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

PKB Mahesh1*, Wasantha Gunathunga2, Mahendra Arnold3, SB Munasinghe3, Sinha De Silva4
1Office of the Regional Director of Health Services, Colombo, Sri Lanka; 2Department of Community Medicine, Faculty of Medicine, University of Colombo, Sri Lanka; 3Department of Mathematics, University of Ruhuna, Sri Lanka; 4Postgraduate 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.
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

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 coefficients 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


