Title: Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses

Increasing the speed for completing a systematic review is needed to keep up to date with the literature. Could automatic data extraction from ClinicalTrials.gov provide an important step in speeding up the process of evidence synthesis?

Tom Fanshawe and Rafael Perera

EBM Methods Verdict on: Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses. J Clin Epid 2019; 105:92-100.

How can we conduct systematic reviews more quickly? An assessment of information in the PROSPERO registry to 2014, Borah et al. (1) estimated the average time to carry out a systematic review to be over a year. This time frame has remained broadly unchanged in comparison with an estimate published eighteen years earlier (2). Because of the speed with which some clinical disciplines generate new evidence, there is a danger that new systematic reviews become out of date within the time taken for them to be completed and published, and as a consequence their findings appear too slowly to influence practice.

Lifting data from electronic articles or data repositories automatically using computer software, rather than extracting it painstakingly by hand, has been gradually growing in popularity amongst systematic reviewers (3). The move to automation seems appealing and a natural progression from carrying out literature searches electronically. It could also free up researcher time to concentrate on the tasks that are not easy to automate, such as writing and disseminating the review itself.

Automated data extraction in the systematic review process is itself the subject of a 2015 review (4). This article found that among the relatively few systematic reviews that used automated methods of extraction, the number of data items that could be retrieved was limited. There were also no standardized methods of extraction currently in use. These findings raised questions of whether appropriate software could be developed for data extraction, and whether the time taken to complete a systematic review could be substantially reduced.

In a recent paper, Pradhan et al. (5) produced the 'EXACT' application (available online at <http://bio-nlp.org/EXACT/>) with the aim of speeding up the process of data extraction. The application links to the results page of trials listed on clinicaltrials.gov and automatically downloads the data provided into a tab-delimited (.tsv) results file.

The authors compared the results obtained using their EXACT software with those extracted manually and via the published papers of the respective trials from systematic reviews of three drugs (Simeprevir, Trametinib and Vortioxetine). Their results indicate that 83% of estimates obtained from

the published papers could be reproduced using data from EXACT, in less than half the time. Observed discrepancies were largely caused by numerical information being reported differently in published articles from the reports at clinicaltrials.gov.

The findings of the article by Pradhan et al. illustrate the difficulties that can arise in automating data tasks: reporting to clinicaltrials.gov follows a standard format but in practice the data are often incomplete or inadequately recorded. Since January 2006, all major trial sponsors have completed 29,377 eligible trials, of which 45% are currently missing results (source <https://trialstracker.ebmdatalab.net>). As the authors acknowledge, extracting information from this data source is likely to be at best a partial solution, with many eligible studies, especially from outside the US, not being listed on the website.

Further irregularities may arise when studies have more complex designs than typical randomized trials of two or more groups. For example, cluster randomized trials and cross-over trials, and analytic considerations such as adjustment for missing data, may not be amenable to standardized formats of reporting. This lack of standardization is just one of several issues relating to automated data extraction that have been raised in a recent overview of the field, (6) and it is difficult to see that software for data extraction alone could deliver much further progress without it.

A further question is whether data extraction is the most important stage at which to focus efforts for speeding up the review process. By the time data extraction is required, the number of papers remaining in the search is often manageable, whereas it can start in the thousands at initial screening. Increasing efficiency in the search may therefore address the step that is most conducive to automation, and the one that researchers may find the most daunting.

EB Methods Verdict:

Automatic data extraction even from a single source is not error proof. Additionally, many of the studies to be included in an evidence synthesis are unlikely to appear in a single source. Other aspects of the evidence synthesis, particularly speeding up study identification/filtering, are more likely to have a more significant impact on increasing their completion speed.

Declaration of interests

Both TF and RP are authors of several Systematic Reviews and have received research funding from the UK NIHR, the MRC, the BHF and Wellcome to carry out some of these reviews. RP has received grants from the UK NIHR, Cancer Research UK, BHF, NIHR School of Primary Care Research and support from the NIHR Oxford Biomedical Research Centre and the NIHR Community Healthcare MedTech and In Vitro Diagnostics Co-operative.

References

1. Borah R, Brown AW, Capers PL, et al Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry *BMJ Open* 2017;7:e012545. doi: 10.1136/bmjopen-2016-012545
2. Allen IE, Olkin I. Estimating Time to Conduct a Meta-analysis From Number of Citations Retrieved. *JAMA*. 1999;282(7):634–635. doi:10.1001/jama.282.7.634
3. Tsafnat G et al. Systematic review automation technologies. *Systematic Reviews* 2014, 3:74

4. Jonnalagadda SR, Goyal P and Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews* 2015;4:78. doi.org/10.1186/s13643-015-0066-7
5. Pradhan R et al. Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses. *J Clin Epidemiol* 2019; 105:92-100. <https://doi.org/10.1016/j.jclinepi.2018.08.023>
6. O'Connor AM et al. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews* (2018) 7:3. DOI 10.1186/s13643-017-0667-4