ARCHIVES OF HELLENIC MEDICINE: ISSN 11-05-3992

APPLIED MEDICAL RESEARCH
ΕΦΑΡΜΟΣΜΕΝΗ ΙΑΤΡΙΚΗ ΕΡΕΥΝΑ

The problem of missing data in randomized control trials
A quick and easy guide

ARCHIVES OF HELLENIC MEDICINE 2021, 38(5):707–710
ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2021, 38(5):707–710

A. Protopapas,[1]
E. Lambrinou[2]

[1]Department of Health Sciences, School of Sciences, European University Cyprus, Nicosia
[2]Department of Nursing, Faculty of Health Sciences, Cyprus University of Technology, Limassol, Cyprus

Το πρόβλημα των ελλειπουσών τιμών στις τυχαιοποιημένες κλινικές δοκιμές. Ένας ταχύς και εύκολος οδηγός

Περίληψη στο τέλος του άρθρου

Key words

MAR
MCAR
Missing data
MNAR
Multiple imputation

Submitted 23.1.2021
Accepted 8.2.2021

1. INTRODUCTION

Evidence-based research in health care has been developed well in recent years. One of the biggest challenges of the researchers is the management of missing data. Missing data is defined as a data value that is not available, and that if it was observed, it would make a difference to the analysis.[1] Missing data may affect the value of the research findings and scientific information provided.[2] It can reduce the statistical power and introduce bias in the estimation of various parameters. The participants who withdrew or were lost from a study may have different characteristics, and can thus diversify the sample, compared with the completely adherent participants. In such case, the study sample may no longer be representative.[3] For example, the range of some of the participant´s characteristics may be changed, such as age, gender, socioeconomic variables or other measurements. There are several causes of missing data; a few may be due to the study design, and others simply due to missing value,[4] especially in the case of the questionnaire, which is commonly used in studies in the medical and nursing sciences. For instance, older people may avoid answering specific questions related to sexual activities or leisure activities.[5] In addition, according to Alm-Roijer and colleagues, some people may have difficulty in understanding the questions, and for that reason do not provide an answer.[6] In addition, the time available may not be enough for the respondent to complete the questionnaire. Each study has a particular design, so there is no universal method for managing missing data. The management of missing data, however, needs to be approached by considering three aspects before choosing the most appropriate way of analysis, namely (a) the proportion of missing data, (b) the mechanism of the missing data, and (c) the pattern of the missing data.[7]

2. MISSING DATA

2.1. The proportion of missing data: How much data is missing?

The quality of the results is directly related to the percentage of missing data. The validity of studies may be threatened when the proportion of missing data is large. A general rule has been reported which says that a percentage of missing data beyond 20% causes serious problems in the validity of the study. In contrast, missing data of less than 5% are reported to cause only minor problems, but there is no common acceptable percentage of missing data.[7,8] Additionally, this rule may be misleading, because if a study involves 1,000 patients and 200 are missing, then the loss accounts for 20%. Finally, the researchers need to take into consideration that a high drop-out rate has an increased risk of type II error ("false negative" findings), without this meaning that a small proportion of missing data cannot cause a bias.[7] In such cases, no management method can be set up to reduce type II error.[9]

2.2. The mechanism of the missing data

The mechanism of the missing data was first introduced by Rubin in 1976. There are three types of mechanisms of missing data: (a) Missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR).[10] The data are MCAR when the missing observation is not related to the intervention of the study and is irrelevant of any other data, such as age and gender. For example, if the intervention is linked to the assessment of the quality of life (QoL), the missing data are irrelevant to the QoL.[9] Missing data in studies that involve measurement may be due to damage to the equipment.[8,11] As another example, work commitments or new illness may be a reason for participant withdrawal.[12] MAR data are related to the observed variables but not the unobserved variables. For example, a question on sexual activity is less likely to be answered by older people, than younger people, who are usually more sexually active.[5] MNAR data are deliberately absent and depend on the observed and non-observed variables.[11] For example, if a trial applies an intervention to improve the QoL, data may be missing when the participant refuses to answer because of side effects from the intervention. The missing data are therefore not related to the intervention.

2.3. The patterns of the missing data

The patterns of missing data may be more understandable when they are shown in a graphic form. The most common classification of patterns is into monotone and non-monotone patterns. In a monotone pattern, the missing data are distributed in such a way that if a participant is missing a response at one point, then responses will be absent at the subsequent points. The monotone pattern is common in longitudinal studies where participants may have died or left the study. In contrast, non-monotone patterns are more common in clinical studies,[7,13] showing, for example, participants who have left the study and participants who just did not answer a specific question.[13] This may give two different patterns, each needing a different approach to management of the missing data.

3. METHODS OF HANDLING MISSING DATA

Methods for handling missing data can be divided into four categories. The first category concerns the deletion methods. Complete case analysis is a widespread method for the handling of missing data. This method deletes all observations which include missing data, limiting the analysis to only the observations in which a complete data set is present. This method can be used in cases where the missing data are MCAR. Unfortunately, this cannot be considered as the "intention to treat" method. A major disadvantage is the reduction in sample size that decreases the power of the study. In addition, even a small percentage of missing data may cause biased estimates.[8,13]

The second category includes the weighting methods, which may be used along with deletion methods. The missing data and the data observed are weighted differently. Using this method, the bias may be reduced, but the variance is increased, and for this reason, the precision of estimates may be reduced.[13]

The third strategy for handling missing data is by use of single imputation and value replacement methods. The various forms of the single imputation method are mean and median imputation, regression imputation, hot and deck imputation, worst case analysis, and the most popular, the last observation carried forward (LOCF) method.[13] This method forwards the last observed value to fill in the space of the missing data, considering that the result does not change, even if the participant has left the study. This method is not recommended, because it may lead to falsely statistically significant results.[1,8] Gjeilo and colleagues, using the SF-36 questionnaire, replaced the missing data with the mean scores, as recommended in the SF-36 scoring algorithm.[14]

The fourth category includes more advanced methods,

such as maximum likelihood estimation and the multiple imputation. These methods are considered state of the art for handling missing data.[15] They give unbiased estimates of data for MAR and MCAR.[16]

Multiple imputation uses predictive factors of the variables with missing data. It replaces any missing data with the range of plausible predictable values. This method can handle missing data under MAR which have monotone and non-monotone patterns. It can also handle missing data under MNAR after modification.[17] Maximum likelihood gives the estimate of the maximum probability using all the available data (completed and missing). For example, it gives the value of the parameter which, among all the possible values of the parameter, is the most probable based on the specific sample.[16] This method has been used by Olsen and colleagues to reduce the bias of missing data, thus including all study patients for analysis.[18]

When the method of handling missing values has been chosen, sensitivity analysis must be conducted to assess how "strong" the main approach is. Sensitivity analysis evaluates the robustness of the results and assesses how the effects of changes are affected. In other words, it uses various different approaches to evaluate the strength of an assessment, in order to identify the results that are most dependent on questionable or unfounded assumptions.[19]

4. CONCLUSIONS

In summary, in research studies, it is very important to address the issue of missing data, which may lead to limited statistical validity and erroneous conclusions. In clinical trials, if the dropout rate is high, no statistical method may be appropriate for replacing the missing values when following the intention to treat principle. Single imputation methods are best avoided, and advanced methods, such as maximum likelihood estimation and multiple imputation, in combination with sensitivity analysis, are preferable. Authors should report missing values and discuss the causes and the handling methods they have used. Of course, the best way to handle missing data is to prevent them during the design and execution of the study.

ΠΕΡΙΛΗΨΗ

Το πρόβλημα των ελλειπουσών τιμών στις τυχαιοποιημένες κλινικές δοκιμές. Ένας ταχύς και εύκολος οδηγός

Α. ΠΡΩΤΟΠΑΠΑΣ,[1] Ε. ΛΑΜΠΡΙΝΟΥ[2]

[1]Τμήμα Επιστημών Υγείας, Σχολή Θετικών Επιστημών, Ευρωπαϊκό Πανεπιστήμιο Κύπρου, Λευκωσία,
[2]Τμήμα Νοσηλευτικής, Σχολή Επιστημών Υγείας, Τεχνολογικό Πανεπιστήμιο Κύπρου, Λεμεσός, Κύπρος

Αρχεία Ελληνικής Ιατρικής 2021, 38(5):707–710

Ένα από τα πλέον συνήθη προβλήματα που αντιμετωπίζουν οι ερευνητές στις τυχαιοποιημένες κλινικές δοκιμές είναι οι ελλείπουσες τιμές λόγω πρόωρης αποχώρησης ασθενών ή ασθενών που χάθηκαν πριν από την ολοκλήρωση της μελέτης. Τα δεδομένα που λείπουν ενδέχεται να μειώσουν τη στατιστική ισχύ και να προκαλέσουν σφάλμα στην εκτίμηση των αποτελεσμάτων, ενώ μπορεί επίσης να επηρεάσουν τη συνολική στατιστική εγκυρότητα της μελέτης, οδηγώντας τους ερευνητές σε εσφαλμένες εκτιμήσεις γενικότερα. Για την αντιμετώπιση των δεδομένων που λείπουν θα πρέπει να προσεγγιστούν από τρεις πτυχές πριν από την επιλογή της κατάλληλης μεθόδου επίλυσης: (α) το ποσοστό των δεδομένων τα οποία λείπουν, (β) ο μηχανισμός των δεδομένων που λείπουν και (γ) το μοτίβο των δεδομένων τα οποία λείπουν. Όσον αφορά στην επίλυση του εν λόγω προβλήματος, προτιμώνται προηγμένες μέθοδοι, όπως η μέγιστη πιθανοφάνεια και ο πολλαπλός καταλογισμός, σε συνδυασμό με την ανάλυση ευαισθησίας. Δεν υπάρχει κάποια καθολική μέθοδος για τη διαχείριση των δεδομένων που λείπουν, αλλά υπάρχουν πολλές δημοφιλείς μέθοδοι οι οποίες είναι ανεπαρκείς και οι ερευνητές θα πρέπει να τις γνωρίζουν.

Λέξεις ευρετηρίου: Ελλείπουσες τιμές, MAR, MCAR, MNAR, Πολλαπλός καταλογισμός

References

1. LITTLE RJ, D'AGOSTINO R, COHEN ML, DICKERSIN K, EMERSON SS, FARRAR JT ET AL. The prevention and treatment of missing data in clinical trials. *N Engl J Med* 2012, 367:1355–1360
2. ROBERTS MB, SULLIVAN MC, WINCHESTER SB. Examining solutions to missing data in longitudinal nursing research. *J Spec Pediatr Nurs* 2017, 22: 10.1111/jspn.12179
3. KANG H. The prevention and handling of the missing data. *Korean J Anesthesiol* 2013, 64:402–406
4. HORTON NJ, KLEINMAN KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007, 61:79–90
5. ÅRESTEDT K, SAVEMAN BI, JOHANSSON P, BLOMQVIST K. Social support and its association with health-related quality of life among older patients with chronic heart failure. *Eur J Cardiovasc Nurs* 2013, 12:69–77
6. ALM-ROIJER C, STAGMO M, UDÉN G, ERHARDT L. Better knowledge improves adherence to lifestyle changes and medication in patients with coronary heart disease. *Eur J Cardiovasc Nurs* 2004, 3:321–330
7. DONG Y, PENG CYJ. Principled missing data methods for researchers. *Springerplus* 2013, 2:222
8. DZIURA JD, POST LA, ZHAO Q, FU Z, PEDUZZI P. Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale J Biol Med* 2013, 86:343–358
9. ARMIJO-OLIVO S, WARREN S, MAGEE DJ. Intention to treat analysis, compliance, drop-outs and how to deal with missing data in clinical research: A review. *Phys Ther Rev* 2009, 14:36–49
10. RUBIN DB. Inference and missing data. *Biometrika* 1976, 63:581–592
11. UENAL H, MAYER B, DU PREL JB. Choosing appropriate methods for missing data in medical research: A decision algorithm on methods for missing data. *J Appl Quant Methods* 2014, 9:10–21
12. ALHARBI M, GALLAGHER R, KIRKNESS A, SIBBRITT D, TOFLER G. Long-term outcomes from Healthy Eating and Exercise Lifestyle Program for overweight people with heart disease and diabetes. *Eur J Cardiovasc Nurs* 2016, 15:91–99
13. HAUKOOS JS, NEWGARD CD. Advanced statistics: Missing data in clinical research – part 1: An introduction and conceptual framework. *Acad Emerg Med* 2007, 14:662–668
14. GJEILO KH, WAHBA A, KLEPSTAD P, LYDERSEN S, STENSETH R. Recovery patterns and health-related quality of life in older patients undergoing cardiac surgery: A prospective study. *Eur J Cardiovasc Nurs* 2012, 11:322–330
15. SCHAFER JL, GRAHAM JW. Missing data: Our view of the state of the art. *Psychol Methods* 2002, 7:147–177
16. BARALDI AN, ENDERS CK. An introduction to modern missing data analyses. *J Sch Psychol* 2010, 48:5–37
17. FARIA R, GOMES M, EPSTEIN D, WHITE IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics* 2014, 32:1157–1170
18. OLSEN SJ, FRIDLUND B, EIDE LS, HUFTHAMMER KO, KUIPER KK, NORDREHAUG JE ET AL. Changes in self-reported health and quality of life in octogenarian patients one month after transcatheter aortic valve implantation. *Eur J Cardiovasc Nurs* 2017, 16:79–87
19. THABANE L, MBUAGBAW L, ZHANG S, SAMAAN Z, MARCUCCI M, YE C ET AL. A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Med Res Methodol* 2013, 13:92

Corresponding author:

A. Protopapas, Department of Health Sciences, School of Sciences, European University Cyprus, 6 Diogenous street, 2404 Nicosia, Cyprus
e-mail: a.protopapas@euc.ac.cy