 or 42.

Results

Table 4 shows that the results across the MI-based analyses were comparable, yielding hazard ratios that suggested that atrial fibrillation before transplantation was significantly associated with increased risk of stroke (by approximately 40%). This estimated increase in risk did not vary with the number of imputations or across MAR models. A complete case analysis, however, provided an estimated risk (modestly attenuated at 1.26) with 95% confidence interval that contained one, suggesting a nonsignificant association and thus, changing the inference. Our final model is model 1, because it is the most inclusive with respect to auxiliary variables. Alternatively, we could have applied Rubin’s rules and averaged results across models 1, 4, and 5 to reflect our uncertainty among MAR models.

Discussion

The most common approach to handling missing outcome or covariate data is to perform a complete case analysis. Many analysts do not even realize that they are excluding subjects with partial information from their analysis, because this exclusion is the default approach for software packages. Thus, the analyst may input a dataset of eligible subjects only to have the analysis performed on the subset with complete information. Validity of the results from such an approach relies on the data being MCAR, an assumption that rarely holds in practice. Consequently, analyses may result in biased and inefficient estimates of relationships of interest. For this reason, investigators must be aware of alternative methods that rely on more realistic assumptions regarding missingness of data. Ideally, ML approaches, which are described in the work by Little and Rubin (6), should be used to address missing data issues, because they have desirable statistical properties under the more flexible

Table 4. Results for death-censored ischemic stroke

Model	Analytic Approach	Imputation Model ^a	Hazard Ratio	Lower 95% Confidence Interval	Upper 95% Confidence Interval	P Value
1	MI $m=8^b$ (sample size $n=62,692$)	Variables in analytic model ^c plus auxiliary variables ^d	1.36	1.10	1.68	<0.01
2	MI $m=42^b$ (sample size $n=62,692$)	Variables in analytic model ^c plus auxiliary variables ^d	1.36	1.10	1.68	<0.01
3	Complete case ^e (sample size $n=36,107$)	—	1.26	0.95	1.70	0.10
4	MI $m=8^b$ (sample size $n=62,692$)	Variables in analytic model ^c plus no auxiliary variables ^d	1.36	1.10	1.68	<0.01
5	MI $m=8^b$ (sample size $n=62,692$)	Variables in analytic model ^c plus BMI at listing and time since BMI was measured	1.36	1.10	1.69	<0.01

MI, multiple imputation; m , number of imputations.

^aImputation model also included censoring indicator and time to event.

^bCox proportional hazards model with multiple imputation to handle missing data (m imputed datasets).

^cAnalytic model included age, sex, race, BMI at transplant, cause of ESRD, dialysis vintage, dialysis modality, skilled nursing facility use indicator, number of hospital days, number of non-nephrology clinic visits, previous solid organ transplant, comorbidities, patient blood type, panel reactive antibody, donor age, donor sex, transplant type, number of HLA mismatches, and cold ischemia time.

^dAuxiliary variables were BMI at listing, time since BMI was measured, weight at listing, and height at listing.

^eComplete case analysis: Cox proportional hazards model on the dataset where observations with at least one covariate missing were removed.

assumption that data are MAR. Unfortunately, a barrier to using ML-based methods is the lack of accessible software. MI-based analyses, like ML-based methods, also produce statistically valid results under the same flexible assumption of MAR, although they are generally less efficient than ML-based estimates. An advantage of using MI over ML, however, is that there is accessible software that is becoming increasingly user-friendly. A barrier to using MI is the number of choices that the analyst must make. In this paper, we have made the analyst aware of a number of issues that we believe are the most critical to performing an MI-based analysis appropriately. Perhaps the most important of these decisions involves the missing data mechanism. More specifically, the analyst needs to rely on *a priori* knowledge to decide whether MAR is plausible theoretically and if so, which variables should be included so that MAR can reasonably hold in practice.

Although we have touched on the basic issues involved in an MI analysis, there are other complexities posed by MI that are beyond the scope of this introductory paper, but we encourage the interested reader to explore them. These issues include how to do imputations in the presence of derived variables, such as interaction and higher-order terms, and how to model the missing data mechanism under the NMAR condition. For these topics and more, we refer the reader to the recent book by van Buuren (36) as well as the book by Little and Rubin (6) on missing data. Other topics relate to the specific type of analysis being performed. For example, White and Royston (28) discuss issues that arise when fitting a Cox proportional hazards model in the presence of missing covariates. Schafer (46) discusses specific challenges that arise when imputing in the presence of clustering, which occurs in longitudinal studies.

Although use of MI has been increasing since its introduction in 1987 (27), complete case analysis is still the most common approach to handling missing data. Its continued use is partly because of the complexities of implementing MI. Many investigators, however, express discomfort at the idea of “making up” data to fill in missing values. Indeed, some of these values might not even be coherent with the original data format. We remind the reader, however, that these filled-in values are not to be viewed as true values. The idea behind MI is to use the imputed values as a device to estimate quantities of interest by retaining observations in a way that preserves the interrelationships among key variables. Thus, the imputed values themselves do not matter. Moreover, desirable properties of MI are that it accounts for the uncertainty of the imputation process and furthermore, that it can also account for the uncertainty of various assumptions on which the imputation process is based.

Conclusion

How data are analyzed in the presence of missing data can greatly affect findings. The most common approach used to handling missing data—complete case analysis—relies on unrealistic assumptions, yielding potentially biased estimates. We encourage analysts to use missing data methods that rely on more flexible and realistic assumptions. Specifically, we propose the use of MI methods. Although sound in theory, however, MI can present challenges in practice. We provided basic guidelines for addressing these issues and illustrated the use of MI on a dataset used for kidney research.

Acknowledgments

M.E.M.-R. and M.D. were supported by Grant ME-1303-5989 from the Patient Centered Outcomes Research Institute entitled “The

Handling of Missing Data Induced by Time-Varying Covariates in Comparative Effectiveness Research Involving HIV Patients.”

Disclosures

In the past year, W.C.W. has served as a scientific advisor to Amgen, Keryx, Medgenics, and Medtronic. M.E.M.-R. and M.D. have no disclosures to report.

References

- Fleming TR: Addressing missing data in clinical trials. *Ann Intern Med* 154: 113–117, 2011
- US Renal Data System: *USRDS 2012 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States*, Bethesda, MD, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2012
- Scientific Registry for Transplant Recipients: Available at: <http://www.srtr.org>. Accessed December 17, 2013
- Greenland S, Finkle WD: A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 142: 1255–1264, 1995
- Klebanoff MA, Cole SR: Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol* 168: 355–357, 2008
- Little RJA, Rubin DB: *Statistical Analysis with Missing Data*, New York, Wiley, 1987
- Allison PD: *Missing Data*, Thousand Oaks, CA, Sage, 2002
- Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, Moons KGM, Geerlings MI: Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *J Clin Epidemiol* 63: 728–736, 2010
- Jones MP: Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc* 91: 222–230, 1996
- van der Heijden GJMG, Donders ART, Stijnen T, Moons KGM: Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *J Clin Epidemiol* 59: 1102–1109, 2006
- Fitzmaurice GM, Laird NM, Ware JH: *Applied Longitudinal Analysis*, New York, Wiley, 2004
- Matilde Sanchez M, Chen X: Choosing the analysis population in non-inferiority studies: Per protocol or intent-to-treat. *Stat Med* 25: 1169–1181, 2006
- Fitzmaurice GM, Laird NM, Ware JH: *Applied Longitudinal Analysis*, Hoboken, NJ, Wiley, 2011
- Revicki DA, Gold K, Buckman D, Chan K, Kallich JD, Woolley JM: Imputing physical health status scores missing owing to mortality: Results of a simulation comparing multiple techniques. *Med Care* 39: 61–71, 2001
- Prentice RL: Covariate measurement errors and parameter estimation in a failure time regression-model. *Biometrika* 69: 331–342, 1982
- Tsiatis AA, Davidian M: A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* 88: 447–458, 2001
- Ibrahim JG, Chen MH, Lipsitz SR, Herring AH: Missing-data methods for generalized linear models: A comparative review. *J Am Stat Assoc* 100: 332–346, 2005
- Chen HY, Little RJA: Proportional hazards regression with missing covariates. *J Am Stat Assoc* 94: 896–908, 1999
- Lin DY, Ying Z: Cox regression with incomplete covariate measurements. *J Am Stat Assoc* 88: 1341–1349, 1993
- Paik MC, Tsai WY: On using the Cox proportional hazards model with missing covariates. *Biometrika* 84: 579–593, 1997
- Zhou HB, Pepe MS: Auxiliary covariate data in failure time regression. *Biometrika* 82: 139–149, 1995
- Qi LH, Wang YF, He YL: A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Stat Med* 29: 2592–2604, 2010
- Robins JM, Rotnitzky A, Zhao LP: Estimation of regression-coefficients when some regressors are not always observed. *J Am Stat Assoc* 89: 846–866, 1994
- Robins JM, Rotnitzky A, Zhao LP: Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 90: 106–121, 1995
- Wang CY, Chen HY: Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics* 57: 414–419, 2001
- Xu Q, Paik MC, Luo XD, Tsai WY: Reweighting estimators for cox regression with missing covariates. *J Am Stat Assoc* 104: 1155–1167, 2009
- Rubin DB: *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley, 1987
- White IR, Royston P: Imputing missing covariate values for the Cox model. *Stat Med* 28: 1982–1998, 2009
- Xie F, Paik MC: Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics* 53: 1538–1546, 1997
- Barzi F, Woodward M: Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol* 160: 34–45, 2004
- van Buuren S, Boshuizen HC, Knook DL: Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 18: 681–694, 1999
- Ibrahim JG, Chen MH, Lipsitz SR: Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 88: 551–564, 2001
- Ibrahim JG, Lipsitz SR: Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics* 52: 1071–1078, 1996
- Ibrahim JG, Lipsitz SR: Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J R Stat Soc Series B Stat Methodol* 61: 173–190, 1999
- Horton NJ, Kleinman KP: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 61: 79–90, 2007
- van Buuren S: *Flexible Imputation of Missing Data*, Boca Raton, FL, CRC Press, 2012
- van Buuren S: Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 16: 219–242, 2007
- Rubin DB: Multiple imputation after 18+ years. *J Am Stat Assoc* 91: 473–489, 1996
- Collins LM, Schafer JL, Kam CM: A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 6: 330–351, 2001
- Ibrahim JG, Lipsitz SR, Horton N: Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *J R Stat Soc Ser C Appl Stat* 50: 361–373, 2001
- Graham JW: Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models, Technical Report No. 01-48, University Park, PA, The Pennsylvania State University, The Methodology Center, 2001
- Meng XL: Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 9: 538–558, 1994
- Bodner T: What improves with increased missing data imputations? *Struct Equ Modeling* 15: 651–675, 2008
- White IR, Royston P, Wood AM: Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 30: 377–399, 2011
- Lenihan CR, Montez-Rath ME, Scandling JD, Turakhia MP, Winkelmayer WC: Outcomes after kidney transplantation of patients previously diagnosed with atrial fibrillation. *Am J Transplant* 13: 1566–1575, 2013
- Schafer JL: Imputation of Missing Covariates under a Multivariate Linear Mixed Model, Technical Report No. 97-04, University Park, PA, Department of Statistics, The Pennsylvania State University, 1997

Published online ahead of print. Publication date available at www.cjasn.org.