Missing Data: What Are You Missing?

Craig D. Newgard, MD, MPH
Jason S. Haukoos, MD, MS
Roger J. Lewis, MD, PhD

Society for Academic Emergency Medicine Annual Meeting
San Francisco, CA
May 2006

INTRODUCTION

Missing data are ubiquitous in clinical research. While many researchers consider missing data a nuisance, ignoring missing data is fraught with potential complications. Naïve methods for handling missing data can introduce substantial bias, reduce precision of estimates, and reduce study power, any of which may lead to invalid study conclusions.

MECHANISMS OF MISSING DATA

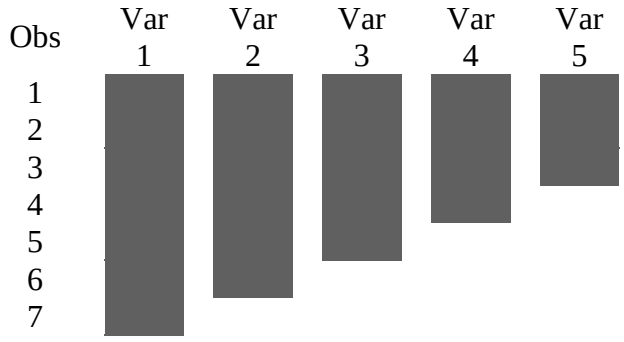
Missing Completely At Random (MCAR): when the probability a value is missing is independent of all other observed and unobserved characteristics of the sample.

Missing At Random (MAR): when missing values can be completely explained by observed values in the sample (i.e., independent of unobserved values). Most valid missing data procedures require data to be missing by MCAR or MAR mechanisms.

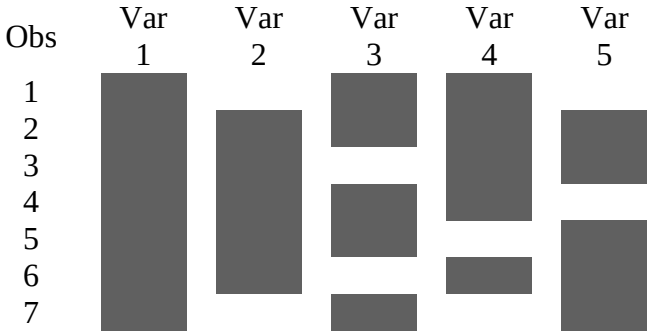
Missing Not At Random (MNAR): when missingness is related to missing values rather than observed values. It is not possible to distinguish between MAR and MNAR mechanisms.

PATTERNS OF MISSING DATA

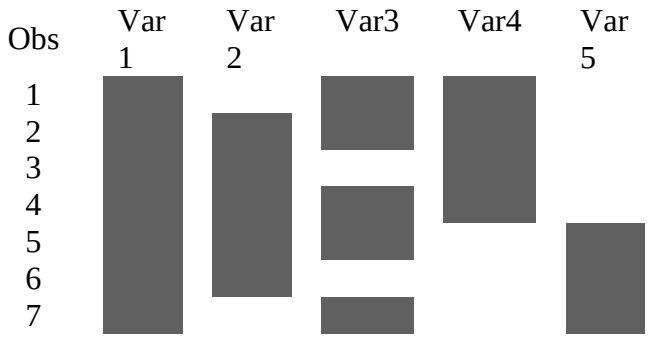
Monotone



Non-monotone (general)



Two variables never jointly observed (variables 4 and 5)



METHODS FOR HANDLING MISSING DATA

Complete Case Analysis: excludes all observations with missing values for any variable(s) of interest (e.g., for multivariable analyses), thus limiting the analysis to those observations for which all values are observed, often resulting in biased estimates and loss of precision.

Available Case Analysis: complete case analysis for univariate comparisons.

Weighted Complete Case Analysis: weighting of complete cases to adjust for nonresponse and to correct for potential bias (at expense of efficiency), generally used in surveys where cases of nonresponse have no available data.

Simple (Single) Imputation: imputing one plausible value for each missing variable within a dataset, then conducting the analyses as if all data were originally observed. Simple imputation generally results in inappropriately small variances and may produce biased results. There are several types of simple imputation, including:

- *Mean and median imputation*: replacing missing values with the mean or median values from the observed data - can lead to biased parameter estimates because missing values are replaced with values at the center of the distribution for a given variable.

-*Regression imputation*: involves replacing each missing value with a single predicted value from a regression analysis in which fully observed cases are used to determine the predicted values - underestimates the variance because no residual error is assumed around the regression line.

-*Stochastic regression imputation*: replaces missing values with the predicted value from a regression analysis plus its residual error - incorporates uncertainty into the predicted value, thus improving upon the primary limitation of simple regression imputation.

-*Hot deck imputation*: a non-parametric method of matching cases with missing values to observations with observed values for the same variable. Works well in certain settings, though still has a potential for bias, need for variance correction, and need for having an adequate number of complete cases for matching.

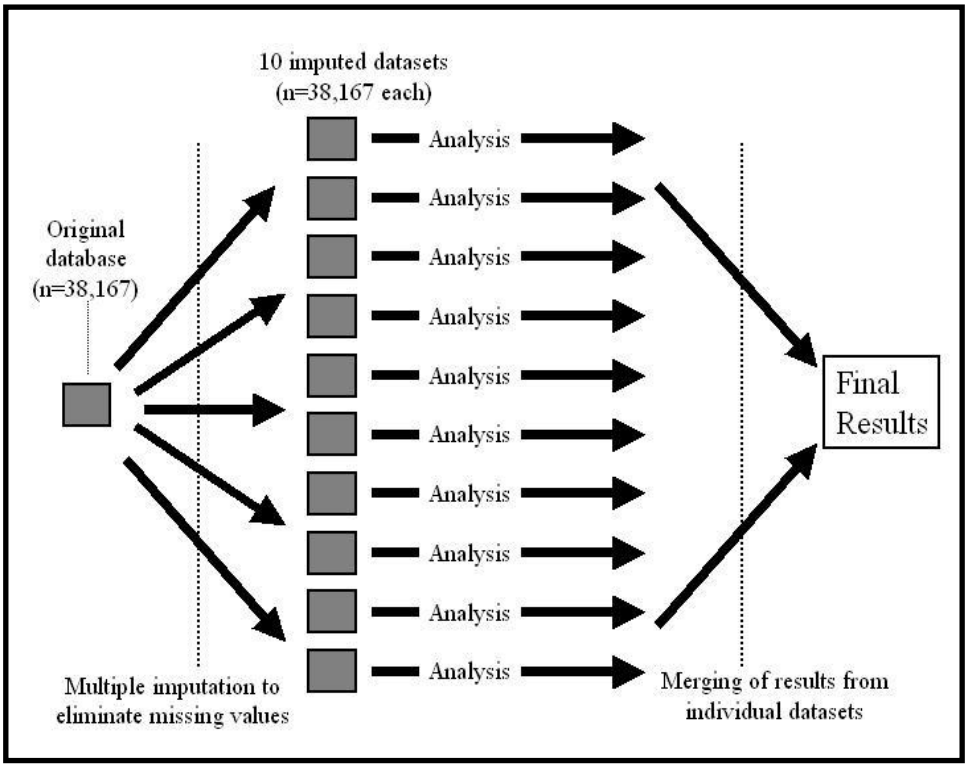
-*Cold deck imputation*: replacing missing values with a value(s) from an external data source using methodology similar to the hot deck technique. Data from the external source may differ systematically from the primary dataset, adding additional bias to estimates.

-*Last observation carry forward*: used during longitudinal repeated measures research - involves imputing missing follow-up measures with the last recorded value for a respondent. Although simple, this method assumes (often mistakenly) that the value remains unchanged after the patient drops out.

-*Worst-case analysis*: imputes the worst-case value for observations with missing values, most commonly for missing outcome data.

Multiple Imputation (MI):

MI is an extension of simple imputation, where each missing value is replaced by a set of $m > 1$ simulated values (generally 5-10) that exist in m complete datasets. Each of the m datasets is then analyzed by standard analytic methods, and the results from each dataset are combined using a standard set of rules that appropriately account for the uncertainty (i.e., variance) in the MI process.



Assumptions:
-MCAR or MAR mechanism for missingness
-Multivariate normal distribution for variables
-Adequate sample size

The Data Model: The first step in MI is to generate a probability model that relates the complete data set Y (consisting of missing values Y_{mis} and observed values Y_{obs}) to a set of parameters. Using Bayesian methods, observed values are used to generate a predictive distribution for missing values, after which multiple random "draws" are made (equivalent to the m number of complete data sets generated), thus producing the multiple imputations for each subject originally missing data.

Selection of variables for the data model:
-target variables: variables to be included in subsequent analyses, including the outcome measure
-auxiliary variables: variables highly predictive of target variables or important in explaining the missingness of variables
-sample design variables: variables important in sampling design (e.g., clusters and strata)

Determining the number of imputed values/datasets to generate: selection of the number of imputations (m) is based on the desired "relative efficiency" of MI estimates and is approximated using the formula:

$$= (1 + \lambda/m)^{-1}$$

where λ is the rate of missing data. In almost all instances, $m = 10$ provides a high rate of efficiency.

	Proportion missing data (λ)				
# imputations (m)	**10%**	**30%**	**50%**	**70%**	**90%**
3	.97	.91	.86	.81	.77
5	.98	.94	.91	.88	.85
10	.99	.97	.95	.93	.92
20	1.00	.99	.98	.97	.96
30	1.00	.99	.98	.98	.97

Analyzing Multiply Imputed Data: Once the m imputed datasets have been created, the data in each dataset can be analyzed using standard statistical procedures.

Parameter estimates (Q) are averaged over the m number of datasets:

$$\text{ave}Q_m = m^{-1}\sum Qi$$

The within-imputation variance (U_m) is represented by the average of the m complete data variances:

$$\text{ave}U_m = m^{-1}\sum U_i$$

The between-imputation variance (B_m) is:

$$B_m = (m - 1)^{-1} \sum (Q_i - \text{ave}Q_m)^2$$

The total variance (T_m) is generated by combining both the within-imputation and between-imputation variances,

$$T_m = \text{ave}U_m + (1 + m^{-1})B_m$$

An interval estimate (e.g., 95% confidence interval) is generated by:

$$\text{ave}Q_m \pm t_v(\alpha/2)T_m^{1/2}$$

where $t_v(\alpha/2)$ represents the upper $100(\alpha/2)$ percentage point of the student *t* distribution with *v* degrees of freedom:

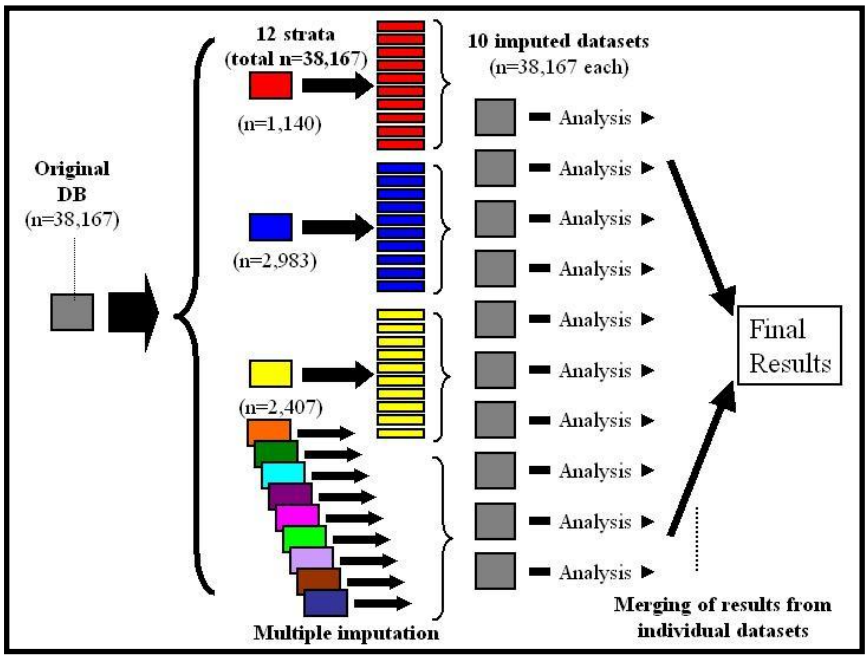
$$v = (m - 1)\ [1 + \text{ave}U_m/(1 + m^{-1})B_m]^2$$

Special Situations with Multiple Imputation:

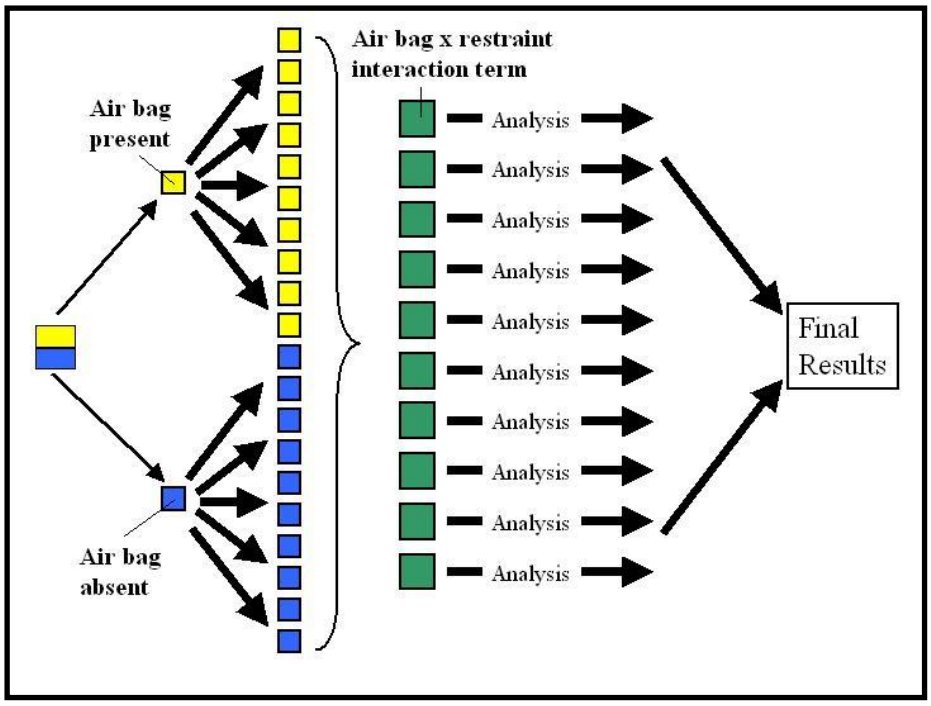
-Imputer versus Analyst

-Coding "missing"

-Probability samples:



-Interaction terms (parallel chains of MI):



-Split sample MI

MI versus Maximum likelihood (ML):

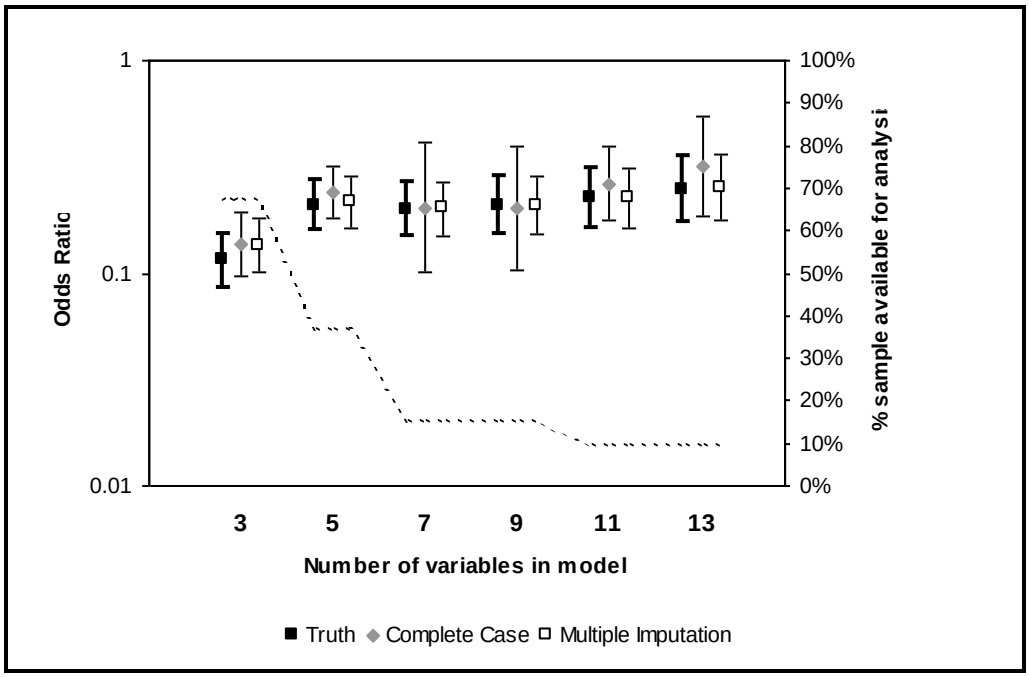
-ML methods are model specific
-ML is computationally intensive
-ML accounts for missing values during the analysis (one step)
-ML is slightly more efficient than MI, though the difference is generally not noticeable
-software using ML may be less able to integrate auxiliary variables, endangering the MAR assumption

HOW MUCH IS TOO MUCH MISSING DATA?

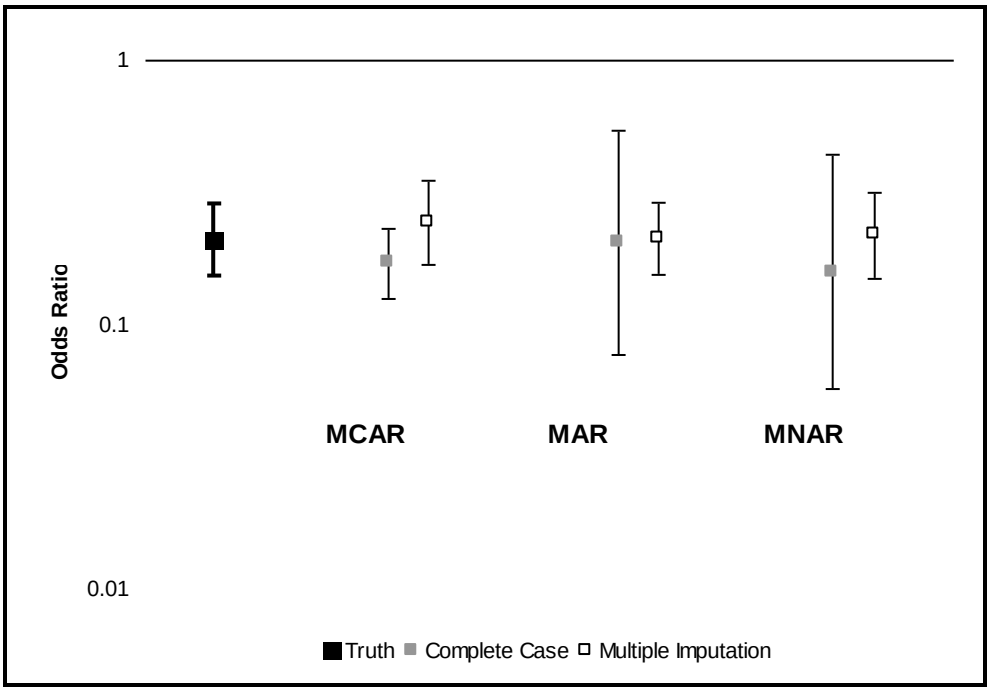
-difficult to generate guidelines for what constitutes a "small portion" of missing data
-small amounts (e.g., 3%) of missing data may still generate substantial bias and reduce precision
-there is no proportion of missing data under which valid results and preservation of study power can be guaranteed

EXAMPLES: National Automotive Sampling System Crashworthiness Data System (NASS CDS) – a probability sample of all drivers ≥ 16 years involved in MVCs with passenger vehicles or light trucks during the years 1995-2003 (n = 38,167). For the majority of examples, the same 10-variable logistic regression model (outcome = Injury Severity Score ≥ 16) was used. The figures demonstrate how various missing data factors affect one categorical (RESTRAINT USE) and one continuous (DELTA V) variable included in such models.

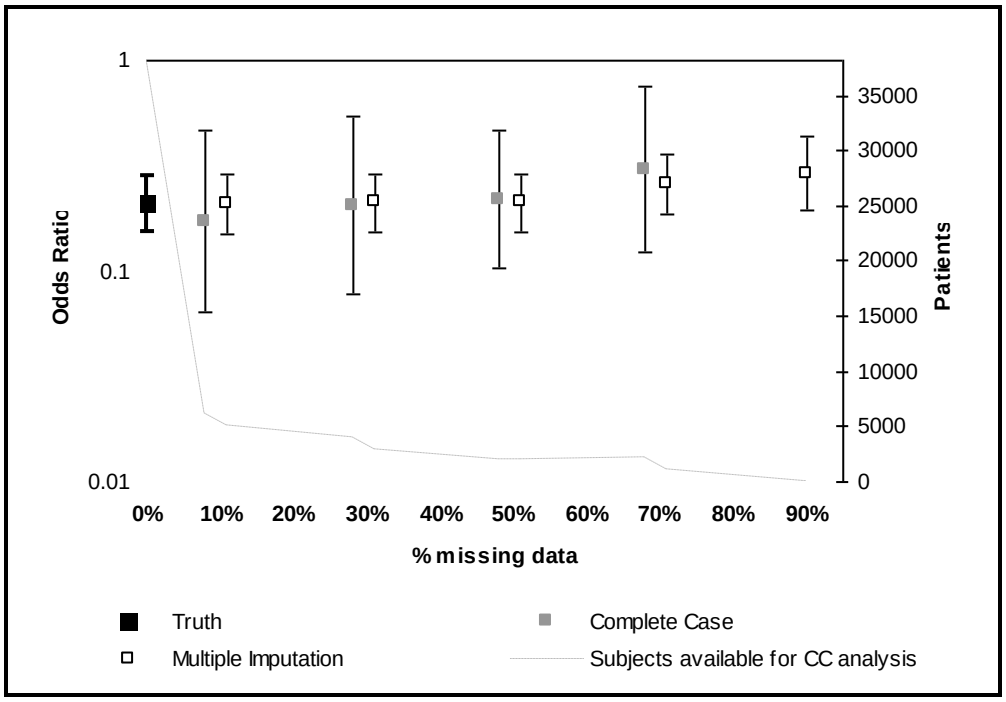
Example 1. Effect of increasing the number of variables in a multivariable logistic regression model on results for RESTRAINT USE (MAR pattern, baseline rate of restraint missing = 30%).



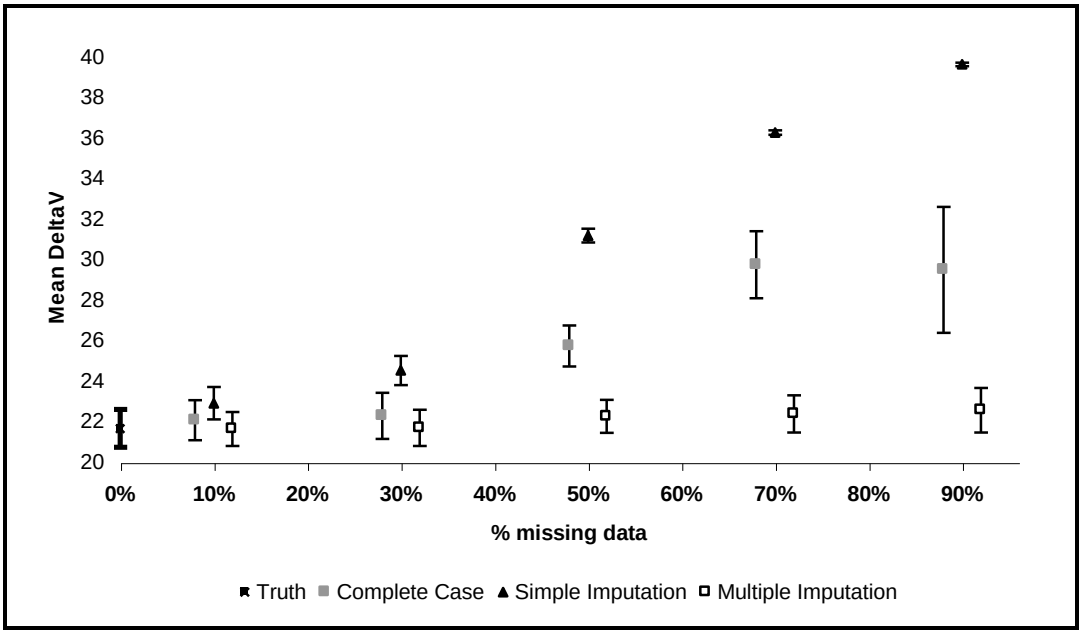
Example 2. The effect of different underlying mechanisms for missing values on study results for RESTRAINT USE in a multivariable logistic regression model (baseline rate of restraint missing = 30%).



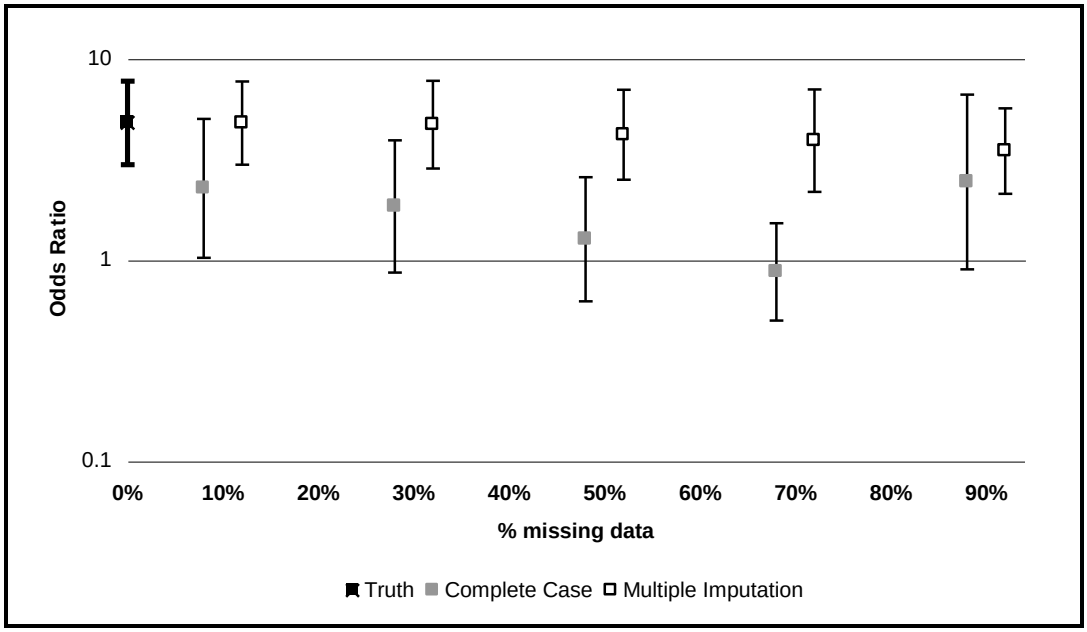
Example 3. The effect of increasing the proportion of missing data on study results for RESTRAINT USE in a multivariable logistic regression model (MAR pattern).



Example 4. The effect of increasing the proportion of missing data on study results for mean DELTA V (MAR pattern, univariate analysis).



Example 5. The effect of increasing the proportion of missing data for restraint use on results for a separate covariate (LATERAL IMPACT) with a fixed proportion of missing data (4%) in a multivariable logistic regression model (MAR pattern).



USEFUL IMPUTATION INFO:

The imputation listserv:
http://lists.utsouthwestern.edu/mailman/listinfo/impute

Joseph Schafer's list of multiple imputation software (including free NORM download and S-Plus macros):
http://www.stat.psu.edu/~jls/misoftwa.html

Sample of software that offer multiple imputation procedures and methods to analyze/combine results:
-SAS v8 and v9 (PROC MI and PROC MIANALYZE)
-IVEware (free SAS-callable software: http://www.isr.umich.edu/src/smp/ive/)
-Multiple Imputation for Chained Equations (MICE) for S-Plus and R
-macros for S-Plus (Joseph Schafer)
-NORM (free stand-alone windows package)
-SPSS
-LISREL
-SOLAS

Additional software that do not perform MI, but can analyze multiply imputed data:
-SUDAAN v9
-Stata

CONCLUSIONS

- Missing data is a major problem in clinical research.
- Ignoring missing values can increase bias, reduce study power, and reduce precision of estimates, sometimes drastically.
- Inappropriate methods for handling missing values (e.g., complete case analysis and simple imputation) can potentially invalidate study results.
- Provided certain assumptions are met, multiple imputation is one method for generating valid estimates and preserving study power, while appropriately accounting for the uncertainty in the imputation process.

SELECT REFERENCES:

Texts:

1. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1989.

2. Allison PD. (2001) Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

3. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

4. Schafer JL. Analysis of incomplete multivariate data. Boca Raton, Florida: Chapman & Hall/CRC, 1997.

SAS Multiple imputation coding:

5. Yuan YC. Multiple imputation for missing data: concepts and new development. P267-25. SAS Institute Inc.

IVEware background and multiple imputation coding:

6. Raghunathan TE, Lepkowski , Van Hoewyk J, Solenberger PW. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;27:85-95.

7. Raghunathan TE, Solenberger PW, Van Hoewyk JV. IVEware: Imputation and variance estimation software user guide. Survey Methodology Program, Institute for Social Research, University of Michigan, March 2002.

Additional missing data references:

8. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychological Methods. 2001;6:317-329.

9. Schafer JL. Multiple imputation: a primer. Statistical Methods in Medical Research. 1999;8:3-15.

10. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001;6:330-351.

11. Reiter JP, Raghunathan TE. Multiple imputation for missing data in surveys with complex sample designs. Technical Report, Institute for Social Research, University of Michigan, Ann Arbor, MI. 2002.

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

13. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician. 2001;55:244-254.