Summarising good practice guidelines for data extraction for systematic reviews and meta-analysis

Kathryn S Taylor, Kamal R Mahtani, Jeffrey K Aronson

Nuffield Department of Primary Care Health Sciences, University of Oxford, OX2 6GG, United Kingdom

Correspondence to: K Taylor kathryn.taylor@phc.ox.ac.uk

Word count: 1500

Data extraction is the process of a systematic review that occurs between identifying eligible studies and analysing the data, whether it be a qualitative synthesis or a quantitative synthesis with a meta-analysis. The aims of data extraction are to obtain information about the included studies in terms of the characteristics of each study and its population and, for quantitative synthesis, to collect the necessary data to carry out meta-analysis. In systematic reviews, information about the included studies will also be required to conduct risk of bias assessments, but these data are not the focus of this article.

Following good practice when extracting data will help make the process efficient and reduce the risk of errors and bias. Failure to follow good practice risks basing the analysis on poor quality data, and therefore providing poor quality inputs, which will result in poor quality outputs, with unreliable conclusions and invalid study findings. In computer science, this is known as 'garbage in, garbage out' (GIGO) or 'rubbish in, rubbish out' (RIRO). Furthermore, providing insufficient information about the included studies for readers to be able assess the generalisability of the findings from a systematic review will undermine the value of the pooled analysis. Such failures will cause your systematic review and meta-analysis to be less useful than it ought to be.

Some guidelines for data extraction are formal, including those described in the Cochrane Handbook for Systematic Reviews of Interventions,[1] the Cochrane Handbook for Diagnostic Test Accuracy (DTA) Reviews,[2,3] the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines for systematic reviews and their protocols,[4-7] and other sources, e.g.[8,9] Formal guidelines are complemented with informal advice in the form of examples and videos on how to avoid possible pitfalls and guidance on how to carry out data extraction more efficiently, e.g.[10-12]
Guidelines for data extraction involve recommendations for:

- Duplication
- Anticipation
- Organisation
- Documentation

Duplication

Ideally, at least two reviewers should extract data independently,[1,2,9-12] particularly outcome data,[1] as data extraction by only one person can generate errors.[1,13] Data will be extracted from the same sources into identical data extraction forms. If time or resources prevent independent dual extraction, one reviewer should extract the full data and another should independently check the extracted data

for both accuracy and completeness.[8] In rapid or restricted reviews, an acceptable level of verification of the data extraction by the first reviewer may be achieved by a second reviewer extracting a random sample of data.[14] Then before comparing the extracted data and seeking a consensus, the extent to which coded (categorical) data extracted by two different reviewers are consistent may be measured using kappa statistics,[1,2,12,15] or Fleis kappa statistics, when more than two people have extracted the data.[16] Formal comparisons are not routine in Cochrane Reviews, and the Cochrane Handbook recommends that if agreement is to be formally assessed, it should focus only on key outcomes or risk of bias assessments.[1]

Anticipation

Reviewers should anticipate a number of problems that may arise during data extraction. The study protocol should pre-specify how these problems will be addressed:

(i) *Disagreement between reviewers when extracting data*. Some differences in extracted data are simply due to human error, and such conflicts can be easily resolved. Conflicts and questions about clinical issues, about which data to extract, or whether the relevant data have been reported can be addressed by involving both clinicians and methodologists in data extraction.[3,12] The protocol should set out the strategy for resolving disagreements between reviewers, using consensus and, if necessary, arbitration by another reviewer. If arbitration fails, the study authors should be contacted for clarification. If that is unsuccessful, the disagreement should be documented and reported.[1,6,7]

(ii) *Outcome data being reported in different ways, which are not necessarily suitable for meta-analysis*. Many resources are available for helping with data extraction, involving various methods and equations to transform reported data or make estimates.[1,2,10] The protocol may acknowledge this by stating that any estimates made and their justification will be documented and reported.

(iii) *Including estimates and alternative data.* It is also important to anticipate the roles that extracted data will play in the analysis. Studies should be highlighted when multiple sets of outcome data are reported or when estimates have been made in extracting outcome data.[9] Clearly identifying these studies during the data extraction phase will ensure that the studies can be quickly identified later, during the data analysis phase.

(iv) *Risk of double counting patients*. Some studies involve multiple reports, but the study should be the unit of interest.[1] Tracking down multiple reports and ensuring that patients are not double counted may require good detective skills.

(v) *Risk of human error, inconsistency, and subjectivity when extracting data*. The protocol should state whether data extraction was independent and carried out in duplicate, if a

standardised data extraction form was used, and whether it was piloted. The protocol should also state any special instruction, e.g. only extracting pre-specified eligibility criteria.[1,2,6-9,11,12]

(vi) *Ambiguous or incomplete data*. Authors should be contacted to seek clarification about data and make enquiries about the availability of unreported data.[1,2,9] The process of confirming and obtaining data from authors should be pre-specified[6,7] including the number of attempts that will be made to make contact, who will be contacted (e.g. the first author), and what form the data request will take. Asking for data that are likely to be readily available will reduce the risk of authors offering data with preconditions.

(vii) *Extracting the right amount of data*. Time and resources are wasted extracting data that will not be analysed, such as the language of the publication and the journal name when other extracted data (first author, title, and year) adequately identify the publication. Which study characteristics are extracted will depend on the aim of the systematic review.[16] For example, if the prevalence of a disease is important and is known to vary across cities, the country and city should be extracted. Any assumptions and simplifications should be listed in the protocol.[6,7] The protocol should allow some flexibility for alternative analyses by not over-aggregating data e.g. collecting data on smoking status in categories “smoker/ex-smoker/never smoked” instead of ”smoker/non-smoker”.[11]

Organisation

Guidelines recommend that the process of extracting data should be well organised. This involves having a clear plan, which should feature in the protocol, stating who will extract the data, the actual data that will be extracted, details about the use, development, and piloting of a standardised data extraction form,[1,6-9] and having good data management procedures,[10] including backing up files frequently.[11] Standardised data extraction forms can provide consistency in a systematic review, while at the same time reducing biases and improving validity and reliability. It may be possible to re-use a form from another review.[12] It is recommended that the data extraction form is piloted and that reviewers receive training in advance[1,2,12] and instructions should be given with extraction forms (e.g. about codes and definitions used in the form) to reduce subjectivity and to ensure consistency.[1,2,12] It is recommended that instructions be integrated into the extraction form, so that they are seen each time data are extracted, rather than having instructions in a separate instruction document, which may be ignored or forgotten.[2] Data extraction forms may be paper-based or electronic or involve sophisticated data systems. Each approach will have advantages and disadvantages.[1,11,17] For example, using a paper-based form does not require internet access or software skills, but using an electronic

extraction form facilitates data analysis. Data systems, while costly, can provide online data storage and automated comparisons between data that have been independently extracted.

Documentation

Data extraction procedures and pre-analysis calculations should be well documented[9,10] and based on "good bookkeeping".[5,10] Having good documentation supports accurate reporting, transparency, and the ability to scrutinise and replicate the analysis. Reporting guidelines for systematic reviews and protocols of systematic reviews are provided by PRISMA.[4-7] In cases where data are derived from multiple reports, documenting the source of each data item will facilitate the process of resolving disagreements with other reviewers, by enabling the source of conflict to be quickly identified.[10]

Table 1. Summarising guidelines for extracting data for systematic reviews and meta-analysis

Issue	**Recommendations**
Duplication	Dual independent data extraction or verification of single extraction of all studies or, in rapid or restricted reviews, a random sample.
Anticipation	Anticipate potential problems during data extraction – conflicts, ambiguities, inconsistencies, missing data, and risk of making errors Outline solutions in the study protocol
Organisation	Having a clear comprehensive plan that allows flexibility Based on good data management A well designed data extraction form Pilot the data extraction form Reviewers trained or given detailed instructions Have clinical and methodological reviewers
Documentation	Based on good bookkeeping Comprehensive Accurate Transparent

Data extraction is both time consuming and error-prone, and automation of data extraction is still in its infancy.[1,18] Following both formal and informal guidelines for good practice (Table 1) will make the process efficient and reduce the risk of errors and bias when extracting data. This will contribute towards ensuring that systematic reviews and meta-analyses are conducted to a high standard.

Acknowledgements

This research was supported by the National Institute for Health Research Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Contributions

KT and KM conceived the idea of the series of which this is one part. KT wrote the first draft of the manuscript. All authors revised the manuscript and agreed the final version.

Competing interests

Dr Mahtani and Dr Aronson were Associate Editors of BMJ Evidence Medicine at the time of submission.

References

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.1 (updated September 2020). Cochrane, 2020. Chapter 5. Collecting data. Available from www.training.cochrane.org/handbook. [assessed 14 Oct 2020]

2. Leeflang MM, Davenport CF, Takwoini Y, Deeks JJ. Collecting study characteristics. Lesson 5.1. Cochrane Collaboration DTA Author Online Learning Materials. The Cochrane Collaboration, October 2014. Videocast (30 slides, 23 min, sound, colour). Available at https://methods.cochrane.org/sdt/dta-author-training-online-learning [accessed 14 Oct 2020].

3. Leeflang MM, Takwoini Y, Davenport CF, FDeeks JJ. Collecting study results. Lesson 5.2. Cochrane Collaboration DTA Online Learning Materials. The Cochrane Collaboration, February 2015. Videocast (23 slides, 21 min, sound, colour). Available at https://methods.cochrane.org/sdt/dta-author-training-online-learning [accessed 14 Oct 2020].

4. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535. Published 2009 Jul 21. doi:10.1136/bmj.b2535

5. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration *BMJ* 2009; 339 :b2700

6. Moher D, Shamsee L, Clare M, *et al*; PRISMA Group. Preferred reporting items for systematic reviews and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews* 2015;4:1

7. Shamseer L, Moher D, Clarke M, *et al*, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015 Jan 2;349(jan02 1):g7647.

8. Centre for Reviews and Dissemination, University of York. *Systematic reviews: CRD guidance for undertaking reviews in health care*. Chapter 1. Core principles and methods for conducting a systematic review of health interventions. 2009. CRD, University of York.

9. Collaboration for Environmental Evidence. 2018. *Guidelines and Standards for Evidence synthesis in Environmental Management*. Version 5.0. (AS Pullin, GK Frampton, B Livoreil & G Petrokofsky, Eds) www.environmentalevidence.org/information-for-authors. Section 7. Data coding and data extraction [accessed 14 Oct 2020]

10. Taylor K. Data extraction tips for meta-analysis. 2019-2020. https://www.cebm.ox.ac.uk/resources/data-extraction-tips-meta-analysis [accessed 14 Oct 2020]

11. Dalhousie University. Systematic Reviews: A how-to guide. Making the most of your data. 2020. https://dal.ca.libguides.com/systematicreviews/dataextraction#s-lg-box-9032274 [accessed 14 Oct 2020]

12. Keenan C. Campbell Collaboration UK & Ireland. Top tips: data extraction. 2018 http://meta-evidence.co.uk/data-extraction/ [accessed 14 Oct 2020]

13. Buscemi N, Hartling L, Vandermeer B, *et al*. Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of clinical epidemiology*, 2006; 59(7): 697-703. doi:10.1016/j.jclinepi.2005.11.010

14. Plüddemann A, Aronson JK, Onakpoya I, *et al*. Redefining rapid reviews: a flexible framework for restricted systematic reviews. *BMJ Evid Based Med*. 2018 Dec;23(6):201-203. doi: 10.1136/bmjebm-2018-110990. Epub 2018 Jun 27. PMID: 29950313.

15. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.

16. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378. doi:10.1037/h0031619

17. Li TJ, Vedula SS, Hadar N, Parkin C, *et al*. Innovations in data collection, management, and archiving for systematic reviews. *Annals of Internal Medicine* 2015; 162: 287-294.

18. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews* 2015; **4**: 78.