ORIGINAL ARTICLES

Missing data as a validity threat for medical and healthcare education research: Problems and solutions

Myrah Stockdale[1], Kenneth Royal*[2]

[1] College of Pharmacy and Health Sciences, Campbell University, Buies Creek, NC, USA
[2] College of Veterinary Medicine, Department of Clinical Sciences, North Carolina State University, Raleigh, NC, USA

Received: May 12, 2016 **Accepted:** June 14, 2016 **Online Published:** June 24, 2016
DOI: 10.5430/ijh.v2n2p67 **URL:** http://dx.doi.org/10.5430/ijh.v2n2p67

ABSTRACT

At present, the problem of missing data has received virtually no attention by medical and healthcare education researchers. This is a significant problem for the education research community because when missing data are disregarded or handled inappropriately it can result in serious validity threats. This article discusses the problem of missing data in the context of medical and healthcare education research and recommends appropriate methods for handling missing data.

Key Words: Medical education, Healthcare education, Missing data, Education research, Research methods

1. INTRODUCTION

Success for healthcare professionals working in academia (*e.g.*, promotion and tenure decisions) often depend on one's performance in the area of research. However, the exact definition of research often varies across institutions. A recent trend across virtually all healthcare fields has been an increase in the number of professionals conducting educational research.[1] Educational research, is defined by the American Educational Research Association[2] as, "the scientific field of study that examines education and learning processes and the human attributes, interactions, organizations, and institutions that shape educational outcomes." Healthcare professionals have recognized the importance of education research, especially as it pertains to improved training and improved patient outcomes. This is in part due to the public's investment in medical training. Walsh[3] notes most medical education training programs are subsidized by national governments. In the United States alone it was reported that over $11.5 billion of federal funding (from Medicare and Medicade) was spent on graduate medical education (GME) in the year 2010,[1] and approximately $13.6 billion in 2012.[5] According to Henderson,[6] more than 40 states and local governments also contribute approximately $5 billion per year.

Clearly, the health professions value education and the increasing commitment to education research by healthcare professionals evidences this support. The problem, however, is most individuals do not have formal training in the field of education, have little familiarity with educational research methods, and often overlook issues of critical importance for quality education scholarship, especially when analyzing education data. One issue that is commonly overlooked involves the appropriate handling of missing data. Data may be missing or incomplete for a number of reasons. In the case of surveys, for example, some participants may inadvertently skip questions, choose not to respond, grow fatigued and abruptly withdraw from participation, provide unintelligible

*Correspondence: Kenneth Royal; Email: kdroyal2@ncsu.edu; Address: College of Veterinary Medicine, Department of Clinical Sciences, North Carolina State University, Raleigh, NC, USA.

answers, run out of time to respond, and so on. In the case of educational assessments, missing data often occur due to issues involving speediness, language, reading comprehension, carelessness, failure to return to a previously skipped item, and so on.

At present, the problem of missing data has received virtually no attention by healthcare researchers conducting education research. This is a significant problem for the healthcare and education research community because when missing data are disregarded or handled inappropriately it can result in serious validity threats. We suspect the problem of missing data has largely gone unrecognized primarily because most statistical software packages will readily calculate statistics in the presence of missing data.

One problem is that most data analysis procedures and statistical software packages are not designed to handle missing data. Therefore, when most statistical software programs encounter missing cases the missing data are either excluded or the entire row of data in which a response(s) was missing is dropped from the analysis. Dropping participants with missing data may lead to systematic differences between groups ultimately leading to biased results,[7] decreased power,[8] and differential accuracy of results.[9] Of course, there is always the risk of making inappropriate inferences about one's results as well.[9–11] Thus, the purpose of this article is to discuss the problem of missing data in the context of healthcare research and recommend appropriate methods for handling missing data.

2. NATURE OF MISSING DATA

Data can be missing in a variety of patterns.[12] Little and Rubin[13] codified the terms missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR) to describe three mechanisms of missing data. Table 1 provides a summary overview of each of the mechanisms including an example and methods classifying data missingness.

Table 1. Data Mechanisms

Data Mechanism	Definition	Example	Determination
Missing Completely at Random (MCAR)	The propensity for a data point to be missing is completely random	A student flips a coin to decide whether to complete the course evaluation.	Little's MCAR test
Missing at Random (MAR)	The propensity for a data point to be missing is not related to the missing data, but it is conditional on another variable.	Male students are more likely to refuse to complete course evaluations, but it does not depend on their level of course satisfaction.	Test for interactions between observed variables. Significant interaction would be indicative of MNAR data.
Missing Not at Random (MNAR)	The propensity for a data point to be missing is not random.	Students with disabilities are more likely to refuse to complete course evaluations.	

If data are systematically missing and the values are dependent on unknown (or unobserved) factors they are categorized as MNAR.[14] For example, if students with test anxiety issues or documented disabilities were more likely to have missing responses, their data would be MNAR because there is no way of estimating participants' response values (*e.g.*, whether a participant would likely respond correctly or incorrectly on an assessment question) with any accuracy or certainty from the existing data and information.

If systematically missing data could reasonably be considered dependent exclusively on observed factors or collected information, the data are categorized as MAR. For example, if nonresponse to a survey item (or items) on an end-of-course evaluation was a result of student gender (*e.g.*, female students were less likely to respond) and there was also no difference in the responses (*e.g.*, females were not more likely to rate the class higher or lower than males) we would conclude that the values were MAR because a researcher could conclude that the responses can be estimated accurately based strictly on an observed variable (or set of variables), and that there are no other confounding variables needed for accurate estimation.

If missing course evaluation scores could not reasonably be explained by observed data (*e.g.* gender, grade in course, *etc*.) or other reasonable, but unobserved, theories (*e.g.*, students that are required to take the course rank the course lower than those that take it for other reasons), then the data are MCAR Continuing with the MAR example, if we were to discover that gender, or other observed variables, did not play a role in student's participation in the evaluation and no reasonable theories could explain the missing-ness, the responses could be categorized as MCAR. Little's MCAR test can indicate if

observed data can predict the missing data,[13] but it is up to the researcher to thoroughly investigate possible theoretical explanations for missing data before deeming it MCAR.

MNAR, MAR, and MCAR data mechanisms do not make assumptions about whether or not the data are missing due to omission or nonresponse. Educational tests used in classroom assessments often have multiple test forms and are administered under time constraints, such as the duration of one class period. Thus, speediness (or "speededness" as discussed by some researchers) could explain missing data for students, particularly on items that appear in the latter part of the assessment. This speediness factor is often overlooked when determining data mechanisms for missing data and employing data handling procedures. In any event, once the data mechanism has been selected, decisions regarding the treatment and handling of missing data should be made.

There are times when data are incorrectly assumed to belong to one data mechanism when it is indeed another. Table 2 presents the effects of incorrect classification. Misclassification of missing data mechanisms can lead to incorrect parameter estimates and inaccurate results. Coding nonresponse observations as incorrect is palatable for free-response items, but less so for multiple choice items.[15]

Table 2. Effects resulting from incorrect treatment of data mechanisms

Actual Data Mechanism	Treated as MCAR	Treated as MAR	Treated as MNAR
Missing Completely at Random (MCAR)		Little's MCAR test should catch this error.	Unnecessary complexity is introduced into data handling procedures
Missing at Random (MAR)	Oversimplification that will reduce generalizability of results		
Missing Not at Random (MNAR)	Missing data process modeled incorrectly resulting in inaccurate parameter estimates		

3. STRATEGIES FOR HANDLING MISSING DATA

Strategies for handling missing data can be classified into three categories: 1) available case methods; 2) single imputation methods; and 3) model-based imputation methods. Available case methods includes techniques that rely on discarding a portion of the cases.[16,17] Single imputation methods and model-based imputation methods both involve replacing missing data with imputed values. The distinction between single and model-based imputation methods is essentially based on the complexity of the imputation process.

3.1 Available case methods

The most commonly applied available case methods for missing data are listwise deletion and pairwise deletion. Both methods require the data to be, at minimum, MAR. Listwise deletion removes all data for a case that has one or more missing values. Deletion methods require the sample to still be representative of the population after the removal of cases and that statistical power be adequate for hypotheses (see Figure 1). Pairwise deletion only removes the specific missing values from the analysis rather than the entire case. This method is rarely recommended, because although pairwise deletion uses all information with each analysis, it does not allow for comparison of analyses because different samples are used in each analysis.

3.2 Single imputation methods

Instead of discarding missing data, single imputation methods replace missing data points with a single fixed value. These naive methods generally assume that the data are MCAR. Single imputation methods typically are a form of constant replacement or regression imputation. There are other less commonly used single imputation methods such as hot deck imputation, imputation from an unconditional distribution, and last observation carried forward,[13] but these methods are rarely recommended or used by most researchers. Constant replacement methods substitute a missing data point with a simple fixed estimate of the missing value, such as the mean or median. Replacing missing data points with a constant can diminish variability of data which, in turn, can lead to biased estimates of error variances and covariances.[18] Regression-based single imputation replaces missing data points with predicted scores based on observed scores on other variables. Although this method is simple to employ, it could distort the underlying distribution. Another downside to single-imputation methods is that if the proportion of missing data are fairly high, then error variance could be significantly miscalculated due to imputing a single value.[17]

3.3 Model-based imputation

Model-based imputation methods replace a missing value with one or more imputed values from a predictive distribu-

tion that models the underlying data loss mechanism. Multiple imputation (MI), full information maximum likelihood (FIML) and maximum likelihood expectation-maximization (EM) imputation are the most common model-based imputation methods. MI uses simulation to produce multiple sets of randomized but plausible data points (as specified by the researcher) that will replace the missing data points. The plausible data sets are then averaged and the resulting data points replace the missing data. This allows for several different values for each missing data point.[18] This is an advantage over mean imputation because it allows for more realistic variances (and errors) over imputation of the same value which reduces variability. For example, using the following ten observations, Original Observations: 2, 3, 5, X, 7, X, 1, X, 8, and 12 (where X denotes missing data)

$$x = 5.42, \sigma = 3.87, \sigma^2 = 14.95 \quad (1)$$

if the missing (X) values were deleted, the already small sample would be reduced by 30%, possibly affecting the researcher's ability to generalize the results (which would already be difficult with 10 responses).

Using mean imputation, the observations would then become: Mean Imputation Observations: 2, 3, 5, 5.42, 7, 5.42, 1, 5.42, 8, and 12

$$x = 5.43, \sigma = 3.16, \sigma^2 = 9.96 \quad (2)$$

The three missing values (X) replaced with the mean (5.42) significantly reduced the variance (σ^2) by approximately 44%, and implies greater measurement precision than actually exists. With MI, the results could look like this: Multiple Imputation Observations: 2, 3, 5, 6.38, 7, 3.91, 1, 10.56, 8, and 12

$$x = 5.89, \sigma = 3.61, \sigma^2 = 13.02 \quad (3)$$

In this example, the resulting mean is slightly higher (8%) and the variance is reduced by 12%, and the standard deviation (σ) remains the same. This example of MI clearly shows an advantage over single imputation because it allowed for greater response variation and more realistic results. However, because each simulated data set is random, the results change with each additional data set making it difficult to replicate these results.

EM is an iterative procedure which uses other variables to impute an expected value, then checks this value against whether that is the most likely value. This procedure in iterative. If the researcher plans on using factor analytic approaches or regression, EM imputations are better than single variable imputations because they preserve the relationships between variables. If more than 5% of data are missing, the standard error can be underestimated by EM.[19] Maximum likelihood methods are typically preferred over MI because they are more efficient, produce consistent results, require less decisions, and there are no model conflicts.[20]

4. RECOMMENDATIONS

Identifying and handling missing data is an integral part of the data cleaning process. Producing a summary table with the number of responses by item is recommended to begin discerning the nature and extent of missing data (*i.e.*, recognizing patterns or systematisms), the next logical step. Most statistical software packages will easily produce a summary table. Detecting MCAR data can be done using Little's MCAR test.[13,16] Detecting MAR data can be a bit more difficult because statistical tests cannot always be relied upon. However, an interaction analysis could be helpful for diagnosing MAR and MNAR data. The presence of statistically significant interaction with observed variables implies that data is MNAR. A priori assumptions, hypotheses, and theory can be useful in guiding the pattern recognition process for missing data. Furthermore, researchers should recognize if speediness (*i.e.*, time limits imposed) could be a factor in the systematism of missing data.

The appropriateness of a missing data method is contextual and depends on the missing data mechanism.[16] MCAR and MAR data can be handled with either case available or imputation methods, if proper conditions exist (see Figure 1). Once data handling methods have been applied, any method of analysis can be used with the resulting data set as if it were complete. Healthcare researchers must recognize the possible effects of data mechanism misclassification for their own research and conclusions (see Table 2).

Sample size issues are also a consideration for healthcare researchers. There is no hard-and-fast proportion of missing data that can be "ignored". The number of missing cases as a proportion of the total number of cases is often more important than the actual number of missing cases, especially in large data sets. The proportion of missing data, as pointed out in the above discussion of EM, can affect standard errors or other distributional measures. In contrast, researchers working with smaller data sets (usually samples less than 100) need to pay close attention to the actual number of missing cases, as even a small number of missing cases could cause problems for analysis and interpretation. As was demonstrated in the listwise deletion, mean imputation, and MI example above, having three (or 30%) of the ten observa-

tions missing resulted in distributional measures (*i.e.*, mean, standard deviation, and variance) that were significantly impacted by the choice of the data handling procedure. Thus, suffice it to say that any amount of missing data can affect the validity and generalizability of results.

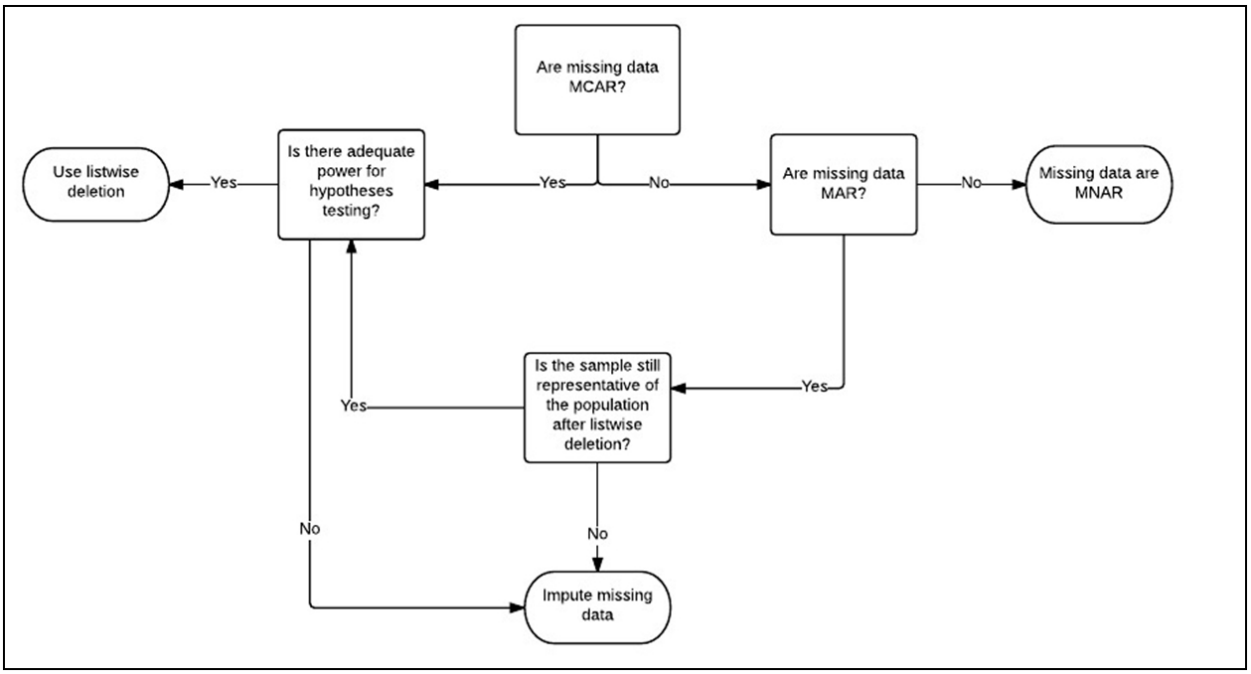

Figure 1. Decision process for missing data imputation and listwise deletion (adapted from Cheema, 2014)

5. Conclusion

In conclusion, handling missing data requires thoughtful consideration by healthcare researchers. Although most statistical programs will readily calculate statistics in the presence of missing data (usually using listwise deletion methods), it is important to recognize the appropriateness of the methods being used and how they can impact the validity and generalizability of results. This issue can become particularly significant when data are missing systematically or the results are biased by case available methods. Healthcare researchers are encouraged to examine the nature of missing data and apply only those methods that are appropriate.

Conflicts of Interest Disclosure

The authors declare no conflict of interest.

References

[1] Kalet A. The state of medical education research. AMA J Ethics. 2007; 9(4): 285-9. http://dx.doi.org/10.1001/virtualmentor.2007.9.4.medu2-0704

[2] AERA. What is education research? Available from: http://www.aera.net/EducationResearch/WhatisEducationResearch/tabid/13453/Default.aspx. Updated 2016. Accessed June/02, 2016.

[3] Walsh K. Medical education: The case for investment. Afr Health Sci. 2014; 14(2): 472-4. PMid: 25320600. http://dx.doi.org/10.4314/ahs.v14i2.26

[4] Dower C. Health policy brief: Graduate medical education. Health Aff. 2012. Available from:http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=73.

[5] Ling L. The basics of GME finance for program directors. Accreditation Council for Graduate Medical Education. 2015. Available from:http://www.acgme.org/Portals/0/PDFs/2015%20AEC/Presentations/PC001/PC001g_Financial.pdf.

[6] Henderson TN. Medicaid direct and indirect graduate medical education payments: A 50-state survey. Association of American Medical Colleges. 2010. Available from:https://members.aamc.org/eweb/upload/Medicaid%20Direct_Indirect%20GME%20Payments%20Survey%202010.pdf.

[7] Puma MJ, Olsen RB, Bell SH, et al. What to do when data are missing in group randomized controlled trials. NCEE 2009-0049. National Center for Education Evaluation and Regional Assistance. Institute of Education Sciences, U.S. Department of Education 2009.

[8] Schmidt FL, Hunter JE, Urry VW. "Statistical power in criterion-related validation studies": Correction to schmidt, hunter, and urry. J Appl Psychol. 1979; 64(1): 65. http://dx.doi.org/10.1037/0021-9010.61.4.473

[9] Roth PL, Campion JE, Jones SD. The impact of four missing data techniques on validity estimates in human resource management. J Bus Psychol. 1996; 11(1): 101-12. http://dx.doi.org/10.1111/j.1744-6570.1994.tb01736.x

[10] Royal KD, Hecker KG. Rater errors in clinical performance assessments. J Vet Med Educ. 2016; 43(1): 5-8. http://dx.doi.org/10.3138/jvme.0715-112R

[11] Royal KD, Guskey TR. Does mathematical precision ensure valid grades? What every veterinary medical educator should know. J Vet Med Educ. 2015; 42(3): 242-4. http://dx.doi.org/10.3138/jvme.0115-005R1

[12] Roth PL. Missing data: A conceptual review for applied psychologists. Pers Psychol. 1994; 47(3): 537-60. http://dx.doi.org/10.1111/j.1744-6570.1994.tb01736.x

[13] Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.

[14] Newgard CD, Lewis RJ. Missing data: How to best account for what is not known. JAMA. 2015; 314(9): 940. http://dx.doi.org/10.1001/jama.2015.10516

[15] Mislevy RJ, Stocking ML. A consumer's guide to LOGIST and BILOG. Appl Psychol Meas. 1989; 13(1): 57-75. http://dx.doi.org/10.1177/014662168901300106

[16] Cheema JR. A review of missing data handling methods in education research. Rev Educ Res. 2014; 84(4): 487. http://dx.doi.org/10.3102/0034654314532697

[17] Vriens M, Melton E. Managing missing data. Market Res. 2002; 14(3).

[18] Gemici S, Bednarz A, Lim P. A primer for handling missing values in the analysis of education and training data. Int J Train Res. 2012; 10(3): 233-50. http://dx.doi.org/10.5172/ijtr.2012.10.3.233

[19] Graham JW. Missing data analysis: Making it work in the real world. Ann Rev Psychol. 2009; 60(1).

[20] Allison PD. Handling missing data by maximum likelihood. SAS Global Forum 2012. Available from: http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.