Data Extraction from Included Studies

Kwi Moon and Shripada Rao

Abstract Accurate data extraction and their synthesis form the basis of appropriate conclusions of a systematic review. Systematic reviewers should extract ALL data relevant to the review question, not just the outcome data. Data to be extracted include baseline characteristics of study participants, information related to study methodology and outcomes and other relevant information. If published articles have given the results using figures instead of actual numbers, specialised software that convert images to pixel values may be utilised to obtain the actual data values. Tools such as Plot Digitizer, WebPlotDigitizer, Engauge, Dexter, Ycasd and GetData Graph Digitizer can be used for this purpose. When unable to extract data from available reports or to seek clarifications, the reviewers could contact the original investigators. Data extraction should be performed using pre-piloted forms independently by at least two reviewers to ensure accuracy. A high level of diligence is required to minimise errors during the stage of data extraction.

Keywords Comparisons · Covidence · Descriptive · Entry · Error · Interventions · Outcomes · Participant · Source

K. Moon (✉)
Department of Pharmacy, Perth Children's Hospital, 15 Hospital Ave, Nedlands, Perth, WA 6009, Australia
e-mail: kwi.moon@health.wa.gov.au

S. Rao
Neonatal Directorate, Perth Children's Hospital, 15 Hospital Ave, Nedlands, Perth, WA 6009, Australia
e-mail: shripada.rao@health.wa.gov.au

K. Moon · S. Rao
School of Medicine, University of Western Australia, Perth, WA 6009, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_6

Introduction

Data extraction is an important aspect of a systematic review because accurate data and their synthesis form the basis of appropriate conclusions (Li et al. 2015). Data collected for systematic reviews should be accurate, complete, and accessible for future updates of the review and data sharing (https://training.cochrane.org/handbook/current/chapter-05#section-5-1). It is essential to pilot the data collection form before beginning data entry. The completed data collection forms should be provided to the editors of journals or the Cochrane review group upon request.

Which Data to Extract

Systematic reviewers should extract ALL data relevant to the review question, and not just the outcome data.

a. *Descriptive data of individual included studies:* Information on authors, settings, study design, characteristics of participants, details of the intervention, outcomes, sample size, funding source and the reason for inclusion or exclusion should be collected. Such detailed extraction and reporting of the descriptive data will enable clinicians to establish the generalizability of the results. Descriptive data are also important to the reviewer and enables them to understand and explore heterogeneity (Munn et al. 2014).
b. *Outcome data:* The outcome data should be collected separately for each outcome. The details should include the raw numbers (numerator and denominator), statistical measures such as relative risks, odds ratios, weighted means, standard deviations and confidence intervals. A standardised approach should be used while extracting data. Examples of data collection templates are available from organisations such as Cochrane Collaboration, Joanna Briggs Institute (JBI) and BMJ group.
c. *Data to assess the risk of bias:* It is vital to collect information to assess the risk of bias in the included studies. For example, while conducting a systematic review of RCTs, it is essential to gather information on methods used for generation of random sequences, allocation concealment, blinding, completeness of follow up and any other sources of bias. Studies with a high risk of bias decrease their internal validity leading to erroneous conclusions. The details have been covered in the chapter titled “Assessment of Risk of Bias”.

How to Minimise Errors in Data Extraction

Errors in data extraction can alter the results and conclusions of the review, and hence utmost diligence is required during this stage.

Jones et al. retrospectively repeated the data extraction in all systematic reviews conducted by the Cochrane Cystic Fibrosis and Genetic Disorders Group using the same articles that were used by the original Cochrane reviewers (Jones et al. 2005). They reported that errors were found in 20 of 34 reviews, including incorrect calculations made when converting data in primary articles into data required for the review and misinterpretation of data that were reported in the primary article (Jones et al. 2005). In another study, Carroll et al. (2013) evaluated differences in the data extracted by three different systematic reviews comparing total hip arthroplasty versus hemiarthroplasty in osteoarthritis. The authors reported that 8–42% of the data differences between the reviews resulted from the selection of alternative reported data, while 8–17% of the differences resulted from data errors. They concluded that systematic reviewers should use double-data extraction to minimise error and make every effort to clarify or explain their choice of data (Carroll et al. 2013). Buscemi et al. found that single data extraction resulted in more errors than double data extraction (relative difference: 21.7%, P = .019) (2006). Mathes et al. identified six studies that had addressed the issue of errors in data extraction (2017). They found a high rate of extraction errors (up to 50%), and often the errors influenced effect estimates.

The Institute of Medicine (IOM) recommends that review authors should, "*at a minimum, use two or more independent researchers to extract quantitative and other critical data from each study*" (Eden et al. 2011). The Cochrane Handbook also makes similar recommendations (Higgins et al. 2019). Any disagreements between authors are to be resolved by discussion among all authors or by consulting a senior author.

Sources to Obtain Data

Reports from included studies are the major source of data for systematic reviews. Such reports may be published or unpublished (e.g. Journal articles, conference abstracts, dissertation, and online clinical trial registries). It is important to be aware that conference abstracts may have preliminary findings only. Sometimes outcome data may be given only as figures in the published manuscripts. Specialised software that converts images to pixel values may be utilised to obtain more accurate data values (Vucic et al. 2015). Tools such as Plot Digitizer, WebPlotDigitizer, Engauge, Dexter, ycasd and GetData Graph Digitizer can be used for this purpose. The software takes an image of a figure and then digitising the data points off the figure using the axes and scales set by the users (https://training.cochrane.org/handbook/current/chapter-05#section-5-5-8).

When unable to extract information from available reports or to seek clarifications, the reviewers need to contact the original investigators. Young and Hopewell (2011) reported that email correspondence with authors resulted in the greatest response (Young and Hopewell 2011). The Cochrane handbook recommends that obtaining unpublished data is highly desirable and potentially increases precision and minimises the impact of reporting biases (https://training.cochrane.org/handbook/current/chapter-05#section-5-2-3). **It is vital to pay special attention to 'errata' from published studies.** Hauptman et al. reviewed the frequency and significance of published errata in 20 general medicine and cardiovascular journals (median impact factor 5.52) over 18 months. They found that 557 articles were associated with errata reports (overall errata report rate 4.2 per 100). At least one significant error that materially altered data interpretation was present in 24.2% of articles with errata (Hauptman et al. 2014).

Where to Enter the Data

Data collection forms promote standardised approach for data extraction and address the review question/assessment criteria directly, providing a clear summary. Furthermore, the forms create a historical record of data collection and decisions made throughout the review process, including the final statistical data for meta-analyses. Depending on the author's preferences, data collection forms can be electronic (e.g. Microsoft Excel) or paper-based. A generic template may be used at the beginning for testing and then updated by reviewers to ensure that the form meets their needs. ***Covidence*** is primary screening and data extraction tool recommended by Cochrane for Cochrane authors (https://www.covidence.org/home). It allows authors to upload search results, screen abstracts and full text, complete data collection, conduct risk of bias assessment and export data into Revman or Excel. For complex systematic reviews, **EPPI-Reviewer** is useful for data collection and other aspects of a systematic review. The Covidence and EPPI-Reviewer can be accessed free of cost by the Cochrane reviewers. For independent systematic reviewers, they are subscription-based. Example of data collection items to be included in the systematic review is shown in Table 1.

Automating Data Extraction

Manual extraction of the data is slow, costly and subject to human error (Bui et al. 2016). Automating or semi-automating this step has the potential to decrease the time taken to complete systematic reviews and thus decrease the time lag for research evidence to be translated into clinical practice (Jonnalagadda et al. 2015). Natural language processing (NLP), including text mining, involves information extraction, which is the discovery by computer of information by automatically extracting

Table 1 Data collection items for studies included in the systematic review

- **Name of data extractor/s, date of data extraction**
- **Source**: Journal name, year of publication; Conference name, year
- **Setting**, **Country**
- **Title** of the article
- **Inclusion and exclusion criteria**
- **Study design**: RCT, cluster RCT, case-control, cohort, others
- **Years of conduct**
- **Duration of follow up**

Methods

- For a generation of random sequence numbers; concealment of allocation sequence, blinding
- For statistical analysis

Participants: Baseline characteristics (e.g. age, sex, weight, comorbidity, socioeconomic status)

Intervention: Details of intervention (e.g. drug dose, frequency, route, and duration)

Control: Details (e.g. no intervention, placebo, standard care)

Description of co-interventions

Outcomes

For each pre-specified outcome (e.g. mortality, morbidity) in the systematic review:

- Definition, Timing of measurements
- Adverse outcomes

Results

For each group, and for each outcome at each time point:

- Number of participants assigned
- Number of participants included in the analysis
- Number of participants who withdrew or excluded
- Number who were lost to follow-up
- Summary data for each group (e.g. 2 × 2 table for dichotomous data; means and standard deviations for continuous data)
- Between-group effect size estimates (e.g. risk ratio, odds ratio, mean difference)

Miscellaneous

- Study authors' conclusions
- Correspondence required for clarification
- Comments from the study authors or by the review authors
- Funding source
- Authors' potential conflicts of interest

information from different written resources (Hearst 1999). NLP techniques have been used to automate the extraction of genomic and clinical information from biomedical literature. Similarly, automation of the data extraction step of the systematic review process through NLP may decrease the time necessary to complete a systematic review (Jonnalagadda et al. 2015). In a recent study, Bui et al. developed a computer system that used machine learning and natural language processing approaches to generate summaries of full-text scientific publications (Bui et al. 2016) automatically. The summaries at the sentence and fragment levels were evaluated in finding common clinical SR data elements such as sample size, group size, and PICO

values. They compared the computer-generated summaries with human-written summaries (title and abstract) in terms of the presence of necessary information for the data extraction as presented in the Cochrane review's study characteristics tables. They found that at the sentence level, the computer-generated summaries covered more information than humans do for systematic reviews. They concluded that machine learning and natural language processing are promising approaches to the development of such an extractive summarisation system (Bui et al. 2016). In the long run, these new approaches to evidence synthesis, which use human effort and machine automation in mutually reinforcing ways, can enhance the feasibility and sustainability of "*living systematic reviews*" (Thomas et al. 2017).

Conclusions

Data extraction is a critical step while conducting a systematic review. A high level of diligence is required to minimise errors during this stage.

References

Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarisation system to aid data extraction from full text in systematic review development. J Biomed Inform. 2016;64:265–72.

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59 (7):697–703.

Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Res Notes. 2013;6:539.

Eden J, Levit L, Berg A, Morton S. committee on standards for systematic reviews of comparative effectiveness research; Institute of Medicine. In: Finding what works in health care: standards for systematic reviews. Washington, DC: The National Academies Press; 2011.

Hauptman PJ, Armbrecht ES, Chibnall JT, Guild C, Timm JP, Rich MW. Errata in medical publications. Am J Med. 2014;127(8):779–85.

Hearst MA. Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics; College Park, Maryland, 1034679: Association for Computational Linguistics; 1999. pp 3–10.

Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons; 2019.

Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005;58(7):741–2.

Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4:78.

Li T, Vedula SS, Hadar N, Parkin C, Lau J, Dickersin K. Innovations in data collection, management, and archiving for systematic reviews. Ann Intern Med. 2015;162(4):287–94.

Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152.
Munn Z, Tufanaru C, Aromataris E. JBI's systematic reviews: data extraction and synthesis. Am J Nursing. 2014;114(7):49–54.
Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. 2017;91:31–7.
Vucic K, Jelicic Kadic A, Puljak L. Survey of Cochrane protocols found methods for data extraction from figures not mentioned or unclear. J Clin Epidemiol. 2015;68(10):1161–4.
Young T, Hopewell S. Methods for obtaining unpublished data. Cochrane Database Syst Rev. 2011(11):Mr000027.