Special Communication

Missing Data and Multiple Imputation

Peter Cummings, MD, MPH

Missing data can result in biased estimates of the association between an exposure *X* and an outcome *Y*. Even in the absence of bias, missing data can hurt precision, resulting in wider confidence intervals. Analysts should examine the missing data pattern and try to determine the causes of the missingness. Modern software has simplified multiple imputation of missing data and the analysis of multiply imputed data to the point where this method should be part of any analyst's toolkit. Multiple imputation will often, but not always, reduce bias and increase precision compared with complete-case analysis. Some exceptions to this rule are noted in this review. When describing study results, authors should disclose the amount of missing data and other details. Investigators should consider how to minimize missing data when planning a study.

JAMA Pediatr. 2013;167(7):656-661. doi:10.1001/jamapediatrics.2013.1329
Published online May 13, 2013.

Author Affiliations: Department of Epidemiology and Harborview Injury Prevention and Research Center, University of Washington, Seattle (Cummings).

Corresponding Author: Peter Cummings, MD, MPH, Department of Epidemiology, University of Washington, 250 Grandview Dr, Bishop, CA 93514 (peterc@uw.edu).

Suppose we wish to estimate the association between an exposure *X* (a treatment or risk factor) and an outcome *Y* (death or systolic blood pressure), accounting for a confounding factor *Z* (age or smoking history). For example, we may want to estimate an adjusted risk ratio, rate ratio, or odds ratio; an adjusted risk difference or rate difference; or an adjusted difference in means. Common analysis methods simply omit records with any missing values for *X*, *Y*, or *Z*, an approach often called complete-case analysis. Estimated associations based on complete records only may be biased.[1,2] Even if the estimated association is unbiased, omitting records can reduce precision; the confidence interval (CI) may be wider compared with an interval based on all records.[2-4]

To illustrate these problems and the role of multiple imputation, I will use hypothetical data for drivers in 2000 consecutive motor vehicle crashes (**Table 1**). In slow crashes 50% were belted, and in fast crashes 25% were belted. The risk of death for unbelted drivers was 10% in slow crashes and 40% in fast crashes. Assuming that speed was the only confounder, the risk ratio for death of belted compared with unbelted drivers, adjusted for speed, was 0.500 (95% CI, 0.380-0.658).

Data Missing at Random

Imagine that the true data, unknown to us, are in Table 1, but the available data are in **Table 2**. Cross-tabulations of Table 2 data can be used to study several questions: (1) How many records have some missing data? (2) Are there systematic relationships between known values of some variables and missingness of other variables? (3) Is missingness of one variable related to missingness of another? In addition, we can create a variable that indicates for each subject whether there is missingness of *any* variable needed for the analysis (exposure, outcome, or confounders) and use tabulations to see how the known values of each variable is related to this *any missing* variable.[5] For longitudinal data, missingness over time can be assessed.[6,7] For continuous data, missingness can be studied by using means, graphs, and cross-tabulations of suitable categories.

In Table 2, speed is missing for 44% of records and seat belt use for 17%, and a complete-case estimate of the speed-adjusted association of belt use with death will omit 52% of the records (**Table 3**). No data are missing for those who died. Among survivors, crash speed is missing for 50% regardless of whether seat belts were used or not used or use is missing, and belt use is missing for 20% regardless of whether speed was slow, fast, or missing (**Table 4**).

Because data are missing only for some survivors, the risk of death is biased upward in the 956 records with complete data compared with all 2000 records. This upward bias differs across levels of speed and belt use (Table 1 and Table 2). Consequently, the speed-adjusted risk ratio for death based on complete cases only, comparing belted with unbelted drivers, is biased: 0.571 (95% CI, 0.442-0.737) instead of 0.500. Because only Table 2 data are available, an analyst would not know that 0.571 is a biased estimate.

Data are said to be *missing at random* (MAR) if the probability that a value is missing may depend on observed values in the data but not additionally on the missing value itself. Despite its name, an MAR process need not be random. For example, Table 2 data were fabricated from Table 1 without any randomness. In Table 2, the probability that speed or belt use is missing depends only on known outcome values; neither is missing among the dead, and among survivors speed is missing for 50% and belt use for 20%. Among survivors in Table 2, the missingness of speed and belt use is unrelated to the missing values themselves; therefore these values are MAR.[4] The MAR designation implies that, within subgroups formed by known data values, missing values of a variable are not systematically different from known values.[3,8]

There are limitations to the cross-tabulation approach I have described for studying missing data. We might suspect from Table 3 and Table 4 that these data *could be* MAR. However, real

Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017

Table 1. Hypothetical Data for a Cohort Study of Drivers in 2000 Motor Vehicle Crashes With No Missing Data

Crash Speed	Seat Belt Use	Driver Death	No. of Drivers	Risk of Death	Risk Ratio for Death in Belted vs Unbelted Drivers
Slow	Yes	Yes	40	.05	.05/.10 = 0.5
Slow	Yes	No	760		
Slow	No	Yes	80	.10	
Slow	No	No	720		
Fast	Yes	Yes	20	.20	.20/.40 = 0.5
Fast	Yes	No	80		
Fast	No	Yes	120	.40	
Fast	No	No	180		

Table 2. Hypothetical Data for the 2000 Drivers in Table 1 With Data Missing at Random[a]

Crash Speed	Seat Belt Use	Driver Death	No. of Drivers	Risk of Death	Risk Ratio for Death for Belted vs Unbelted Drivers
Slow	Yes	Yes	40	.116	.116/.217 = 0.535
Slow	Yes	No	304		
Slow	No	Yes	80	.217	
Slow	No	No	288		
Fast	Yes	Yes	20	.385	.385/.625 = 0.615
Fast	Yes	No	32		
Fast	No	Yes	120	.625	
Fast	No	No	72		
Slow	Missing	No	148		
Fast	Missing	No	26		
Missing	Yes	No	336		
Missing	No	No	360		
Missing	Missing	No	174		

[a] Among survivors, data on speed are missing at random for 50% of the records, and data on seat belt use are missing at random for 20%.

Table 3. Missing Data Patterns for Data Missing at Random in Table 2[a]

		No. of Driver Records, No. (%)	
Crash Speed	Seat Belt Use	All 2000	All 1740 Survivors
Known	Known	956 (48)	696 (40)
Missing	Known	696 (35)	696 (40)
Known	Missing	174 (9)	174 (10)
Missing	Missing	174 (9)	174 (10)

[a] The outcome, death or survival, was always known.

Table 4. Cross-tabulation of Speed Data With Seat Belt Use Data for the 1740 Surviving Drivers in the Missing at Random Data of Table 2[a]

	Seat Belt Use		
Crash Speed	Yes	No	Data Missing
Slow			
Count	304	288	148
Row %	41%	39%	20%
Column %	45%	40%	43%
Fast			
Count.	32	72	26
Row %	25%	55%	20%
Column %	5%	10%	7%
Missing			
Count	336	360	174
Row %	39%	41%	20%
Column %	50%	50%	50%

[a] The row percentages are the same (20%) in the last column, and the column percentages are the same (50%) in the last row.

data can have many variables, making it difficult to understand tabular summaries. Sampling variation in real data can make it hard to discern a possible MAR pattern. By comparing Tables 1 and 2, we can see that missing-data mechanism is MAR, but for real data the missing values are unknown, so we cannot be sure whether they are MAR or whether the complete-case risk ratio is biased. Belief that data are MAR may be best supported by knowledge about the missing data mechanism; for example, an assumption of MAR may be justified if data are missing because of the sampling design.

Multiple Imputation to Reduce Bias in MAR Data

If data are MAR, complete-case analysis may be biased, though not necessarily, as illustrated by data in Table 2.[1,9] Multiple imputation can reduce or correct this bias and is justified by statistical theory and simulation studies.[5,8,10-12] Given exposure *X*, outcome *Y*, and confounder *Z*, multiple imputation proceeds in 3 steps. First, the distribution of any missing values is estimated for each combination of known *X*, *Y*, and *Z* values. The estimation step should include variables that might be used in later analyses (exposures, outcomes, possible confounders, interactions) and any other variables possibly related to the missing values.[3,10,13-15] Second, a value is randomly chosen

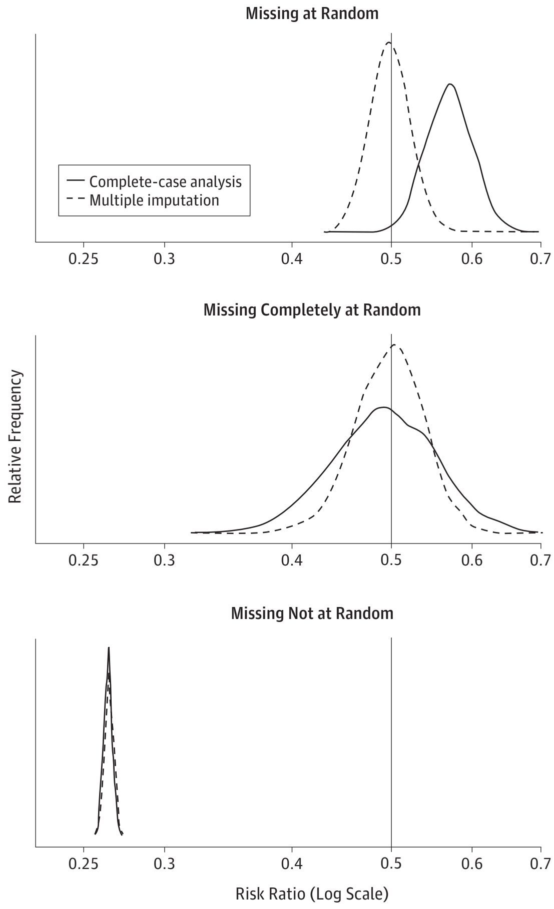
Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017

from the estimated distributions to replace each missing value in each record. These replacement (imputed) values can vary from record to record, even between records with the same known values of *X*, *Y*, or *Z*. Third, this process is repeated several times, resulting in multiple sets of data, each the same regarding known values but different regarding imputed values.

The *X*-*Y* association, adjusted for *Z*, is then estimated in each set of data using conventional statistical methods and the average of these is the pooled estimate. For ratio measures, such as the risk ratio, we average the logarithms of the ratios and then exponentiate. The pooled variance is the average of the variances for each estimate plus an additional quantity for the variance between the estimates.[8,10,12,16] This combined variance accounts for uncertainty about imputed values. Some software packages have commands that simplify description and estimation across the multiple data sets.

To illustrate the benefits of multiple imputation, I started with Table 1 data and created 2000 data sets with speed data MAR for 50% and seat belt use data MAR for 20% of surviving drivers. The adjusted risk ratio for seat belt use was estimated in each data set by using complete-case analysis. Multiple imputation was used to create 25 sets of imputed values for each data set and to estimate 2000 pooled risk ratios using known and imputed values. The average adjusted risk ratio for death comparing belt users with nonusers was 0.573 from complete-case analysis of the 2000 data sets and 0.501 after multiple imputation, close to the true risk ratio of 0.500 (**Figure 1**).

Figure 1. Distribution of Risk Ratio Estimates From Complete-Case Analyses (Solid Line) and After Multiple Imputation (Dashed Line) in 2000 Simulations Using Data From Table 1


Missing at Random

Missing Completely at Random

Missing Not at Random

Data sets were created with data missing at random for 50% of speed data and 20% of seat belt use data among drivers who survived, data missing completely at random for 25% of speed data and 20% of seat belt data among all drivers, and data missing not at random for both seat belt use and death among 30 drivers who used seat belts and died. The true adjusted risk ratio for death among belted vs unbelted drivers was 0.500, shown by a vertical line. The distributions of the risk ratio estimates were smoothed by using a kernel density method.

Other Methods

There are other methods for dealing with missing data that are supported by statistical theory, such as maximum likelihood methods and weighting,[2,4,12,17-19] but multiple imputation is often easier to invoke and more widely applicable. Using a missing data indicator can produce biased estimates.[1,2,20] Single imputation of missing data may result in biased estimates and CIs that are too narrow because it treats imputed values as if they were known.[2,11,16]

Describing the Missing Data

Because missing data are a potential source of bias, an article reporting on a study should include details about missing data. For a randomized trial or cohort study, it is customary to describe data in a table with column headings for exposure categories and rows for other variables (including the outcome) with cell counts and column percentages. This format (**Table 5**) can be used to describe the data in Table 2; column percentages are used only for known data, allowing the reader to easily compare belted and unbelted drivers regarding known values.

If multiple imputation is used, Table 5 fails to describe the data analyzed after imputation. I applied multiple imputation to Table 2 data, creating 100 sets of imputed values with an average adjusted seat belt risk ratio of 0.501. After imputation, the descriptive table could show the rounded mean cell counts and average column percentages based on the known and imputed data (**Table 6**), as well as information about missing data. My preference is for Table 6 if multiple imputation is used; perhaps Table 5 could be published as an online supplement in this scenario.

In addition to Table 5 or 6, the article could describe the missingness as follows: "Among 2000 drivers eligible for analysis, speed values were missing for 870 (44%) and belt use information for 348 (17%); 1044 (52%) of the records had some missing data. Only survivors had missing data." The type of imputation (hot-deck imputation,[21] multivariate normal,[3] chained equations[22]), the variables used for imputation, and the number of imputed data sets can be described in 2 or 3 sentences. In the discussion section of an article, I like to acknowledge that missing data are a potential source of bias. If some data are missing, bias is usually possible and any method of analysis will rest on assumptions that cannot be tested; there is no free lunch. Authors can write something like, "We used multiple imputation to reduce any bias due to missing data, but our estimate of the *X*-*Y* association could still be biased if missingness depended not only on the variables we used to impute missing values but also on the missing values themselves."

Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017

Table 5. Descriptive Table for a Hypothetical Cohort Study of Seat Belt Use and Driver Death Using Data From Table 2[a]

	Seat Belt Use		
Characteristic	Yes (n = 732)	No (n = 920)	Data Missing (n = 348)
Speed			
Slow, No. (%)	344 (87)	368 (66)	148
Fast, No. (%)	52 (13)	192 (34)	26
Missing, No.	336	360	174
Driver death			
Yes, No. (%)	60 (8)	200 (22)	0
No, No. (%)	672 (92)	720 (78)	348
Missing, No.	0	0	0

[a] Data for 2000 consecutive crashes, with speed data missing at random for 50% and seat belt use missing at random for 20% of drivers who survived. Cell counts and column percentages are shown for known values. For missing data only cell counts are shown.

Table 6. Descriptive Table for a Hypothetical Cohort Study of Seat Belt Use and Driver Death Using Data From Table 2 After Multiple Imputation[a]

	Seat Belt Use		
Characteristic	Yes (n = 900)	No (n = 1100)	Data Missing (n = 348)
Crash speed			
Slow, No. (%)	800 (89)	799 (73)	148
Fast, No. (%)	100 (11)	301 (27)	26
Missing, No.	336	360	174
Driver death			
Yes, No. (%)	60 (7)	200 (18)	0
No, No. (%)	840 (93)	900 (82)	348
Missing, No.	0	0	0

[a] Rounded average cell counts and column percentages are shown based on results from 100 imputations for the missing values. For missing values, the actual cell counts are shown without percentages.

Data Missing Completely at Random

Sometimes missing information is unrelated to any data values, known or unknown; this is called *missing completely at random* (MCAR), a special instance of MAR. Complete-case analysis will not be biased in MCAR data because, aside from sampling variability, records with known values are not systematically different from all records. The hypothesis that data are MCAR can be tested.[12] In Table 2, known values of seat belt use are related to missingness, violating the MCAR assumption; 46% of records with belt use coded yes had missing data compared with 39% with belt use coded no (*P* = .006 for the difference in proportions). As with any hypothesis test, large *P* values cannot prove that missingness is truly MCAR. These *P* values will tend to be large if the study sample or the proportion of missing data is small.[23,24]

The data in **Table 7** are MCAR. Sixty percent of the records have complete data, and the complete-case adjusted risk ratio for seat belt use is 0.500 (95% CI, 0.351-0.713). This CI is 30% wider than that based on the full data in Table 1, 0.380-0.658. In 2000 simulations of this MCAR mechanism (with data on speed missing for a random 25% of records and data on belt use missing for 20%), the complete-case seat belt risk ratio averaged 0.501, and the multiply imputed risk ratio averaged 0.502; neither method showed appreciable bias. However, the complete-case risk ratios had a wider distribution and a wider average 95% CI (0.351-0.715) than did risk ratios from multiple imputation (average 95% CI, 0.367-0.689) (Figure 1 and **Figure 2**). Use of multiple imputation when data are MCAR will usually, but not always, produce a CI that is the same as or smaller than the CI from complete-case analysis.[5,9]

Data Missing Not at Random

In Table 1, 60 belted drivers died. Among these 60, let us change both belt use and survival data to missing for 20 drivers who crashed at slow speeds and 10 who crashed at fast speeds. Then only 30 of 2000 records (1.5%) will have missing data, but a complete-case analysis will estimate a biased seat belt risk ratio of 0.264 (95% CI, 0.181-0.385). This missingness depends on missing (and therefore unknown) values, a mechanism called *missing not at random* (MNAR). This term is not terribly clear because the bias is not due to a lack of randomness but arises because the missingness depends on the unknown missing values.

I simulated 2000 sets of data by starting with Table 1 and changing data on both belt use and death to missing randomly for 30 of the 60 drivers who were belted and died. The average complete-case seat belt risk ratio was 0.264, and after multiple imputation it was 0.265. The distributions of the risk ratios from these methods were alike (Figure 1) and their CIs hardly differed. Multiple imputation was useless in this example, but not harmful, because the known data lacked information about the missing values, resulting in clueless imputations. In this extreme example, a small amount of missing data produced substantial bias, and multiple imputation was of no help.

Efficiency of Multiple Imputation vs Complete-Case Analysis

Multiple imputation will not always narrow CIs compared with complete-case analysis.[5] In the MCAR data in Table 7, speed data were missing independently of data on belt use. Imagine instead that both speed and belt use were missing jointly for half the records in each row of Table 1. These simultaneously missing data would be MCAR, so risk ratio estimates would be unbiased with either method. However, no record with missing speed would have known information about belt use, and vice versa. Consequently, the imputation process is uninformed and the CI from multiple imputation will be the same as the interval from complete-case analysis. The degree to which multiple imputation will narrow the CI increases as the amount of missing data for *Z* increases when *X* is known; White and Carlin provide useful details.[5]

Bias of Multiple Imputation vs Complete-Case Analysis

If missingness is related to values of exposure *X* or confounder *Z* but not to values of outcome *Y*, then complete-case estimates of the *X*-*Y* association will not be biased provided that the estimating method includes the *X* and *Z* variables (**Table 8**)[4,5,9,25] (in addition, Robert J. Glynn, ScD, PhD, and Nan M. Laird, PhD, unpublished manuscript, 1985) Imagine that a true linear association exists between body weight and a health outcome and that this is estimated by using regression. Even if the heavier subjects refuse to divulge their weight,

Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017

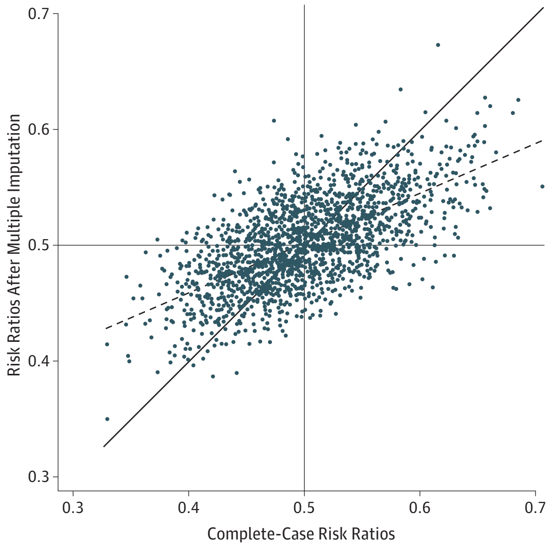
Table 7. Hypothetical Data for Drivers in 2000 Motor Vehicle Crashes[a]

Crash Speed	Seat Belt Use	Driver Death	No. of Drivers	Risk of Death	Risk Ratio for Death for Belted vs Unbelted Drivers
Slow	Yes	Yes	24	.050	.050/.100 = 0.5
Slow	Yes	No	456		
Slow	No	Yes	48	.100	
Slow	No	No	432		
Fast	Yes	Yes	12	.200	.200/.400 = 0.5
Fast	Yes	No	48		
Fast	No	Yes	72	.400	
Fast	No	No	108		
Slow	Missing	Yes	18	.075	?[b]
Slow	Missing	No	222		
Fast	Missing	Yes	21	.350	
Fast	Missing	No	39		
Missing	Yes	Yes	12	.067	.067/.182 = 0.37
Missing	Yes	No	168		
Missing	No	Yes	40	.182	
Missing	No	No	180		
Missing	Missing	Yes	13	.130	?[b]
Missing	Missing	No	87		

[a] Data about speed are missing completely at random (MCAR) for 25% of the records, and data about seat belt use are MCAR for 20%.

[b] Risk ratios comparing belted with unbelted drivers cannot be estimated because of missing data concerning seat belt use.

Figure 2. Scatterplot From 2000 Simulations of Table 1 Data



Data for speed were missing completely at random (MCAR) in 25% of the records, and data about seat belt use were MCAR for 20%. Vertical and horizontal lines indicate the true risk ratio of 0.500. The solid diagonal line indicates identical values for both risk ratios, and the dashed diagonal line is from linear regression, with complete-case risk ratios as the explanatory variable and multiple-imputation risk ratios as the outcome. This regression line and the 2000 plotted points both show that the risk ratios produced by multiple imputation are usually closer to 0.500 than those produced by complete-case analysis.

an MNAR mechanism, the linear association between weight and health could still be estimated correctly from the known data. Although estimated associations from regression will be unbiased, the estimated mean of *X* or *Z* usually will be biased if missingness is related to their true values.

If missingness depends only on values of *Y*, complete-case analysis will be biased except for associations estimated by the odds ratio[4,5,25] (as well as Robert J. Glynn, ScD, PhD, and Nan M. Laird, PhD, unpublished manuscript, 1985). Potential bias in complete-case analysis depends not on whether missing data are MAR or MNAR but on the relationship between the variable that is related to the missingness and the variable with missing values (Table 8). Multiple imputation is biased if the missing data are not MAR except when the odds ratio is used to estimate the *X-Y* association.[5,25]

It will often be unclear whether the probability of missingness is related to *X*, *Z*, *Y*, or some combination. One study[5] of simulated data offers evidence that bias after multiple imputation is usually less, compared with complete-case analysis, when missingness is related to values of both *X* and *Y*.

Summary

Because missing data may bias estimated associations, analysts should examine the missing data pattern and consider the possible causes of the missingness. Authors should disclose the amount of missing data and useful details about the missingness.

Any analysis of missing data will usually require assumptions that cannot be verified. When the proportion of missing data is less than 10%, multiple imputation may offer little advantage.[26] I used to opt for complete-case analysis in that situation, but new software has made multiple imputation so easy that I now consider it the default method and view complete-case analysis as requiring special justification. In many—perhaps most?—situations, multiple imputation should reduce bias, increase study power, or at least do no harm.[1-3,5,12,14,27] It should be part of the

Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017

Table 8. Some Missing Data Mechanisms in Relation to Bias in Estimates of Association[a] From Complete-Case Analysis or Multiple Imputation[b]

Variables: Exposure *X*, Confounder *Z*, and Outcome *Y*			Analysis Method[c]	
Variables With Missing Data	Missingness Related to Values of This Variable	Missing Data Mechanism	Complete Case	Multiple Imputation
Any combination	Nothing	MCAR	Unbiased	Unbiased
X	*Y*	MAR	Biased[d]	Unbiased
Y[e]	*Y*	MNAR	Biased[d,f]	Biased[d,f]
X	*X*	MNAR	Unbiased	Biased[d]
Y	*X*	MAR	Unbiased	Unbiased
Z	*X*	MAR	Unbiased	Unbiased
X	*X* and *Y*	MNAR	Biased[g]	Biased

Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

[a] Estimates of association include risk ratios, rate ratios, odds ratios, risk differences, rate differences, and differences in means.

[b] This table is based chiefly on results from White and Carlin[5] and Westreich.[25]

[c] The analytic method (regression model, etc.) includes *X*, *Z*, and *Y*, using suitable transformations of *X* and *Z*, and the effect of *X* on *Y* is not modified by *Z*.

[d] Odds ratios are unbiased.

[e] Missing outcomes related to the value of *Y* are common in case-control studies, in which the fraction of missing cases is usually much smaller than the fraction of missing controls.

[f] All association estimates are unbiased if *X* has no true effect on *Y*.

[g] Odds ratios are unbiased if missingness of *X* and *Y* are independent of each other.

statistical toolbox for anyone who estimates associations. Much practical advice, with helpful references, is provided by White et al.[22]

Missing data may result in biased estimates no matter what analytic method we adopt. For that reason, investigators should try to minimize missing data when planning a study.[1,28,29]

ARTICLE INFORMATION

Accepted for Publication: October 30, 2012.

Published Online: May 13, 2013. doi:10.1001/jamapediatrics.2013.1329.

Conflict of Interest Disclosures: None reported.

Additional Contributions: Thomas D. Koepsell, MD, MPH, and Barbara McKnight, PhD, provided helpful advice on early drafts of this article.

REFERENCES

1. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol*. 1991;134(8):895-907.

2. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol*. 1995;142(12):1255-1264.

3. Schafer JL. *Analysis of Incomplete Multivariate Data.* New York, NY: Chapman & Hall; 1997.

4. Little RJA, Rubin DB. *Statistical Analysis With Missing Data.* 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.

5. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*. 2010;29(28):2920-2931.

6. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology*. 2002;13(3):347-355.

7. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol*. 2010;172(4):478-487.

8. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res*. 2007;16(3):199-218.

9. Little RJA. Regression with missing X's: a review. *J Am Stat Assoc*. 1992;87(12):1227-1237.

10. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8(1):3-15.

11. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087-1091.

12. Enders CK. *Applied Missing Data Analysis.* New York, NY: Guilford Press; 2010.

13. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.

14. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59(10):1092-1101.

15. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393.

16. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med*. 1991;10(4):585-598.

17. Raghunathan TE. What do we do with missing data? some options for analysis of incomplete data. *Annu Rev Public Health*. 2004;25:99-117.

18. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data [published online January 10, 2011]. *Stat Methods Med Res*. doi:10.1177/0962280210395740.

19. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Stat Methods Med Res*. 2013;22(1):14-30. doi:10.1177/0962280211403597.

20. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*. 2012;184(11):1265-1269.

21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley & Sons; 1987.

22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.

23. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485.

24. Cummings P, Koepsell TD. *P* values vs estimates of association with confidence intervals. *Arch Pediatr Adolesc Med*. 2010;164(2):193-196.

25. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology*. 2012;23(1):159-164.

26. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol*. 2004;160(1):34-45.

27. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res*. 2007;16(3):277-298.

28. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? *Stat Methods Med Res*. 2006;15(3):213-234.

29. Fleming TR. Addressing missing data in clinical trials. *Ann Intern Med*. 2011;154(2):113-117.

Downloaded From: http://jamanetwork.com/ by a Washington University - St Louis User on 08/21/2017