PLOS ONE

RESEARCH ARTICLE

Improving data sharing in research with context-free encoded missing data

Marieke P. Hoevenaar-Blom[1]*, Juliette Guillemont[2], Tiia Ngandu[3], Cathrien R. L. Beishuizen[1], Nicola Coley[4,5], Eric P. Moll van Charante[6], Sandrine Andrieu[4,5], Miia Kivipelto[3,7,8,9], Hilkka Soininen[7,10], Carol Brayne[11], Yannick Meiller[12☯], Edo Richard[1,13☯]

1 Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands, **2** INSERM, University of Toulouse, Toulouse, France, **3** Chronic Disease Prevention Unit, National Institute for Health and Welfare, Helsinki, Finland, **4** Department of Epidemiology and Public Health, Toulouse University Hospital, Toulouse, France, **5** INSERM, University of Toulouse UMR1027, Toulouse, France, **6** Department of General Practice, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands, **7** Institute of Clinical Medicine/Neurology, University of Eastern Finland, Kuopio, Finland, **8** Aging Research Center, Karolinska Institutet/ Stockholm University, Stockholm, Sweden, **9** Karolinska Institutet Center for Alzheimer Research, Stockholm, Sweden, **10** Neurocenter, neurology, Kuopio University Hospital, Kuopio, Finland, **11** Department of Public Health and Primary Care, Cambridge Institute of Public Health, University of Cambridge, United Kingdom, **12** Department of Information and Operations Management, ESCP Europe, Paris, France, **13** Department of Neurology, Radboud University Medical Center, Nijmegen, The Netherlands

☯ These authors contributed equally to this work.
* m.p.hoevenaarblom@amc.uva.nl

Citation: Hoevenaar-Blom MP, Guillemont J, Ngandu T, Beishuizen CRL, Coley N, Moll van Charante EP, et al. (2017) Improving data sharing in research with context-free encoded missing data. PLoS ONE 12(9): e0182362. https://doi.org/10.1371/journal.pone.0182362

Editor: Christian Gluud, Copenhagen University Hospital, DENMARK

Received: January 24, 2017

Accepted: July 17, 2017

Published: September 12, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: HATICE (www.hatice.eu) is a collaborative project co-funded by the European Union's Seventh Framework Program (FP7, 2007-2013), under grant agreement No 305374. The research leading to these results has also been funded by the "Multimodal preventive trials for Alzheimer´s Disease: towards multinational strategies-programme: MIND-AD", Academy of

Abstract

Lack of attention to missing data in research may result in biased results, loss of power and reduced generalizability. Registering reasons for missing values at the time of data collection, or—in the case of sharing existing data—before making data available to other teams, can save time and efforts, improve scientific value and help to prevent erroneous assumptions and biased results. To ensure that encoding of missing data is sufficient to understand the reason why data are missing, it should ideally be context-free. Therefore, 11 context-free codes of missing data were carefully designed based on three completed randomized controlled clinical trials and tested in a new randomized controlled clinical trial by an international team consisting of clinical researchers and epidemiologists with extended experience in designing and conducting trials and an Information System expert. These codes can be divided into missing due to participant and/or participation characteristics (n = 6), missing by design (n = 4), and due to a procedural error (n = 1). Broad implementation of context-free missing data encoding may enhance the possibilities of data sharing and pooling, thus allowing more powerful analyses using existing data.

Introduction

Missing data are often unavoidable in research, despite all efforts to reduce their occurrence in study design and conduct. Lack of attention to this important area may result in biased results,

Finland (291803) and VTR, Kuopio University Hospital (5772815). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

loss of power and reduced generalizability. This can seriously compromise inferences from clinical trials and observational studies.[1]

Knowing why data are missing is important to determine the most appropriate way to handle them in the analyses. The encoding of missing data should ideally be context-free—i.e. the code itself is sufficient to understand the reason why data are missing. This makes it easier to determine whether data are missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR).[2] The information gained is particularly useful when assessing the need for various types of sensitivity analyses (if any) and when separating clearly plausible MCAR data from the rest of missing data. In the latter case this may produce more simple missing data patterns that need to be subjected to multiple imputation or alternative and equally valid methods. This again may imply that more simple methods could be used. Examples of such a situation are that family history of CVD could not be answered because of broken contact with family or that a box of questionnaires got lost. However, in the worst-case scenario, if the number of missing data is large and information on the reason why data is missing is lacking, collected data may lose their scientific value, leading to 'research waste'.[3] Registering this information at the time of data collection, or—in the case of sharing existing data—before making data available to other teams, can therefore save time and efforts, improve scientific value and help to prevent erroneous assumptions and biased results.[4]

The current trend of data sharing and open access, often involving large datasets from different countries, increases the risk of incorrect handling of missing data since there is no link between the researchers performing the collection and those analyzing the data. To the best of our knowledge, there are no clear methods for conveying the reasons for missing data, despite a large body of literature on how to prevent and analyze missing data.[5] Therefore, we developed a list of context-free codes of missing data and used them in a project to pool three existing datasets from three countries as well as for a new, international randomized controlled clinical trial[6].

Materials and methods

The 'Prevention of dementia by intensive vascular care (PreDIVA, ISRCTN 29711771)', 'Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER, NCT 01041989)' and the 'Multidomain Alzheimer Preventive Trial (MAPT, NCT 00672685)' are recently completed large randomized controlled clinical trials with a total of over 6400 participants.[7–9] The 'Healthy Ageing Through Internet Counselling in the Elderly (HATICE) randomized controlled clinical trial' is an ongoing study on the effect of a multidomain internet intervention on cardiovascular risk factors, in over 2700 participants in three countries.[6]

To identify missing values, all variables in the three completed trials[7–9] were evaluated by an international team collaborating in the HATICE consortium, consisting of clinical researchers and epidemiologists with extended experience in designing and conducting trials and an Information System expert. The individual research teams of each trial first listed all situations that led to missing values in their study. Next, the Information System expert merged all missing situations into one list of missing-categories (S1 File). In a consensus meeting, the international team agreed on the most important missing-categories, taking into account their external applicability. For pooling datasets, an additional missing data code was created for variables that were not collected by at least one of the other studies. We used numerical codes to accommodate analyses in most statistical packages. To avoid confusing missing data with non-missing data we used codes with 6 digits and starting with '9'; e.g. 930000 for 'not applicable' (NA) and 931000 for 'not applicable due to conditional value' (NAC). Before pooling data from the three trials in an online platform specifically designed for the purpose, we converted

Table 1. Categories of missing data.

Source	Category of missing data	Examples	Abbreviation	Type[a]
Participant and/ or participation characteristic	Assessed but the participant does not know	Family history could not be answered, because of broken contact	ASSU	MCAR
	Assessed but the participant was not able to provide the information	Disability preventing a physical test	ASSD	MNAR[c]
	Refusal	Participant does not want to tell his weight because he is embarrassed	ASSR	MNAR
	Not applicable[b]	"Do you feel disabled in doing volunteer work" in case the participant is not engaged in any volunteering	NA	MNAR[c]
	The visit has been missed	In case of a missed visit, all variables for this visit are missing	MISS	MAR/ MNAR[d]
	Dropout	In case of dropout, variables subsequent to the date of dropout are missing	DROP	MAR/ MNAR[d]
By design	Not assessed, variable not in the study	Only applicable for data pooling	NASS	MCAR
	Not applicable because of conditional variable[b]	Date of birth of siblings if participant does not have siblings [e]	NAC	MNAR[c]
	Due to random subsampling	Expensive measurement only performed in random subsample. For others these values are missing.	RS	MCAR
	Answer/value not available yet	Blood sample was collected though not analyzed yet	NAV	MCAR
Procedural error	Not assessed/ registered, by mistake	Box of questionnaires got lost	ERR	MCAR

[a] MCAR: missing completely at random, MAR: missing at random or MNAR: missing not at random. The types are an indication of the most common scenarios fitting to this category (see Discussion section).

[b] Often, the question whether it is applicable (for instance 'do you take medication') is not included. In this case NA has to be filled in manually. However, if this question is asked and therefore the conditional variables can be skipped, a digital questionnaire can fill out the NAC category automatically.

[c] In 'ASSD' and 'NA' categories the fact that the value is missing depends on the reason why it is missing, so this fits the definition of MNAR. However, how to handle this in the analyses should be decided on a case to case basis.

[d] We advise for these categories to make subcategories, specific to the study (see Discussion section).

[e] In a digital questionnaire it is possible that the conditional questions are automatically skipped so the participant does not have to deal with the questions that are not applicable to their situation. To inform the data analyst that the variable is deliberately skipped the NAC value will be automatically filled out.

https://doi.org/10.1371/journal.pone.0182362.t001

the original missing data encoding of every dataset into the encodings represented in Table 1. To establish the applicability of the encodings on a trial that was not used to develop the encodings, they were implemented at the data collection stage of the currently ongoing HATICE trial[6].

Technical details on the context-free data encoding have been published previously (S2 File).[4]

Results

We identified 11 different types of missing data (Table 1). These can be divided into the following categories: missing due to participant and/or participation characteristics (n = 6), missing by design (n = 4), and due a procedural error (n = 1). The 11 missing encodings were sufficient to recode all missing data in the three completed trials[7–9] and the HATICE trial.[6]

Discussion

To initiate a systematic approach for context-free missing data encoding, we described 11 separate missing codes that could be classified in three categories: missing due to participant and/ or participation characteristics (n = 6), missing by design (n = 4), and due a procedural error (n = 1).

Clearly, a careful balance is needed between accuracy (determined by the number of missing data categories) and the validity of the information. Consequently, the missing data categories that we identified, cannot be used one on one to determine whether data are MCAR, MAR or MNAR. For instance, in the ASSU category (asked but participant does not know the answer) not knowing the answer could be independent of observable and unobservable parameters of interest, and as such be MCAR. However, if the outcome is cognitive function, not knowing is probably informative and MNAR applies. To account for all possible scenarios, the categories may need to be further subdivided. However, too many missing data categories may be confusing for the person filling out the assessments, particularly if this person is a participant. This may jeopardize the validity of the information. The missing data encoding cannot cover the nuances that can be explained in free text. Missing data encoding and free text can co-exist. Especially in big studies, free texts are difficult to take into account and the missing encodings have most of their value.

For the MISS (visit missed) and DROP (dropout) categories, which are generally filled out by the researchers, subcategories are recommended. Current common practice is to have a separate variable for reasons for dropout which can be combined with the system missing variables to decide on analytical techniques. One could choose to integrate the reasons for dropout (or missed visits) in the missing encodings. This would require the MISS (visit missed) and DROP (dropout) categories to be divided into subcategories, specific to the study. For instance a code 911000 for dropout because deceased, a code 912000 for dropout because of adverse effects of treatment, etc. As these categories are registered already in most studies, no further confusion is expected from this approach.

A major strength of our approach is the combination of expertise from information specialists, clinical researchers and epidemiologists. Both from an information systems perspective and an epidemiological perspective, our efforts can be a starting point for adopting these encodings as well as further developing categories applicable to specific situations/ domains. Current existing standard classifications/ nomenclatures/ terminologies are lacking a system for missing data encoding. Our encodings can, for instance, easily be adopted in existing standard Case Report Forms such as those in CDASH (Clinical Data Acquisition Standards Harmonization) of CDISC (Clinical Data Interchange Standards Consortium) thereby contributing to their mission to enable data sharing[10]. The issue of missing data is relevant for all domains using data intensively. Our work has focused on healthcare-related research, but can be applied to other branches of research, after appropriate validation. When different studies apply the same missing encodings, recoding for data pooling will be reduced in the future. Whether a higher level of granularity in missing encodings can prevent biased results, loss of power and reduced generalizability will have to be further investigated.

Conclusions

Missing data can rarely be fully avoided, but not knowing why data are missing can be avoided. Capturing information on the reason for missing data values at the moment of data collection reduces the loss of relevant information and thereby the need for assumptions in the analysis phase. Broad implementation of context-free missing data encoding may enhance the possibilities of data sharing and pooling, thus allowing more powerful analyses using existing data.

Supporting information

S1 File. Missing data encodings in dataset.
(CSV)

S2 File. Conference proceeding on missing data encodings for information specialists. Meiller Y, Guillemont J, Beishuizen CR, Richard E, Kivipelto M, Andrieu S. An IS Approach for Handling Missing Data in Collaborative Medical Research. Twenty-second Americas Conference on Information Systems; San Diego2016.
(PDF)

Acknowledgments

The authors would like to thank the 'Prevention of dementia by intensive vascular care' (pre-DIVA) team, the 'Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability' (FINGER) team and the 'Multidomain Alzheimer Preventive Trial' (MAPT) team.

Author Contributions

Conceptualization: Marieke P. Hoevenaar-Blom, Juliette Guillemont, Tiia Ngandu, Cathrien R. L. Beishuizen, Nicola Coley, Eric P. Moll van Charante, Sandrine Andrieu, Miia Kivipelto, Hilkka Soininen, Carol Brayne, Yannick Meiller, Edo Richard.

Data curation: Juliette Guillemont, Tiia Ngandu, Cathrien R. L. Beishuizen.

Formal analysis: Marieke P. Hoevenaar-Blom, Yannick Meiller.

Funding acquisition: Edo Richard.

Investigation: Yannick Meiller.

Methodology: Marieke P. Hoevenaar-Blom, Juliette Guillemont.

Resources: Eric P. Moll van Charante, Sandrine Andrieu, Hilkka Soininen, Carol Brayne.

Supervision: Sandrine Andrieu, Miia Kivipelto, Yannick Meiller, Edo Richard.

Validation: Marieke P. Hoevenaar-Blom, Cathrien R. L. Beishuizen, Nicola Coley, Eric P. Moll van Charante, Yannick Meiller.

Visualization: Marieke P. Hoevenaar-Blom.

Writing – original draft: Marieke P. Hoevenaar-Blom, Cathrien R. L. Beishuizen, Yannick Meiller, Edo Richard.

Writing – review & editing: Marieke P. Hoevenaar-Blom, Juliette Guillemont, Tiia Ngandu, Cathrien R. L. Beishuizen, Nicola Coley, Eric P. Moll van Charante, Sandrine Andrieu, Miia Kivipelto, Hilkka Soininen, Carol Brayne, Yannick Meiller, Edo Richard.

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(14):1355–60. https://doi.org/10.1056/NEJMsr1203730 PMID: 23034025
2. Little RJ. Methods for handling missing values in clinical trials. J Rheumatol. 1999; 26(8):1654–6. PMID: 10451057
3. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009; 374(9683):86–9. https://doi.org/10.1016/S0140-6736(09)60329-9 PMID: 19525005
4. Meiller Y, Guillemont J, Beishuizen CR, Richard E, Kivipelto M, Andrieu S. An IS Approach for Handling Missing Data in Collaborative Medical Research. Twenty-second Americas Conference on Information Systems; San Diego2016. https://www.researchgate.net/publication/304024535_An_IS_Approach_for_Handling_Missing_Data_in_Collaborative_Medical_Research date last accessed: 7 June 2017
5. National Research Council Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials. Press TNA, editor. Washington DC -USA: National Academy of

Sciences; 2010. https://www.ncbi.nlm.nih.gov/books/NBK209904/ date last accessed: 15 February 2017

6. Richard E, Jongstra S, Soininen H, Brayne C, Moll van Charante EP, Meiller Y, et al. Healthy Ageing Through Internet Counselling in the Elderly: the HATICE randomised controlled trial for the prevention of cardiovascular disease and cognitive impairment. BMJ Open. 2016; 6(6):e010806. https://doi.org/10.1136/bmjopen-2015-010806 PMID: 27288376
7. Ngandu T, Lehtisalo J, Solomon A, Levalahti E, Ahtiluoto S, Antikainen R, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. Lancet. 2015; 385(9984):2255–63. https://doi.org/10.1016/S0140-6736(15)60461-5 PMID: 25771249
8. Vellas B, Carrie I, Gillette-Guyonnet S, Touchon J, Dantoine T, Dartigues JF, et al. Mapt Study: A Multidomain Approach for Preventing Alzheimer's Disease: Design and Baseline Data. J Prev Alzheimers Dis. 2014; 1(1):13–22. PMID: 26594639
9. Moll van Charante EP, Richard E, Eurelings LS, van Dalen JW, Ligthart SA, van Bussel EF, et al. Effectiveness of a 6-year multidomain vascular care intervention to prevent dementia (preDIVA): a cluster-randomised controlled trial. Lancet. 2016 20; 388(10046):797–805. https://doi.org/10.1016/S0140-6736(16)30950-3 Epub 2016 Jul 26. PMID: 27474376
10. CDISC. Clinical Data Acquisition Standards Harmonization (CDASH) https://www.cdisc.org/standards/foundational/cdash date last accessed: 8 April 2017