Missing Data in Orthopaedic Research

Keith D Baldwin, MD, MSPT, MPH, Pamela Ohman-Strickland, PhD

Abstract

Missing data can be a frustrating problem in orthopaedic research. Many statistical programs employ a list-wise deletion by default to eliminate missing data. This elimination may be helpful to conduct the analysis but its use without first considering the type of data, and the possible mechanisms of the "missing ness" of data can lead to a loss of power, or to biased estimates. Orthopaedists without a statistical background should have at least a superficial knowledge of how to deal with missing data. This article represents a review of different methods of dealing with missing data and experimental demonstration of weaknesses and strengths of these methods. The goal of this article is to provide orthopaedists with a limited statistical background a basic understanding of the effects of missing data, alternatives to list-wise deletion, and ways of utilizing each technique when different mechanisms of missing data exist. In addition, this article explores the new options available in SAS versions 8.1 and higher to deal with missing data without using macros.

Introduction

Missing data presents a significant challenge to orthopaedic researchers of all different calibers. The central conundrum of missing data is to produce estimates that mimic the non-existent or missing data, while keeping the uncertainty inherent in missing values.[1] For researchers with a significant budget, and ample time, the solution is to hire a brilliant biostatistician to identify the mechanism (if any) for the missing data, and provide a model for the "missingness" of the data, that will provide suitable estimates accounting for possible biases. Typically, in a large clinical trial, the observations with missing values will simply be deleted (listwise deletion) to avoid bias, unless there is evidence that there is some mechanism to the "missingness" of the data that is relevant to the drug or procedure under investigation (a non-ignorable mechanism).

Address for Correspondence:

Keith D. Baldwin, MD, MPH, MSPT;
Hospital of the University of Pennsylvania, Department of Orthopaedic Surgery
3400 Spruce Street, 2 Silverstein Building
Philadelphia, PA 19104
Phone: 1-215-662-3340
Fax: 1-215-349-5890
Email: keith.baldwin@uphs.upenn.edu

Keith Baldwin is a Resident in the Department of Orthopaedic Surgery at the University of Pennsylvania
Pamela Ohman-Strickland is a Associate Professor of Biostatistics at the UMDNJ School for Public Health in Piscataway, NJ

Missing data in smaller studies can be more problematic. These studies include pilot studies, research that is done by medical students, or unfunded research by housestaff or attending physicians. Typically, these missing observations will be dealt with by eliminating them, as this is the default in most statistical programs. Listwise deletion in this setting may still be a reliable method as it is relatively unbiased provided the data are missing completely at random (MCAR). However, if the study is small, or the amount of missing data is above 5 to 10% of the total data set, significant loss of power is possible.[2]

The goal of this review is to identify a few methods, that are accessible to medical students, residents, and attending physicians who do not have a statistician readily available. These methods in general will be easily applicable, but will depend on basic assumptions that are illustrated below.

Missing completely at random (MCAR): This is the most powerful assumption. Suppose there are two variables X and Y. Consider values for Y. The values if missing are MCAR if the probability of missing data on Y is unrelated to the value of Y and unrelated to the value of any other variable in the data set.[2]

Missing at random (MAR): This assumption is that the probability of missing a variable Y is unrelated to the value of Y, but may be related to the value of one or more variables in the data set.[2] This assumption is weaker, and impossible to test (because the data are missing).[2] For example, consider a case where it was desired to determine if hospital X treated more patients daily than hospital Y. If the probability of missing data on the number of patients treated depended on which hospital they were in, but within

each hospital, the probability of missing data on number of patients was unrelated to the number of patients treated, MAR would be satisfied.[2] In this case, the missing data is a result of bad record keeping that is unrelated to the value of the number of patients treated.

Ignorable missing data: This condition is satisfied if first, the data is MAR or MCAR and second, the mechanism governing the pattern of missing data is unrelated to the mechanism of effect to be measured (in the case illustrated above, the probability of missing a case is unrelated to the number of patients treated in that case).[2]

Non-ignorable: Usually this occurs in cases where the missing data are not MAR. The mechanism of missing the data must be modeled. In the previous example, if it was discovered that hospital Y systematically underreported low census days, this would be an example of non-ignorable missing data.[2] If a non-ignorable mechanism for missing data is suspected, a biostatistician (preferably one with interest in this area) should be consulted.

Methods

List-wise Deletion:

This method is the simplest method available. Most statistical packages utilize this as the default method of calculating regression estimates and other test statistics. This is probably the best method to use if: 1. The data are MCAR, and either 2. The data set is large, or 3. The amount of missing data is rather small relative to the total amount of data.[2,3] If the second or third condition is not satisfied then the power of the estimates can be seriously reduced.[2,3] If the data is MCAR, the smaller sub sample will simply be a random sub sample of the original population.[2] If the original sample is normally distributed, the estimates calculated will be relatively unbiased. However, the standard errors will be larger, due to the smaller sample size.[2] In list-wise deletion no data is imputed, thus, any error that is built in is inherent to the sampling methods or violations of the MCAR assumption. List-wise deletion is less trustworthy for data that are MAR, unless the data that are MAR are the independent variables.[4] In this case list wise deletion is very robust to violations of MAR of independent variables in a regression analysis.[2] The dependant variable however must be MCAR, and other variables must be MCAR with respect to the dependant variable. Suppose in our hospital census example the probability of missing data on the number of patients treated depended on which hospital they were treated in. However, within each hospital, as long as the probability of missing data on number of patients treated was unrelated to the number of patients treated, regression estimates based on list-wise deletion (the smaller data set) would be unbiased. List-wise deletion is also best for variables where the probability of the data being missing is dependent on the value of that variable (not the dependant). E.g., if hospital X, and hospital Y underreported slow census days, but that under reporting was uniform, list-wise deletion would be appropriate. Therefore -in summary, list-wise deletion is appropriate where:

1. The data are MCAR, or at least MCAR with respect to the dependant variable.
2. The data set is large and/or the amount of missing data is small relative to the total sample.
3. It is undesirable to have to model the missing data.

Pair-wise Deletion:

This method is mentioned briefly here, because it can produce some confusing situations when attempting to determine an appropriate model in regression analysis, in addition to producing problems with bias. Pair-wise deletion is a very enticing option because it allows one to use more of the data from the initial data set than does list wise deletion. In list-wise deletion, if one parameter is missing, that entire observation (all variables) is deleted. In pair-wise deletion, all of the data are used for all of the analyses for which they can be used. So for computation of a covariance matrix, all cases of X and Y are considered; whereas for list wise deletion, only the cases with full data would have been considered.[4] The problem with this method is that if there is any violation of MCAR, there can be serious problems with bias.[2] So the decreased standard errors that are observed with pair-wise deletion (because more information is used) are not worth the price that is paid in terms of biased estimates. As a result, if any deletion is used, list-wise deletion is recommended because of the aforementioned reasons, in addition to the decrease in confusion when using automated mechanisms of model building. In addition, pair-wise deletion can make selection of linear models with computer programs confusing.

Dummy variable adjustment:

This method makes use of a dummy variable to represent the fact that there is missing data for a certain variable (e.g rather than male and female, there would be 3 groups, male, female and missing) This method, though appealing, turns out to have

serious problems with bias, and can be very confusing, even under the best of circumstances (MCAR).[2,5,6]

Weighting techniques:

This method is commonly used when unit non-response exists. Unit non-response is said to occur when, for some reason, some individual fails to return a survey, answer a research question, or in some other way is recruited, but does not participate. This method assumes that there is demographic information present about the subject that would allow for comparison with others. Then, a subject with similar demographic characteristics is assigned a weight of 2 to account for the fact that a demographically similar subject's data were missing. This method eliminates all variability between demographically similar data points, which can affect standard errors adversely.[3] In addition, this technique does not account for stratification of samples, which could make weighting cumbersome.

Maximum Likelihood:

Simple imputation allows for calculation of estimates based on maximum likelihood. The expectation maximization (EM) algorithm is a commonly used statistical method for obtaining regression estimates in the presence of missing data. This approach in essence creates a complete data set by "filling in" the missing values. Analysis is then conducted using this complete data set. Technically, the method is comprised of two steps, an expectation step, and a maximization step. The two steps are repeated many times until eventually they converge to the maximum likelihood estimates.[2] The expectation (E) step consists of imputing of the missing values using information from the observed data and previous estimates of the regression parameters via regression. The maximization step estimates new values for the regression parameters and variance covariance matrix using the completed data set from the previous E-step. The process is then started with the E-step, such that parameters have been estimated using list-wise deletion. The iterative process continues until the estimated parameters are barely changing from one iteration to the next. The advantage of this method is that it uses all of the available predictors for imputing missing data. Software is available to implement the EM algorithm and, hence, simple imputation. However, this software does not typically take into account that the data are imputed, hence the standard errors are inappropriately low, and the test statistics are inappropriately high.[4] Unfortunately, the techniques used to adjust the standard errors for "filling in" the missing data are not available in most "user-friendly" software, and would likely require the assistance of a statistician who is familiar with this software.[2] If simple imputation is used without making such adjustments, it will lead to a higher chance of making a Type I error.

Multiple Imputation:

Multiple imputation is a process that uses random error to increase the standard errors relative to those calculated via simple imputation. The observed data set is completed multiple times with different plausible values for the missing data. The range of standard errors from the multiple completed data sets is used to estimate the appropriate increase in the standard errors due to the uncertainty of the missing values. Among the upsides of multiple imputation are that it produces unbiased estimates when the data are MAR.[2] The model can even be used modified, presumably through consultation with a statistician, to model incomplete data with non-ignorable mechanisms. Multiple imputation is available with conventional software that almost anyone can use, in many different situations.[4] Multiple imputation is the procedure of choice in many situations, such as when the dependent variable is MAR, or when there is a large amount of missing data, or a relatively small data set (That is to say, most situations where the list-wise deletion would be unacceptable and collecting more data would not be feasible). The major drawback of multiple imputation is that, because it uses a random component, slightly different estimates will be obtained every time it is used.[4] One must specify if multiple imputation is employed so researchers who re-run your data will not be surprised when their results differ from the original results.

In order to use multiple imputation, data should be MAR. The model created should be the final model used in analysis, that is to say, all transformation, and interactions should be identified prior to imputation.[8]

Procedurally, one of the simplest and most widely used programs that has a multiple imputation procedure is SAS (versions 8.1 and higher, Copyright (c) 2003-2004 by SAS Institute Inc., Cary, NC, USA). This procedure can be run by the following set of generic code (The Y value must be included):

PROC MI data=<dataset> out= <output data set name>
var <variable1 variable2….variableN>;
run;

This procedure uses estimates from the EM algorithm as starting values, does a default of 200 "burn in" iterations, and then 100 iterations (default) between each imputation. There is a default of 5 data sets created (so if the original data set had 100 observations, the output data set would have 5 data sets x 100 observations = 500 total observations). The output data set also has a variable _imputation_ that would have a value of 1-5 in the default setting to indicate which data set. A more complicated set of code is listed below with some of the more useful options in PROC MI:

```
PROC MI data=<dataset> out= <output data set name> seed=<any number>
nimpute=<number of imputations (data sets created)>
minimum=<number var1 number var2…number varN>
maximum=<number var1 number var2…number varN>
round= <what each variable should be rounded to>;
var <variable1 variable2….variableN>;
MCMC nbiter=<number "burn in iterations prior to 1> niter= <number of
iterations between imputations>;
run;
```

The seed option in the first line allows duplication of the same parameter estimates. The nimpute option in line two dictates the number of data sets that are created. Note, that if you make nimpute=0 then the MI procedure is simply an EM algorithm using the maximum likelihood method previously mentioned. The "minimum" and "maximum" statements in lines 3 and 4 protect against hidden extrapolation (making estimates outside of the range of your data set). The round option allows the imputed values to be rounded in any way specified. The MCMC statement specifies that a Markov chain Monte Carlo method is being used to impute the values. It is the default, and this statement has in its options to specify the number of "burn in" iterations prior to the first, and the number of iterations between imputations. If more are specified, the likelihood of convergence of estimates is increased, and the chances of statistical dependence are decreased although not assured, some experts believe that the default is sufficient in most cases.[4] Under this statement, other useful options are time series plots, and autocorrelation plots. These options are a good start for the appropriate usage of PROC MI. The output for PROC MI will specify the method of imputation (MCMC at default), how the initial estimates were found, the number of imputations, the number of "burn in" iterations, the number of iterations between imputations, and the seed for the random number generator. In addition, the MI output provides information on the pattern of the missing data, a covariance matrix of EM estimates, as well as increase in variance, parameter estimates, fraction of missing data, and relative efficiency.

Some astute readers may note, however, that by using PROC MI, multiple data sets are created. How can a single set of parameter estimates be obtained when there is random variability inherent to PROC MI? The answer is PROC MIANALYZE. First, the output data set must be sorted by _imputation_;

```
PROC SORT data=<output data set name>;
by _imputation_;
run;
```

Next, PROC REG must be used to combine the regression estimates and covariances into one data set.

```
PROC REG data=<output data set name> outset=<regression coefficient output
data set name> covout;
model <Y>=<X1 X2 X3….XN>;
by _imputation_;
run;
```

So in essence, the usual regression statement is run for each of the five (at default) data sets, to get regression parameters. The covout option at the end of the first line tells SAS to include the covariance matrix in the data set. Next, PROC MIANALYZE will combine them.

```
PROC MIANALYZE data=< regression coefficient output data set name>;
var <intercept variable 1 variable 2….variableN>
run;
```

It is possible to calculate the number of imputations needed, but this calculation is beyond the scope of this article. More imputations translates into more efficiency. Efficiency should be high, but if it is 95% or greater, typically adding more imputations will provide little benefit

One word of warning when using multiple imputation with this or any other procedure is that multiple imputation does not account for interactions or non linearities in data sets. It is advised that prior to doing multiple imputation, a regression with list wise deletion is carried out, with diagnostics. If interaction is detected, then a variable should be

Method	AGE b1 (SE)	AGE p-value	ISS* b2(SE)	ISS * p-value	Gender b3 (SE)	Gender p-value	Age*Gender b4 (SE)	Age*gender p-value	N†	R^2
Complete	-0.033 (0.007)	**<0.0001**	0.062 (0.016)	**0.0001**	-1.906 (0.665)	**0.0049**	0.022 (0.008)	**0.010**	128	0.605
Listwise deletion	-0.024 (0.009)	**0.0071**	0.049 (0.018)	**0.0098**	-1.452 (0.859)	0.0958	0.015 (0.011)	0.1533	69	0.672
Multiple imputation	-0.030 (0.008)	**0.0002**	0.041 (0.020)	0.0563	-1.349 (0.729)	0.0661	0.0145 (0.009)	0.1108	640	0.601
Max Likelihood	-0.030 (0.0007)	**<0.0001**	0.053 (0.002)	**<0.0001**	-1.427 (0.071)	**<0.0001**	0.016 (0.0009)	**<0.0001**	128	0.616

Table I : Select Regression Parameters, standard errors and, P-values of parameters with different methods for Missing Completely at Random (MCAR)data Age*Gender indicates the interaction term between Age and Gender

* ISS=Injury severity score

† Multiple imputation has N=640 because 5 data sets x 128 observations

generated for the product of the two terms interacting.[2] If non- linearities are detected, the suitable transformation should be carried out prior to imputation.[2] After these methods are carried out, imputation can proceed as before.

In summary, multiple imputation is recommended as a method when:

1. The data are MAR, especially if the dependant variable is MAR.
2. The data set is small, or the amount of missing data is large relative to the size of the data set.
3. List wise deletion is unacceptable for some other reason.

Results

Example 1: (MCAR)

The following example illustrates PROC MI and MIANALYZE versus a complete data set, versus listwise deletion and maximum likelihood with the same missing data. The "full" data set was a set of 128 observations from the NTDB (American College of Surgeons 2004). The objective was to fit an ordinary least squares regression line in order to adjust cost (charges) of internal fixation versus partial hip replacement as a treatment for femoral neck fracture for age, gender, length of stay (LOS), injury severity (ISS) and comorbidities. When the complete data set was investigated with diagnostics, the regression equation was found to have a non-constancy of error variance. A box cox procedure indicated that a log transformation would provide an adequate correction for the non-constancy of error variance. In addition, a significant interaction was found between age and gender. A new variable was created to describe that interaction (agegen). Table I demonstrates the behavior of the data when the data are MCAR (random deletion by ID number of every 6^{th} observation. Note that standard errors for simple imputation (maximum likelihood) are lower than for other methods. This is because this method tends to bias towards the mean, which produces narrower confidence intervals, and lower p values (table I). Table I shows that Listwise deletion often performs in a superior fashion to Multiple Imputation with MCAR data. This is because with MCAR data, Listwise deletion tends to be a random subset of the whole data set, and so the loss of power is less important unless the amount of missing data is large.

Example 2 (MAR):

Consider a case where the data was not MCAR. Suppose for some reason hospitals reported charges for females less often than for males, but that the reporting of the charges did not depend on the value of charges. In this case, hospital charges (the output variable) are MAR. For this data set, the same elimination rules are used as in the prior example for the other variables, but now for gender=female, each fourth point of charges is missing, whereas for males, the same values as previously are missing (randomly one every 13). Table II summarizes the results of this experiment. Note that Gender has become non-significant with listwise deletion. This example demonstrates that when data is MAR, listwise deletion may not perform as well as multiple imputation in certain circumstances. In addition, note that the R squared is biased upward with listwise deletion. Recall that R squared compares the sum of squares for regression (SSR) to the sum of squares error (SSE). It is plausible that because of random elimination of data points the SSE may have been decreased (by chance the data farther from the mean were eliminated)

Method	AGE b1 (SE)	AGE p-value	ISS* b2(SE)	ISS * p-value	Gender b3 (SE)	Gender p-value	Age*Gender b4 (SE)	Age*gender p-value	N†	R^2
Complete	-0.033 (0.007)	**<0.0001**	0.062 (0.016)	**0.0001**	-1.906 (0.665)	**0.0049**	0.022 (0.008)	**0.010**	128	0.606
Listwise deletion	-0.022 (0.008)	**0.0104**	0.072 (0.020)	**0.0008**	-1.598 (0.859)	0.0686	0.017 (0.011)	0.1069	59	0.713
Multiple imputation	-0.028 (0.008)	**0.0005**	0.076 (0.021)	**0.0021**	-1.554 (0.709)	**0.0299**	0.017 (0.008)	0.0531	640	0.614
Max Likelihood	-0.028 (0.0007)	**<0.0001**	0.078 (0.002)	**<0.0001**	-1.556 (0.070)	**<0.0001**	0.0173 (0.003)	**<0.0001**	128	0.625

Table II: Select Regression Parameters, standard errors and, P-values of parameters with different methods with Missing at Random (MAR) data Age*Gender indicates the interaction term between Age and Gender

* ISS=Injury severity score

† Multiple imputation has N=640 because 5 data sets x 128 observations

Example 3 (Non-Ignorable):

Finally, consider a data set where the mechanism of missing data is non-ignorable. If the charges (dependant variable) are in the top quartile, the last 4 out of every 5 are eliminated. If the charges are in the second to top quartile the last 3 out of every 5 are eliminated. If the charges are in the third quartile the last 2 out of every 5 is eliminated, all of the data in the last quartile are retained. The other missing data are the same as the first set. Since the data are missing according to the value of the output variable, this mechanism is non-ignorable. The results for the complete data set, list-wise deletion, multiple imputation and maximum likelihood are summarized in table III. The power problems of list-wise deletion are accentuated by the larger amount of missing data in this example. In addition, serious problem with bias are introduced in the estimates, in all methods, especially for gender, and the interaction term between age and gender. Problems like this will be typical if the mechanism is non-ignorable. In this case, the mechanism of missing data would need to be modeled.

Discussion

Missing data will continue to be an issue in orthopaedic research as long as it exists. The absolute best solution to missing data is not to have any. If there is missing data in the data set, and reconciliation of this missing data is irreconcilable, it is important to recognize whether the "missingness" is ignorable, and finally if it is random. Most often listwise deletion is appropriate. It is however important to recognize when it is not, and have tools to deal with this situation should it arise. When there is some suspected mechanism that has led to missing data or when there is a longitudinal or a more complex model needed for the data we strongly recommend consulting a statistician. SAS 8.1 or higher provides programming options by which even non-statisticians can easily perform these functions. We have been able to demonstrate through the examples presented in this article that listwise deletion is robust in most cases, but if the amount of missing data is large, significant power can be lost by using it. Multiple imputation is useful if the amount of missing data is large relative to the total n, or if the data is not missing completely at random. Non-ignorable missing data must be appropriately dealt with if biased estimates are to be avoided. Except in special circumstances, pairwise deletion, weighting, dummy variables, and simple imputation should be avoided because of problems with biasing estimates.

Method	AGE b1 (SE)	AGE p-value	ISS b2 (SE)	ISS p-value	Gender b3(SE)	Gender b5 p-value	Age*Gender b6 (SE)	Age*gender p-value	N	R^2
Complete	-0.033 (0.007)	**<0.0001**	0.062 (0.016)	**0.0001**	-1.906 (0.665)	**0.0049**	0.022 (0.008)	**0.010**	128	0.606
Listwise deletion	-0.022 (0.010)	**0.0394**	-0.013 (0.037)	0.7168	-0.580 (1.078)	0.5936	0.007 (0.013)	0.6195	49	0.631
Multiple Imputation	-0.022 (0.011)	0.0824	-0.008 (0.036)	0.8419	-0.790 (1.088)	0.4869	0.009 0.013)	0.5001	640	0.540
Maximum Likelihood	-0.024 (0.0007)	**<0.0001**	-0.022 (0.001)	**<0.0001**	-1.042 (0.065)	**<0.0001**	0.013 (0.0008)	**<0.0001**	128	0.562

Table III: Select Regression Parameters, standard errors and, P-values of parameters with different methods with non-ignorable missing data Age*Gender indicates the interaction term between Age and Gender

* ISS=Injury severity score

† Multiple imputation has N=640 because 5 data sets x 128 observations

Acknowledgments

This paper has been presented in the Biostatistics journal club at the School of Public Health at the University of Medicine and Dentistry of New Jersey May 2005, no sources of outside funding were used to complete this study, the authors are not consultants for SAS, and have no conflicts of interest

.**References**

1. Schafer, J. Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/ CRC 2002.
2. Allison, P. Missing Data. Thousand Oaks, CA: Sage Publications 2002.
3. Patrician, P. Multiple imputation for missing data. Research in Nursing and Health 2002; 38: 76-84.
4. Allison, P. Missing data techniques for structural equation modeling. Journal of Abnormal Psychology 2003; 112(4): 545–557.
5. Jones, M. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association 1996; 91: 222–230.
6. Vach W. Logistic Regression with Missing Values in Covariates. New York: Springer-Verlag 1994.
7. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annual Review of Public Health 2004; 25: 99-117.
8. Rubin, D. Multiple imputation for non-response in surveys. New york: Wiley 1987.
9. Ferketich, S; Verran, J. Analysis issues in outcomes research: examining the effectiveness of nursing practice. Washington, DC: .S. Department of Health and Human Services 1992.