Systematic searching: practical ideas for improving results

Edited by Paul Levay and Jenny Craven

9 Evidence surveillance to keep up to date with new research

James Thomas, Anna Noel-Storr and Steve McDonald

9.1 Overview of the topic

Research is being published at an ever-increasing rate and it is becoming more and more difficult for systematic reviewers to find research in a timely way and keep existing reviews updated as new studies are published.[1] This is a particular problem for organizations which maintain libraries of systematic reviews, such as the Cochrane and Campbell Collaborations, as the more systematic reviews they publish, the greater the burden of maintenance. It is also a challenge for guideline-producing organizations which, for pragmatic reasons, typically invest significant resources and effort in one-off periodic updates without knowing whether the evidence base has changed or has actually changed so rapidly that more frequent updating would have been warranted. Previous work has shown that systematic reviews can date very quickly – with some out of date as soon as they are published [2] – and it is becoming clear that our current methods of research curation are wasteful of societal investment in research, and risk resulting in suboptimal outcomes.[3]

This chapter is concerned with this problem of 'data deluge' and the need to maintain a better surveillance of research in order to keep abreast of new developments. It is thus related to work on *Living Systematic Reviews* (LSRs), that are 'continually updated, incorporating relevant new evidence as it becomes available'.[4, 5] The chapter outlines developments in automation technologies that are already making the systematic review process more efficient and then focuses on the way that global research curation systems are organized. The chapter suggests that new approaches are needed in order to support the production of evidence syntheses in efficient and timely ways. Case studies 9.1 and 9.2 explain how these new developments are being put into practice to realise these benefits.

9.2 Discussion

New ways of working that integrate and capitalize on automation are necessary to tackle the growing burden of identifying and synthesizing research. New technologies – which range from the mundane (such as identifying duplicates in bibliographic records) to full Artificial Intelligence (AI) systems – are under constant development and are already assisting various aspects of the evidence curation (see box) and discovery process. It is possible to break these new tools down into two broad categories:

1. Tools which can make existing manual processes more efficient.
2. Tools which aim to change the way systematic reviews are carried out in more fundamental ways by linking tools together and changing the sequencing of activities.

The following section examines the potential for automation to assist in existing processes, and the section after that considers the potential for linking these tools into integrated surveillance systems for LSRs and other types of living evidence, such as guidelines.

What is 'curation'?
A key concept in this chapter is 'evidence curation'. 'Curation' as an idea has been around a long time, and concerns the activities necessary to manage, sort and arrange information. We see current methods for finding and processing evidence as inadequate, since they result in so much repetitive and avoidable work in systematic reviews. We will therefore use the word 'curation' throughout to denote the work necessary to organize research information in a way that facilitates its easy discovery and reuse. In this sense, 'curation' also involves an acknowledgement that research data has value, and this value needs to be protected through fit-for-purpose management.

9.3 How automation can make existing processes more efficient

The systematic review process follows what is now a well-trodden path which, once the research questions have been agreed and the review team established, usually includes the following steps:

- Developing the search strategy
- Searching databases and other sources, and downloading the results
- De-duplication of records
- Screening records for eligibility
- Retrieving the full text of potentially eligible records
- Screening full text reports
- Checking bibliographies and citation indexes
- Data extraction and quality assessment
- Synthesis
- Write-up of final reports.

This list emphasizes the retrieval and curation tasks that occur early in the review process. The development of enabling tools has tended to focus on these tasks since many are repetitive and time-consuming (e.g. reference screening) and also more amenable to machine assistance. We will now consider each of the stages in terms of the available tools and their readiness for use in reviews.

9.3.1 Developing the search strategy

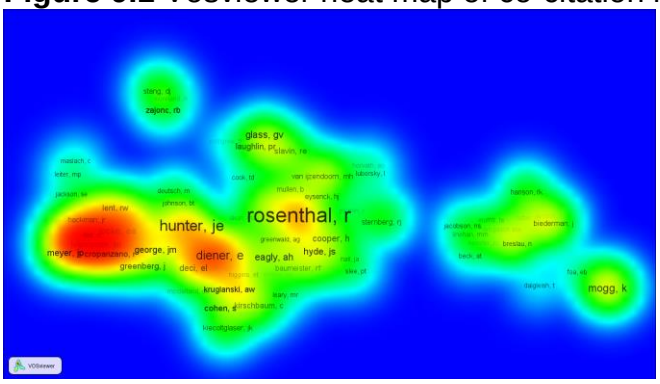
As Chapter 7 shows, there has been a proliferation of tools to help with developing search strategies. Previous work has shown that the way that search strategies are developed can impact on the recall of the search (e.g. [6]), and in a case-study, Stansfield and colleagues [7] describe how text analytic software is able to assist in five ways:

1. improving the precision of the search and so reducing manual effort in screening results.
2. identifying search terms which can improve the sensitivity and reliability of the search.
3. assisting with the way that searches can be translated between databases (for example the Polyglot Search Translator http://crebp-sra.com/#/polyglot).

4. searching and screening within an integrated system [8].
5. developing search strategies objectively (see also [9]).

There is interest in using relationships between citations for identifying other relevant studies (e.g. [10]), and tools such as Google Scholar (https://scholar.google.co.uk), Microsoft Academic (https://academic.microsoft.com), Scopus (https://www.scopus.com) and Web of Science (http://wokinfo.com) all offer tools to support this. An important issue to bear in mind – which can be highlighted using new technology – is the care required when following citation trails. Research is often carried out by communities of researchers and if they fail to cite the work of other groups working in the same field, then this kind of pearl growing will not find all relevant pockets of research. A heat map of a citation network of systematic reviews in psychology (Figure 9.1) shows graphically how related research does not always connect. Information specialists should be mindful of this, utilizing tools to help visualize and identify these kinds of disconnect to ensure that search methods do not rely exclusively on specific sources and approaches.

Figure 9.1 Vosviewer heat map of co-citation relationships in psychology



9.3.2 Searching databases and other sources, and downloading the results

New technologies for searching databases are emerging all the time, and Chapter 7 has highlighted some of the main tools being developed for searching PubMed and MEDLINE (e.g. PubReMiner). Tool development for this dataset is particularly common because of its size and the fact that it is freely accessible. Automation tools rely on Application Programming Interfaces (APIs) which can enable systematic review tools to access databases 'behind the scenes'. For example, users can search PubMed from within EPPI-Reviewer without needing to visit PubMed directly, and Cochrane's Central Register of Studies (which sits behind the public CENTRAL database) also searches sources such as Embase and ClinicalTrials.gov at the programming, rather than user, interface level. Although API access to commercial databases is more complicated because of the issues involved in retaining and reusing references it is becoming more commonplace.

Being able to search databases from within systematic review tools can save the user having to download multiple text files and then reload them for screening. The process of ensuring compatible formats (e.g. RIS) can also be more reliable using this route. However, it is when we consider the need for automated searches to facilitate surveillance that API access to databases becomes particularly important. Here, we need regular searches to be run, and relevant references retrieved, automatically with minimal user interaction. While some sources, such as PubMed, are relatively unproblematic, others do not offer API access at all, limiting the scope for full implementation of an automated surveillance process at this time.

9.3.3 De-duplication of records

Searching multiple sources inevitably results in the retrieval of duplicate references. Although bibliographic software tools like EndNote can minimize much of the workload in identifying duplicates, eliminating them altogether is not a simple task. Records entered manually by different providers can contain slightly different information. Some research has been carried out into developing and evaluating de-duplication algorithms (e.g. [11, 12]), but it would be fair to say that further evaluations are required, including increasing the availability of 'gold standard' datasets on which to evaluate new algorithms.

9.3.4 Screening records for eligibility

The ability of text mining and machine learning tools to assist in the screening of records for eligibility has been outlined in Chapter 7. There are two approaches to highlight here in the context of evidence surveillance. First, the use of 'active learning' – whereby the machine learns the review's inclusion criteria in an iterative fashion alongside human screening [13] – is still of relevance when we are updating an existing review. We naturally already have all the screening data generated when the review was originally written and from any subsequent updates. This means that it is possible to build a machine learning classifier to rank the relevance of newly retrieved references at the outset of an update and, because it can utilize more data than are typically available when a new review is conducted, it is likely to perform better too.

Second, the use of study-type machine learning classifiers is now becoming more widespread. These are classifiers which have been trained from large quantities of data (and so are highly accurate) that are able to classify papers according to the type of study they describe. While not all study types are yet covered, there are some high-performing classifiers now which can distinguish between records which do, and do not, describe randomized controlled trials (RCTs) with a high degree of confidence [14, 15]. Since many systematic reviews I the health sector only include RCTs, being able to eliminate the records which are very unlikely to be describing RCTs can be an efficient way of reducing manual effort.

9.3.5 Full-text document retrieval and screening

Many tools, including generic bibliographic software such as EndNote, now offer the facility to identify and download full-text reports based on citation records. The Digital Object Identifier (DOI) system has transformed this task, and the main barrier to full automation is the fact that many documents lie behind journal subscription paywalls. Services such as CrossRef are making the identification of these documents more efficient, even if the last phase of document recovery needs to be undertaken manually.

9.3.6 Checking bibliographies and citation indexes

Current tools that allow reviewers to undertake citation searching – and examine both the references cited in a given paper, and those that cite it – have two main weaknesses. First, since no tool is truly comprehensive it may be necessary to use multiple services such as Web of Science, Scopus and Google Scholar. Second, extracting the bibliography from a paper and the automated linking of references to one another is challenging; tools often fail to list, for example, all the papers in a given bibliography. That said, the OpenCitation (http://opencitations.net) and CrossRef (https://www.crossref.org) initiatives are gaining ground, and these systems are becoming increasingly comprehensive. If major publishers, such as Elsevier (https://opencitations.wordpress.com/2017/11/24/elsevier-references-dominate-those-that-are-not-open-at-crossref), sign up to the consortium, then the case for making increased use of citation networks will become even more compelling [16] (the warning above about relying too heavily on citation trails notwithstanding).

9.3.7 Data extraction and quality assessment

As we move through the review process, the number and maturity of tools for deployment decreases. There is still significant potential to reduce manual workload but it is much more difficult to build an automated data extraction system than it is to build one for screening search results. Every report of a study is unique and identifying even apparently simple information, such as the number of study participants, is quite a challenge. The ExaCT tool aims to extract information from clinical trial reports [17], and the RobotReviewer tool can automate the risk of bias assessment for RCTs [18].

9.3.8 Synthesis and write-up of final reports

While writing the report is currently beyond the capacity of even the most advanced systematic review AI system, some tools do aim to write text around the results of a statistical meta-analysis. Slightly more prosaically, it is recognized that many sections of a systematic review report standard processes; and standardized wording might be entirely appropriate. Currently, support for the needs of LSRs, where information about new studies automatically appear in the right part of the report, is scant, with few alternatives available beyond those for standard systematic reviews.

9.4 Creating surveillance systems with how automation tools

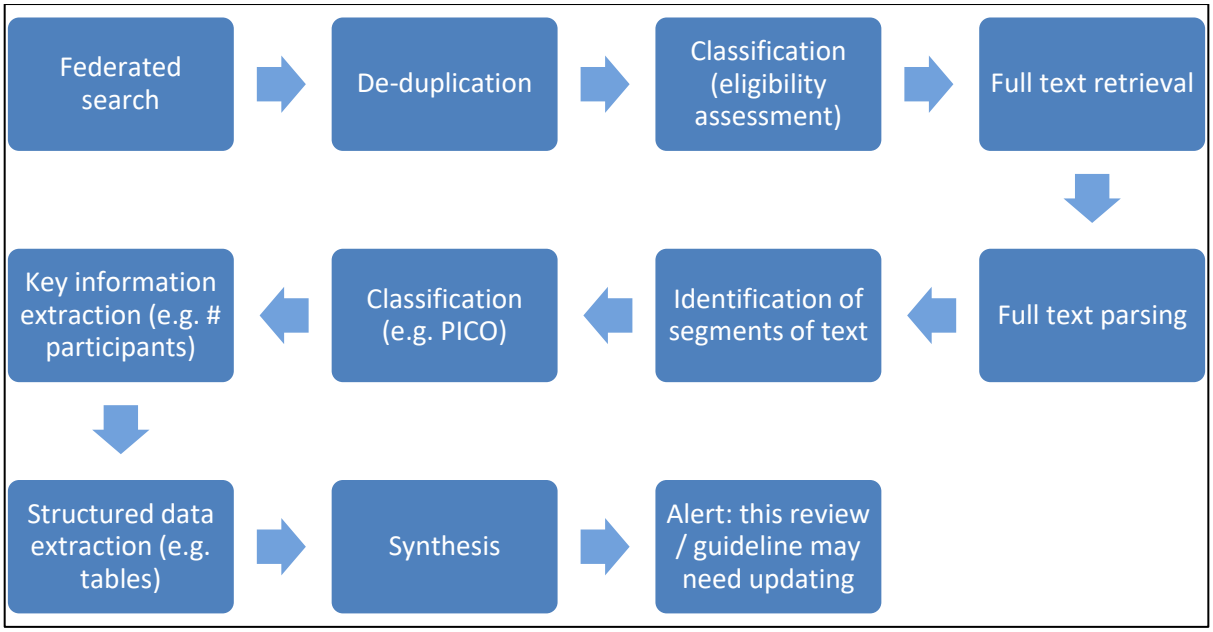
Until now, this chapter has considered how individual automation technologies can improve the efficiency of existing tasks and support an LSR workflow.

However, the real potential of these technologies, and the promise to bring about fundamental changes to the way is evidence is identified, lies in joining them together.

Consider the workflow depicted in Figure 9.2. This shows the key steps involved in the information retrieval and curation aspects of the review in a continuous process from a 'federated search' through to 'synthesis'. Outlined here is a 'living' process which maintains and enables a constant surveillance of the evidence as it becomes available whether as, for example, published papers or registrations in trials registries. One critical difference between the 'micro' systems described above and the workflow in Figure 9.2 is that a full surveillance system operates across wide areas of research, rather than individual reviews; it is concerned with the up-front identification of evidence at scale, rather than the one-off processes which accompany the traditional systematic review process. Instead of reviewers having to periodically interrogate multiple sources to identify whether any relevant evidence has been published, there is now the promise of a system that can prospectively signal when new evidence is available.

Conceptualizing evidence discovery as an up-front process across a domain brings economies of scale. For example, the need for sensitive searches leads to the same studies being examined multiple times by people doing different but related systematic reviews across the globe; for reviews aiming to find RCTs, the same assessments are being made numerous times. Currently, there is no system for ensuring that these assessments do not need to be made the next time a given study is retrieved, and so systematic reviewers frequently duplicate one another’s work. If however, knowledge generated about a study (whether automated or manual) contributed to a domain-wide map of the evidence base, we would know that if a study is relevant for a particular review then it is also *not* relevant for many other reviews. Thus, a full surveillance system uses information about a study’s relevance to answer one question to know simultaneously that it is not relevant for others.

Figure 9.2 Living Systematic Review and evidence surveillance workflow



One vision for an integrated system is that it involves automated:

- searches of major databases
- de-duplication
- full text retrieval
- assignment of new research to appropriate reviews
- data extraction
- synthesis.

While the synthesis is unlikely to be as reliable as a human-controlled process, it may be sufficient to indicate whether the 'new' study is likely to change the review's conclusions. For example, if an automated system is able to estimate the size of a study – and the direction of its effect – it might be able to alert authors as to the likely impact of the new study in a given review. As the description of tools above has indicated though, the current state of tool development, the limited availability of APIs and the difficulty of obtaining full-text content held behind paywalls, means that a fully automated system is currently unachievable.

Full automation may not be a necessary, or even desirable, goal for an effective evidence surveillance system, if moving towards an integrated process (that minimizes the duplication of effort and identifies relevant studies up-front) brings significant efficiencies on its own. Critical to the success of such a system are effective tools to ensure knowledge is shared between reviews and not lost; and a good interaction between human and machine effort to maximize the benefit that is gained from human input and minimize the expenditure of human effort on tasks which can be automated. Case studies 9.1 and 9.2 describe how two evidence surveillance systems are being created to put these innovative ideas into practice.

Case study 9.1 Cochrane Evidence Pipeline

Introduction

The Cochrane Collaboration is an international collaboration of clinicians, researchers, patients, policymakers and others who are concerned with evidence-based decision-making to improve people's health. Central to the work of the collaboration is the production of rigorous systematic reviews which are published in an online library. There are over 7,500 systematic reviews, a significant effort on the part of Cochrane's many thousands of volunteer contributors. Ensuring these reviews are up-to-date so they continue to be useful for decision makers is an even more formidable challenge.

Duplication of effort is a problem across Cochrane, with individual reviewers and review teams using different systems for their reviews. Cumulatively, they have probably screened more than 40 million references over the past 20 years to find RCTs (mostly) for inclusion in Cochrane reviews; yet there have probably been little more than 1 million RCTs conducted so far in human history. It would be much more efficient to find and collate all RCTs and for reviewers to look in this limited pool, than to continually pour over the same records, making the same decisions about a reference's eligibility.

This is the problem that the Cochrane Evidence Pipeline has been designed to solve. As Figure 9.3 shows, research enters the pipeline at the top of the graphic and ends in the Cochrane Register of Studies (CRS) ready for inclusion in systematic reviews. In between are a number of stages which enrich the study record with additional information. Critically though, this is a mixture of human and machine effort: no part of the process is expected to be fully automated.

At the start of the process, research reports enter the Pipeline, typically these are references to articles indexed in bibliographic databases. Of increasing importance is Cochrane's centralized search service, in which monthly searches of key databases feed a constant stream of new research into the system. Not all databases have an API though, so research enters the pipeline through a mixture of automated and manual searches. Importantly though, the aim is for sensitive searches to be carried out up-front, outside the process of any individual review.

What type of study is this?

The first piece of data enrichment to be applied is the type of study that the report describes. Systematic reviews commonly rely on particular study types to answer specific questions, so identifying the right types of study is more important for a database like the Cochrane Register of Studies than a more generic database, such as MEDLINE. The Pipeline is currently focused on the identification of RCTs and has a machine learning classifier which is able to distinguish between RCTs and non-RCTs accurately.[15, 19] This classifier was built with data from Cochrane Crowd (http://crowd.cochrane.org/index.html), a citizen-science platform where volunteers help with specific data curation activities (called 'micro-tasks'), including classifying abstracts according to whether they describe RCTs or

not. An agreement algorithm ensures a high degree of collective accuracy with sensitivity and precision both exceeding 99%. Research is first classified by the machine; those records that the machine classifier rates as *possibly* describing RCTs are then examined by Cochrane Crowd, and a final determination made as to their eligibility. There is a pleasing symbiosis to this machine-Crowd interaction: the initial machine algorithm that was derived from the efforts of the Crowd is now deployed to reduce the number of abstracts that the Crowd needs to assess. As the dataset examined by the Crowd grows, so the accuracy of the machine algorithm improves in a virtuous circle that is freeing up Crowd resource for other tasks. Other machine classifiers include one to identify systematic reviews and economic evaluations.

This work in identifying RCTs is already saving reviewers' time and reducing duplication of effort. An increasingly large dataset of previously assessed records is being accumulated against which new search results can be checked. A Cochrane Information Specialist can upload their search results to the database and have them automatically matched against existing records. A record already assessed as not describing an RCT can be discarded without further manual checks. Remaining records can then be assessed by the machine-Crowd service, leaving very few for review authors to examine.[20]

What is this study about?
While identifying the study type is already saving significant reviewer effort, the LSR process depends on identifying studies relevant to *specific* reviews. The next item in the pipeline – the enrichment of data by identifying PICO (Population, Intervention, Comparison, and Outcome) characteristics – aims to achieve this. PICO precisely describes the scope of a review and is often used as the organizing framework for search strategies. If it were possible to classify research studies according to their PICO, it would be possible to pre-allocate them to reviews, thus not only facilitating the updating of existing reviews, but identifying where there might be gaps in the synthesized evidence. As described in Chapter 8, the Cochrane 'linked data' project has developed a model which encapsulates the PICO structure of clinical trials and a standardized vocabulary to describe their elements in detail.[21] Developing the standardized vocabulary is a significant effort and an ongoing task: it is linked to terms used in other major thesauri (e.g. SNOMED CT, MedRA and RxNorm) and contains many hundreds of thousands of terms.

Figure 9.3 Cochrane Evidence Pipeline
See the final published chapter for a copy of Figure 9.3.

It is already possible to automate the identification of the broad area of a study (e.g. the article is about treatments for heart disease or injuries) but being able to detect the detailed PICO terms (e.g. the article is about men aged over 65 at high risk of heart attack) is a challenging task for automation alone, and so human and machine are again working together

on this problem. At the time of writing, many thousands of Cochrane reviews and their included studies have had their PICO manually classified by Cochrane information specialists, thus providing valuable training material for machine learning. In addition, Cochrane Crowd has recently launched a new task which involves identifying specific concepts in abstracts and associating them with PICO elements. Alongside this, machine learning work is analysing the training data and attempting to automate the PICO predictions. Overall accuracy is approaching 50%, which may not yet be sufficient for global roll-out, but does represent good progress, considering how challenging this classification task is (bearing in mind there are hundreds of thousands of terms for the machine to learn from very little training data).[22]

What are the data?
The final piece of metadata enrichment in the Evidence Pipeline is related to data extraction. Here we have focused mainly on one of the most time-consuming and error-prone aspects of the process, namely the extraction of numeric data. There are two tools under development: one for extracting structured data from tables in PDF documents, and the other for extracting numeric data from graphs. Neither tool aims to be completely automated, but both aim to reduce manual effort by providing some automation, for example by having users check results of the machine automation and undertake aspects of the process which would be unreliable if left fully automated.

After the various machine and human processes have enhanced the study record, it is saved in the CRS database and is ready for use in reviews. The ultimate aim is for authors to receive automated alerts when new content is available in their area.

Case study 9.2 The Human Behaviour-Change Project

The Human Behaviour-Change Project (http://www.humanbehaviourchange.org) shares some characteristics with the Cochrane Evidence Pipeline but aims to take the automation one stage further into the synthesis itself. It is funded by a grant from the Wellcome Trust, and is being carried out by a consortium of behaviour change experts from UCL (principal investigator: Professor Susan Michie), and computer scientists from IBM Research, Dublin and UCL.[23]

The project is concerned with the discipline of behaviour change, since changing people's behaviour is key across a range of social, health and environmental challenges. It is asking the overarching question: which interventions are effective (for whom, in which circumstances) for achieving behaviour change? In order to help researchers and decision-makers to answer this question it is developing a system that will produce recommendations for potentially useful interventions based on the extant behaviour change literature.

The system thus requires some of the same components as the above Cochrane Evidence Pipeline: it needs a system for locating and keeping up to date with research evidence, processes to organize the research, and a system for synthesizing the evidence and making recommendations to users. The first component for locating research is similar to that employed in the Cochrane Evidence Pipeline: a continuous feed of research. Again, in common with the previous example, an ontology is needed to organize the literature and in this case, an *ontology of behaviour change interventions* is being developed and widely consulted upon. Research papers are being annotated according to this ontology in order to provide training data for machine learning systems.

This annotation work is designed with future automation in mind and involves selecting specific pieces of text in research reports and associating these snippets of text with a given classification in the ontology (e.g. a section in the description of the intervention which communicates to the reader which type of behaviour change intervention is being used). As an example, one behaviour change technique involves the 'self-monitoring of behaviour'. This classification has been applied to the following text from three studies:

- 'messages are interactive and prompt users to track smoking, report on cravings, and provide smoking status'.
- 'examples of behavioural treatment strategies used include providing personalized feedback about use of alcohol including comparison of personal level of use to peer norms'.
- 'participants learned about their smoking habit by writing down time of day, situation, and perceived need for every cigarette smoked before smoking cessation'.

Based on these snippets of text (among many others), machine learning systems can then learn to recognize similar pieces of text and make the association with, in this case, the type of behaviour change intervention being described. The project also aims to automate the process of recognizing and extracting results from tables in documents. This is necessary to automate the extraction of accurate statistics describing study findings.

The final component of the project is the system for generating inferences and recommendations for users which are generated by the system reading across the entire field of literature. In order to automate the synthesis of research in this way, the system will need to have achieved a very high level of reliability in terms of the detailed classification and extraction of results from studies. It will then synthesize the results across studies in order to generate recommendations using methods which are similar to those employed in existing methods for network meta-analysis (e.g. [24]).

9.5 Future directions

Case studies 9.1 and 9.2 are both part of a vision for a future of evidence curation and synthesis workflow, which changes the current process of evidence discovery in quite fundamental ways.

First, the task of locating evidence is taken outside the process of individual reviews. Instead, research is identified as soon as it is published (or registered in registries of trials) and the record enhanced with sufficient information to know in which systematic review(s) it might be relevant.

Second, the classification systems being developed and applied go beyond the usual indexing that takes place in some databases (e.g. MeSH and Emtree tagging) because they are more structured and *use orientated* i.e. the domain models for PICO and behaviour change allow the research to be described in ways which precisely describe the characteristics required for synthesis. In the Cochrane Evidence Pipeline, the aim is for the search for relevant studies to start and end with the specification of a PICO question. When the PICO for a given review has been searched for, the system should locate a small number of studies with high precision *and* recall, greatly truncating the current lengthy search and screen process.

The case studies also point to changes in working practices necessary to make them a reality. Both systems rely on the up-front identification of research and seek to eliminate (or at least reduce) the current duplication of effort across review teams. Implementing systems that will achieve these gains in productivity requires both technical interoperability and organizational willingness to invest in systems which facilitate data sharing. At an organizational (and inter-organizational) level there needs to be investment in the up-front organization of research. If the systems are to be sustainable, and not require more human effort than can reasonably be made available, tasks which are amenable to automation will need to be automated. If automation efficiencies are to be realized, then specific training data sets will also need to be created, requiring both a commitment to this model of research curation, and investment in human annotation. The term 'evidence ecosystem' is quickly gaining currency, as it aims to encapsulate this new dynamic of a global system of linked data which has been organized to facilitate its reuse.

The most obvious gains in efficiency appear to come from improving the global discovery and curation of research for use in systematic reviews and decision-making. However, case study 9.2 also points to a world where the synthesis of research itself may also change in the future. AI systems, which digest vast quantities of information and make recommendations for practice, are becoming commonplace and systems which digest data and suggest treatments, diagnoses and prognoses (e.g. IBM Watson Oncology (https://www.ibm.com/watson/health/oncology-and-genomics/oncology)) are being deployed in practice settings. Despite their possible pivotal and game-changing role, these systems are rarely evaluated in traditional RCTs, and one important issue for future research is to develop methodologies and regulatory expectations for this new generation of decision aids.

9.6 Conclusion

This chapter has examined moves towards the automation of systematic reviews and the developments underway to support living systematic reviews and guidelines. These processes can be seen as strategic shifts towards evidence surveillance systems which change traditional systematic review methods, moving some elements of study identification outside the processes of any individual review. Few automation systems operate entirely autonomously, and most require human intervention in order to achieve acceptable levels of performance. Realizing a new dynamic 'ecosystem' of research evidence will require organizational investment, a willingness to share data, and a more strategic and use-orientated way of understanding evidence curation.

9.7 Suggestions for further reading

Akl EA, Meerpohl JJ, Elliott J, et al. (2017) Living systematic reviews: 4. Living guideline recommendations. *Journal of Clinical Epidemiology* 91.

Elliott JH, Synnot A, Turner T, et al. (2017) Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*.

Jonnalagadda SR, Goyal P and Huffman MD (2015) Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, Systematic Reviews 4(1): 78.

Michie S, Thomas J, Johnston M, et al. (2017) The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science*, Implementation Science 12(1): 121.

O'Mara-Eves A, Thomas J, McNaught J, et al. (2015) Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 4(1): 5.

Simmonds M, Salanti G, McKenzie J, et al. (2017) Living systematic reviews: 3. Statistical methods for updating meta-analyses. *Journal of Clinical Epidemiology*.

Thomas J, Noel-Storr A, Marshall I, et al. (2017) Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology* 91: 31–37.

Tsafnat G, Dunn A, Glasziou P, et al. (2013) The automation of systematic reviews. *Bmj* 346(jan10 1): f139–f139.

Tsafnat G, Glasziou P, Choong MK, et al. (2014) Systematic review automation technologies. *Systematic reviews* 3(1): 74.

Wallace BC, Trikalinos TA, Lau J, et al. (2010) Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 11(1): 55.

9.8 References

[1] Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLoS Med* 2010; 7: e1000326.

[2] Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S MD. Article Annals of Internal Medicine How Quickly Do Systematic Reviews Go Out of Date ? A Survival Analysis. *Ann Intern Med* 2007; 147: 224–33.

[3] Macleod MR, Michie S, Roberts I, et al. Biomedical research: Increasing value, reducing waste. *Lancet* 2014; 383: 101–104.

[4] Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction—the why, what, when, and how. *J Clin Epidemiol*. Epub ahead of print 2017. DOI: 10.1016/j.jclinepi.2017.08.010.

[5] Elliott JH, Turner T, Clavisi O, et al. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Med* 2014; 11: e1001603.

[6] Hausner E, Guddat C, Hermanns T, et al. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *J Clin Epidemiol* 2016; 77: 118–124.

[7] Stansfield C, O'Mara-Eves A, Thomas J. Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. *Res Synth Methods* 2017; 1–11.

[8] Mergel GD, Silveira MS, da Silva TS. A method to support search string building in systematic literature reviews through visual text mining. In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing - SAC '15*, pp. 1594–1601.

[9] Hausner E, Guddat C, Hermanns T, et al. Development of search strategies for systematic reviews: Validation showed the noninferiority of the objective approach. *J Clin Epidemiol* 2015; 68: 191–199.

[10] Octaviano FR, Felizardo KR, Maldonado JC, et al. Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? *Empir Softw Eng* 2015; 20: 1898–1917.

[11] Jiang Y, Lin C, Meng W, et al. Rule-based deduplication of article records from bibliographic databases. *Database* 2014; 2014: 1–7.

[12] Rathbone J, Carter M, Hoffmann T, et al. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. *Syst Rev* 2015; 4: 6.

[13] O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015; 4: 5.

[14] Cohen a. M, Smalheiser NR, McDonagh MS, et al. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *J Am Med Informatics Assoc* 2015; 1–11.

[15] Marshall I, Storr AN, Kuiper J, et al. Machine Learning for Identifying Randomized Controlled Trials: an evaluation and practitioner's guide. *Res Synth Methods* 2018; 1–13.

[16] Choong MK, Galgani F, Dunn AG, et al. Automatic evidence retrieval for systematic reviews. *J Med Internet Res* 2014; 16: e223.

[17] Kiritchenko S, de Bruijn B, Carini S, et al. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak* 2010; 10: 56.

[18] Marshall I, Kuiper J, Wallace BC. RobotReviewer : evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Informatics Assoc* 2015; 1–10.

[19] Wallace BC, Noel-Storr A, Marshall IJ, et al. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J Am Med Informatics Assoc* 2017; 0: 1–4.

[20] Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol* 2017; 91: 31–37.

[21] Mavergames C, Oliver S, Becker L. Systematic reviews as an interface to the web of (trial) data: Using PICO as an ontology for knowledge synthesis in evidence-based healthcare research. *CEUR Workshop Proc* 2013; 994: 22–26.

[22] Singh G, Marshall IJ, Thomas J, et al. A Neural Candidate-Selector Architecture for Automatic Structured Clinical Text Annotation. In: *CIKM, Singapore*. 2017.

[23] Michie S, Thomas J, Johnston M, et al. The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implement Sci* 2017; 12: 121.

[24] Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. *Syst Rev* 2014; 3: 109.