Continued Medical Education

A simplified eye opener on managing missing data and in evaluation of non-response bias in medical research

PKB Mahesh[1*], Wasantha Gunathunga[2], Mahendra Arnold[1], SB Munasinghe[3], Sinha De Silva[4]

[1]Office of the Regional Director of Health Services, Colombo, Sri Lanka; [2]Department of Community Medicine, Faculty of Medicine, University of Colombo, Sri Lanka; [3]Department of Mathematics, University of Ruhuna, Sri Lanka; [4]Postgraduate Institute of Medicine, University of Colombo, Sri Lanka

Correspondence: buddhikamaheshpk@gmail.com https://orcid.org/0000-0002-9037-5142

DOI: https://doi.org/10.4038/jccpsl.v24i2.8147

Received on: 07 April 2018

Accepted on: 28 June 2018

Abstract

Using correct methods for prevention, analysis and treatment of missing data is essential in preserving the validity of scientific research. In spite of this, issues related to missing data and non-response bias are found to be inadequately discussed in medical research. Facts related to the 'missingness', such as justifying the missing data as missing-completely-at-random, missing-at-random and not-missing-at-random often confuse many of the medical researchers who are non-statisticians. This article focuses on the essential components related to missing data.

Missing data impose serious negative effects on medical research. Yet, this issue has not been given adequate emphasis even in clinical trials (1). The way response rates were calculated, how the analysis was done for non-response bias and how the missing values were treated are not mentioned even in many peer reviewed articles (2). Even though statistically sound literature is available on missing values and non-response, most of those articles focus on segmented aspects. The objective of this article is to discuss all the elements of missing data and their implications on medical research in a simplified manner, so that it can be understood by medical researchers who are non-statisticians.

Missing data and non-response: are they synonymous?

Missing data include facts that are not available but would have been useful if they were available (1). There has been no consensus on the definition of the term 'non-response'. However in literature, many attempts have been made to define the scope of non-response. One commonly used definition is "the degree to which a researcher does not succeed in obtaining the co-operation of all potential respondents" (3). Another explanation depicts that non-response encompasses three facets: non-coverage of units during the sample selection stage, unit-non response during the recruitment stage and item non-response during the data collection stage (3). Unit non-response is when no data is available from a participant. Item-non-response is when unit provides information but only some of the variables are missing (4-6). Hence, though it literally seems that non-response is one sub-unit of missing data, all elements of missing-data are encompassed within the scope of non-response.

Further classifications based on 'missingness' of missing data

Based on the randomness, missing data are classified as; missing completely at random (MCAR), missing at random (MAR) and not missing at random

(NMAR) (5, 7). MCAR means that the missingness is completely unsystematic and that observed data represents a random subset of the hypothetically complete data. As an example, consider a patient in a study moving to another country midway through the study. The missing values are MCAR if the reason for this movement is unrelated to other variables in the study. MAR means although there can be a systematic difference between missing data and observed data, this difference is related to the other variables, but not the underlying values of the incomplete variable. As an example, consider a patient undergoing a test and when the value of the test is above a certain cut-off, he participates in another test. The second test values are MAR, as missingness is entirely determined by the values of the first test. Finally, NMAR means if the missing data is systematically related to the hypothetical values that are missing. In other words, if a systematic difference exists even after adjusting for the observed variables, the missing data are then said to be NMAR (8-9). As an example, consider a study in which blood pressure measurements are among the variables of interest. If some patients do not attend the clinic due to severe symptoms, missing blood pressure values can be assumed as MNAR. When it is due to MCAR and MAR, the missingness is said to be ignorable (10).

Missingness is further elaborated with another example, where a questionnaire is given to a set of parent attendees of a child vaccination clinic. A question is being asked about the satisfaction of clients on the clinic services and several missing values could be found for this response. If there is no difference between the participants whose response is missing and with all participants, it is then MCAR. If the missingness can be explained by the education level and gender of the participants, it is then MAR. If the missingness is dependent on the satisfaction itself (i.e. if unsatisfied clients did not return the questionnaire as an example), it is then NMAR.

Implications of non-response in medical research

In research, response rate is calculated as follows (11):

$$\text{Response rate} = \frac{\text{Number of reporting units from which data was collected}}{\text{Total eligible reporting units}}$$

In other words, it is the proportion who participated in the research out of those who were eligible (12). In medical research, in order to enrol a participant to a study, informed consent is regarded as a must (13). Due to this, study units that were not contacted, who were not able to participate or who refused are included under the category of non-responders (14-15). Response rate is an important indicator of the quality of research (16). Even though there is no consensus, a response rate of 60% has been commonly used as the cut off for a reasonable response rate (12).

Sample size calculation in research studies refer to the number of participants whose data should be available during the data analysis stage (17). Having a correct sample size is needed for reducing type I as well as type II errors in research (18). The power of a study reflects the likelihood of detecting a difference, if such a difference truly exists (19). Furthermore, 'underpowered' studies would have got 'false negative' associations (20). In summary, having a lesser sample size would provide inaccurate findings. On the other hand having an unnecessarily larger sample size would raise ethical and economic issues (21-22). Hence, meticulous calculation of the sample size is necessary. Non-response leads to reduction of the sample size, hence reduction in the accuracy of findings (23). Following the calculation of a sample size, an adjustment is made by dividing it with the response rate (i.e. 1 – non-response rate), to estimate the sample size needed during the data collection stage (17).

Non-response bias occurs when there is a systematic difference between responders and non-responders (24). Non-response bias is a type of systematic error, which would lead to erroneous findings irrespective of the sample size (25). Its magnitude depends on the non-response rate as well as the systematic difference between the responders versus non-responders (4). Non-response bias is not the converse of response bias (24). Response bias is said to occur when there is a systematic difference in the way participants respond (24). The risk of non-response bias may be reflected by the response rate. However, non-response bias is not totally evident from the response rates (26). A response rate of 90% may be due to non-response bias, whereas a response rate of 10% may not be due to it (27). Hence, there is a limited extent by which the non-response bias can be minimized by increasing the response rate (16).

Management of non-response

The usual methods adopted for managing unit non-response include: non-response prevention, analysis of response-non-response behaviour and adjustments

for non-response (14). Many steps can be included in the prevention of non-response (1). These may depend on the study design and research question being studied. Arbitrarily, these steps may be categorized under the headings of: participant related factors, investigators related factors, study-tool related factors and data-collection related factors. Examples such as selection of participants with more potential of being responders and training of investigators are relevant for the first two factors. Clarity and easiness of the study tool are examples for the third. Utilizing comfortable and feasible ways of data collection for participants and anticipation of a feasible response-rate at the beginning are examples for the fourth.

In evaluating the unit non response, methods used for evaluation of the non-response bias include: arranging a follow-up study by contacting initial non-responders, comparing the non-respondents and respondents using the data available in the sampling frame, comparison of survey results with data obtained from other sources, comparison using external data sources and comparison of early versus late respondents (12). Auxiliary variables or data that are available prior to sampling are very much helpful in the non-response evaluation (14, 16). In unit non-response, the missingness due to MCAR can be explored by development of a regression model describing the influence of each independent variable with the status of participation (i.e. being a respondent or a non-respondent) (16). If the regression co-efficients are found to be non-significant, it then points towards MCAR.

Similarly, for item-nonresponse, regression models can be developed using the status of missingness as the indicator variable (i.e. "1" for missing and "0" for not missing) for each variable and regressing on the outcomes (10). Similar to the unit-response mentioned, non-significant co-efficients would reflect MCAR. Based on this principle, tests such as Little's test are used (10, 28).

In compensating for the unit non-response, several linear as well as rank-based weighing techniques have been used in literature. These include using propensity weighing scores, iterative proportional fitting scores and Heckman method (16, 29). Many of these techniques assume that the participatory units have a certain probability of responding, rather than units being straightforward respondents or non-respondents (stochastic rather than non-stochastic) (30).

When it is item non-response, several management strategies can be used (1, 14). These include complete case analysis in which records with missing data are excluded in the form of either 'list-wise' or 'pair-wise' deletions (31). Other strategies include simple imputation methods (i.e. last observation carried forward) and estimating-equation methods in which weighing techniques are used and statistical model methods (i.e. maximum likelihood, Bayesian methods and multiple imputation methods). In imputation, values are assigned for the missing variables and a complete dataset is made available. There are several types of imputation methods such as class-method imputation, regression imputation and multivariate imputation (32). In multiple imputation which is a three-step strategy, multiple plausible values are created for the missing data, several completed datasets are created and subsequently the results are combined (33). Yet, the MAR assumption is made in imputation.

The MAR assumption is a justification of the analysis and not an inherent property of the dataset. As an example, it is justifiable to use MAR assumption, if predictive variables of missing data are included in the imputation models (9). When the missing data is NMAR, the analysis then must include several additional steps. Possible options include using sensitivity analysis with MAR assumption and with NMAR assumption (34).

Several software and associated packages have options for the management of missing values. A few examples include Amelia II, Hmisc, ICE/STATA, IVEware, MICE/STATA, LogXact, SAS PROC MI, S-Plus, SOLAS, R and SPSS33.

In summary, evaluation of any potential non-response bias and utilization of appropriate missing-data treatment methods should receive adequate attention in medical research.

References

1. Little RJ, D'agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* 2012; 367(14): 1355-1360.

2. Bennett C, Khangura S, Brehaut JC, Graham ID, Moher D, Potter BK, Grimshaw JM. Reporting guidelines for survey research: an analysis of published guidance and reporting practices. *PLOS Medicine* 2011; 8(8): e1001069.

3. Barriball KL & While AE. Non-response in survey research: a methodological discussion and development of an explanatory model. *Journal of Advanced Nursing* 1999; 30(3): 677-686.

4. MacDonald SE, Newburn Cook CV, Schopflocher D, Richter S. Addressing nonresponse bias in postal surveys. *Public Health Nursing* 2009; 26(1): 95-105.

5. De Leeuw ED, Hox JJ, Huisman M. Prevention and treatment of item nonresponse. *Journal of Official Statistics* 2003; 19: 153-176.

6. Yan T, Curtin R. The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Res*earch 2010; 22(4): 535-551.

7. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputations. *International Journal of Research & Method in Education* 2016; 39(1): 19-37.

8. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology* 2014; 43(4): 1336-1339.

9. Sterner JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* 2009; 339(7713): 157-160.

10. Pigott TD. A review of methods for missing data. *Educational Research and Evaluation* 2001; 7(4): 353-383.

11. Halpern SD, Asch DA, Shaked A, Stock PG, Blumberg E. Standard definitions: final dispositions of case codes and outcome rates for surveys. *American Journal of Transplantation* 2005; 5(6): 1319-1325.

12. Johnson TP & Wislar JS. Response rates and nonresponse errors in surveys. *JAMA* 2012; 307(17): 1805-1806.

13. Evans JG, Beck P. Informed consent in medical research. *Clinical Medicine* 2002; 2: 267-272.

14. Cobben F. *Nonresponse in sample surveys: methods for analysis and adjustment.* The Hague: Netherlands, 2009.
Available from: http://www.cbs.nl/NR/rdonlyres/2C300D9D-C65D-4B44-B7F3-%0A377BB6CEA066/0/2009x11cobben.pdf.

15. Daly JM, Jones JK, Gereau PL, Levy BT. Nonresponse error in mail surveys: top ten problems. *Nursing Research and Practice* 2011.

16. Wittwer R, Hubrich S. Nonresponse in household surveys: a survey of nonrespondents from the repeated cross-sectional study "mobility in cities–SrV" in Germany. *Transportation Research Procedia* 2015; 11: 66-84.

17. Suresh KP, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences.* 2012; 5(1): 7.

18. Maggard MA, O'connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? *Surgery* 2003; 134(2): 275-279.

19. Ayeni O, Dickson L, Ignacy TA, Thoma A. A systematic review of power and sample size reporting in randomized controlled trials within plastic surgery. *Plastic and Reconstructive Surgery* 2012; 130(1): 78e-86e.

20. Sjögren P & Hedström L. Sample size determination and statistical power in randomized controlled trials. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontics* 2010; 109(5): 652-653.

21. Kelly PJ, Webster AC, Craig JC. How many patients do we need for a clinical trial? Demystifying sample size calculations. *Nephrology* 2010; 15(8): 725-731.

22. Malone HE, Nicholl H, Coyne I. Fundamentals of estimating sample size. *Nurse Researcher* 2016; 23(5): 21.

23. Lahaut VM, Jansen HA, Van de Mheen D, Garretsen HF. Non-response bias in a sample survey on alcohol consumption. *Alcohol and Alcoholism* 2002; 37(3): 256-260.

24. Sedgwick P. Non-response bias versus response bias. *British Medical Journal* 2014; 348.

25. Malone H, Nicholl H, Tracey C. Awareness and minimisation of systematic bias in research. *British Journal of Nursing* 2014; 23(5): 279-282.

26. Groves RM, Couper MP, Presser S, Singer E, Tourangeau R, Acosta GP, Nelson L. Experiments in producing nonresponse bias. *International Journal of Public Opinion Quarterly* 2006; 70(5): 720-736.

27. Halpern SD, Asch DA. Commentary: improving response rates to mailed surveys: what do we learn from randomized controlled trials? *International Journal of Epidemiology* 2003; 32(4): 637-638.

28. Little RJ. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 1988; 83(404): 1198-1202.

29. Sales AE, Plomondon ME, Magid DJ, Spertus JA, Rumsfeld JS. Assessing response bias from missing quality of life data: the Heckman method. *Health and Quality of Life Outcomes* 2004; 2(1): 49.

30. Jones JM. Propensity to respond and nonresponse bias. *Metron* 2008; 66(1): 51-73.

31. Kang H. The prevention and handling of the missing data. *Korean Journal of Anaesthesiol*ogy 2013; 64(5): 402-406.

32. Brick JM, Kalton G. Handling missing data in survey research. *Statistical Methods in Medical Research* 1996; 5(3): 215-238.

33. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 2007; 61(1): 79-90.

34. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputations under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007; 16(3): 259-275.