Opportunities for computer support for systematic reviewing - a gap analysis

Linh Hoang[0000-0002-5565-1586] and Jodi Schneider[0000-0002-5098-5667]

University of Illinois at Urbana-Champaign, Champaign IL 61820, USA
{lhoang2,jodi}@illinois.edu

Abstract. Systematic review is a type of literature review designed to synthesize all available evidence on a given question. Systematic reviews require significant time and effort, which has led to the continuing development of computer support. This paper seeks to identify the gaps and opportunities for computer support. By interviewing experienced systematic reviewers from diverse fields, we identify the technical problems and challenges reviewers face in conducting a systematic review and their current uses of computer support. We propose potential research directions for how computer support could help to speed the systematic review process while retaining or improving review quality.

Keywords: Systematic review, Meta-analysis, Gap analysis, Interview study

1 Introduction

A systematic review is a type of literature review designed to provide all available evidence on a given question. Systematic reviews can support translation of research into practice, when the underlying research has concordant findings; and they can also draw attention to gaps in the evidence, such as discordant findings that need further investigation. Despite their importance, systematic reviews require great amount of human effort: a mean of 67 weeks from deposit of a protocol to publication of the review [1], with a mean of 1000 hours of person time [2]. To address this, informatics and methodology researchers are working to minimize the effort required to complete systematic reviews [3]. Already, several commercial software packages have been designed as end-to-end tools to support the reviewing process. Dissemination of tools and methods is an ongoing effort (for instance by the Medical Library Association and by Cochrane [4]) and there are some large-scale efforts to transform the production of systematic reviews (e.g. Cochrane's Project Transform [5]). However, the gap between reviewers' current practices and existing computer support is not well understood. Through interviews with systematic reviewers, we seek to identify the gaps between the computer support available and what reviewers actually use, at a Research I university without an academic medical center. Our two main research questions are:
R1. What technical problems and challenges do reviewers face in conducting a systematic review?
R2. What current computer support technology are reviewers using?

2 Background

The process of conducting a systematic review includes a series of steps designed to locate and synthesize all available evidence on a specific research question. Figure 1 shows the typical steps as described in [6]: after identifying relevant studies, reviewers extract data from these studies and evaluate and interpret the evidence. While the typical methodological challenges [7] and methodologies are well-documented (e.g. PRISMA[1], Cochrane Handbook[2], among many others), review takes varied forms [8]. Since the reliability of a systematic review hinges on the completeness of the information used, a systematic search [9] is of key importance, though this work can come at a cost: In a typical systematic review, over 2000 abstracts need to be reviewed in order to find 15 relevant studies [10].

This intense cost in time and effort has led to the development of computer support tools. Previous survey research has found that reviewers typically use software such as EndNote, Reference Manager, RefWorks, and Excel to manage references [11]. Some commercial products are designed as end-to-end support tools: DistillerSR[3] and Covidence[4] primarily provide an integrated environment for data capture and management, for tasks such as harvesting search results from databases, screening studies, and providing questionnaires for manual data extraction. Another end-to-end tool, EPPI-Reviewer[5], provides (and continues to develop) advanced features such as automatic term reorganization, and document clustering and classification, using machine learning and data mining. The Systematic Review Toolbox[6] collects and describes relevant tools. Currently, research prototype systems are in development to support or automate each of the steps shown in Figure 1. A 2014 review [3] listed fourteen tasks that could potentially be automated, and identified more than 10 applications being developed to assist different phases of the review process, including search engines (Quick Clinical and Metta); and data extraction support tools using machine learning and natural language processing (ExaCT and RobotReviewer). Yet to understand whether and how reviewers are actually using these tools, and how well the current and emerging technologies fulfill reviewers' requirements, a gap analysis is needed.

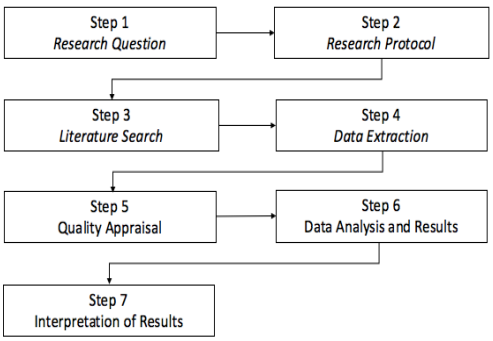

Fig. 1. Steps in a systematic review process according to [6]

[1] Preferred Reporting Items for Systematic Reviews http://prisma-statement.org/
[2] http://training.cochrane.org/handbook
[3] https://www.evidencepartners.com/
[4] https://www.covidence.org/
[5] https://eppi.ioe.ac.uk/cms/
[6] http://systematicreviewtools.com/

3 Methods

We conducted a series of semi-structured interviews with 16 systematic reviewers who had co-authored at least one published systematic review. We used interviews in order to investigate the technologies reviewers use and why, based on our interviewees' detailed explanations [12]. Potential interviewees were initially identified by searching for "systematic review" in publication databases (e.g., Scopus, limited by affiliation) and university websites. Our email invitation and our interviews both ended by asking who else we should consider interviewing; this led us to add publications with "meta-analysis" in the title. After a number of interviews, we focused our recruiting on maximizing the diversity of interviewees fields and career stage since, for instance, faculty were far less likely to accept interview invitations than graduate students.

Our data analysis was rooted in thematic analysis [13]. We recorded and transcribed interviews, then coded transcripts using ATLAS.ti 7, starting with 4 preliminary codes related to our research questions: systematic reviews in practice; difficulties and challenges; current technology support; and opinions and suggestions about technology support. We iteratively coded transcript segments, allowing more specific sub-codes to emerge within the initial coding framework. After several rounds, we identified two themes and collected sub-coded data to support these two themes described next.

4 Interview Analysis

We first report interviewee demographics and then describe two prominent themes:
(1) Technical challenges in the current practice of conducting a systematic review.
(2) Limitation of technological support in the systematic review process.

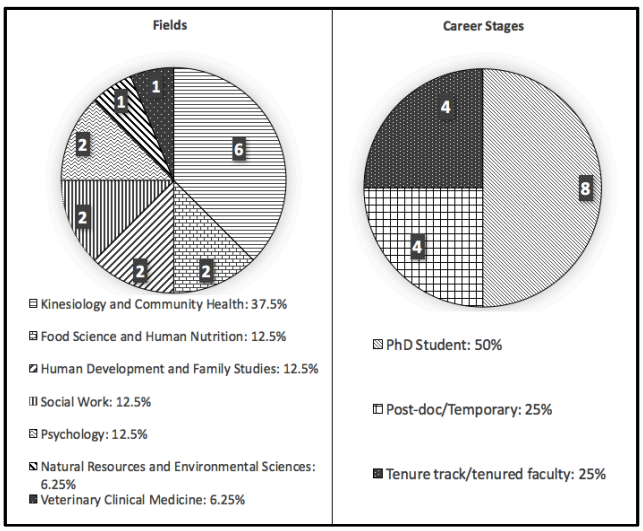

Fig 2. Interviewees' Demographic Information

4.1 Interviewee Demographics

Our interview study comprised sixteen interviewees associated with a Research I university without an academic medical center. Figure 2 summarizes interviewees' fields and positions. Overall our 16 interviewees had published 25 reviews, 55% had published 1 systematic review and the remaining 45% had published at least 2 systematic reviews. Half (50%) of our interviewees were also actively working on a new systematic review project.

4.2 Interview Results

Theme 1: Technical problems and challenges in the current practice of conducting a systematic review

Our interviewees described multiple technological problems and challenges they face in the current practice of systematic reviewing, summarized in Table 1.

Table 1. List of technical problems and challenges in each phase

Phase	Technical problems and challenges
Research Protocol	• Lack of a collaborative platform
Literature Searching	• Lack of comprehensive search strategy • Varied vocabulary • Database coverage • Manual screening process
Data Extraction	• Manual data extracting process

Problems and challenges in the Research Protocol phase:

The lack of a collaborative platform: During the research protocol phase, reviewers needed to decide how to communicate with each other and how to share data. Many of our interviewees were hampered by the lack of a collaborative platform for sharing data and communicating with the team. The most common methods for sharing data were either to manually copy data and send it to other team members or to copy into a cloud storage space (e.g. Box or Dropbox) that the review team could access at the same time. Reviewers often described spending a huge amount of time figuring out how to share their work together during the process.

"I then have the students go back to that link where we can find the full text and save a copy of the full text in a folder that we use on Box." (P15)

The team size averaged five people for a small to medium systematic review and the average time to complete was about one year. This raised a real difficulty of how team members communicated efficiently during the reviewing process.

"The other two studies we had everybody in the same department, and we were meeting on a weekly basis. So, it was kind of easy to coordinate meetings, and to motivate each other to keep going. With this particular one, where you have three or more institutions involved and everybody having different time commitments, I think it will take longer." (P18)

This communication problem not only exists across reviewers from different institutions, but also between the review team members who regularly meet face-to-face throughout the project.

"That actually is very difficult to get everyone on the same page to have them all understand what's going on." (P16)

While multiple software is used throughout a systematic review, even in the same step, most of this software works separately. The lack of streamlined connections between this software creates a serious threat to data integration. It leads to a potential data loss problem when users try to transfer data from one software to another or when multiple users working on the same articles at the same time.

"When we got down to have 34 articles left. My advisor and I were both reading articles so if I could've went in and logged in and seen, all right he's read these first 10. I could've probably read those too and then we could've talked about them. But if he was updating one and I was working on opposite ones then we weren't kind of getting to a point where we should sync together." (P8)

Problems and challenges in the Literature Searching phase: The literature search is one of the most difficult and time-consuming steps in the whole review process.

The lack of comprehensive search strategy: Our interviewees often start the literature search without a consistent, well-designed strategy. Interviewees normally start by identifying simple keywords which they use to pull out "potentially relevant" studies from online databases. Then they quickly scan through the initial search results—normally up to thousands of papers for the first round—to determine whether the studies may actually be relevant. Interviewees keep revising the search algorithms by using alternative keywords and repeat the "search - screen" tasks multiple times until they "feel" that all studies are captured correctly.

"I remember one time when I selected some kind of keywords… Traditionally, I applied maybe the first attempt and then I would read some of the titles to see what I missed something that worrisome, my keywords. So, I need to do some refine of the algorithm. Maybe 4-5 times." (P7)

Determining whether all potential relevant studies have been captured is also another concern. Oftentimes, the reviewers' biggest fear is that they do not know whether they got everything from the search.

"The searching, I never knew if I got everything or I did it right, I never knew if my search terms are good enough. I could have had search terms that never turned anything out, I never knew if I was searching the right databases." (P3)

It takes time, and requires adequate knowledge of the review topic, for reviewers to figure out which terms should be used in their searches, and to identify useful keywords and variants.

Varied vocabulary: Interviewees repeatedly mentioned that it is common for people from different fields, or even in the same field but doing research from different angles, to use different terminologies to describe the same thing within the same narrow topic. Thus, the process often requires extra time for reviewers to read through the search results, identify alternative keywords, and then revise the search terms accordingly.

"People use different phases to describe one same thing. So, when you started to search for related studies, you only start with one phrase, and then you realize that they use different phrases, so you need to revise all the time in order to search all of the relevant studies." (P5)

Varied vocabulary is a core and enduring issue in library science that especially impacts multi-disciplinary work [14]. Articles may mention the keyword reviewers search on

without talking about the same topic or analyzing it in a way reviewers find relevant. Not all interviewees seem familiar with the techniques of using a controlled vocabulary such as MeSH as an efficient alternative to keyword search.

"Different fields looked at the same construct in different ways, and so you had to make sure you were capturing it by using all the keywords possible for that construct you're interested in investigating." (P6)

Database coverage: Another long-recognized problem, database coverage and information scattering [15], also poses challenges according to our interviewees.

"We have the problem of all the journals that are not available. Even smaller journals that are not available when we're doing the search terms. So, good studies were just not coming up. They're not indexed in the places where we're searching, so we're not finding them…" (P15)

Manual screening process: One of the most critical problems in the screening phase is its manual nature, and the lack of trusted technical support for this process. Our interviewees typically export search results from online databases to an application (usually EndNote) in order to perform screening tasks including title, abstract, and full-text screening. The average number of studies screened in each review as reported by our interviewees is approximately 4000 studies. Due to the large number of search results, the screening process is considered one of the longest steps to complete.

"I think once you're screening, you're screening everything. You go title, then you go abstracts, then you go full studies. That takes a lot of time and a lot of understanding, and your part about everything and how everything's interconnected. ... Sometimes you have to be very strategic because titles may not necessarily really imply, looking at it, so maybe you hold off on taking it on." (P6)

"The quality checking and the inclusion/exclusion criteria, the abstract, pulling out…that stuff is so time-consuming when you do it by hand and there is no need for that to be done by hand, but that's the only option you have." (P2)

Limitations of current support tools will be discussed in more detail under Theme 2.

Problems and challenges in the Data Extraction phase:

Manual data extraction process: Data extraction is another time-consuming, highly manual phase, in part because data is normally extracted separately by at least two reviewers, who then come to an agreement about which information and how much of it should be used for the subsequent synthesizing stage.

"It's totally a manual process and you need to be very careful because the results will be published in public. So, I think each of the article I need to go through maybe 3-5 times." (P7)

"What I did was I extracted the data. I actually manually entered it on an Excel, whatever data is available from the studies." (P12)

Our interviewees describe being accustomed to reading the studies manually by themselves without any advanced technology support. Some interviewees even printed out articles and performed the data extraction manually with pen and highlighter.

"I downloaded all the papers for, well, it probably took a long time. I'm not very good about reading on the computer, so I actually printed them all out. I started organizing them just briefly by heading, like the subject. … Then essentially, I went through and I read every one. While I was reading them, I took notes for myself. Just on a paper, on each of them." (P10)

Theme 2: The limitations of technology support

The second theme that emerged from the interviews is the limitations of technology support in systematic review. Despite a large amount of commercial software, we notice a gap between the technology available and what our interviewees are using. Figure 3 summarizes the technology support our interviewees used in the review process. The most common applications that are used in the process are Excel and EndNote, which are not designed specifically for systematic review, and in some ways, they do not meet all reviewers' requirements.

Excel was the most common software package our interviewees reported using. Excel was especially used in the data extraction process and sometimes for copying and pasting reference lists during searching and screening. Excel is popular among our interviewees because it allows users to organize data in tabular format. Nevertheless, the software is not specifically designed for bibliographic purposes. It has limitations especially for organizing a large number of publications. The most commonly used version of Excel (version 2013) is an offline application, which leads to data integration problems when users need to export data from one package to another manually.

> *"This is what it looks like [showing the Excel spreadsheet used for [screening]. This is horrible. We ended up color-coding it. It's horrible. It's based off what we included or exclude. This is the abstract. That is the title. We ended up with putting the abstracts in and then need to dump everything in the Excel file." (P2)*

EndNote[7], the second most popular systematic review support software among our interviewees, has somewhat similar data integration issues, but is specifically designed for publishing and managing bibliographies. It is popular in the systematic review community because of the tool's affordances for performing screening tasks. However, our interviewees also reported a number of significant problems when using EndNote. Losing data seems to be the most serious problem.

> *"I'm not sure how reliable it [EndNote] is. The very first obstacle you have is that ... I don't know, sometimes I feel I have 7,000, and then the next day, I have 6,500, and I'm like where did my 500 go? If we do the search with [EndNote] the same exact terms, I won't find the same 7,000. That is one of my main fears." (P9)*

Moreover, reviewers reported that it was not straightforward to share an EndNote library (especially version 6 backwards) between collaborators. In order to do so, users either needed to use EndNote Web in an online environment, or to export the EndNote library file locally and then copy it for other team members. This is inconvenient

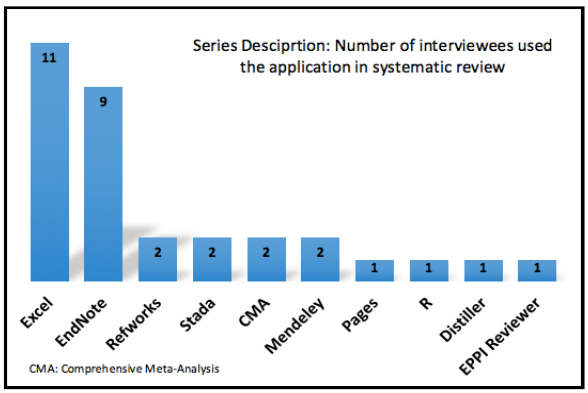

Fig.3. Current applications used by reviewers

[7] http://endnote.com/

and time-consuming, and especially within a large group of reviewers and with a large number of studies, data loss became a common problem.

"We use EndNote Web, it ended up deleting all the information at some point because we were sharing it. One day we woke up, there was nothing on it." (P9)

Some reviewers are aware of or use meta-analysis software such as Stata[8], R[9], and Comprehensive Meta-Analysis[10]. Even though some software is specifically designed to support the systematic review process, they are not popular with our interviewees. Only two of our 16 interviewees mentioned using more advanced end-to-end applications. Particularly, P17[11] mentioned using EPPI-Reviewer for a published review and P18 mentioned using Distiller for a future review.

Steep learning curve was one reason mentioned for software avoidance.

"I used formulas to convert them into mean and standard deviation. I could have used Stata or whatever but I did it manually. It would take more time to figure out how to do in that software rather than do it manually." (P12)

The size of review also impacted the decision to use software.

"For the smaller sample of studies, when we had between 10 and 15, it was doable manually. But when we're looking at 130 studies, I'm hoping that Distiller is going to cut short some time, at least by half if not more. Just so that we can get this process completed in a timely fashion." (P18)

5 Gap analysis, discussion and future work

Our interview results indicate a gap between the technology support available and what technology is being used by our reviewers. Despite the existence of advanced technology support (e.g. end-to-end applications or automation programs discussed in the background section), most steps are still done manually, making the review process more time-consuming and inefficient than it needs to be. There seem to be four potential explanations: first, reviewers might not be aware of these technologies; second, reviewers might have limited access to these technologies due to cost; third, reviewers might be stymied by actual or perceived learning curve and may prefer simple, familiar tools that require less training (e.g. the preferences of using such tools like EndNote and Excel); and fourth, tool features and availability may have changed since interviewees started their reviews.[12]

We also acknowledge the limitations of our study. Our conclusions about methodological problems may not be applicable for the whole population of reviewers since we had a small sample that may not have been representative beyond the large

[8] Data Analysis and Statistical Software http://www.stata.com/

[9] The R Project for Statistical Computing https://www.r-project.org

[10] Comprehensive Meta-Analysis https://www.meta-analysis.com

[11] We report 16 interviews with systematic reviewers; we exclude from our report 2 interviews with librarians who support and conduct systematic reviews.

[12] The three end-to-end systematic review tools were commercially available circa 2002 (EPPI-Reviewer v2), 2010 (Distliler), and 2013 (Covidence). For comparison, the oldest systematic review authored by our interviewees was published in 2012.

Research I university where we conducted our work. Future work should seek an even more diverse sample, with the awareness that multiple aspects may impact reviewers' practices and propensity towards computer support.

These interview results open up multiple directions for future research. Automation is not the only opportunity. Facilitating communication between team members could help make reviewing faster and more efficient, because according to interviewees, the more they communicated, the faster review tasks could be done. Bridging between low-tech and high-tech solutions, or integrating smaller tools into a custom pipeline might also help. Dissemination work is also needed, especially beyond the clinical medicine community, to help reviewers become more familiar with these existing applications as well as with co-evolving methodologies. A comprehensive review of what applications are available with detailed analysis of their costs, availabilities, feature advantages and disadvantages could help. One of our interviewees specifically called for cross-pollination between evidence synthesis methodologies in different fields.

Trust and accountability of software is another area that needs further development. For instance, automatic data extraction is one of the newest focus areas for systematic review automation research, often involving machine learning and natural language processing. However, our interview findings show that, once they reach the data extraction step, our interviewees prefer to read and extract data from included studies themselves. Being able to check machine results in a natural way (such as RobotReviewer's inline annotation of extracted data [16]) could help reviewers gain trust and identify further development needs for specific software. The ability to experiment with tools and observe their results is likely to increase reviewers' acceptance of new technologies [17].

Further research is also needed to address the enduring vocabulary and scatter problems which heavily impact the systematic review community (e.g. retrieval of ~2000 references in order to find ~15 relevant studies [9]). One underexplored approach is to use science mapping tools, ranging from visualization to automatic citation network generation. Another idea is to develop a vocabulary mapping mechanism, which could collect keywords from scientific studies across fields, then identify the term definitions in order to map related terms with the same meanings together. By doing that, once reviewers search for studies that include a term, the system would be able to identify which other terms potentially have the same meaning. We believe these future directions for computer support could help to speed the systematic review process while retaining or improving review quality.

Acknowledgments
We would like to show our gratitude to all of the interview participants for sharing their experiences and also the pearls of wisdom that allowed us to complete this study. We would also like to thank our colleagues Lori Kendall and Peter Darch for discussions about qualitative research methodologies; Susan Lafferty who provided expertise that greatly assisted in the IRB process; and Katrina Felon for comments that greatly improved the manuscript. The research leading to these results has received funding

from the National Library of Medicine: "Text Mining Pipeline to Accelerate Systematic Reviews in Evidence-based Medicine" (R01LM010817).

References

1. Borah, R., Brown, A. W., Capers, P. L., Kaiser, K. A.: Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 7(2), e012545 (2017).
2. Allen, I. E., & Olkin, I.: Estimating time to conduct a meta-analysis from number of citations retrieved. Journal of the American Medical Association 282(7), 634-635 (1999).
3. Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., Coiera, E.: Systematic review automation technologies. Systematic Reviews 3(1), 74-88 (2014).
4. Turner, T., Green, S., Tovey, D., McDonald, S., Soares-Weiser, K., Pestridge, C., Elliott, J.: Producing Cochrane systematic reviews—a qualitative study of current approaches and opportunities for innovation and improvement. Systematic Reviews 6(1), 147-157 (2017).
5. Thomas, J., Noel-Storr, A., Elliott, J.: Human and machine effort in Project Transform: How intersecting technologies will help us to identify studies reliably, efficiently and at scale. Cochrane Methods Supplement 1, 37-41 (2015).
6. Wright, R. W., Brand, R. A., Dunn, W., Spindler, K. P.: How to write a systematic review? Clinical Orthopedics and Related Research 455, 23-29 (2007)
7. Anderson, N. K., Jayaratne, Y. S.: Methodological challenges when performing a systematic review. European Journal of Orthodontics 37(3), 248-250 (2015).
8. Grant, M.J. and Booth, A.: A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal 26(2), 91-108 (2009).
9. Bartels, E.M.: How to perform a systematic search. Best Practice & Research Clinical Rheumatology 27(2), 295-306 (2013).
10. Ross-White, A., Godfrey, C.: Is there an optimum number needed to retrieve to justify inclusion of a database in a systematic review search? Health Information & Libraries Journal 34(3), 217-224 (2017).
11. Lorenzetti, D.L., Ghali, W.A.: Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. BMC Medical Research Methodology 13(1), 141-145 (2013).
12. DiCicco-Bloom, B., Crabtree, B.F.: The qualitative research interview. Medical Education 40(4), 314-321 (2006).
13. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology 3(2), 77-101 (2006).
14. Svenonius, E.: The intellectual foundation of information organization. MIT Press, Cambridge (2000)
15. Sutton, S.: Encyclopedia of library and information sciences. Taylor and Francis, Abingdon (2009)
16. Marshall, I.J., Kuiper, J., Wallace, B.C.: RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association 23(1), 193-201 (2015).
17. Thomas J.M: Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? OA Evidence-Based Medicine 1(2), 12-17 (2013).