Timely, Efficient, and Living Systematic Reviews: Opportunities in International Development

Birte Snilstveit[1], Jennifer Stevenson[1], Ian Shemilt[2], Mike Clarke[3], Emmanuel Jimenez[1], James Thomas[2]

March 2018

Centre of Excellence for Development Impact and Learning

International Initiative for Impact Evaluation

UKaid from the British people

[1]International Initiative for Impact Evaluation (3ie)
[2]EPPI-Centre, UCL Department of Social Science
[3]Queen's University, Belfast

Suggested Citation: Snilstveit B, Stevenson J, Shemilt I, Clarke M, Jimenez E, Thomas J, 2018 Timely, Efficient and Living Systematic Reviews: Opportunities in International Development CEDIL Inception Paper 7: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by UKAID through DFID. The mission of the centre is to develop and promote new impact evaluation methods in international development.

Corresponding Author: Ms Birte Snilstveit, email: bsnilstveit@3ieimpact.org

Table of Contents

Abstract

Systematic reviews provide reliable summaries of available evidence on the effects of programmes to improve people's lives. We know that timely availability of evidence is a key factor influencing evidence use, but the time-lag between new study results becoming available and their integration into new or updated systematic reviews is typically measured in years, due to lags in study publication, coupled with a resource and time-intensive review process. Moreover, in a sector where new evidence is being produced on a frequent basis, this time-lag means reviews can rapidly become out of date.

In the health sector, systematic review methodologists, knowledge managers, information scientists and computer scientists have developed technology-enabled approaches to help produce systematic reviews more efficiently and within shorter time frames. One key goal is to develop the methods and infrastructure for "Living Systematic Reviews" – systematic reviews that are updated as soon as new evidence emerge. At the core of this idea is a "continuous evidence surveillance" approach, underpinned by improved production efficiency, which enables reviews to be rapidly updated as new eligible studies become available, helping their conclusions to remain current and meet identified end-user needs (Mavergames and Elliott 2016).

Drawing on the developments in other sectors, this paper explores the opportunities for more efficient and timely review production in international development, including, but not limited to "living systematic reviews". In doing so it is addressing three objectives: (1) Reviewing the "state-of-the-art" approaches to producing more efficient and timely systematic reviews, including, but not limited to "living systematic reviews"; (2) Assessing applicability and feasibility of applying these approaches to syntheses in international development; and (3) Identifying a "roadmap" to rapid (and living) systematic reviews in international development.

Section 1

Introduction

Systematic reviews provide reliable summaries of evidence on the effects of programmes to improve people's lives and are said to provide 'the most reliable and comprehensive statement about what works' (Petrosino et al., 2001). In the recent decade several hundred systematic reviews addressing questions of relevance to international development have been commissioned, including by the Department for International Development (DFID) in the UK, the International Initiative for Impact Evaluation (3ie) and the United States Agency for International Development (USAID) (White and Waddington, 2012).

However, if systematic reviews are to be useful and used they need to answer questions of relevance for policy-makers and be conducted within a timeframe that is acceptable to decision-makers (Whitty, 2015). But the time-lag between new study results becoming available and their integration into new or updated systematic reviews is typically measured in years, due to lags in study publication, coupled with the resource and time intensive review process.

The value proposition of systematic reviews has primarily been focused on the rigor and transparency of the methodology. To date, there has been relatively less focus on producing systematic reviews in the sector within significantly shorter time frames. Thus decision makers and commissioners are increasingly turning to other, more rapid approaches, such as rapid reviews.

In other fields, primarily biomedicine and clinical health, interdisciplinary collaboration between systematic review methodologists, knowledge managers, information scientists and computer scientists have made significant progress in developing approaches which help produce reviews that are both rapid and systematic (e.g.: Marshall et al., 2015; Millard et al., 2016; Schűnemann & Moja, 2015; Shemilt et al., 2016; Thomas et al., 2017; Tsafnat et al., 2014; Tsertsvadze et al., 2015). The uptake and development of such approaches has to date been limited in the international development sector.

This paper explores opportunities for more efficient and timely review production in international development, including, but not limited to "living systematic reviews". The next section outlines the specific objectives and methods of the paper in brief, followed by a section discussing the main reasons for the time and labour intensive nature of systematic reviews in the sector. Section four discusses different approaches to reducing the time required for producing reviews, while section five reviews recent developments from other fields to produce rapid systematic reviews. In section six we discuss the applicability and feasibility of applying some of the "state-of-the-art" rapid systematic review approaches to international development, drawing on worked examples. This is followed by conclusions and a road map for further research and development initiatives to realise the vision of living systematic reviews, and an outline of recommendations of how CEDIL and DFID may contribute to such efforts.

Section 2

Objectives and Methods

The overall aim of this paper is to provide an overview of the potential opportunities for producing systematic reviews in international development more quickly and efficiently, including the idea of "living systematic reviews". We also identify research and development needs for applying new methodologies to achieve rapid and living systematic reviews at scale in the sector. The aim of doing so is also to consider how the resource requirements of systematic reviews can be reduced over time. In doing so we address three main objectives:

(1) To review the "state-of-the-art" approaches to producing more efficient and timely systematic reviews, including, but not limited to 'living systematic reviews';

(2) To assess applicability and feasibility of applying these approaches to syntheses in international development; and

(3) To provide an outline of a "roadmap" to rapid (and living) systematic reviews in international development.

To address these objectives we use a combination of literature review, worked examples, feasibility testing and modelling approaches. We used targeted searching and snowballing to identify methodological literature on rapid systematic reviews, and living systematic reviews and innovative methods being used to produce systematic reviews in fields other than international development. We aim to identify three types of studies for the methods review: systematic reviews of innovative systematic review methods, evaluations of the effectiveness of such methods, and papers presenting new ideas and innovations at early stages of development. Finally, to assess their applicability to international development, we test several of these methods in ongoing 3ie systematic reviews and evidence gap map projects. The specific methods for these tests are discussed within the relevant sections.

Challenges in the Current Model of Systematic Review Production

The challenges presented in this section draw on the extensive experience of the authors in producing and managing the production of systematic reviews in the international development sector. It also draws upon the literature from other sectors where systematic reviews are increasingly used including environmental management, crime and justice and education.

3.1 Timeliness

A systematic review of barriers and facilitators of evidence use by policy makers found that one of the most frequently reported barriers to evidence use was the lack of timely availability of research outputs (Oliver et al., 2014). Decision-makers are often faced with time-sensitive policy questions, so if systematic reviews are to be useful and used they need to answer relevant questions and be conducted within a timeframe that is acceptable to decision-makers (Oliver et al., 2014; Whitty, 2015).

However, the labour intensive nature of systematic reviews and the linear fashion in which they are typically conducted means they normally take 18-24 months to complete, sometimes longer. This relatively long production time means that, at least among some commissioners, they are seen as less relevant for informing decision-making. By the time the results are available decisions may already have been made, or commissioners have moved on to other roles.

The long time frame for producing systematic reviews also presents challenges in ensuring they reflect the most recent evidence at the time of publication. Over the last decade there has been a rapid growth in the rate of publication of new studies in international development, with several hundred papers published every year (Cameron et al., 2016). The search for studies is typically conducted at an early stage of the systematic review production process, meaning that by the time a review is published it may already be 12 months or more since the search was concluded. The high publication rate of primary studies and time lag from search to review publication, increases the likelihood that reviews do not include the most up to date evidence as new studies have become available since the search was completed.

3.2 Inefficiencies and labour intensive nature of tasks

Conducting systematic reviews is a labour-intensive process. Table 1 (appendix 1) summarise the main tasks and level of effort for each task for a medium-sized review. It highlights that conducting a systematic review involves sifting through a large number of citations and full texts, extracting significant amounts of data, calculating effect sizes and critically appraising the methods of included studies. A highly trained screener may be able to screen through 750 citations a day, or screen around 100 full text papers. Experienced coders may be able to handle a couple of studies per day, but for complex reviews with associated qualitative studies, one study per day is typically more realistic. Once data has been extracted researchers need to calculate effect sizes, and while some of this can be done using templates applying commonly used formula, inconsistent reporting of data means a significant portion of effect sizes need to be calculated manually. All of these tasks are typically conducted by two reviewers in duplicate. So considerable time and resources are invested before reviewers are able to start with the analysis.

The table also highlights that the tasks associated with search, screening, data extraction, critical appraisal and effect size calculation are the most labour intensive tasks of a systematic review. This is probably similar for other disciplines, but features of the international development literature means the workload is even more extensive. For example, the search and screening tasks is a key step in the production of a systematic review (Higgins & Green, 2011; Kugley et al., 2017; Waddington et al., 2012), but it also consumes significant resource.

This task has been made easier in health because abstracts have become better structured and studies are more appropriately indexed using controlled vocabulary, for example the Medical Subject Headings (MeSH) terms in MEDLINE (Coletti & Bleich, 2001). But, in the international development literature, as in many other social sciences, the use of structured titles and abstracts is less widespread, databases are often poorly indexed, without corresponding controlled vocabulary and studies are scattered across a range of databases, many of which do not support complex search strings or use controlled vocabulary.

The tendency in international development to publish in the grey literature only adds to the labour intensive nature of the search process. For example, research outputs by Non-Government Organisations (NGOs) and international organisations may only be published on their websites and databases only. To identify such reports researchers have to manually search a large number of websites and organisational databases.

Moreover, it is widespread practice in development economics to publish working papers in well-read series such as the World Bank Policy Research Paper and National Bureau of Economics Research (NBER). These working papers are not typically indexed on academic databases, and are often only available through university websites and smaller databases of grey literature, which often have more limited search functionalities.

These features of the literature result in reviewers retrieving a large number of irrelevant search hits where the title and abstracts need to be manually screened to identify a small number of studies. To illustrate this point, a brief review of all systematic reviews published by the International Development Coordinating Group (IDCG) of the Campbell Collaboration revealed that reviewers screened the title and abstract of between 756 and 109,017 search hits, with an average of 19,810 per review (Campbell Collaboration Library, 2017). This is despite an average number of included studies of 48. On average therefore, reviewers need to screen 1095 irrelevant titles and abstracts to identify a relevant paper.

Separate groups of reviewers exhaust extensive resources undertaking the same search and screening tasks, for example, identifying whether a study can be classified as an impact evaluation or whether it takes place in a low-or middle-income country. Similarly, in systematic reviews with overlapping inclusion criteria, different review teams may spend significant time collecting the same data from the same studies, from basic descriptive data on the intervention, outcome and study design, to appraisal of risk of bias, extraction of effect size data and calculation of standardised effects.

There is no coordination or collaboration of efforts around these most labour-intensive tasks of evidence synthesis (e.g.: search, screening and data extraction). Review organisations like the Campbell Collaboration and the Cochrane Collaboration coordinate the *editorial process* and *publication* of systematic reviews, partially as a way of ensuring quality, but also to avoid duplication. But despite the significant resources going to labour intensive systematic review tasks such as search, screening and data extraction, there is no collaboration to share data and avoid duplication of such tasks. As a result there is significant duplication of labour among review teams. In a context of limited human and financial resources this represents a missed opportunity for significant time and efficiency gains.

Finally, the lack of an integrated platform which covers all stages of the review process creates inefficiencies. Conducting a systematic review involves performing a number of micro-tasks, accumulating large quantities of different types of data, with often complex structures. Methodological developments and the diverse range of types of evidence in a review to address complexity, including qualitative and cost data alongside quantitative outcome data (Petticrew et al., 2013), make the data management task even more complex.

There is a range of software and tools which reviewers can use to support different stages of the systematic review process. For example, Refworks and Endnote are tailored for reference management and can be used to support the search and screening process, Microsoft Excel and Access can be used for data extraction, and R and STATA can be used for meta-analysis. However, none of these tools were developed specifically for systematic reviews and therefore have some limitations when used for this purpose. Moreover, just as Kohl et al (2018) find a lack of a tool which is suitable for all the review stages in the context of the environmental sector, there is no integrated platform specifically tailored to the complexities of the international development literature covering all stages of review production and management[1]. Reviewers therefore typically rely on several different platforms to produce and manage a review, making review management more labour intensive and increasing the potential for error and loss of data.

[1] Kohl et al.'s (2018) review of the functionality of 22 existing software for the production of systematic reviews in the context of the environmental sector demonstrated the lack of a product that sufficiently covers all review stages.

Section 4

Rapid Reviews versus Rapid Systematic Reviews

In response to the challenges outlined above there is a rapidly expanding literature on approaches to address the timeliness and efficiency of evidence synthesis. It is beyond the scope of this paper to provide an overview and critical appraisal of this literature. However, in this section we summarise the characteristics of the main two broad categories of responses: efforts to develop rapid review approaches by applying *methodological modifications*, and efforts to expedite systematic reviews by modifying the way in which the systematic review methodology is implemented, as outlined in more detail below.

4.1 Rapid Reviews: Timely Evidence Synthesis through Methodological Shortcuts

To address time sensitive policy questions researchers have developed a range of rapid review methodologies to answer time sensitive policy questions, including rapid reviews or rapid evidence assessments (REAs) (Hartling et al., 2015; Ganann et al., 2010). There is no single definition of a rapid review and recent reviews of study methods have highlighted the variation in rapid review methods (Featherstone et al., 2015; Hartling et al., 2015; Khangura et al., 2012; Tricco et al., 2015). However, such approaches typically involve adjusting methods used in traditional systematic reviews and adopt one or more *shortcuts* to give more timely answers to urgent questions (Schünemann & Moja, 2015).

Common shortcuts include addressing more narrowly focused questions; limiting the number of sources consulted in the search; limiting the time frame and language of the search; using only one reviewer for screening, full text review, appraisal and data extraction; and presenting results as a narrative summary rather than conducting a formal statistical synthesis (Ganann et al., 2010; Featherstone et al., 2015; Tricco et al., 2015). The reported production time for rapid reviews vary significantly, ranging from 5 minutes to 12 months, with most reviews typically taking between 1 and 6 months (Hartling et al., 2015; Featherstone et al., 2015; Tricco et al., 2015). One study found that the rapid reviews which took longer were also those with fewer shortcuts (Hartling et al., 2015).

The lack of standards around the methodology for REAs and other rapid synthesis products means that the quality of rapid reviews varies (Featherstone et al., 2015). While adjustments to traditional systematic review methodology can reduce production time, it may come at the expense of methodological quality, bias and lower confidence in review findings. Methodological shortcuts such as limiting the literature search, single screening and simplified critical appraisal can introduce bias in the findings of reviews, for example by missing papers published in the grey literature and introducing human error (Gannan et al., 2010). There may also be limitations in the types of questions rapid reviews can address, more uncertainty attached to findings of rapid reviews, and less detailed information and recommendations regarding implementation (Hartling et al., 2015).[2]

4.2 Rapid Systematic Reviews: Expediting the Systematic Review Process without methodological modifications

Another broad approach to producing evidence synthesis in a more timely and efficient manner is to focus on producing systematic reviews, but doing them differently. This approach is based on the rationale that while systematic reviews are time, labour and cost intensive (Tsertsvadze et al., 2015), there is nothing inherent with systematic reviews which makes it necessarily so (Schünemann & Moja, 2015). Thus, systematic review methodologists are working to develop methods to help expedite systematic reviews (Marshall et al., 2015; Millard et al., 2016; Schűnemann & Moja, 2015; Shemilt et al., 2016; Tsafnat et al., 2014; Tsertsvadze et al., 2015).

Such efforts combine improvements to the review processes (streamlining), with the use of innovative technology (automation or semi-automation of resource intensive tasks) and optimal use of human resources (increasing the size and expertise of review teams) (Tsertsvadze et al., 2015). For example, tasks such as study identification, data extraction and critical appraisal can be completed by different people simultaneously, rather than sequentially (Tsertsvadze et al., 2015). This approach is aided by technology – by using text mining that orders studies by the likelihood of inclusion in the reviews, reviewers can identify the most relevant citations first. While some reviewers continue screening through the studies that are less likely to be included, others obtain full text papers, screen these and proceed with data extraction and quality appraisal.

The goal of these efforts is to produce more efficient and timely reviews without sacrificing rigour (Schűnemann & Moja, 2015). Rapid systematic reviews may involve the same amount of work, but conducted within a shorter time frame (ibid). A key difference between rapid reviews and rapid systematic reviews is therefore that the latter does not imply modification of the systematic review methodology, but rather a modification in the way in which the methodology is implemented (Tsertsvadze et al., 2015). Few studies have provided head to head comparisons of rapid reviews and systematic reviews. It is therefore not clear to what extent rapid review methods lead to different findings from those following traditional systematic review methods (Tricco et al., 2015). Future research should assess the effect of shortcuts on the reliability and usefulness of review findings. However the reminder of this paper will focus on the potential for modifying the way in which systematic review methodology is implemented.

Section 5

Increasing the Efficiency of Review Production and Moving Towards a Constantly Updated Evidence Base

To address the data deluge problem that systematic reviews are increasingly experiencing, some organisations are examining the potential of evolving the review workflow to make it more efficient. Much recent discussion has centred on the production of “living” systematic reviews – summaries of evidence which are continually updated as soon as new evidence becomes available (Elliot et al., 2014; Elliot et al., 2017; Thomas et al., 2017).

Living systematic reviews moves the update of existing systematic reviews away from one-off, labour intensive and static sets of tasks, towards dynamic and ongoing review processes to produce up-to-date evidence summaries that are available online (Elliot et al., 2014). They require the continual monitoring of the publication of new studies, immediate coding and appraisal once new studies are identified and rapid incorporation into the existing review.

Living systematic reviews are considered most useful in fields where evidence is rapidly emerging and thus where standard systematic review reports would quickly become out of date (Elliot et al., 2017). They are therefore not necessary for every policy topic, leaving a space for standard updates of systematic reviews. To date, they have only been trialled in the health sector, primarily as part of Cochrane’s Project Transform (Cochrane Collaboration, 2017).

5.1. Producing living systematic reviews and improving timeliness and efficiency of standard SR production

The development and use of new processes and technologies enable the production of both living and ‘static’ systematic reviews, as they facilitate tasks to be done more rapidly and efficiently, with greater automation (Thomas et al., 2017). The following discussion can therefore be applied to the production of both living systematic reviews as well as a means to improve the efficiency and timeliness of the production of ‘static’ systematic reviews. Broadly speaking, such efforts can be separated into four domains: people, processes, technology and infrastructure. We describe these four domains below and present some examples of how advances in these areas have been applied in other sectors beyond international development.

5.1.1. People

In terms of personnel, systematic reviews have often been conducted by teams of people with different expertise. For example, information specialists have long been recognised as essential, given the complexity of constructing and conducting a systematic search. Once studies have been identified, however, it is not unusual for many of the latter review tasks to be undertaken by a small team, often by two or three individuals (Elliott et al., 2014).

An alternative way of conducting a review is to break review activities down into "micro-tasks" (see table 1, Appendix) to which many people with a wide range of skills and experience contribute. A larger authorship team that includes appropriate substantive, statistical and methodological experts, information specialists and sufficient dedicated research assistance can achieve a better division of labour and allow labour-intensive tasks to be completed in parallel (Tsertsvadze et al., 2015). By having multiple tasks done in parallel, the review can be completed sooner, and often with less effort, by individuals with greater specialisation in each task.

For example, sifting through thousands of citations to find relevant research is extremely time-consuming. This task, traditionally carried out by a small team of reviewers, might be accomplished just as accurately by a much larger team, even if they individually possess less specialist skills. This approach is often known as "crowdsourcing", and it has been used to good effect in many areas of science (Lee et al., 2017; Ranard et al., 2013).

Crowdsourcing in systematic reviews is in an early stage of implementation, but it is already showing promise. For example, the "Cochrane Crowd" platform hosts a range of micro-tasks which enable people of varying degrees of experience to contribute to individual reviews or to maintaining databases (Thomas et al., 2017). Every month Cochrane searches a range of databases to identify new randomised controlled trials (RCTs). The Cochrane Crowd has now screened over 2 million citations and identified those which are and those which are not RCTs. This process promises to make the task of RCT identification for Cochrane reviews much more efficient and also reduces duplication of effort. An algorithm, combining the responses from multiple volunteers and with different level of expertise is working behind the scenes to ensure that the accuracy of crowd assessments is very high, exceeding 99% in recall and accuracy (Thomas et al., 2015). Evaluations testing the efficacy of crowdsourcing of tasks in systematic reviews suggest a large potential for workload reduction, without any effect on accuracy (Mortensen et al., 2017; Nama et al., 2017; Noel-Storr et al., 2014). However, to date crowdsourcing of systematic review tasks has been limited to the health sector and there is a need to test the feasibility and accuracy of this approach in sectors like international development.

5.1.2. Better process management

Better process management – by changing how the systematic review tasks are organised – has the potential to improve both the timeliness and efficiency of systematic review production. Key components of better process management are process parallelisation, streamlined workflows and active project management. Compared to the other domains discussed in this section, the opportunities discussed here can be taken up immediately by review teams and are largely cost neutral.

Process parallelisation is one approach promoted to improve timeliness of systematic reviews (Tsertsvadze et al., 2015). The tasks involved in producing a systematic review are conducted in different stages, from protocol development, searching, screening, critical appraisal, data extraction, synthesis and dissemination. Each stage includes a number of discrete micro-tasks. Performing the tasks associated with most stages relies on some progress having been made at earlier stages, but there is still significant potential for process parallelisation to improve the timeliness of review production (Tsertsvadze et al., 2015).

For example, as soon as a share of records have been screened, potentially includable studies can be moved to full text screening. As soon as a decision has been made to include a study the data extraction and critical appraisal tasks can start. Depending on the size and complexity of the review it might also be sensible to split these tasks between reviewers with different levels of expertise. If a review includes quasi-experimental studies using complex statistical techniques, the data extraction may be split so that a relatively junior reviewer extracts descriptive information, while a more senior reviewer focuses on extracting statistical information and risk of bias assessment.

The increasing availability of bespoke systematic review platforms such as EPPI-Reviewer, Covidence, Rayyan and DistillerSR can help facilitate improved process management. The use of such tools helps support process parallelisation and streamlined workflows for example by allowing easy allocation of tasks to larger groups of people, including tasks at different stages of the review, and allows automatic comparison of decisions made by different reviewers. These platforms can also reduce errors which easily occur when data are moved between different software and digital platforms. For example, from text files into bibliographic software for screening, then into word processing or spreadsheet software for data extraction, and then into a statistical analysis package for meta-analysis.

However, most existing tools are developed primarily for systematic reviews of clinical and public health literature, and do not for example easily allow for the complex data structures and multiple and linked publications commonly found in the international development literature. In addition, only Rayyan is open access and freely available, providing a barrier to access, in particular for review teams working with limited budgets. Therefore, while bespoke systematic review platforms facilitate improved process management, reviewers conducting international development systematic reviews still typically rely on several different platforms, tools and software to produce and manage a review. Most reviews published by the International Development

Coordinating Group (IDCG), for example, cite the use of a combination of tools, such as Excel, R and EPPI-Reviewer.

Finally, a systematic review, as any other study, is a research *project*. As such, underpinning any efforts to improved process management (and addressing timeliness and efficiency more generally) should be a strong and integrated approach to project management. Freely available project management tools, like Asana and Teamwork, can provide an efficient digital environment for managing large teams, based at multiple locations and handling a large number of very diverse tasks in a transparent and efficient way.

5.1.3. Technologies

In recent years there has been an expansion of innovation and collaborative research projects do develop tools that use technology to aid the production of systematic reviews, including by automating and semi-automating different review tasks. This is highlighted by the Systematic Review Toolbox, a freely available catalogue of tools to support systematic reviews (Marshall and Brereton, 2015), which currently lists 143 different software tools[3] available to support different stages of the review process.

Tsafnat and colleagues (2013) present a vision of the future where each systematic reviews can be created in real-time at the push of a button, consistent with the vision of 'Living Systematic Reviews' (Elliot et al., 2014). These 'living' reviews take the form of a computer programme which automatically performs all review tasks, from retrieving new trials, to critical appraisal, data extraction, meta-analysis and report production. Their survey of informatics systems to support automation of systematic review tasks identified tools tailored for all stages of review production, from protocol development to report writing (Tsafnat et al., 2014).

Despite extensive innovation to automate systematic review tasks, there is still limited take up of such innovations by review authors (Ouzzani et al., 2016; Thomas 2013; Tsafnat et al., 2013). There are likely a number of reasons explaining this low take up, including the level of technical skill required to use many tools, the lack of independent reliability and functionality testing (Ouzzani et al., 2016; Thomas 2013), the fragmented nature of much of the activity in this field (Elliot et al., 2013), with most tools developed as stand-alone software supporting specific tasks (Ouzzani et al., 2016).

Significant research and development is still needed to realise the vision of a fully automated review (Thomas et al., 2017; Tsafnat et al., 2014). Nevertheless, tools using automation and semi-automation for conducting labour intensive review tasks, especially at the early stages of the review, are ready to use (Thomas et al., 2017) and have the potential for significant workload savings. It is beyond the scope of this paper to provide an exhaustive review of all existing tools and tasks with opportunities for technology enabled workload reduction and automation. But table 1 provides an

[3] As of June 2018. SR Toolbox, available at: http://systematicreviewtools.com/

example of tasks where researchers have developed and tested different approaches, using technology to reduce workload. We discuss some of the more promising opportunities, based on current developments, below.

Because of the scale of manual effort required, considerable research and development activity has focused on assessing the potential to improve screening efficiency through the use of emerging automation technologies (O'Mara-Eves et al., 2015). Often, the use of technology at this stage of a systematic review will involve the machine "learning" how to apply the inclusion / exclusion criteria of the review, based on human interaction. A process known as "active learning" is usually followed, whereby human screening efforts are fed back to the machine at regular intervals, resulting in a reordering of the ordering of citations according to likelihood of inclusion, based on learning from the human generated screening decisions.

Table 1: Examples of technology enabled tasks

Review stage	Task (author)
Search	Snowballing (Choong et al 2014) Developing search terms (Hausner et al 2016; Stansfield et al., 2017) De-duplication (Rathbone et al., 2015)
Screening/ study classification	Study classification (Bekhuis and Demner-Fushman 2012; Marshall et al., 2018; Wallace et al., 2010), Study classification and crowd-sourcing (Wallace et al., 2017) Inclusion/ exclusion prediction applied to title/ abstract screening (Gates et al., 2018; Hempel et al., 2012; Howard et al., 2016; Li et al., 2016) Crowd-sourcing screening tasks (Mortensen et al., 2017; Nama et al., 2017; Noel-Storr et al 2014)
Data extraction	Text classification (Bui et al., 2016) Automated data extraction (Hsu et al., 2011; Kiritchenko et al., 2010)
Critical appraisal	Automated risk of bias assessment (Marshall et al., 2015; Millard et al., 2016)
Report production	Automated write up based on extracted data (Torres and Adams 2017)

A systematic review of evidence from simulation studies assessing the use of text mining in study identification suggest potential for significant workload savings (O'Mara-Eves et al., 2015). Their findings suggest a reduction in screening workload of between 30 and 70 per cent could be possible, although this may be accompanied by a loss of 5 per cent of relevant studies. They conclude the use of text mining to eliminate studies should be considered promising, but require more testing, specifically in international development which is a more complex area with limited application to date. Since the review was published, the use of text mining for study identification has become increasingly widespread and it is now offered as a feature on EPPI-reviewer. There is a need for further testing to identify if, and at what threshold, it can be used to automatically exclude studies. Nevertheless, being able to use it to prioritise studies for screening enable reviewers to identify studies for inclusion earlier in the screening process, providing an opportunity for process parallelisation and potentially more rapid review completion.

Duplicate removal is another area where semi-automation is already being used to reduce the need for manual labour. Because many databases with varying amounts of overlap are included in a comprehensive search strategy, it is inevitable that the final search results include duplicate records. Identifying duplicate bibliographic records manually is challenging and time consuming, but significant advances have been made to semi-automate this process. A number of reference management tools and systematic review platforms now include this as a standard feature, including EndNote and EPPI-reviewer, and it is relatively widely used by systematic review authors. There are a number of duplicate removal tools, and it is not clear which is most effective. A recent evaluation comparing the sensitivity and specificity of EndNote and the Systematic Review Assistant-Deduplication Module found the latter to be superior on both criteria, identifying over 40 per cent more duplicate records (Rathbone et al., 2015). Thus, a review and comparison of deduplication applications would be useful in identifying those with the highest workload saving.

Retrieval of full text papers for assessment is another example of an area that holds great promise for reducing work load through the use of technology. Once records have been identified as of potential interest, the full texts need to be retrieved. This was once a lengthy process involving reviewers visiting a range of libraries and obtaining photocopies of papers, but the widespread use of DOI data in bibliographic records and the availability of services such as CrossRef and OpenCitation is making the identification of full text reports much more efficient (Thomas et al., 2017). As not all papers are easily discoverable and there is a need to navigate different access options, there will continue to be a role for people in this process, but technology is already producing workload reductions.

Finally, in addition to using machine learning based on screening data within a single review, more advanced automation using machine learning to build ‘review independent’ classifiers is also starting to emerge (Marshall et al., 2018; Thomas et al., 2017; Wallace et al., 2017). For example, the data produced by Cochrane Crowd has been used to develop a ‘RCT classifier’ which when applied to a new set of citations to identify, excluded 60-80 per cent of irrelevant records with a sensitivity of more than 99 per cent (Wallace et al., 2017). These types of classifiers rely on machine learning to classify citations according to the probability of it being an RCT, using high-quality screening data produced by humans (Thomas et al., 2017). Thus, a key criteria for creating such a model is availability of training data.

Data extraction and critical appraisal is another area with significant efforts, and some promise for semi-automation to reduce workloads in the short to medium term. However, a recent systematic review assessing automation and semi-automation of data extraction (Jonnalagadda et al., 2015) concludes there is a need for significant further development in this area. They identified 26 studies from the health literature that had attempted to automate extraction of at least one type of data, for example country, sample size or intervention details. But just over half of the possible types of data that could be extracted have been subject to attempts at automation. Most tools are focused on a limited number of data types, with the most comprehensive attempt covering seven out of the 52 possible types they looked at. The results of these attempts to automate data extraction nevertheless appear promising, with most data being extracted with a mean sensitivity and positive predictive value of over 70 per cent.

Perhaps the most extensive effort to develop a tool for data extraction is RobotReviewer[4] – a tool developed to assess risk of bias in clinical trials and extract statements supporting the assessment. RobotReviewer is based on machine learning and was developed using data from over the risk of bias assessment in 5,400 systematic reviews, and from the pdfs of over 12,000 of the clinical trials included in these reviews (Marshall et al., 2015). The initial evaluation found the tool and associated algorithm to be less accurate than the human risk of bias assessment, but by automatically extracting data that is then reviewed by authors the tool can still reduce author workload (Marshall et al., 2015). A similar tool to RobotReviewer could be developed for international development, but this would require a standardised risk of bias tool and training data in the form of annotated pdfs associated with the data structure of such a standardised tool.

Study identification and data extraction are the most labour intensive tasks within a systematic review. Investing in developing and adapting tools aiming to reduce time and workload requirements of the tasks provides the greatest opportunity for producing more efficient and timely reviews. However, as demonstrated by table 1 opportunities to use technology to achieve time and workload reductions are being developed and tested for all stages of review production.

5.1.4. Infrastructure

Finally, the changes to working practices, processes and advances in technology are combining to facilitate the creation of a new infrastructure (sometimes described as an "ecosystem") of connected services which aim to support the dynamic identification and use of research evidence. Here the aim is to reduce duplication and connect data, held in many different locations, so that assessments made for one purpose (e.g. a specific review) can be utilised without the work needing to be carried out again, as was described in section 2. It also includes moving some of the work outside the scope of any given systematic review and into a more "upstream" centralised model, where new research is identified and moved into the system in a more usable format as soon as it is published.

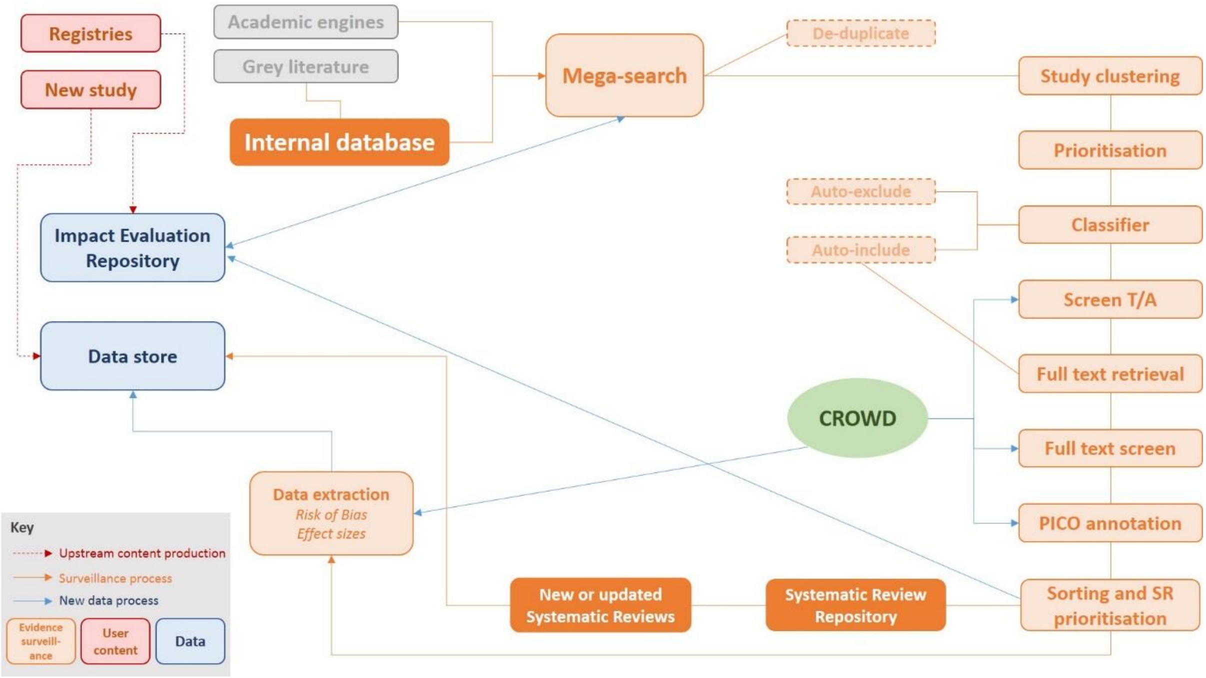
The above changes in personnel, processes and technologies are proceeding rapidly, and the health care field is already developing many components of this connected "ecosystem", which promises a step change in society's ability to utilise research evidence. The Cochrane Collaboration has pioneered this approach through Project Transform, an ongoing effort to develop a range of linked services, including an

[4] www.robotreviewer.net

evidence pipeline (a system of continuous evidence surveillance), Task Exchange, Cochrane Crowd, Linked Data and piloting of living systematic reviews (Cochrane, 2017).

There are opportunities for the field of international development to learn and benefit from many of these advances, but it will require significant investment to build a similar infrastructure. Figure 1 provides an example of how such a system may look like, drawing on relevant existing resources such as the 3ie Impact Evaluation Repository, and incorporating new features like a mega search, evidence pipeline and a data store (Cochrane, 2017; Cohen et al., 2010). It suggests a two pronged approach to developing a research information management system suitable for the 21st century.

Figure 1: Integrated evidence infrastructure



Firstly, 'upstream' efforts to ensure new research is 'synthesis ready' by implementing changes to study reporting, publication formats and indexing. Such efforts would aim to improve discoverability (and avoiding issues like publication and outcome reporting bias) by ensuring all new impact evaluations are pre-registered in a central registry, linked to an open database indexed by study with links or cross-reference to all structured data and publications associated with that study (Goldacre et al., 2016). Data would be made ready for use in systematic reviews by making study data and results available in an open access data store for easy export into software for analysis, making manual data extraction redundant. The usability of data and study information would need to be improved by broad based take-up and enforcement of study reporting guidelines.

Secondly, it suggests the development of a centralised system for dealing with the 'backlog' of studies which are already published but have yet to be identified and/ or curated in a way that can easily feed into systematic reviews. To achieve this objective as efficiently as possible, the system would draw on the best available technology, concepts of citizen science, improved workflows and collaborative ways of working. For example, it would include an evidence pipeline (Cohen et al., 2010) with features such as a topic based mega search, with screening and study identification aided by technological applications such as de-duplication, machine learning for citation prioritisation, an impact evaluation classifier for auto-exclusion of citations below a certain threshold and technology enabled data extraction. New studies and data identified through this process would feed into the existing 3ie Impact Evaluation Repository and an open access data store. To reduce any further duplication of data extraction we also suggest a coordinated effort to get researchers involved in completed and ongoing systematic reviews to deposit their data in the data store (eg: Ip et al., 2016; Li et al., 2015; Wolfenden et al., 2016).

To facilitate the linking and sharing of data in this way would require investment in research and development. But two largely cultural shifts would also need to be in place: first, agreed protocols and structures for sharing data; and second, a culture and incentive system that supports and rewards the sharing, rather than hiding, of research data (Goldacre et al., 2016; Thomas et al., 2017; Wolfenden et al., 2016). In relation to the first point, semantic web technologies and protocols for "linked data" are well advanced and ideal for the sharing of data for systematic reviews. Also necessary are agreed vocabularies and sets of term lists which enable reviewers to classify research in the same way. The health field has seen many advances in this area, identifying both problems and opportunities that others might learn from (Mavergames et al., 2013).

The changing of academic cultures to support and reward the sharing of data may proceed hand-in-hand, as the increased efficiency and benefits that flow from data reuse become apparent. In some fields, such as astronomy, the sharing of data is routine, and increased expectations for transparency from regulators and funders like the Bill and Melinda Gates Foundation, the Department for International Development (DFID) and 3ie means that data from new studies is increasingly becoming open access.

Section 6

Applicability and Feasibility of Timely, Efficient and Living Systematic Reviews in International Development

Opportunities to improve review timeliness and efficiency through changes in production models have largely been explored in systematic reviews of clinical and public health interventions (Miwa et al., 2014). There are a number of differences between international development reviews, and the literature they draw on, which may limit applicability and feasibility of implementing such tools in international development. We undertook two small studies to test feasibility of using machine learning approaches for study identification, as well as an evaluation of an adaptation of the screening process to start exploring if an how changes to review production may be applied to international development. In this section we present the results of these exercises.

6.1 Testing the Use of Text Mining Technologies for Study Inclusion: 3ie Systematic Review Database

The first study used machine learning to develop a classifier to identify systematic reviews of interventions in international development. To do this, we made use of existing screening data from the search and screening process for the 3ie Systematic Review Repository (SRR) (Snilstveit et al., 2012). The procedures for identifying studies is similar to that of identifying studies for a systematic review, and new searches are undertaken on an annual basis to keep the SRR content current. The process is labour intensive and developing more efficient ways of doing so would free up resources for other activities, such as better curation.

To test the applicability of using machine learning to reduce the screening workload we developed a dataset of 1240 study records, each with a decision on whether or not it met the inclusion and exclusion criteria for the Systematic Review Repository (SRR). We used this existing screening data to develop a classifier using the machine-learning function in EPPI-Reviewer 4 to determine a probability score for study inclusion in, or exclusion. The classifier assigned each study in the dataset into a ten per cent probability score interval, using text mining to classify records based on words included in the title and abstract.

The classifier was retrospectively applied to the whole dataset and we simulated decisions on inclusion, exclusion or to be manually screened by a reviewer, using defined cut-off points. This allowed us to identify potential for workload reduction through semi-automation of screening decisions. We evaluate the performance of the classifier according to the precision and recall of automated inclusion/ exclusion decisions.[5]

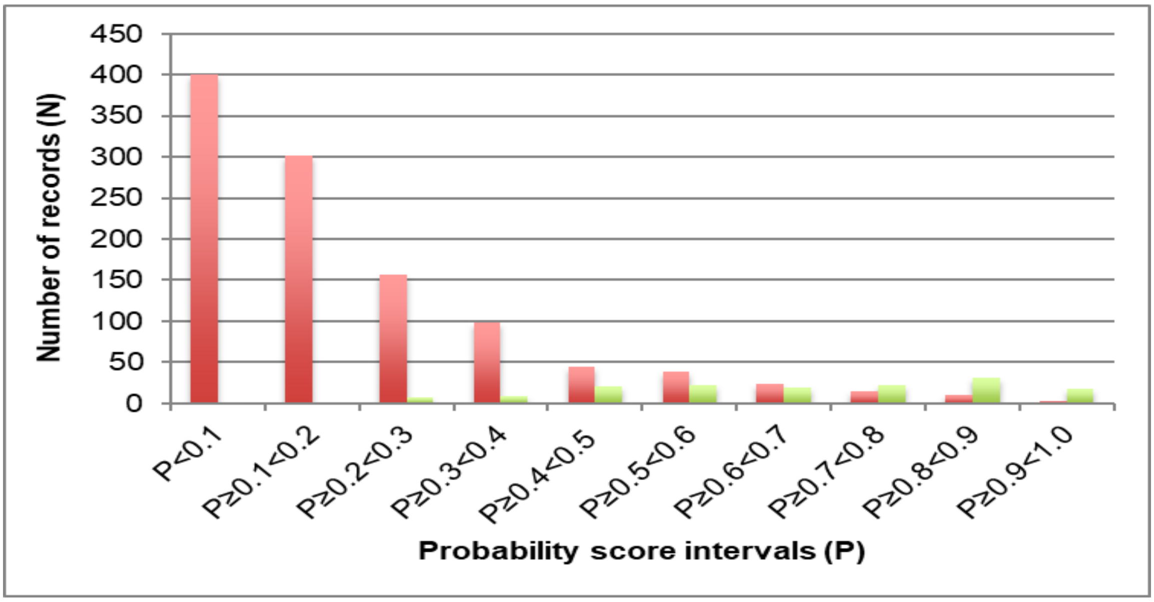
The results of the application of the classifier are presented in Figure 1. The green bars represent included studies and the red bars represent excluded studies according to the original screening decisions. The intervals show how many included and excluded records were sorted into each probability interval by the machine-learning classifier.

Based on the rates of recall and precision, we decided on a cut-off point of auto-exclusion when a study had less than 20 per cent probability of inclusion, and auto-inclusion for full-text screening when a study had more than 90 per cent probability of inclusion. Any study with greater than 20 per cent and less than 90 per cent probability of inclusion was manually screened. Using these functions, the classifier automatically excluded 702 out of 1240 studies and automatically included 18 studies, with just one false negative and two false positives. This is a recall of 99.32 per cent and precision of

[5] The proportion of correctly identified included studies among all included by the classifier, and the proportion of all real includes correctly identified by the classifier.

98.65 per cent. Based on this simulation using semi-automated decisions on inclusion and exclusion, we see a reduction in screening workload of 58.3 per cent. The goal of the Systematic Review Repository (SRR) is to index all systematic reviews in international development. While the results suggest a chance of missing a few studies by applying the classifier, we made the decision that the risk was so small and the workload reductions so substantial that we applied the model to our last search update. Doing so resulted in over 80 per cent reduction in the number of records we had to screen manually, so improving on what we saw in the simulation applied to a smaller data set.

Figure 1: Results from the 3ie Systematic Review Database Classifier



Source: Authors data from unpublished study.

6.2 Testing the Use of Text Mining Technologies for Study Inclusion: Systematic Review of Education

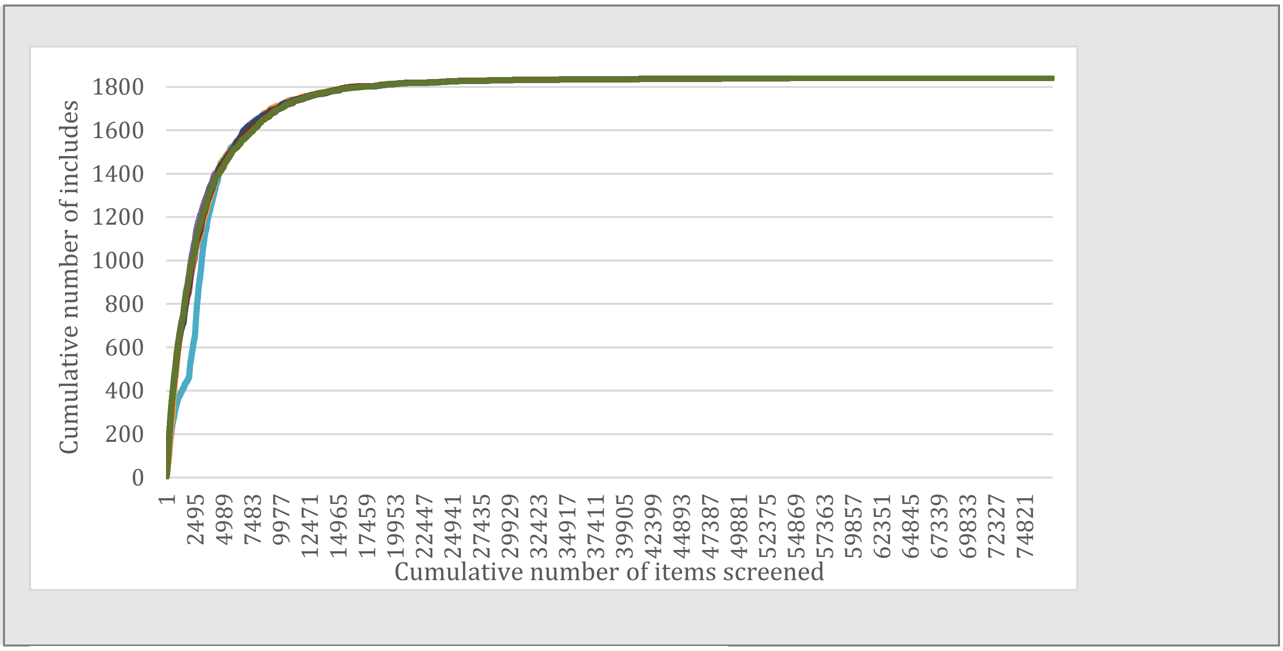
To examine the potential of using an "active learning" process to identify impact evaluations in international development, we conducted a retrospective simulation study using data from a completed systematic review on access and quality of education in L&MICs (Snilstveit et al., 2016). The study was undertaken using EPPI Reviewer 4 for reference management and study identification. The review is a good candidate for using machine learning because it provides a large data set of nearly 80,000 citations needed to be screened manually, identifying 1841 potentially relevant records for inclusion at the title/ abstract stage.

We began the process by selecting five relevant, and five irrelevant studies at random from the lists of included and excluded studies respectively, and used these to build the machine learning model. After ranking the citations in order of likely relevance, we then selected the top 25 studies and added them to the set we selected at random. We then rebuilt the machine learning model and cycled through steps 2-4 until all citations had been screened. Since the performance of the process might vary, depending on which citations were sampled randomly at the outset, we performed the simulation process 10 times, with different randomly selected citations each time.

The results of the active learning simulations are illustrated in Figure 2. This shows screening progression through the citations along the horizontal axis x, and the cumulative number of relevant citations identified along the vertical axis y. If we had screened the citations at random, we would expect to see a diagonal line on the graph, from the bottom left to the top right. It is clear, however, that the machine learning has changed the order in which relevant studies are found, with the majority of the 1841 relevant citations being identified much earlier in the process than they would otherwise have been. Indeed, all 1841 citations were "identified" after approximately 50 per cent of the citations were screened.

If the reviewers had been able to take advantage of this process in their original review, they could have avoided screening nearly 40,000 citations, without any loss to recall. As Figure 2 implies, even larger workload savings might be possible, were reviewers happy to retrieve only 98 per cent or 95 per of relevant studies.

Figure 2: Results of Active Learning Simulation



Source: Screening data from Snilstveit et al., (2015).

6.3 Worked Example: Reducing screening workload in a Payment for Environmental Services

Several of the authors of this paper are involved in an ongoing systematic review of the effects of Payment for Environmental Services where a secondary objective was to implement changes to the review production and produce a draft review within a shorter time frame. The review combines a number of innovative approaches discussed in this paper, including a highly qualified large team, streamlined workflows and semi-automation.

The team is composed of several co-principal investigators who collectively have extensive experience of undertaking mixed-methods reviews in international development; a statistician, a senior substantive expert on payment for environmental services, an expert on the integration of cost data into systematic reviews, and an experienced group of research assistants.

We have used this review as an opportunity to test and compare the reliability of various screening methods within one review. At the title and abstract stage, we tested and compared the timing and results of two different screening approaches incorporating machine learning in EPPI reviewer 4:

1. Full independent double screening undertaken manually but using the priority screening function to order results by probability of inclusion, based on a training set of screening (15 per cent of the search results). The priority screening function orders the results by the likelihood of inclusion based on the words in the title and abstract of the included and excluded papers from the training set, using machine learning text mining technology;

2. Single screening using the priority screening function with a "safety first" approach (an option to mark unclear studies for review by a second screener) (Shemilt et al., 2016);

The first approach using independent double screening is considered best practice according to systematic review guidelines, and represents the ‘business as usual’ comparison scenario. The search returned 4769 papers for screening at title and abstract. We set up a duplicate review project in EPPI-reviewer and the two approaches were applied simultaneously to the same set of study records by independent reviewers with a similar level of experience and training.

Comparison of the first and second approaches shows very similar numbers in terms of number of studies included for full-text screening. Twenty studies were put through for screening by a second screener during the second approach, creating a minimal amount of double screening work.

More importantly, the single screener with a "safety first" approach resulted in a 97.5 per cent recall rate. That is, the single screener approach correctly included 97.5 per cent of the 40 papers eventually included in the review from this screening stage. The approach missed only one paper, which was a paper associated with another included paper. The missed paper would have been identified through targeted searches at a later stage of the search process. In sum, using a single screener with "safety first", in combination with a priority screening approach based on machine learning resulted in approximately 50 per cent reduction in title and abstract screening workload as compared to independent double screening.

Section 7

Potential Challenges in Applying (semi-) automation technologies to SRs in International Development

The examples presented in the previous section suggest that the use of text mining and active learning has great potential for reducing workload in study identification – one of the most labour intensive tasks in a systematic review. The technology is available to use, and we encourage greater uptake and evaluation within new evidence review and synthesis projects. In this section we discuss a number of sector specific challenges that need to be addressed to fully exploit the potential for technological innovation, largely relying on some form of text mining to (semi-) automate review tasks, across all stages of and fully realise the potential of rapid systematic reviews.

Machine learning technologies rely to some extent on similarity of terms across the textual data they rely on as well as on a representative set of training data from which to learn from (O'Mara-Eves et al., 2015). In the social sciences, there is greater heterogeneity in the structure of papers than in clinical health. Articles are frequently 20 pages or more, with long appendices or supplementary materials in separate documents. Titles and abstracts, which typically provide the input for machine learning, do not usually follow a standard format and often do not include clear descriptions of key study concepts such as intervention, outcomes, population, geographical location and study design. This contrasts with papers in the clinical health literature which are typically shorter and more structured. In addition, even when key concepts are clearly described, there is generally greater heterogeneity and complexity of terminology. This heterogeneity may result in reduced accuracy of machine learning tools in identifying relevant studies, for example resulting from hasty generalisations from biased training sets (O'Mara-Eves et al., ibid).

Moreover, in development economics and in political sciences, where a large share of impact evaluations in international development originate, there is a tradition of publishing results in one or more working papers before a journal version is published. Different results from the same study may be published in different journal articles. Some of these may be duplicates or near duplicate papers, but there are often discrepancies between documents, with for example some details, outcomes or samples not reported in the journal paper. The result is that systematic reviews need to rely on several papers reporting on a single study, with reviewers making decisions about which data to use. Procedures for handling different reports from the same study will be outlined in the study protocol, but an element of detective work and considerable time identifying differences is often required. Until the issue of duplicate papers and reporting of results across different publications has been addressed, study identification will likely remain a semi-automated process, with machine-human interactions in identifying linked and duplicate reports.

The issues that can present challenges in using machine learning for study identification also apply to efforts to use such technology for data extraction and risk of bias assessment. Limitations in study reporting is a major issue in reviewing the literature. Even manually extracting data and critically appraising studies present challenges because study reports are longer and more unstructured, and key information is often missing. For example, key statistical information such as sample size and standard deviations are often not reported clearly. Intervention, outcome, participant eligibility criteria and methods are often vague and with details added throughout the paper, rather than in a separate section. In addition, impact evaluations in the sector typically provide a greater number of outcome measures, sub-samples, follow up periods and analysis specifications, meaning that reviewers have to make decisions on which data to extract and use. While how to handle such decisions are typically pre-specified in the study protocol, an element of reviewer judgement and discussion is often required.

Tools that rely on machine learning techniques to facilitate semi-automation of labour intensive tasks, rely on training data to perform these tasks. To date most of the technological development and testing of automated and semi-automated approaches have been limited to biomedical and clinical health. While algorithms may be relatively easily transferred to other domains, there is a need for sector specific training data. RobotReviewer for example was developed based on over 12,000 pdfs with risk of bias appraisal of health focused RCTs (Marshall et al., 2015). It was possible to create this large data set because of standardised use of the Cochrane risk of bias tool across systematic reviews. While there is commonality across reviews, data extraction and risk of bias is not yet standardised in the same way. It is probably feasible to develop a data set which could be used for machine learning purposes, but it would require some effort in merging data sets from different review groups.

Section 8

Discussion and Conclusions

Just as in other sectors of public policy, efforts to promote the use of evidence to inform decision making in international development is facing an information management problem. The current model of systematic review production relies largely on labour and time intensive manual processes conducted by disparate review groups. It results in significant duplication of efforts, and the resources available are not sufficient to deliver high quality systematic reviews at the scale and time frame needed to meet demand.

In other sectors, principally health, systematic reviewers, information specialists and computer scientists are investing significant efforts in developing methods that are both rapid and systematic. This is based on the idea that there is nothing inherent in the systematic review approach preventing it from being completed more rapidly and efficiently. These efforts can be broadly separated into the four domains of people, processes, technology and infrastructure.

As described above, changes to the first two domains, such as larger study teams with appropriate balance of junior reviewers and specialists, adopting better process management, can be implemented in a largely cost neutral manner. When combined with readily available technology such as semi-automated duplicate removal and priority screening based on machine learning it is feasible to conduct a medium-sized systematic review in 6 months, instead of the typical time frame of 12-24 months. While some training may be required, we suggest a medium-term goal should be for this to become the new norm.

However, as highlighted above, innovative technology has the potential to significantly reduce the time and effort required to produce systematic reviews. A large number of tools and applications are available, but these require further development and adaptation to the international development literature. This should be combined with research and independent evaluation to assess effectiveness compared to more traditional approaches. To realise the vision of living systematic reviews would require investment in a more centralised, integrated evidence infrastructure as a public good.

For this to become a reality in a way that is sustainable there should also be supported by upstream efforts to ensure studies are published and reported in a format that is “synthesis ready”. This will include promoting the uptake of trial registries (which also link to all papers published from any single study) and promoting and enforcing the use of reporting guidelines such as CONSORT (or adapted version thereof). Future automation will also be aided by standardisation of data schema and terminology used to describe key study features such as interventions, study designs and outcomes (e.g. drawing on efforts such as COMET, 2011).

Below we provide a “roadmap” for how the evidence movement in international development may approach the production, reporting, organisation and publication of systematic reviews and impact evaluation in a way that facilitates more rapid and efficient systematic reviews in the short-to-medium term, and the longer term vision of living systematic reviews.

Rapid and efficient systematic reviews in short-to-medium term

What can researchers do now to produce more rapid systematic reviews?

There are number of existing tools and steps which can be implemented immediately, but which are currently not widely adopted by teams conducting international development systematic reviews.

Firstly, search and screening efforts can be significantly reduced by using a platform which includes a machine learning feature and effective duplicate removal. Systematic review platforms like EPPI reviewer includes machine learning functionality, and most reference management software includes a duplicate removal functionality.

Secondly, when putting together their study team principal investigators and funders should deploy human resources in the most efficient way. This means having sufficient substantive, statistical, methods and information specialist skills, as well as sufficient research assistant capacity to handle the most labour intensive, entry level tasks. Doing so will also allow study teams to adopt process parallelisation, where tasks are completed in parallel by different people, rather than consecutively, reducing the overall time period required to complete the review.

Finally, when adopting these steps, we suggest researchers report workload savings and consider implementing simulation studies within existing projects to evaluate the effects of technology enabled changes to workflow and tasks to facilitate evidence informed decisions about potential adaptations to current review practice.

Achieving the longer term vision of living systematic reviews

Investment in an integrated evidence infrastructure

To facilitate greater workload reductions, reduce inefficiencies from duplication and allow for rapid and living systematic reviews, we suggest review groups and research funders invest in developing research synthesis infrastructure, in line with figure 1 above:

Develop a system of "continuous evidence surveillance" to facilitate rapid and continuously updated syntheses of evidence. Such a pipeline could be modelled on the Cochrane Evidence Pipeline and combine the use of technology (text mining), process improvement (regular instead of one-off searches) and people (crowd sourcing) in an ongoing effort to provide a continuously updated collection of impact evaluations of international development interventions.

A starting point for such a pipeline already exists in the 3ie Impact Evaluation Repository (IER), which aims to provide a repository of all existing impact evaluations in international development (Cameron et al., 2016).

Develop a data store for sharing of data extracted from systematic reviews to avoid duplicate data extraction from different review teams. Such a platform would complement efforts to publish impact evaluation data sets (for example through the 3ie dataverse).

Invest in research to evaluate existing tools. There is no systematic review of the availability and efficacy existing tools, and no published evaluations of the effectiveness of using such tools in international development.

Develop new/ adapt existing technology tailored to the idiosyncrasies of international development literature. Doing so will require investment to tailor technology to a new literature and developing training data for machine learning. There is also a need to develop new sector specific tools, such as a fully integrated user interface for review production.

Explore the use of crowd sourcing of different review tasks. Doing so would involve developing training and quality assurance procedures, recruiting volunteers and developing a platform for completing the tasks. In the first instance we suggest a pilot to assess the feasibility of this approach in the sector, including the ease of recruiting volunteers and the types of tasks that may be crowdsourced.

Standardise risk of bias assessment and data extraction. Developing such standardised tools is an essential criteria for applying (semi-) automation tools to data extraction, and for facilitating data sharing.

Develop a standardised intervention taxonomy and core outcome sets to facilitate discoverability and data sharing. International development covers a very broad range of interventions and outcomes, and getting conceptual agreement among the broad range of stakeholders involved in each sector may be challenging. Nevertheless, such agreement would aid both the discoverability, use and interpretation of new research and we recommend this to be a medium term goal. We suggest drawing on lessons from the COMET initiative (2011) and other similar exercises to conduct an initial pilot exercise to identify the best approach to standardise terminology in this way.

Upstream efforts to produce and report primary studies that are synthesis ready

Ensure new studies are 'synthesis ready'. This means addressing deficiencies in how studies are currently conducted, reported and published. Doing so necessitates a change in culture, development of new guidelines and better enforcement of existing guidelines (eg: Nosek et al., 2015; International Initiative for Impact Evaluation [3ie], 2018).

Encourage and enforce study pre-registration. Uptake of study registration is currently relatively low, and funders and journal editors in particular can help address this through making pre-registration a criteria for publication and funding.

Section 9

Recommendations for DFID and CEDIL

The main implications of the road map above is the need for investment in research and development to realise the vision of living systematic reviews. So our main recommendation is that DFID and CEDIL consider supporting the development of an integrated evidence infrastructure as outlined above, either in part or its entirety.

In addition, both DFID and CEDIL as funders and key stakeholders in research production in the sector can contribute by encouraging compliance with reporting guidelines and research transparency standards, and enforcing this by making compliance a criteria for funding. In addition, there is also a role to play in developing frameworks for standardised data schema, including intervention and outcome typologies.

Finally, when funding systematic reviews and other mapping products we recommend that DFID and CEDIL consider incorporating the production of worked examples and independent evaluation of promising new technology.

References

3ie, 2018. 3ie Research Transparency Policy. New Delhi: International Initiative for Impact Evaluation.

Bekhuis, T. and Demner-Fushman, D. 2012. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif Intell Med.* 55(3), p.197-207.

Bui, D.D.A., Del Fiol, G. and Jonnalagadda, S. 2016. PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics*, 61, p.141-148

Cameron, D. B., Mishra, A. and Brown, A. N., 2016. The growth of impact evaluation for international development: how much have we learned? *Journal of Development Effectiveness*, 8 (1), p. 1-22.

Campbell Collaboration, 2017. *The Campbell Collaboration Online Library,* Viewed 10 October 2017, < https://campbellcollaboration.org/library.html>

Choong, M.K., Galgani, F., Dunn, A.G. and Tsafnat, G. 2014. Automatic evidence retrieval for systematic reviews. *J. Med. Internet Res.;* Oct 1:16 (10).

Cochrane Collaboration, 2017, September 5. *Project Transform – Making high-quality evidence synthesis quicker and easier*, viewed 5 January 2018, <https://community.cochrane.org/news/project-transform-making-high-quality-evidence-synthesis-quicker-and-easier>

Cohen, A.M., Adams, C.E., Davis, J.M., Yu, C., Yu, P.S., Meng, W., Duggan, L., McDonagh, M. and Smalheiser, N.R., 2010. November. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In *Proceedings of the 1st ACM international Health Informatics Symposium*, p. 376-380.

COMET, 2011. *COMET Initiative*, viewed 7 January 2018 <http://www.comet-initiative.org/>

Coletti, M. H., and Bleich, H. L., 2001. Medical Subject Headings Used to Search the Biomedical Literature. Journal of the American Medical Informatics Association . *JAMIA*, 8(4), p. 317–323.

Crosas, M. 2011. The Dataverse Network: An Open-source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine, viewed 8 January 2018 <http://dlib.org/dlib/january11/crosas/01crosas.html>

Elliot, J.H., Turner, T., Clavisi, O., Thomas, J., Higgins, J.P.T., Mavergames, C. and Gruen, R.L. 2014. Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. *PLOS Medicine*, 11(2).

Elliott, J.H., Synnot, A., Turner, T., Simmonds, M., Akl, E.A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J. and Tovey, D., 2017. Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of clinical epidemiology*, 91, pp.23-30.

Featherstone, R. M., Dryden, D. M., Foisy, M., Guise, J-M., Mitchell, M. D., Paynter, R. A., Robinson, K. A., Umscheid, C. A., and Hartling, L., 2015. Advancing knowledge of rapid reviews: an analysis of results, conclusions and recommendations from published review articles examining rapid reviews. *Systematic Reviews*, 4(50), p. 1-9.

Ganann, R., Ciliska, D. and Thomas, H. 2010. Expediting systematic reviews: Methods and implications of rapid reviews. *Implementation Science,* 5(56).

Gates, A., Johnson, C. and Hartling, L. 2018. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews 7(45)*

Goldacre, B., and Gray, J., 2016. OpenTrials: towards a collaborative open database of all available information on all clinical trials. *Trials*, 17(164).

Hagen-Zanker, J. and Mallett, R., 2013. How to do a rigorous, evidence-focused literature review in international development A Guidance Note. ODI working paper. London: ODI.

Hartling, L., Guise, J.M., Kato, E., Anderson, J., Aronson, N., Belinson, S., Berliner, E., Dryden, D., Featherstone, R., Foisy, M., Mitchell, M., Motu'apuaka, M., Noorani,, H., Paynter, E., Robinson, K. A., Schoelles, K., Umscheid, C. A. and Whitlock, E., 2015. EPC methods: An exploration of methods and context for the production of rapid reviews. Rockville (MD): Agency for Healthcare Research and Quality (US).

Hausner, E., Guddat, C., Hermanns, T., Lampert, U. and Waffenschmidt, S. 2016. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of Clinical Epidemiology.* Volume 77, p.118-124.

Hempel, S., Shetty, K.D., Shekelle, P.G., et al. 2012. Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations. *Rockville (MD): Agency for Healthcare Research and Quality.* Sep 2012

Higgins, J. and Green, S., 2011. Cochrane Handbook for Systematic Reviews of Interventions (version 5.0.2, updated September 2009). The Cochrane Collaboration.

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R. and Thayer, K. 2016. SWIFT-Review: a text-mining workbench for systematic review. *Systematic Reviews*, *5*(87).

Ip, S., Hadar, N., Keefe, S., Parkin, C., Iovin, R., Balk, E. M., and Lau, J. 2012. A web-based archive of systematic review data. *Systematic Reviews*, 1(1), 15.

Jonnalagadda, S.R., Goyal, P. and Huffman, M. D. 2015. Automating data extraction in systematic reviews: A systematic review. *Systematic Reviews*, 1(16).

Khangura, S., Konnyu, K., Cushman, R., Grimshaw, J. & Moher, D. 2012. Evidence summaries: The evolution of a rapid review approach. *Systematic Reviews,* 1(10).

King, G., 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods and Research*, 36, p. 173–199.

Kohl, C., McIntosh, E.J., Unger, S., Haddaway, N.R., Kecke, S., Schiemann, J. and Wilhelm, R. 2018. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environmental Evidence*, 7(8), https://doi.org/10.1186/s13750-018-0115-5

Kugley S, Wade A, Thomas J, Mahood Q, Jørgensen AMK, Hammerstrøm K, and Sathe N., 2007. Searching for studies: a guide to information retrieval for Campbell systematic reviews. Oslo: The Campbell Collaboration.

Lee, Y.J., Arida, J.A. and Donovan, H.S., 2017. The application of crowdsourcing approaches to cancer research: a systematic review. *Cancer medicine*, 6(11), p. 2595-2605.

Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M. H., ND Liu, H. 2016. A Text-Mining Framework for Supporting Systematic Reviews. *American Journal of Information Management*, *1*(1), p.1–9.

Marshall, C., and Brereton, P., 2013. Tools to support systematic literature reviews in software engineering: A mapping study. In 2013 *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement,* pp. 296-299. IEEE.

Marshall, I., Kuiper, J. and Wallace, B.C., 2015. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials*. Journal of the American Medical Informatics Association,* 1(10).

Marshall, I.J., Kuiper, J. and Wallace, B.C. 2016. RobotReviewer: Evaluation of a System for Automatically Assessing Bias in Clinical Trials. *Journal of the American Medical Informatics Association (JAMIA)*, 23 (1), p.193-201.

Marshall, I.J., Noel-Storr, A.H., Kuiper, J., Thomas, J., and Wallace, B.C. 2018. Machine Learning for Identifying Randomized Controlled Trials: an evaluation and practitioner's guide. *Research Synthesis Methods*, [In press].

Mavergames, C., Oliver, S. and Becker, L., 2013. 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 Proceedings 994*, p. 22–26.

Mavergames C., and Elliott, J., 2016. Living Systematic Reviews: towards real-time evidence for health-care decision-making [Blog Post]. BMJ Clinical Evidence Blog, 10 May 2016. Retrieved: 10 July 2018. Retrieved from: https://blogs.bmj.com/ce/2016/05/12/living-systematic-reviews-towards-real-time-evidence-for-health-care-decision-making/

Miwa, M., Thomas, J., O'Mara-Eves, A. and Ananiadou, S., 2014. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51, p. 242-253.

Millard, L.A.C., Flach, P.A. and Higgins, J.P.T., 2016. Machine learning to assist risk-of-bias assessments in systematic reviews. *International Journal of Epidemiology,* p. 266-277.

Mortensen, M.L., Adam, G.P., Trikalinos, T.A., Kraska, T. and Wallace, B.C., 2017. An exploration of crowdsourcing citation screening for systematic reviews. *Research Synthesis Methods*, 1(21).

Nama, N., Iliriani, K., Xia, M.Y., Chen, B.P., Zhou, L.L., Pojsupap, S., Kappel, C., O'Hearn, K., Sampson, M., Menon, K. and McNally, J.D., 2017. A pilot validation study of crowdsourcing systematic reviews: update of a searchable database of pediatric clinical trials of high-dose vitamin D. *Translational Pediatrics*, 6(1), p.18-26.

Nixon, H., Cambers, V., Hadley, S. and Hart, T. 2015. *Urban Finance: Rapid Evidence Assessment*. London: Overseas Development Institute.

Noel-Storr, A., Struthers, C., Cullum, S., McShane, R., Creavin, S., Davis, D. and Huckvale, K. 2014. Many hands make light work –or do they? Results of two pilot studies looking at crowdsourcing and The Cochrane Collaboration. *Cochrane Methods*. Suppl 1. pp. 34-35.

Nosek, B. A. et al. 2015. Promoting an open research culture. *Science*, 348(6242), p. 1422-1425.

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

Oliver, K., Innvar, S., Lorenc, T., Woodman, J. and Thomas, J., 2014. A systematic review of barriers to and facilitators of the use of evidence by policymakers. *BMC health services research*, 14(1), p.2.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5(1), 210.

Petrosino, A., Boruch, R.F., Soydan, H., Duggan, L. and Sanchez-Meca, J., 2001. Meeting the challenges of evidence-based policy: The Campbell Collaboration. *The Annals of the American academy of political and social science*, 578(1), p.14-34.

Petticrew, M. and Roberts, H., 2006. *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell Publishing.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P. T., Mayhew, A., Pantoja, T., Shemilt, I. and Sowden, A., 2013. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology,* 66 (11), p. 1230-1243.

Ranard, B. L., Ha, Y. P., Meisel, Z.F., Asch, D. A. Hill, S. S., Becker, L.B., Seymour, A. K. and Merchant, R. M. 2013. Crowdsourcing – harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, p.187-203.

Rathbone, J., Carter, M., Hoffmann, T., et al. 2015. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. *Systematic Reviews*, 4(1), p. 6.

Schünemann, H. J. and Moja, L. 2015. Reviews: Rapid! Rapid! Rapid! ...and systematic. *Systematic Reviews*, 4(4).

Shemilt, I., Khan, A., Park, S. and Thomas, J. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Systematic Reviews, 5(140), p. 1 – 13.

Snilstveit, B, 2012. Systematic reviews: from 'bare bones' reviews to policy relevance. *Journal of Development Effectiveness,* 4(3), p. 388-408.

Snilstveit, B., Eyers, J., Bhavsar A., Gallagher, E., and Stevenson, J. (2014). *3ie database of systematic reviews in international development: search strategy and procedures*. London: International Initiative for Impact Evaluation (3ie).

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T. and Jimenez, E., 2016. *The impact of education programmes on learning and school participation in low- and middle-income countries: a systematic review summary report,* 3ie Systematic Review Summary 7. London: International Initiative for Impact Evaluation (3ie).

Snilstveit, B., Bhatia, R., Rankin, K. and Leach, B., 2017. *3ie evidence gap maps: a starting point for strategic evidence production and use,* 3ie Working Paper 28. New Delhi: International Initiative for Impact Evaluation (3ie).

Stansfield, C., O'Mara-Eves, A. and Thomas, J. 2017. Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. *Res Syn Meth*. 8, p.355–365.

Stevens, C., Irfan, I., Massa, I. and Kennan, J. 2015. *The Impact of Free Trade Agreements between Developed and Developing Countries on Economic Development in Developing Countries: A Rapid Evidence Assessment.* London: Overseas Development Institute.

Synnot, A. Turner, T. and Elliott, J. 2017*. Cochrane living systematic reviews: Interim guidance for pilots* (Version 0.3, 24 April 2017). The Cochrane Collaboration.

Tanner, C., Scalise, E. and Mutema, M., 2015. *Public Overseas Investments: Ensuring Respect for and Protecting Legitimate Land Tenure Rights. A Rapid Evidence Assessment*. London: DAI.

Thomas, J. 2013. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? *OA Evidence-Based Medicine*, Oct 21, 1(2):2.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T. and Elliott, J., 2017. Living systematic reviews: 2. Combining human and machine effort. *Journal of clinical epidemiology,* 91, p.31-37.

Torres, M.T. and Adams, C.E. 2017, RevManHAL: towards automatic text generation in systematic reviews. *Systematic reviews,* 6(27)

Tricco, A. C., Antony, J., Zarin, W., Strifler, L., Ghassemi, M., Ivory, J., Perrier, L., Hutton, B., Moher, D. & Straus, S. E., 2015. A scoping review of rapid review methods. *BMC Med*, 13(1), p. 224.

Tsafnat, G., Dunn, A., Glasziou, P. and Coiera, E. 2013. The automation of systematic reviews. *BMJ*, p. 346.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F. and Coiera, E. 2014. Systematic review automation technologies. *Systematic Reviews*, 3(1), p. 74.

Tsertsvadze, A., Chen, Y.-F., Moher, D., Sutcliffe, P., and McCarthy, N., 2015. How to conduct systematic reviews more expeditiously. *Systematic Reviews*, 4(1), p. 160.

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., Bhavsar A., Eyers, J., Koehlmoos, T. P., Petticrew, M., Valentine, J. C. and Tugwell P., 2012. How to do a good systematic review of effects in international development: a tool kit, *Journal of Development Effectiveness*, 4(3), p. 359-387.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley and C., Schmid, C. H. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics,* 11(55).

Wallace, B. C., Small, K. Brodley, C. E., and Trikalinos, T.A., 2011. Who should label what? Instance allocation in multiple expert active learning. *Proceedings of the 2011 SIAM International Conference on Data Mining*, p. 176 – 187.

Wallace, B.C., Noel-Storr, A., Marshall, I.J., Cohen, A.M., Smalheiser, N.R. and Thomas, J., 2017. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, 24(6), p.1165-1168.

White, H., and Waddington, H., 2012. Why do we care about evidence synthesis? An introduction to the special issue on systematic reviews. *Journal of Development Effectiveness*, 4(3), p.351–358.

Whitty, C. J., 2015. What makes an academic paper useful for health policy? *BMC Medicine*, 13, p. 301.

Wolfenden, L., Grimshaw, J., Williams, C.M. and Yoong, S.L., 2016. Time to consider sharing data extracted from trials included in systematic reviews. *Systematic reviews,* 5(1), p.185.

Yaffe, J., Montgomery, P., Hopewell, S., and Shepard, L. D., 2012. Empty Reviews: A Description and Consideration of Cochrane Systematic Reviews with No Included Studies. *PLOS ONE,* 7(5), p. e36626.

Appendix 1

Table A1: Task Breakdown of a 'Medium-Sized' Systematic Review

	Task	No days
Study design/ Protocol	Develop SR scope for consultation (PICOs)	2
	Stakeholder mapping and developing stakeholder engagement plan	3
	Advisory group consultation and engagement	1
	Develop search strategy	4
	Develop protocol	10
Search and data extraction	Search grey literature and organisational websites	10
	Citation tracking	6
	Contacting experts	2
	Screening at title and abstract (assuming 15,000 hits, 600 screened a day)	25
	Full text retrieval (assuming 300 papers)	4
	Full text screening (assuming 300 papers)	8
	Targeted search for process, qual and cost data (assuming 30 included studies, 4 per day)	8
	Outcome mapping	10
	Data extraction (intervention, process, implementation, context, cost, effect size), assuming 30 papers, 1 per day	30
	Check data extraction (assuming 3 studies per day)	10
	Appraisal qualitative studies (assuming 30 included, 4 per day x 2)	15
	Data extraction (effect size checking + risk of bias, assuming 2 per day x 2)	30
	Cost calculations	3
	Effect size calculation	4
	Data management	3
Analysis and write up	Create table of characteristics	1
	Create RoB graphic. PRISMA and any other graphics for descriptive section	1
	Write up descriptive section	2
	Meta-analysis	5
	Qualitative synthesis	17
	Cost-analysis	3
	GRADE/ Summary of findings table	3
	Write up technical report	10
	Advisory group consultation (draft report)	1.5
	Team management and oversight	10
	Total days	**241.5**

Table A2: Comparison of “Ideal Type" Systematic Reviews of Biomedical/Clinical Interventions and International Development

	Systematic reviews of clinical/ biomedical interventions	Systematic review of effects of socioeconomic development interventions in low and middle income countries
Scope	Typically focused on one intervention (treatment) and a few primary outcomes	Scope typically broad. May be intervention or outcome driven. If the former, will typically look at broad range of outcomes along the causal chain (intermediate, primary and secondary). If outcome focused broad range or interventions.
Type of research question	What is the effect of treatment *a* on outcome *x*?	What is the effect of intervention a on outcomes x, y and z in low and middle income countries? Does the effect vary by different population sub-groups? Does the effect vary by different contexts? What is the effect of a, b, c and d interventions on outcome x? What are the factors that may moderate intervention effects?
Target studies - population	Often a narrowly defined and heterogeneous population, for example by age or condition.	More often than not the population is broad, for example primary and secondary school children in low and middle income countries (L&MICs) or farmers in L&MICs.
Target studies - intervention	A drug or procedure.	Complex socioeconomic intervention, such as agricultural extension, women's empowerment programmes or behaviour change interventions.
Target studies - comparator	Placebo or defined alternate drug or procedure.	Typically business as usual (although most SRs will also include 'factorial' study designs comparing different interventions. Placebo very rare.
Target studies - outcomes	A select few clearly defined outcomes.	Range of broadly defined outcomes, typically accepting different outcome measures. For example, income may be measured using self-reported income, income index, or house hold expenditure.
Target studies - study design	Randomised controlled trials	Experimental designs (Randomised controlled trials and cluster randomised controlled trials), regression discontinuity designs, natural experiments (exogenous geographical variation in the treatment allocation), controlled before and after studies with non-random treatment allocation and statistical controls for selection bias and confounding (including propensity score matching, covariate matching, difference in difference, fixed-effects regressions, Heckman selection model, instrumental variable regression, single difference regression, other multiple regression), studies with non-random

		treatment allocation and post-intervention measures only with statistical controls for selection bias and confounding, interrupted time series
Sources of studies	A select number of large health databases, trial registries like alltrials, Cochrane Central Register of Controlled Trials	Academic databases: A broad range of general social science databases (some of which are very poorly indexed), in addition to a range of more subject specific databases. Grey literature, in particular working papers often make up a relatively large share of studies included in SRs in the sector. There are a few key WP collections (IDEAS RePeC, Social Science Research Network, various World Bank WP series) but it is also common for research institutions to have their own working paper series and research funders will also publish study reports on their websites. The 3ie impact evaluation repository is the most similar to CENTRAL and currently contains over 4000 studies (RCTs and QEDs) of socioeconomic development interventions in low and middle income countries.
Number of titles/abstracts screened	In the 1000s	Often in 10,000s
Number of full texts screened	?	Several hundreds, for large reviews 1000s
Format of study reports	Typically short, structured journal articles.	Study reports, including their titles and abstracts, are often relatively unstructured. Papers, in particular working papers, but also journal articles, are typically 20 pages and often longer. They often contain long appendices with further details and results. Titles may not be clearly related to the content of the study, and study design is not always explicit from abstract.
Reporting standards	Reporting generally concise and comprehensive, partially because of reporting standards such as CONSORT.	Poor reporting a major issue in reviewing the literature. For example, key statistical data such as standard deviations often not reported and it may be difficult to identify appropriate sample sizes. Participant eligibility criteria not always explicit, methods often reported throughout the paper (in addition to in a separate methods section).
Type of data	Statistical outputs often provided in form of means and standard deviations, proportions or ratios.	Statistical outputs are typically reported in the form of regression coefficients, together with standard errors, standard deviations, p-value or confidence intervals. But some studies will also report means, proportions or ratios. So within a single review likely that review authors will have to use a range of formulae to calculate effect sizes depending on data available, frequently having to contact authors for statistical data necessary.

Data structure	Studies typically focus on a couple of key outcomes.	Studies typically provide a large number of outcome measures. It is not uncommon for studies to provide 10+ estimates for different outcomes, sub-samples and follow up periods. Studies often report estimates from different analysis specifications. These complex data structures makes data extraction a laborious process and data management can be challenging as few existing SR platforms provide the flexibility needed to map out and extract such a large number of estimates.
Data collection	Description of PICOS, outcome data and compliance/ fidelity.	Description of PICOS, outcome data and compliance/ fidelity. Detailed description of intervention design, process, implementation, context and costs. Often in the form of qualitative data.
Critical appraisal	Risk of bias or other similar tool.	Risk of bias, using tool adapted to also address risk of bias for different quasi-experimental designs
Analysis	Meta-analysis, Network Meta-analysis, Meta-regression, presentation of standardised effect sizes in MA not possible	Meta-analysis (random effects), Meta-regression, presentation of standardised effect sizes in MA not possible, narrative/ causal chain synthesis of process, implementation and context. Often detailed descriptive analysis of intervention design features.
Report format	Cochrane reports fairly long and technical. Journal articles may be shorter.	Long technical reports similar to those in health. Often also 'user friendly' summary reports
Summary of findings tables	GRADE relatively common	There are concerns GRADE is not appropriate and work on a GRADE extension for complex interventions is under way. Authors often develop their own summary of finding tables adapting GRADE.
Key challenges not mentioned elsewhere		A large number of duplicates or near duplicates, partially stemming from a tradition in economics to publish working papers (often multiple) before journal version published. There are often discrepancies between different documents, with for example some outcomes and/ or samples dropped from the journal paper. Or just simply 'salami' slicing by reporting results from one trial in different papers. The result is that we often rely on several papers reporting on a single study.

Contact Us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDILProgramme