Research in Social and Administrative Pharmacy 18 (2022) 2308–2316

Contents lists available at ScienceDirect

Research in Social and Administrative Pharmacy

journal homepage: www.elsevier.com/locate/rsap

Missing data in surveys: Key concepts, approaches, and applications

Ardalan Mirzaei *, Stephen R. Carter, Asad E. Patanwala, Carl R. Schneider

School of Pharmacy, Faculty of Medicine and Health, University of Sydney, Australia

ARTICLE INFO

Keywords:
Missing data
Research design
Questionnaire design
Research methods
Surveys

ABSTRACT

A recent review of missing data in pharmacy literature has highlighted that a low proportion of studies reported how missing data was handled. In this paper we discuss the concept of missing data in survey research, how missing data is classified, common techniques to account for missingness and how to report on missing data.

The paper provides guidance to mitigate the occurrence of missing data through planning. Considerations include estimating expected missing data, intended vs unintended missing data, survey length, working with electronic surveys, choosing between standard and filtered form questions, forced responses and straight-lining, as well as responses that can generate missingness like "I don't know" and "Not Applicable".

We introduce methods for analysing data with missing values, such as deletion, imputation and likelihood methods. The manuscript provides a framework and flow chart for choosing the appropriate analysis method based on how much missing data is observed and the type of missingness.

Special circumstances involving missing data have been discussed, such as in studies with repeated or cohort measures, factor analysis or as part of data integration. Finally, a checklist of questions are provided for researchers to guide the reporting of the missing data when conducting future research.

Introduction

> *There are two types of people: 1) Those who can extrapolate from missing data.*
>
> - A random T-Shirt I saw.

What is missing data?

Missing data is when an observation has no value assigned to it. For any particular data set, missing data is present in cases where, for any item, an input has not been entered or generated. In surveys, a respondents' response value is not available for it to be taken further for analysis.

There are multiple reasons why surveys can have missing data. For example, respondents may have skipped questions, data encoding caused variables to be counted as null or missing, the internet may have cut out during data gathering with electronic devices, a page of printed information may be missing, or a response item is deemed invalid.

Whether intended or unintended, classifications for missing data have been developed to describe the type of missingness. Classifying missing responses allows for decisions to be made on how to handle missing data and when reporting, how to inform readers of the considerations that were taken to mitigate or minimise missing values.

A recent review into missing data in pharmacy literature highlighted that a low proportion of studies reported on how missing data was handled.[1] A lack of reporting can lead to bias in the interpretation of findings and validity of the research. The aim of this paper is to introduce the concept of missing data, how missing data is categorized as well as introduce common techniques to account for and report on missing data.

Classification of missing data

Before a decision could be made about what to do with the missing data, the type of missingness needs to be characterised. Consideration of missing data requires both subjective and objective analyses. Missing data may be classified according to the degree of randomness with three categories described; Missing at Random (MAR), Missing Completely at Random (MCAR) or ignorable missingness, and, Missing Not at Random (MNAR), also known as non-ignorable missingness.[2,3]

Missing Completely at random

MCAR is when a missing value is not related to any other value in the

* Corresponding author. Room N411, Pharmacy and Bank Building A15, The University of Sydney, Sydney, 2006, NSW, Australia.
E-mail address: ardalan.mirzaei@sydney.edu.au (A. Mirzaei).

https://doi.org/10.1016/j.sapharm.2021.03.009
Received 19 November 2020; Received in revised form 26 February 2021; Accepted 11 March 2021
Available online 19 March 2021

data set.[4,5] Conceptually, data that are MCAR are not usually attributed to a question in the survey or other phenomenon, whether observable or unobservable. Assume for example, a question being asked relates to income and is represented by the letter X_1, while another question relates to occupation and is represented by the letter X_2. In MCAR, the reason for X_1 (income) having a missing response is not because of X_1 (income) or X_2 (occupation) i.e., neither the survey question, nor another confounder is the reason for the missing value. When MCAR is suspected, Little's Test of Missingness can be used to determine whether the missing values meet the specification of MCAR.[6] A significant p-value result indicates that we reject the null hypothesis and assume that a pattern exists to the missing data (not MCAR). Little's Test of Missingness is available in most statistical software packages, either as a direct test or via a macro.

Missing at random

Data that are MAR are missing based on another observable instance, such as an underlying or confounding factor causing respondents to not answer questions. Certain groups may not respond to a question, as a result of an underlying reason. For instance, individuals with high paying jobs may not be inclined to answer questions that relate to finance. This is both theoretically and conceptually true, as research indicates that higher income earners are more likely non-responders of income questions.[7] Using the example from MCAR above where X_1 is income and X_2 is occupation. The reasons why X_1 (income) may not be reported is based on X_2 (occupation), where those with higher paying occupations are less inclined to provide a response.[8] Thus, in the case of MAR, the reason for X_1 having a missing response is based on X_2, another variable.

Missing not at random

Finally, MNAR, or data that contains non-ignorable missingness, are data that do not meet the criteria of either MCAR or MAR. Unlike MCAR and the use of an objective statistical test, subjective analysis is required to ascertain whether data are MNAR. In MAR, there may be a correlation between an observable phenomenon and why data are missing, but not a direct cause. Data that are MNAR, on the other hand, can be attributed to an unobservable factor that is directly affecting the reason that the data values are missing. This can be the question itself being the cause of the missing response, or underlying assumptions.[5] Using another example in a survey of overall health, assume X_1 is a depression related question and X_2 is gender. X_1 (depression) can have a missing response based on X_2 (gender) where men are less likely to talk about depression. This case would be MAR. On the other hand, if it is the level of depression, X_1, that is causing the person to provide a null response, then the missingness is MNAR. This is where the cause of the missingness is the phenomenon that is being evaluated by the item itself, which in this case is X_1.

To summarise the three categories, assume X_1 is the variable with missing responses and X_2 is another variable:

MCAR = Neither X_1 nor X_2, can explain the missingness. Mathematically from Little's Test, "No pattern exists."

MAR = Missingness of X_1 is based on X_2, where X_2 is another variable in the dataset

MNAR = Missingness of X_1 is based on X_1 itself or another phenomenon that is rarely observed. Cannot be attributed to another observable dataset variable

Planning for missing data

Ideally, consideration of how to avoid missing data should be part of the initial survey design, sampling strategy, as well as the data analysis plan. Estimation of the proportion of missing data may be inferred from literature as well as pilot studies. The estimated proportion of missing data obtained allows for improved survey sample calculation.

If participants forget to answer a question or refuse to answer a question, then that information will not be collected. Missing by design is when the researcher has caused a missing value.[9,10] This can happen when a response was provided, but the response was converted to a missing value. In attempting to minimise missing data, some researchers have the misconception "If I force them to answer, then there won't be anything missing." Some surveys are used to gather the perceptions or opinions of participants. In order to form an opinion, you need to have awareness, such as experience.[11] Experiences are especially needed in order to gather participants' perceptions. Participants of a survey might have a broad understanding of a topic, but when you are measuring their perceptions, it might be useful to consider "do they have awareness". In other cases, if the respondent is forced to answer they may display a behaviour known as straight-lining.[12] Straight-lining is when respondents answer identical or a similar response in order to finish a survey. Their provided answers may not be useable as it is not providing a response to the question asked and affected responses are coded as missing.

The number of questions and anticipated time required for survey completion can also influence the amount of data that is missing. This is due to question fatigue or loss of interest by participants with longer surveys. The appropriate survey length can vary, based on circumstances such as whether participants are reimbursed for their time,[13] the health of participants, or environmental factors. Although several studies have demonstrated survey length impacts on response rates, this effect is inconsistent and not demonstrated for all surveys.[14] The planning phase of the survey should balance the need for comprehensiveness versus the risk of reduced participant response, which could result in missing data. In this context, a longer survey in some circumstances could lead to less usable information.

Standard form and filtered form questions

What is to be done when a respondent does not respond to a question? We may ask the standard form question, "Diabetes has a genetic link. Do you agree or disagree", and wait for a response of either "Agree" or "Disagree". However, how is a response to be recorded if a respondent answers "I don't know"? This form of response is a volunteered don't know response.[15] These responses can be seen in telephone or paper based surveys where the respondent and the interviewer can converse.

An alternate form of questioning is the filtered form of questioning. In this example, we may first ask the participants "Diabetes has a genetic link. Do you have an opinion about that?" To which if they respond "Yes" it can then be asked "Do you agree or disagree?" Schuman and Presser demonstrate that differences in responses can be gathered by using either standard form questions and filtered forms.[15] Therefore, if persons have no opinion about a topic, it is inappropriate to force them to respond to a question. It should be noted that filtered form and standard form questions provide different results.[15] In interview based surveys, the interviewer has the ability to ask filtered form questions, however this is difficult in self-administered paper based surveys. An alternative is to use self-administered surveys with the option of branching.

Electronic surveys

Causes of MCAR can arise from participants skipping or forgetting to complete a question. This missingness commonly occurs in a paper-

based survey or in electronic surveys when forced responses[1] before progression is not implemented.[2]

In electronic surveys, we can force each question to be answered before allowing participants to progress through the survey. In paper-based surveys or telephone surveys, they may choose to say, "I don't know" or provide no response. Forcing the participant to have an opinion about a topic where the person cannot form an opinion about may lead to a biased or inaccurate data. Branching to ask filtered form questions in an electronic format is an ideal way to overcome this issue. However, if standard form questions are required for the survey, then consideration should be made to providing other option(s).

Response options that can lead to missingness

In interviewer led questionnaires, item non-responses from participants can be encoding by the interviewer according to a pre-defined data dictionary.[16] However, in self-administered surveys, it is sometimes difficult to specify a non-response. Durand and Lambert suggest the use of "Don't know" options to allow for participants to have "less guessing, reductions in the threatening nature of items, and decreased embarrassment due to erroneous answers to knowledge- oriented items".[17] When providing participants with the option of "I don't know" or "N/A", it is unclear in the cases of ordinal, interval or ratio type data, how such responses should be valued. For example, in a 7-point Likert-type response scale, is "I don't know" coded as a zero or an eight? In either case, the scale changes from 7-point to an 8-point scale. The midpoint is consequently shifted from 3 to 3.5, presenting problems about what the value of "3" represents. If the "neutral" response was coded as "3", then we have artificially shifted its location, without it having any true value to "neutral".

Some researchers encode "I don't know" as a neutral response. Assuming a 7–point bidirectional scale with positive and negative ends, "neutral" would be in the middle. However, encoding an "I don't know" response as a "neutral" response would be inappropriate as they have different meaning. Imagine we are exploring the effect of the direct-to-consumer advertising (DTCA) on perceptions of a drug in different populations. If the question was "TV drug advertisements make me nervous", and the 7-point scale had the option of "Neutral" and an "I don't know". In countries with DCTA we would see a mix of responses as some may have seen an advertisement and have no affective response from it (neutral), whereas others may not have seen an TV advertisement and have no affective response (I don't know). In countries without DTCA, it would be expected to observe a greater number of "I don't know" responses as DTCA is not allowed. Having no opinion about a topic that can be conceptualised as different to not being able to conceptualise it at all.

Therefore, having the option of "I don't know" allows the researcher to explore whether the question is conceptualised by the respondents. The values of "I don't know" could then be converted to a missing value, as a mathematical value for that type of response cannot exist.

A respondent selected response is a deliberate response despite being converted to a missing response for analysis. In the case of where the 'absence of a response' is a 'response', it suggests the data may be MNAR. Therefore, consumer selected null responses can be further examined to determine what value they may have. In surveys, many have the option of "I don't know" or "N/A".

[1] Forced responses in electronic surveys is a parameter which requires every question to be answered before the participant can progress onto the next part, or before saving the data. This presents a problem as it forces a participant to answer a question. However, it has the advantage as it can create a safety check to ensure the questions are being answered and minimise your missing data.

Analysing data with missing values

Data can be missing from two levels, either the variable (item) level or the case (individual) level.[18] The item level non-response is where the data for a particular item is missing for a very high proportion of participants. For example, most respondents may have answered the whole survey, except that many have missed all the items regarding income, particularly if administered in a cohort of high wealth individuals. A non-response on the case level is where information pertaining to a particular respondent is missing. For example, a respondent may have missed a series of questions, but completed the rest of the survey. The information from both levels is taken into consideration to assess the total missing data in the survey and to decide on the analytical approach.

Analytical methods to handle missing data are commonly available in statistical software. It is up to the researcher to understand and use the missing information appropriately. The purpose of this paper is not to restate the existing principles and methods used in missing data analysis, but to guide the initial consideration that need to be made when planning the analyses of data which include missingness. Table 1 reports some of the missing data handling methods. As statistical software has become more accessible, there has been a shift to using imputation and likelihood methods over traditional deletion methods when handling missing data.[19]

Deletion methods

The traditional approach when handling missing values is to "exclude by listwise" or Complete Case Analysis (CCA). CCA is when the entire case that contains any missing data is removed from analysis. It does not matter if all the other items are answered completely, when one of the items for your analysis has a missing value, the entire case is removed prior to any analysis which would include that item. An easy method to apply, however, there is the possibility of introducing bias if the included individuals vary from those excluded.[21] To address bias, a weighted CCA[4] can be considered, with methods such as inverse probability weighting,[21,22] calibration[23] or propensity weighting for nonresponse.[24]

An alternative option is to "exclude by pairwise" or Available Case Analysis. Pairwise deletion does not delete the whole case like CCA, but takes into account the variables that are missing with those that are not missing.[25] For example, if we have 3 variables and each has different number of missing cases, pairwise deletion will take into account variables as 3 groups. Group A will have variables 1 and 2 together, Group B will have variables 2 and 3 together and Group C will have variables 1 and 3 together. Therefore, assuming each variable had 4 unique cases of missingness that means each group will have 8 unique cases of missingness. This contrasts with listwise deletion (CCA), as all cases with a missing value are removed from analysis, meaning each group has 12

Table 1
Possible methods to handling missing data (adapted from Bennett 2001).[20]

Category	Method
Deletion Methods	Complete Case Analysis (CCA)/Listwise Deletion
	Available Case Analysis (ACA)/Pairwise Deletion
Single Imputation	Mean Imputation
	Last Value/Observation Carried Forward (LVFC/LOFC)
	Regression Methods (RM)
	Hot-Deck Imputation
	Cold-Deck Imputation
Multiple Imputation	Multiple Imputation (MI)
Other	Markov-Chain Imputation
	Missing-Indicator Method
Likelihood Methods	Expectation-Maximisation Algorithm
	Full Information Maximum Likelihood
Indicator Methods	Indicator Method Imputation
	Pattern Mixture Models

cases removed. Pairwise deletion or exclusion can also be referred to as available-case analysis or excluding cases analysis-by-analysis. However, in the discipline of psychology, the American Psychological Association (APA) discourages using pairwise and listwise deletion.[26,27]

Imputation methods

Imputation is the process of replacing missing data with substituted values. The option to "replace by mean" or "mean substitution", is a single imputation method that imputes the mean value of the responses for that item into the missing value field. However, depending on the type and form of missing data, this imputation method may or may not be appropriate.

Multiple imputation, as compared to single imputation approaches, estimates a few possible options for the missing value. The number of possible options is selected by the researcher. For example, the researcher may choose a number 'M', and through multiple imputation methods, M possible values are provided to create M possible and complete datasets. These M datasets are then used together to perform statistical analysis and provide a single summary finding.[20]

Hair et al. also discusses other imputation methods such as case substitution, cold deck imputation, and regression imputation.[28]

Likelihood methods

Likelihood methods may also be used when handling missing data from surveys.[29,30] Expectation-Maximisation (EM)[31] and Full Information Maximum likelihood (FIML) are the most widely used of the likelihood methods. These are model-based methods for estimating parameters in the presence of missing data, without requiring the use of prior imputation. Parameters that are required, yet missing, are estimated by maximizing the likely value based on all other observed parameters in the model. The EM algorithm is available in many programs. FIML does require specialised software and it should be noted that FIML assumes multivariate normality, which is not typical of 5-point Likert-type responses used in many surveys. Both methods are simpler to implement than imputation methods and FIML in particular is less susceptible to bias, particularly for latent variable analyses. Extensions of certain FIML programs use robust statistics to manage excessive multivariate kurtosis. FIML is available in commercial products including SAS, STATA MPLUS (https://www.statmodel.com/), AMOS (https://www.statisticssolutions.com/amos/and EQS (http://www.mvsoft.com/eqs60.htm) and also in R.

Choice of method

The percentage of missing data from the item level, the case level and complete survey dictates the different techniques used. It is difficult to have a rule of thumb for missing data. As Little and Rubin mention the "degree of bias and loss of precision depends not only on the fraction of complete cases and pattern of missing data, but also on the extent to which complete and incomplete cases differ, and on the parameters of interest."[4] It can, however, benefit researchers to use an initial framework approach handling missing data, until they have gained sufficient experience to justify choices based on theoretical and empirical foundations. As such, we have provided an example framework to approaching missing data in surveys. A possible consideration is that if there is <5% missing data, then the technique of multiple imputation may not provide much benefit, and thus using a simpler single imputation approach may be appropriate.[26] However, if >10% of the data is missing, then there is more likely to be bias and as such multiple imputation techniques can be used.[20] Once there is >40% missing data, then imputation or likelihood methods can lead to results that are no more than hypothesis generating. Between 5% and 10% missingness is a grey area where the researcher should use a theoretical consideration of the phenomenon of interest, before deciding to delete listwise, impute or use likelihood methods.

Once the type of missing data is ascertained, a decision needs to be made about how to deal with those missing observations. The three different levels of missingness that exist can help determine how to approach the handling of the data. If it can be determined and confirmed that the data is missing under MCAR, then imputations or deletions can be performed with minimal bias.[32] Thinking about the cause of why the data is missing, under MCAR, a respondent may not have answered the question, but this may be either an isolated case or mishandled error in collecting data. It does not reflect the nature of the question being asked. Assuming MCAR means that the missing data is a random sample of the complete data. The traditional approach is to perform a CCA. Removing the data in MCAR situations does not introduce bias, however, it does increase the standard error due to the reduced sample size.[18] The alternative approach would be to perform an imputation in order to estimate a response that the respondent may have answered, had they answered the question.

Handling data that is MAR can generally be imputed similarly to MCAR, but the procedures are more complex. As mentioned above, MAR data is categorized as MAR when other variables are related to the missing information. Therefore, in order to apply imputation techniques those other variables have to be factored into the modelling process. Likelihood methods, particularly FIML can work well with MAR data.[33]

Missing data that is MNAR cannot appropriately be imputed and FIML is not appropriate. As the data that is missing depends on the missingness itself, it becomes difficult to use any other information to impute or infer the value. Most imputation methods result in biased results when dealing with MNAR. Suggestions for handling MNAR data are to utilise questions in order to find a related factor, assess the theoretical consideration for using the question or variable, or consider dropping the variables when performing analysis. Fig. 1 shows the decision process and possible options that could be taken before running analysis. An example in handling missing data in surveys is provided in the supplementary material.

Missing data in special circumstances

Missing data in repeated measures

Of important mention is the missingness in trials and longitudinal studies. During the data gathering phase of these studies, loss of follow up can lead to missing entries. In clinical trials patients may be withdrawn due to side effects or alternative treatments provided. At other times, patient dropout, with no indication of the cause, and this can also lead to missing values. Clinical trials papers and discussions on preventing and handling the missing values have been written at length.[34] Probably due to the nature of clinical trials and their purpose of providing demonstration of effect, the strategies to handle and prevent missingness and the influence of missing values are well considered. As parting of integrating cohorts for epidemiological studies of meta-analysis, Rajula et al. suggests using deletion methods (CCA), single imputation methods, multiple imputation methods or missing indicator methods.[35]

Missing data in factor analysis

Within the discipline of social pharmacy research, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are frequently conducted and missingness is a concern because factor analysis is often performed at the lower end of sample size recommendations. Even though maximum likelihood with list-wise deletion will lead to unbiased factor loadings when the data is MCAR, the representativeness of data is drastically reduced. While FIML methods maximise information from small datasets, Nassiri et al. found that FIML can be adversely affected by convergence problems.[36] McNeish recommends a range of options for small datasets including a specialised two-stage approach, or using

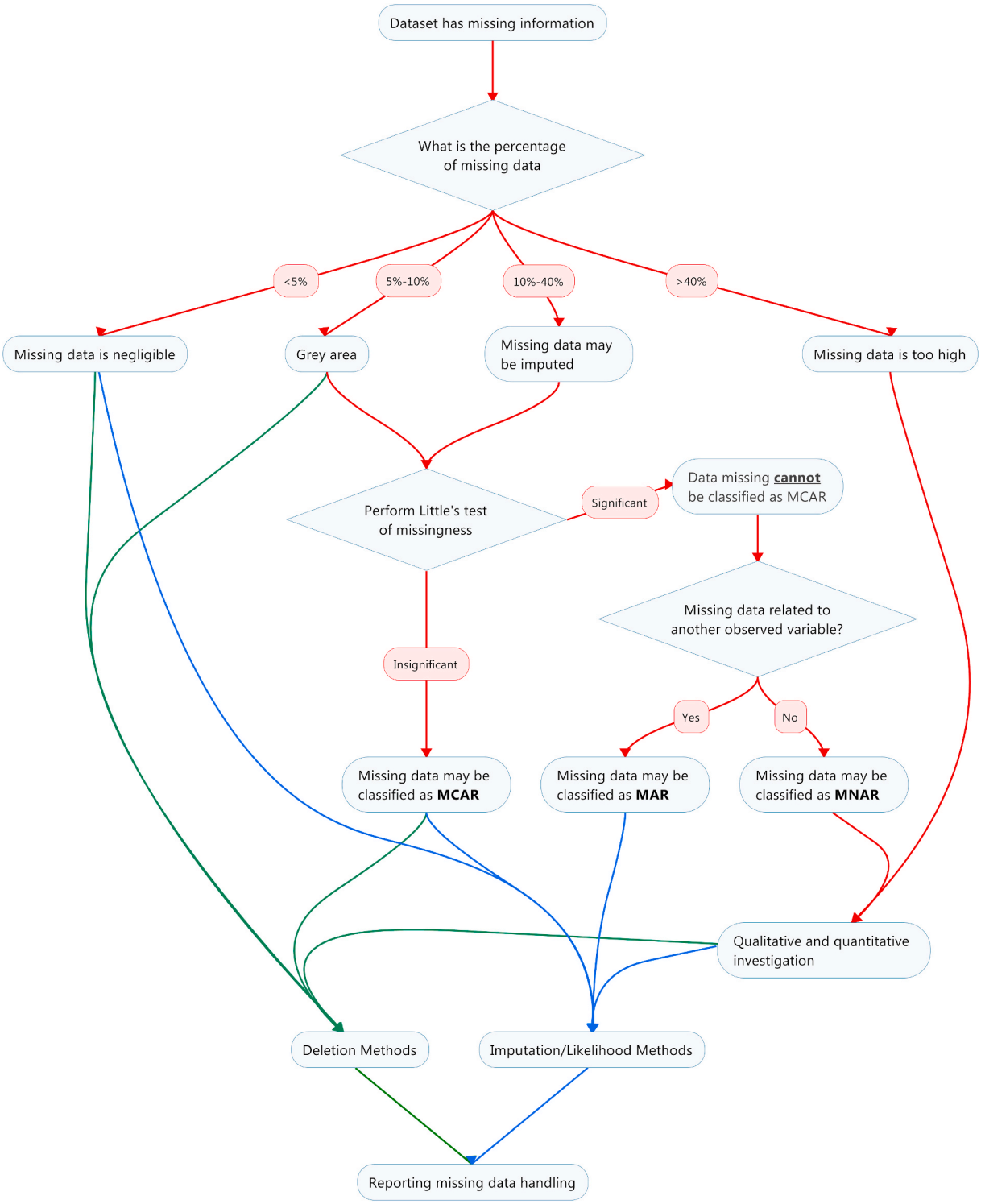

Fig. 1. Algorithm for handling missing data.

multiple imputation prior to performing EFA.[37]

Missing data as part of data integration

Data integration is the process of collecting various datasets and creating a combined dataset.[38] In linking datasets, item level information may be missing for some cases, or may not merge correctly. In building larger datasets by combining cross-sectional surveys, there is the possibility that items were not repeated in surveys, or their label coded names have changed. This creates an issue of missing items. Again, this results in missing data at both the item and case level, with missing data handling required to perform data analysis. An approach to handling missing data as part of record linkage has been explored by Fienberg and Manrique-Vallier.[39] In their paper they describe Baker's work with breast cancer[40] and the use of the Expectation-Maximisation (EM) algorithm. In the field of omics, multiple imputation[41] and Bayesian methods[42] have been used to handle the missing data.

Reporting on missing data

With multiple approaches available, it is important that the method/s of handling missingness be reported. Many fields require specifics in reporting on missing data, and this is left up to the researcher to determine. However, most readers would be interested in the responses to the following questions related to the missingness:

- What is the percentage of missing values?
- How did the missingness develop? Was it respondent allocated, administration related or other?
- How the missingness was classified (MCAR, MAR, MNAR)?
- Was item-level deletion performed and what were the criteria for deletion?
- Was a list-wise or pair-wise deletion performed?
- Was imputation performed? How?
- Was a likelihood method performed? Which method?
- What was the justification for the methods applied?

Missing data reporting should be presented in the results section. Tables reporting values can report the actual number (n) as well as those that are missing. An alternate approach is to report missingness and handling of missing data in the supplementary section for sake of conciseness.

As an example of reporting, the paper from Mirzaei et al., 2019 reports[43] missingness in a study conducting EFA as follows:

> "*Therefore, a strategy was used to reduce the impact of missingness in this study. Items where more than 10% of the respondents answered 'I don't know', were removed from the analysis. The rationale for this was that these items were likely to be poorly conceptualised by respondents and would not be a valid measure of service quality. As such 7 items were excluded from the analysis with the entire 'special services' sub-dimension requiring removal. Missing data were then handled using list-wise deletion.*"

Summary

Missing data needs to be considered throughout the course of survey-based research, from planning through to reporting. This paper has introduced multiple approaches for handling missing survey data and presented a guide for when these approaches should be used. It is essential to consider and report on missing data to accurately report the findings of a survey study.

Authorship statement

Ardalan Mirzaei: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. Stephen R Carter: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision. Asad E Patanwala: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. Carl R Schneider: Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision.

Acknowledgements

We would like to thank Dr Jack Collins for his initial read of the manuscript.

Appendices.

Example: Missing Data in Surveys

The methods and data followed that of the study by Mirzaei et al., 2019. In the study, the authors performed scale validation by subjecting a 61-item questionnaire for psychometric testing. Respondents were consumers waiting for their prescriptions in an Australian metropolitan pharmacy with a Price-Focused Marketing Strategy (PFMS).[43] Respondents completed the questionnaire online using a tablet computer, with forced responses enabled. The items were on a 7-point Likert scale with the option of "I Don't Know". The requirement to include only respondents who had previously obtained medicines from the pharmacy was designed to ensure that respondents had sufficient experience to evaluate elements of service quality.

PFMS pharmacies focus their energy on marketing on price-competitiveness rather than on the provision of services.[44] Therefore, the additional special services that they may offer may not be as recognizable to the participants.

408 participants completed the 61-item scale of which 49 were to be used for analysis. Visualising missing data from the item and case level through heat maps assists with the decision-making process. Initial case level inspection showed that 29 out of the 408 cases had more than 40% missing contributing to about 5% of the missing data overall. Investigation showed that these cases had lost internet connectivity, see Fig. 2. Five other cases also had the same issue and had 20% missing data. Therefore 34 removed from analysis through a complete case analysis.

As part factor analysis and Likert scales, the "I don't know" response was converted to a missing value. A heat map as seen in Fig. 3 showed the spread of missing values. Examination from the item level reported missing data for all the items. A 51% missingness was reported for one item. On closer inspection, 7 items had greater than 10% missing values. A complete case analysis this time would result in only 143 useable cases. Not enough to adequately perform factor analysis.

The item with >40% missingness was removed, and the remaining missing data was used to decide whether to impute. Little's test resulted in a significant value therefore the missingness was not MCAR. For our missing data to be MAR, items that had missing values would be related to other variables. In our dataset, items with a high missing variable demonstrated a conceptual relationship. For example, all items referring to pharmacy special services contained 20% or more missing values. As much as it is important to know that the missing variables are related to another variable, it is does not help that variable also has missing data. Therefore, the missing variables themselves are part of the problem. As no other complete variable could be conceptualised to be related to the items, the data was neither MCAR nor MAR, following, the algorithm above, the data was concluded to be MNAR. Therefore, excluding the items with 10% missingness allowed for a useable complete dataset of 374, with a 230 complete cases (see Fig. 4), sufficient to perform factor analysis.

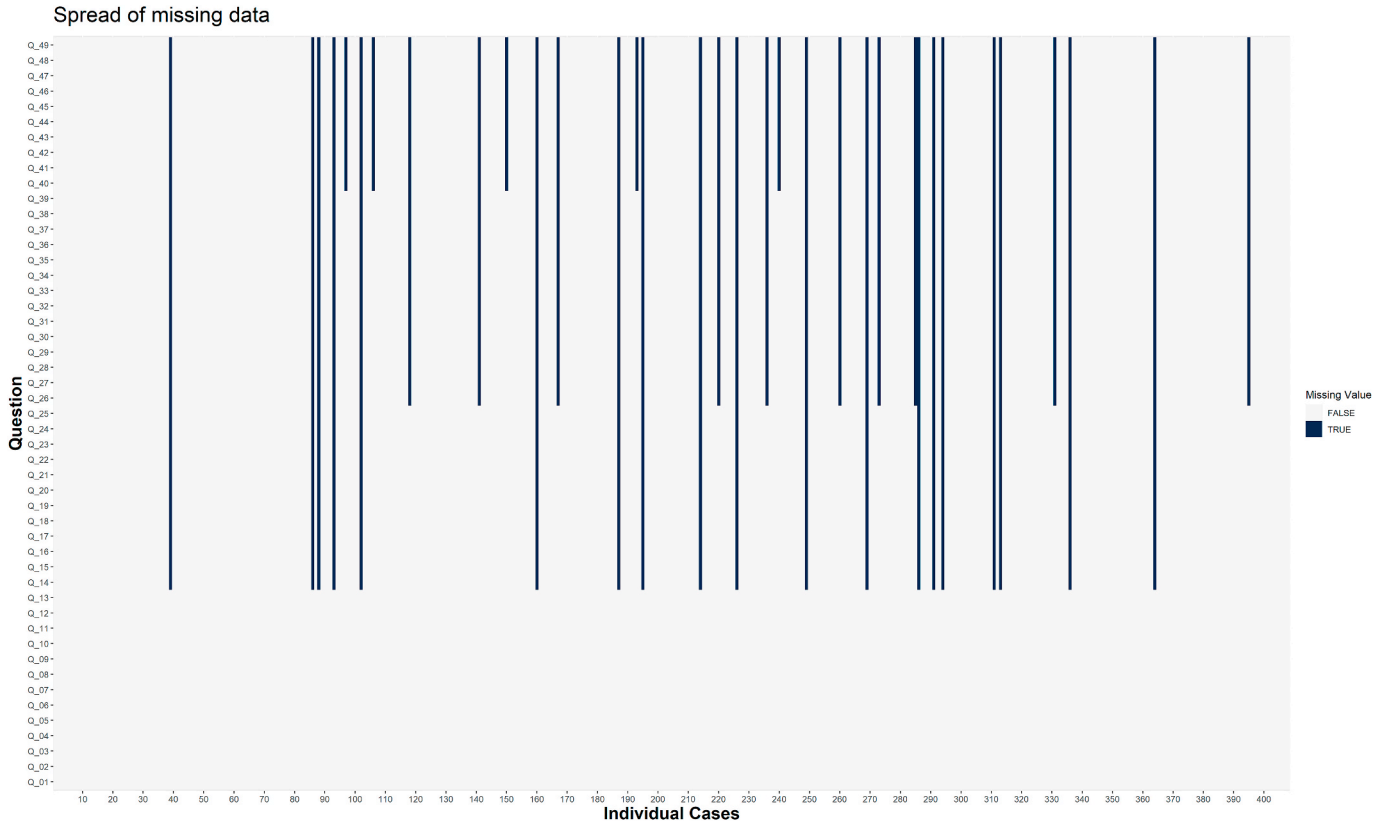

Fig. 2. Spread of missing data at the beginning of the survey. Dark colours represent missing values.

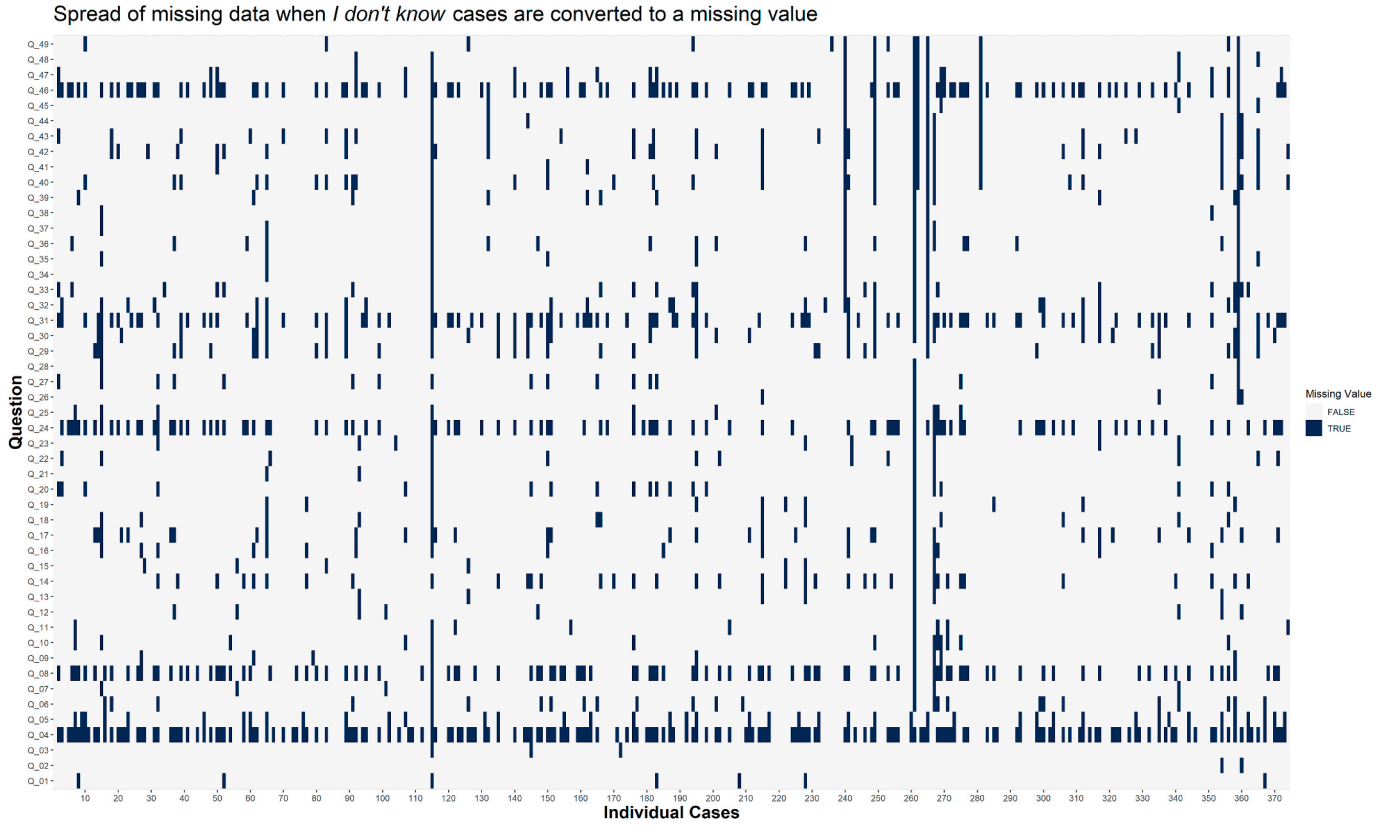

Fig. 3. Spread of missing when I don't know responses are converted to a missing value. Dark colours represent missing values.

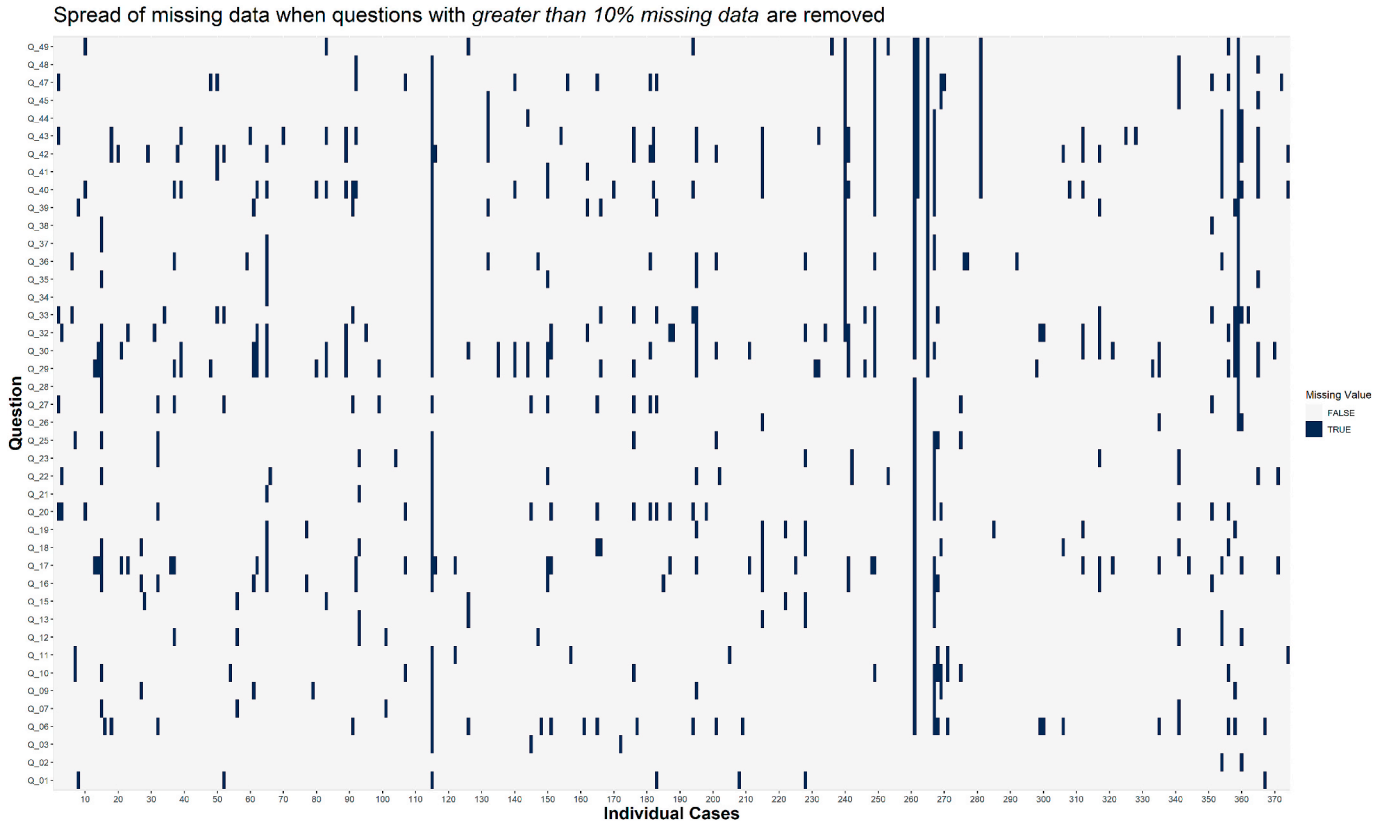

Fig. 4. Spread of missing data when questions with greater than 10% are removed. Dark colours represent missing values.

Having the option of "I Don't Know" gave participants the freedom to be more expressive and allowed researchers to make decisions about handling the missing data more appropriately. A missing indicator method analysis using parallel analysis further revealed a latent construct that existed, though was not relevant to include for the population sampled.

References

1. Narayan SW, Yu Ho K, Penm J, et al. Missing data reporting in clinical pharmacy research. *Am J Health Syst Pharm*. 2019;76:2048–2052.
2. Allison PD. *Missing Data*. vol. 136. Thousand Oaks, CA, US: Sage publications; 2001.
3. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.
4. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2002.
5. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med*. 2007;4:e297.
6. Little RJ. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83:1198–1202.
7. Turrell G. Income non-reporting: implications for health inequalities research. *J Clin Epidemiol*. 2000;54:207–214.
8. Aquilino WS. Telephone versus face-to-face interviewing for household drug use surveys. *Int J Addict*. 1991;27:71–91.
9. Little T, Rhemtulla M. Planned missing data designs for developmental researchers. *Child Dev Perspect*. 2013;7.
10. Pokropek A. Missing by design: planned missing-data designs in social science. *ASK Res Methods*. 2011:81.
11. Rosenberg SW. Opinion formation, theory of. In: Wright JD, ed. *International Encyclopedia of the Social & Behavioral Sciences*. second ed. Oxford: Elsevier; 2015: 243–245.
12. Kim Y, Dykema J, Stevenson J, Black P, Moberg DP. Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc Sci Comput Rev*. 2019;37:214–233.
13. Dirmaier J, Harfst T, Koch U, Schulz H. Incentives increased return rates but did not influence partial nonresponse or treatment outcome in a randomized trial. *J Clin Epidemiol*. 2007;60:1263–1270.
14. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value Health*. 2011;14:1101–1108.
15. Schuman H, Presser S. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Sage; 1981.
16. Laaksonen S. *Missingness, its Reasons and Treatment. Survey Methodology and Missing Data*. Springer; 2018:99–110.
17. Durand RM, Lambert ZV. Don't know responses in surveys: analyses and interpretational consequences. *J Bus Res*. 1988;16:169–188.
18. Dong Y, Peng C-YJ. Principled missing data methods for researchers. *SpringerPlus*. 2013;2:222.
19. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7:147–177.
20. Bennett DA. How can I deal with missing data in my study? *Aust N Z J Publ Health*. 2001;25:464–469.
21. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2011;22:278–295.
22. Seaman S, White I. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Commun Stat Theor Methods*. 2014;43: 3499–3515.
23. Lundström S, Särndal C-E. Calibration as a standard method for treatment of nonresponse. *J Off Stat*. 1999;15:305.
24. Roderick JAL. Survey nonresponse adjustments for estimates of means. *Int Stat Rev/ Rev Int Stat*. 1986;54:139–157.
25. *The SAGE Encyclopedia of Social Science Research Methods*. 2004.
26. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8:3–15.
27. Wilkinson L. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol*. 1999;54:594.
28. Hair JF, Tatham RL, Anderson RE, Black W. *Multivariate Data Analysis*. New Jersey: Pearson Prentice Hall; 2006.
29. Enders CK. A primer on maximum likelihood algorithms available for use with missing data. *Struct Equ Model*. 2001;8:128–141.
30. Lee S-Y, Chiu Y-M. Analysis of multivariate polychoric correlation models with incomplete data. *Br J Math Stat Psychol*. 1990;43:145–154.
31. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*. 1977;39:1–38.
32. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59:1087–1091.
33. Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Struct Equ Model*. 2001;8:430–457.
34. Laird NM. Missing data in longitudinal studies. *Stat Med*. 1988;7:305–315.

35. Rajula HSR, Odintsova V, Manchia M, Fanos V. Overview of federated facility to harmonize, analyze and management of missing data in cohorts. *Appl Sci*. 2019;9:4103.
36. Nassiri V, Lovik A, Molenberghs G, Verbeke G. On using multiple imputation for exploratory factor analysis of incomplete data. *Behav Res Methods*. 2018.
37. McNeish D. Exploratory factor Analysis with small samples and missing data. *J Pers Assess*. 2017;99:637–652.
38. Doan A, Halevy A, Ives Z. 1 - introduction. In: Doan A, Halevy A, Ives Z, eds. *Principles of Data Integration*. Boston: Morgan Kaufmann; 2012:1–18.
39. Fienberg SE, Manrique-Vallier D. Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *Adv Stat Anal*. 2009;93:49–60.
40. Baker SG. A simple EM algorithm for capture-recapture data with categorical covariates. *Biometrics*. 1990:1193–1200.
41. Voillet V, Besse P, Liaubet L, San Cristobal M, González I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinf*. 2016;17:402.
42. Fang Z, Ma T, Tang G, et al. Bayesian integrative model for multi-omics data with missingness. *Bioinformatics*. 2018;34:3801–3808.
43. Mirzaei A, Carter SR, Chen JY, Rittsteuer C, Schneider CR. Development of a questionnaire to measure consumers' perceptions of service quality in community pharmacies. *Res Soc Adm Pharm*. 2019;15:346–357.
44. Mirzaei A, Carter SR, Schneider CR. Marketing activity in the community pharmacy sector–a scoping review. *Res Soc Adm Pharm*. 2018;14:127–137.