Crowdsourcing in health evidence synthesis: the distribution of small parts of the problem

Thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

(PhD by Published Works)

Informatics Research Centre, Henley Business School, University of Reading

Anna Noel-Storr, BA (Hons), MA, MSc

January 2022

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Name: Anna Noel-Storr

Date: 10 January 2022

Signature:

Certificate of readiness to be included in library

I grant powers of discretion to the university librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Acknowledgements

As Helen Keller famously stated: “Alone we can do so little; together we can do so much”. It may be a cliché, but it strongly reflects how I feel about this journey. It would not have happened without the incredible support of many colleagues, friends and fellow researchers. I would particularly like to thank James Thomas, Director of the EPPI-Centre's Reviews Facility for the Department of Health, England, for always providing gentle encouragement and heaps of wisdom. I also would especially like to thank Gordon Dooley, Managing Director of Metaxis Ltd. who has been there from the very beginning, fulfilling my technical demands and only occasionally throwing his hands up in horror at last minute changes to technical specifications. Along with your patience Gordon, your ability to turn often fairly vague ideas into a reality has made much of this work possible.

Whilst the work of this thesis has all been published in the last two years, it reflects a journey that began over ten years ago in a tiny office on the fourth floor of the John Radcliffe Hospital in Oxford. My then boss, Rupert McShane, introduced me to the world of scientific research and gave me an incredible amount of freedom to explore ideas and pursue projects. Those were exciting and enlightening times. When Caroline Struthers joined the team to manage a project on patient and carer engagement in dementia research, bringing with her an unswerving vision and determination to improve public involvement in research, this marked another turning point. Together, we recruited carers of people with dementia to help maintain a register of trials in dementia. It was a short project but one with a lasting impact.

Then, at a conference in Madrid, I remember waiting patiently to talk to eHealth expert Kit Huckvale after a presentation he’d given about a mobile application he’d developed to help reviewers screen the search results for their reviews. As he presented, it occurred to me that his app might enable us to run our first crowdsourcing experiment. Our subsequent study, *Trial Blazers* proved a catalyst for me, not only demonstrating the feasibility of crowdsourcing in this way, but also in terms of proving to myself that I could conduct research, as well as overcome shyness and stand up and present the results.

Julian Elliott, Associate Professor in Evidence Synthesis at Monash University, was to become another person instrumental in this work. He, together with James Thomas, invited me to join a four year ‘game changer’ project we called Project Transform. It was during this time that Cochrane Crowd came into being. Julian’s unswerving belief in it (and in me) gave me the confidence to turn a

vision into a reality. The whole Project Transform team, brilliantly managed by Melissa Murano, really demonstrated the power of collaboration and teamwork. That project will stand out as a highlight of my working life, and I extend a heartfelt thank you to the team: Julian and James, Melissa, Tari Turner, Chris Mavergames, Steve McDonald, Sally Green, Emily Steele, and David Tovey.

Indeed, I have been so fortunate to work with so many truly inspiring, intelligent and dedicated people across Cochrane and beyond. From the Cochrane Dementia and Cognitive Improvement Group: Sue Marcus, Jenny McCleery and Terry Quinn, you have always been so supportive of me; and to my fellow information specialist and data science community, I have learnt so much from you all over the last decade, and no doubt will continue to do so. Especial thanks here to Julie Glanville, Carol Lefebvre, Susanna Wisniewski, Iain Marshall, Ian Shemilt, Byron Wallace, Ruth Foxlee, Robin Featherstone and Jenni Burt.

I would like to also thank my family: my husband, my two children and my mum. I remember my children, who were very young at the time, making bookmarks that said: *Sign up to Cochrane Crowd!* We were about the launch the platform and I had been worried no one would join. There have also been many times since the launch of Cochrane Crowd where work-life balance has been lost. I look back now and realise how fortunate I was to have support and help from home. The pandemic added an additional layer of logistic complexity, with home schooling about photosynthesis one minute to trying to fix some of the problems of evidence synthesis the next.

Thank you also to my supervisors for their support and encouragement: Dr Weizi Li, Associate Professor of Informatics and Digital Health, Henley Business School, University of Reading, and Dr Stephen Gulliver, Associate Professor of Pervasive Informatics, Henley Business School, University of Reading.

And last, but certainly not least, I need to thank the wonderful Cochrane Crowd, a global community of volunteers who help classify the research needed to support informed decision-making about healthcare. This community has amazed me from day one. It has proved to me time and again that working together towards common goals can produce amazing results. We live in unsettling times, yet this community shows me on a daily basis some of the best of what we humans have to offer. None of this work would have been possible without this community.

Dedication

For JGB.

Your life-long passion for learning will always stay with me.

Abstract

Scientific output doubles every nine years. This rising torrent of information has placed the evidence synthesis process under increasing strain, contributing to lengthy production times and impacting the translation of health research into practice and policy. The process of evidence synthesis is extremely resource intensive, often taking small research teams years to complete. Updating reviews as new evidence becomes available, has also proved challenging with many remaining static publications, reporting outdated or even inaccurate information.

A critical stage in the evidence synthesis process is the identification of evidence for inclusion. The advent of bibliographic databases such as PubMed and Embase marked a step-change in information retrieval practices. However, a myriad of problems including poor reporting of primary research, inconsistent indexing, and lack of standardised record formatting, compounded to produce a significant specificity problem in information retrieval for health evidence syntheses. In short, the process is inefficient and wasteful.

Using crowdsourcing for the study identification stages of review production may help to remove this bottleneck. Crowdsourcing is the engagement of a large group of people, usually via the internet, in a problem-solving or idea-generating activity. It can take a range of forms depending on the nature of the problem and the required output. One such crowd model is the crowdsourcing of human computation, or micro, tasks. This involves the manual classification of large data sets that have been broken down into smaller (micro) units and distributed via an open call to willing contributors. The importance of being systematic, and the very rule-driven processes involved in producing robust health evidence, lends itself well to the breaking down of larger tasks to a micro format, and distributing them to anyone with an interest in health and an internet connection.

This applied research aimed to develop, evaluate, and deploy a hybridised model of contribution using crowdsourcing and machine learning within the context of health evidence production. My specific objectives were to investigate the conditions under which each modality (crowd or machine) performed optimally, with a focus on outcome measures related to data quality, efficiency, engagement and capacity. The first three papers (Chapters 2, 3 and 4) form a collection that focus on the identification of reports of randomised trials. Paper 1 looks at the development and evaluation of crowdsourcing this task; Paper 2, at developing and evaluating machine learning capability; and Paper 3 at the performance of a hybrid workflow that uses both components. Papers 4 and 5 are feasibility studies looking at crowd performance when tasked with a different, potentially more

challenging, question and dataset. Systematic reviews are becoming increasingly complex, and evidence based on randomised trials is often not applicable or appropriate. Papers 6, 7 and 8 are set within a COVID-19 context. Paper 7 evaluates a crowd tasked with identifying studies across a range of review question types and under tight time constraints; Paper 8, adopting a similar methodology developed in Paper 2, describes the development and evaluation of a machine learning classifier designed to identify COVID-19 related primary research.

Taken together, this body of work has furthered our understanding of the role crowdsourcing and machine learning can play in the production of health evidence. Specifically, it has contributed new knowledge on the types of tasks suitable as well as methods related to aggregating crowd contributions to achieve high quality data output. In practical terms, crowdsourcing is now implemented into Cochrane review production processes both within the current information retrieval paradigm, in terms of assessing sets of search results retrieved for individual reviews, but also in terms of helping to produce and maintain highly curated repositories of studies as part of Cochrane's Evidence Pipeline. This collection can be leveraged by researchers, academics and practitioners to enable the successful application of such a model across multiple domain areas grappling with information overload.

Contents

List of figures

List of tables

List of abbreviations

AI	Artificial Intelligence
CENTRAL	Cochrane Central Register of Controlled Trials
CCSR	Cochrane COVID-19 Study Register
CCT	Controlled Clinical Trial
CRS	Cochrane Register of Studies
CSS	Centralised Search Service
EBHC	Evidence-based Health Care
FAIR	Findable, Accessible, Interoperable, Reusable [data]
ICMJE	International Committee of Medical Journal Editors
IRMG	Cochrane Information Retrieval Methods Group
MECIR	Methodological Expectations of Cochrane Intervention Reviews
ML	Machine learning
MTurk	Amazon Mechanical Turk
PDF	Portable Document Format
PICO	Population, Intervention, Comparator, Outcome
q-RCT	Quasi-randomised controlled trial
RCT	Randomised controlled trial
S4M	Screen4Me
SR	Systematic review
SVM	Support vector machine
THIS Institute	The Healthcare Improvement Studies Institute

Chapter 1. Introduction

1.1 The research context

The practice of evidence-based healthcare (EBHC) integrates three components: clinical expertise, patient values and preferences, and the best available scientific evidence. Often heralded as the gold standard of scientific evidence, a systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing reliable findings from which conclusions can be drawn and decisions made[1].

Systematic reviews as we conceptualise them today began to appear in the mid-1970s. Their impact has been profound and far-reaching as encapsulated by the story behind the Cochrane logo. The logo depicts the forest plot from an iconic Cochrane systematic review evaluating the effectiveness of corticosteroids given to women about to give birth prematurely. The synthesised evidence demonstrated that the treatment could save the life of the new-born child[2]. Prior to the review, and despite several studies showing the benefit of this intervention, corticosteroids were not routinely used. Numerous successes in EBHC have followed, helping to reduce morbidity and mortality across a broad range of healthcare domains[3,4,5,6].

With decisions affecting people's lives based on systematic review findings, it is critical that they are of high quality. Systematic reviews, as the name implies, should be produced systematically, i.e., according to pre-defined rules and rigorous methods. As with primary research, secondary research of this nature can be affected by bias[7] which can influence or distort the results of the review and render it unreliable, inaccurate, and even harmful. Examples pertinent to evidence synthesis include publication bias, the over-reliance of studies that have been published based on the nature of their results, or outcome reporting bias, the selective reporting of some outcomes but not others. Another potential bias in systematic review research is time-lag bias – the rapid or delayed availability of research findings from primary studies depending on their results.

The methods involved in the synthesis of health evidence in this way have evolved substantially over the last three decades with the aim of reducing risk of bias and minimising statistical imprecision. Cochrane, a leading provider of health-related systematic reviews, has produced the seminal text, *The Cochrane Handbook for Systematic Reviews of Interventions*[8]. This half-a-million-word tome

periodically undergoes major updates to ensure that new methods are adopted. Despite these methodological advances, the production process itself, in terms of the broad stages involved in producing a systematic review, have remained largely unchanged, being frequently conducted in a linear sequence, with one stage completed before the next is begun, generally by small author teams[9]. These key stages are:

- Question formulation
- Search for potentially relevant evidence
- Assessment of potentially relevant evidence
- Appraisal of relevant evidence
- Data extraction
- Statical and/or qualitative synthesis
- Interpretation

Fifty years ago, this research production process was appropriate, and indeed likely the only viable approach. However, the advent of the digital age and with it the semantic web, has brought new opportunities to change this research production paradigm. It has never been easier to access the world's scientific output and be able to share that output within seconds. It has also become easier to work collaboratively as a global community, in real time. Yet despite these technological advances these opportunities have not been realised. The production of secondary research in the form of evidence synthesis such as systematic reviews and meta-analyses has become increasingly challenging. This research is therefore situated within a meta-research context concerning as it does the methods involved in the production of research itself. As described by meta-research methodologists Ioannidis and colleagues:

> As the scientific enterprise has grown in size and diversity, we need empirical evidence on the research process to test and apply interventions that make it more efficient and its results more reliable.[10]

In 2014, Greenhalgh and colleagues published an essay in the *British Medical Journal* entitled: *Evidence-based medicine: a movement in crisis*[11]. In it she described a range of problems – one of which was the notion of 'too much evidence'. Drawing on a bibliometric study conducted by Allen and Harkins in 2005[12], Greenhalgh and colleagues cited one example: "[A] 2005 audit of a 24-hour medical take in an acute hospital, for example, included 18 patients with 44 diagnoses and identified

3679 pages of national guidelines (an estimated 122 hours of reading) relevant to their immediate care"[11]. The problem extends beyond just the overwhelming number of clinical guidelines. Global scientific output doubles every nine years[13]. In the healthcare domain alone, over 4000 new research articles are published every week.

Compounding this issue of exponential output (and in part due to it) is the high level of poor and inconsistent indexing of research[14,15], a lack of conformity by researchers and journal editors in applying appropriate reporting standards[16,17], and the increasing number of new publication channels[18]. These issues directly impact the efficient production of evidence synthesis within the current production paradigm. Sensitive searches, required to reduce the risk of missing potentially eligible studies, often retrieve thousands of results. In a study by Borah and colleagues estimating the time and effort required to produce a systematic review, the number of search results retrieved (based on a sample of 195 published reviews that had been registered in PROSPERO) ranged from 27 to just over 92,000 hits, averaging 2000 hits per review[19]. The mean yield rate (the proportion of the results that were includable studies), calculated by dividing the final number of included studies by the number of hits retrieved post de-duplication, was less than 3%, equating to an appalling level of specificity. Within the context of Cochrane review production alone, it is estimated that in the last twenty years more than 40 million records have been assessed to identify randomised controlled trials (RCTs) for inclusion despite the fact that there have been no more than two million RCTs conducted so far in human history[20].

Methodological filters are a collection of terms appended to a search strategy to help reduce the number of hits retrieved. Systematic reviews based on evidence from randomised controlled trials will likely use a validated methodological filter in the core bibliographic databases they search, such as PubMed/Medline and Embase[21]. Methodological filters have improved search specificity in certain domain areas, but few filters outside of RCT scope have had the same level of validation or achieve an acceptable level of sensitivity (i.e., relevant studies are excluded). This means that for many review question types, for example diagnostic test accuracy or prognostic factor reviews, a methodological filter is not recommended for use as key evidence might be missed and therefore compromise the findings of the review[15,22].

Most reviews are undertaken by small author teams, of around five people[19], many of whom have multiple competing commitments and varying levels of availability[9]. Each team operates effectively within a production silo, beginning the review from scratch with formulating the question and

undertaking the search for relevant evidence. It is therefore unsurprising that the identification of potentially thousands of search results to assess acts as a significant bottleneck early in the review production process. In addition, current guidance recommends that this time-intensive activity be undertaken in duplicate (dual-screening) by members of the author team. Some teams try to alleviate this bottleneck through single-screening of the search results but research indicates this is not reliable and key studies may be missed[23].

In a recent qualitative study conducted by Turner et al., exploring current approaches to producing systematic reviews and opportunities for improvement, several respondents suggested expanding or extending the idea of the author team "beyond a single review, and beyond a single version of a review, to encompass a community that was responsible for the ongoing life of a review as a way of ensuring ongoing consistency and continuity of input"[9]. The current dependence on a single small author team in undertaking all aspects of a review with increasing methodological complexity and the rapidly expanding body of evidence is not sustainable. As one respondent stated: "Teams should be more dynamic; if someone has to drop out of a task, then there should be someone else who can take their place."

1.2 The research problem

The research problem is therefore a meta-research problem concerning the effective production of secondary research in the form of health evidence synthesis. The sheer quantity of research produced has outpaced the traditional review team's capacity to keep up. As the number of systematic reviews published annually continues to grow, many are produced by cutting corners, duplicating effort, and are out of date by the time they are published. In addition, the vast majority are static publications that are never updated to incorporate new evidence. This inability to maintain currency has important ramifications. Reviews are at risk of time-lag bias as results data from negative trials often take substantially longer to publish than evidence from trials reporting positive results. A survival analysis by Shojania et al., identified that significant new evidence was already available for 7% of the reviews at the time of publication and became available for 23% within two years[24]. Taken together, these significant challenges are central to an evidence production process that is under increasing strain. New approaches to the production of evidence are needed. There will be no one single solution that will fix the complex problems of producing robust, reliable and relevant evidence; it will take a cross-discipline, and cross-organisational effort. However, at the core of the research problem described here lies the need for better approaches to managing information, and better organisation of human effort in the production of health evidence synthesis.

1.3 Crowdsourcing

Crowdsourcing is the organised outsourcing of a problem, task, or activity to a large group of people, usually via the internet. The term was first coined by Jeff Howe in 2006 in his well-known *WIRED* article *The Rise of Crowdsourcing*[25]. However, the notion of collective intelligence via communities (also known as: 'the wisdom of the crowd', the 'hive mind', 'swarm intelligence' etc.) dates back centuries. In 1714, the British government set up The Longitude Prize. Here the crowd were tasked with coming up with a way to determine a ship's longitude at sea (determining latitude was far less problematic as this could be found based on the altitude of the sun at noon). A series of monetary rewards were established: the equivalent of £1.3 million would be rewarded to the person or group who produced a method of determining longitude at sea within 1 degree[26].

Crowdsourcing can take several forms depending on the nature of the task and the hoped-for result or output. Several definitions, typologies and frameworks of crowdsourcing exist[27] but one, developed by Brabham and colleagues[28] describes four discrete types of crowdsourcing based on the nature of the problem that needs solving. First, *knowledge discovery and management*, where an organisation tasks the crowd with finding and collecting information into a common location and format; second, the *broadcast search* where the crowd is challenged to solve an empirical problem (e.g. The Longitude Prize described above); third, *peer-vetted creative production* tasks a crowd with creating and selecting creative ideas; and finally, *distributed human intelligence tasking*, where a crowd is tasked with analysing large amounts of information or data that have been decomposed into smaller (micro) units[28].

The last two decades have witnessed a dramatic increase in the use of crowdsourcing across both public and private sectors[29]. Well known examples across each of the types defined by Brahbam include *Threadless*[30], an online community of artists where designs are created and selected to be made available as t-shirts by the community (an example of peer-vetted creative production). The site *HeroX*[31] hosts modern day 'challenges' that "connect everyday problem solvers like you to bring innovative thinking to the world", the broadcast search approach to crowdsourcing. Indeed, The Longitude Prize, described above, also remains a good example of Brabham's broadcast search approach. The Prize is now a £10m prize fund, for a team of innovators who develop a diagnostic test that will conserve antibiotics for future generations[32]. An important and increasingly utilised area in which crowds are engaged, is in emergency response and disaster management for natural hazards such as floods, wildfires, and earthquakes. Here Brabham's knowledge discovery and

management model is most commonly adopted, using platforms such as *Crowdmap*[33]. Crowdmapping, made possible by global positioning system (GPS) technology, aggregates multiple types of crowd-generated inputs to create an up-to-date digital map of a particular event. *OpenStreetMap* is another example of this type of crowdsourcing[34].

Crowdsourcing marketplaces have also emerged. Amazon Mechanical Turk (MTurk) is likely the most prominent such marketplace. It was launched in 2005, initially to assist with the maintenance of its own site but quickly expanded to enable others to post crowd tasks. Here the mode of crowd use is more aligned to Brabham's distributed human intelligence tasking. Businesses (termed Requesters in MTurk) are invited to "break down a manual time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the internet…so internal staff can focus on higher value activities"[35]. A varied range of use cases can be found at any one time browsing the available microtasks on MTurk.

Microworking or microtasking is often accompanied by micropayment (small, piece rate payments), as is the case for MTurk and other similar crowd marketplaces e.g., Clickworker[36] and Minijobz[37]. However, a branch of crowdsourcing that heavily, but not exclusively, utilises both knowledge discovery and management, and distributed human intelligence tasking, is citizen science. Citizen science is the practice of public participation and collaboration in scientific research to increase scientific knowledge[38]. The term is used widely and increasingly often in contemporary discourse regarding public participation in science and research. It can take many forms across the participatory spectrum but is historically most commonly associated with environmental and ecological monitoring activities such as the eBird project[39]. In many projects or initiatives described as citizen science activities, contributors perform either data collection, monitoring activities, or classification tasks. The latter classification tasks are essentially microtasks: small, discrete tasks that cannot be reliably, or entirely, performed by a machine. Such tasks are an example of human computation.

Human computation methods "leverage human processing power to solve problems that are still difficult to solve by using solely computers…While human computation methods could theoretically involve only small numbers of contributors, crowdsourcing approaches leverage the 'wisdom of the crowd' by engaging a high number of online contributors to accomplish tasks that cannot yet be automated, often replacing a traditional workforce."[40] Human computation approaches are therefore ideally suited to situations where the following conditions apply: 1. Large amounts of data

or information are produced and need processing; 2. The amount of information or data that needs processing has outpaced traditional workforce capacity; 3. The data sets can be broken down and distributed in a microtask format; 4. The new microtask can attract contributors to perform it.

The prolific production of scientific output is creating significant bottlenecks within the current health evidence production process (conditions 1 and 2 listed here). Additionally, we have described the systematic and rule-driven activities required to reduce risk of bias and produce robust evidence syntheses. Rule-based activities offer huge potential for reformatting tasks as microtasks (condition 3). And we have touched on contributor incentives (condition 4) in terms of micropayment. Monetary reward is one viable approach to attracting contributors. However, it is not the approach taken by the citizen science movement. Instead, altruism, topic interest, educational aspirations and fun are leading motivators[41,42]. People want opportunities to participate either because they are interested in the aims or goals of the initiative and/or they want to learn about a new topic or gain news skills. In the area of health, a further related motivation may also be pertinent: the initiative may directly relate to the experience of the individual contributor whether as a patient with a particular health condition or as a friend or relative of someone with the condition.

A highly successful citizen science initiative that leverages human computation methods in a health-related area (Alzheimer's disease dementia) is StallCatchers[43]. This initiative meets all conditions described above. In dementia due to Alzheimer's disease, stalls (clogged blood vessels) in the brain reduce blood flow. This is linked to the development of Alzheimer's disease. StallCatchers is an online game that invites contributors to watch video clips showing the brain of mice with Alzheimer's disease. The aim of the game is to identify the 'stalls' in the video clips. The game has proved incredibly popular attracting millions of contributors and demonstrates the potential to harness human effort in this way. This approach produces relevant data needed, whilst eliminating the processing bottlenecks: "In one hour of playing the game, citizen scientists are able to analyze what it takes scientists one week to accomplish in a lab setting"[43]. All essential conditions required for the successful crowdsourcing of a human computation microtask have been met: large quantities of data that required processing have been broken down into a micro format, and made into an appealing game-with-a-purpose with the clear goal-value of gaining a better understanding of a debilitating condition that affects 50 million people around the world. StallCatchers is therefore an excellent example of crowdsourcing a human computation task within a basic science or primary research remit.

1.4 Literature review

Jeff Howe's 2006 article in which he defined crowdsourcing as the act of a company or an institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call, is thought to mark the launch of the modern era of crowdsourcing[24]. As described above, the practice of crowdsourcing grew quickly, with many instantiations across multiple domains. Indeed, its diffuse nature likely impeded its initial development as a coherent field of research. A variety of definitions quickly emerged, and sometimes conflicted with each other, signifying an unstructured and rapid evolution.

In 2012, Estellés-Arolas published a paper entitled *Towards an Integrated Crowdsourcing Definition*[44]. Recognising that the theoretical knowledge base was not yet solid, they sought to produce a single, cohesive, global definition of crowdsourcing that would align with developing typologies, such as Brabham's problem-focussed typology described above[27], or Geiger's example-based taxonomy[45]. Through analysis of multiple existing definitions and extraction of common elements, Estellés-Arolas established the basic characteristics of any crowdsourcing initiative. Eight key characteristics were identified:

(a) There is a clearly defined crowd
(b) There exists a task with a clear goal
(c) The recompense received by the crowd is clear
(d) The crowdsourcer is clearly identified
(e) The compensation to be received by the crowdsourcer is clearly defined
(f) It is an online assigned process of participative type
(g) It uses an open call of variable extent
(h) It uses the internet

The resulting definition, designed to cover any type of crowdsourcing and to reduce the pre-existing semantic confusion, was:

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organisation, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always

> entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken[44].

During this time, citizen science, a form of crowdsourcing that engages the public in scientific projects, was also burgeoning – both in terms of the science of citizen science, and the number and types of applications of it/projects using it. In Follet's 2015 analysis of citizen science based research[46], they assert increased acceptance of this method by the scientific community and describe research into the science of citizen science focussing on three key areas: the methods applicable to citizen science projects, validation techniques, and studies on motivating volunteers. However, almost all early applications of citizen science methods were related to environmental, ecological, or astronomical endeavours, with the vast majority of these being highly visual, image-based tasks. Indeed, crowdsourcing more broadly had had very limited exposure in the field of health.

A systematic review by Ranard and colleagues in 2013 looked specifically at applications of crowdsourcing in the health and medicine domains[47]. They identified only 21 studies reflecting the use of crowdsourcing in health-related research. Within those 21 studies, crowdsourcing was utilised in four main ways: problem solving, data processing, surveillance, and surveying. There was considerable variability in how the methods of crowdsourcing were reported and relatively little by way of robust validation. The conclusion was that the field was in its infancy, and that important questions remained around the quality of the data crowdsourcing provides.

The same year as Renard's review also saw the publication of *The Handbook of Human Computation*[40]. Human computation was another emerging and relevant area of enquiry. The term had been coined by Luis von Ahn in 2008[48] and refers to methods that combine human brainpower with computers to solve problems that neither could solve alone. *The Handbook of Human Computation* brought together experts in the field to cover the foundations of the field, its application domains, techniques and modalities, algorithms and so on. The editor, Pietro Michelucci, was also the founder of the StallCatchers initiative, described above.

Another systematic review, published in 2018, which mapped crowdsourcing applications in health, showed that the use of crowdsourcing across health promotion, health research and health maintenance, had increased substantially[49]. By this stage, data processing was the most frequently

used application of crowd effort, mainly in public health, with none, as yet, having looked at researching crowdsourcing within a health evidence synthesis context.

1.5 Aims and objectives

My overarching research question is *how can crowdsourcing be effectively utilised in the production of health evidence syntheses?* My aim is to establish new knowledge on how crowdsourcing data can be generated and used to its full potential in the context of health evidence synthesis. Within that aim, I am primarily concerned with four main areas of enquiry: (1) quality of the data produced by the crowd, and identifying factors that may affect data quality; (2) efficiency of the crowdsourced processes in comparison to other approaches; (3) engagement of the crowd and factors that might affect recruitment and retention; (4) implementation into evidence production processes; how best to integrate crowd generated data into existing and new processes. Based on these four areas of enquiry, my specific research objectives are:

- **Objective 1:** To evaluate crowd accuracy across a range of crowdsourced microtasks
- **Objective 2:** To evaluate measures of efficiency and consensus across a range of crowdsourced microtasks
- **Objective 3:** To evaluate crowd demographics and engagement across a range of crowdsourced microtasks
- **Objective 4:** To explore use of crowd data for machine learning and human-machine workflows

1.6 Research questions

In response to these objectives, I designed and conducted a range of studies. Within each study I addressed a specific research question. See Table 1.1 for the list of specific research questions for each study, the related research objectives, and the outcome measures evaluated for each research question.

Table 1.1 Research questions

Main research questions ***Related research objective***	**Outcome measures**	**Thesis chapter**
Can a crowd accurately and efficiently identify reports of randomised or quasi-randomised controlled trials? *Objectives 1, 2, 3*	- What is the crowd's accuracy in terms of sensitivity? - What is the crowd's accuracy in terms of specificity? - What is the level of crowd consensus? - What are the demographics of the crowd? - How does the crowd model compare with the previous model of RCT identification?	Chapter 2
What is the accuracy of the machine learning Cochrane RCT Classifier? *Objective 4*	- What is the RCT Classifier's recall? - What is the RCT Classifier precision? - What is missed by the RCT Classifier and why?	Chapter 3
How effective is Cochrane's Centralised Search Service workflow at identifying randomised or quasi-randomised controlled trials? *Objective 4*	- What is the overall performance of the workflow in terms of sensitivity? - How does each component (search, crowd, classifier) within the workflow perform? - What is missed and why? - What the additional considerations for researchers wanting to identify RCTs from CENTRAL?	Chapter 4
How accurately and efficiently can a crowd perform a topic-based assessment for an interventional systematic review? *Objectives 1,2,4*	- What is the crowd's accuracy in terms of sensitivity? - What is the crowd's accuracy in terms of specificity? - What is the level of crowd consensus? - What are the demographics of the crowd?	Chapter 5
Can a crowd accurately and efficiently identify studies for a complex mixed studies systematic review? *Objectives 1,2,3*	- What is the crowd's accuracy in terms of sensitivity? - What is the crowd's accuracy in terms of specificity? - What is the level of crowd consensus? - How replicable are the results? - What impact on accuracy measures does changing the agreement algorithm have? - What are the demographics of the crowd? - What did the crowd think of the task?	Chapter 6
How has Cochrane Crowd handled the response to the COVID-19 pandemic? A case study. *Objectives 1,3*	- Can a crowd identify, and tag human studies related to COVID-19? - What role can a crowd play in the production of Cochrane Rapid Reviews related to COVID-19? - Can crowd generated data help to produce a machine learning classifier to reduce manual screening burden?	Chapter 7
Can a crowd accurately and efficiently identify studies for a for a range of rapid reviews under tight time constraints? *Objectives 1,2,3,4*	- What is the crowd's accuracy in terms of sensitivity? - What is the crowd's accuracy in terms of specificity? - What is the level of crowd consensus? - What was the time-to-task completion for each task? - What was the impact of missed studies review conclusions?	Chapter 8
What is the accuracy of the COVID-19 Classifier? *Objective 4*	- What is the C-19 classifier's sensitivity? - What is the C-19 Classifier's precision? - What is missed by the classifier and why? - What is the workload reduction on manual screening?	Chapter 9

1.7 Research methodology

I designed and conducted a range of empirical studies utilising appropriate quantitative and qualitative study designs. To assess crowd performance in terms of crowd accuracy measures I employed a discriminatory performance approach that sought to compare crowd performance against a gold or reference standard. The two main accuracy measures related to crowd performance are crowd sensitivity and crowd specificity. Crowd sensitivity is the crowd's collective (as opposed to an individual's) ability to correctly identify the class of interest (what is being looked for). Crowd specificity is the crowd's collective ability to correctly identify the items that should be rejected (the non-class of interest). As described below, the agreement algorithm employed for each human computation task plays a critical role in helping to ensure collective accuracy and high-quality data output. It also produces a further measure of performance which I have termed crowd consensus. Similar to a notion of efficiency, crowd consensus is the proportion of the data set processed by the crowd that does not require any further manual input. It is an important measure alongside measures of accuracy.

As well as evaluating crowd performance within specific microtasks, I also sought to explore more broadly the uses, implementation, and impact of crowd-generated data. This is detailed in two main ways. First, in the development of machine learning models trained using crowd-generated data, and second in the development and deployment of evidence production workflows that incorporate crowd (and machine) processes. Many promising innovations are not adopted due the challenges of integrating them into feasible workflows[50,51]. Therefore, a critical aspect of enabling scale-up and widespread adoption lies in either integration of new technology into existing production workflows or in the creation of new workflows.

1.8 Thesis structure

This thesis is made up of ten chapters, eight of which correspond to a research paper, all of which have been published in peer-reviewed journals in the fields of epidemiology, evidence-based healthcare or health informatics. Below is a brief synopsis of those eight chapters and a description of how each part of the investigation connects and contributes to the overall research project.

1.8.1 The relationship between the chapters

The first three papers (Chapters 2, 3 and 4) form a collection that focus on the identification of reports of randomised trials. In Cochrane, over 90% of the systematic reviews produced rely on the

identification and inclusion of randomised or quasi-randomised controlled trials. This therefore represented a valid starting point in terms of evaluating both crowd and machine potential in identifying this particular study design. Chapter 2 looks at the development and evaluation of crowdsourcing this task; Chapter 3, at developing and evaluating machine learning capability; and Chapter 4 at the performance of a hybrid workflow that uses both components. Chapters 5 and 6 are feasibility studies looking at crowd performance when tasked with a different, potentially more challenging, question and dataset. Systematic reviews are becoming increasingly complex where evidence based on randomised trials is often not applicable or appropriate. Chapters 7, 8 and 9 are set within a COVID-19 context. Here, knowledge generated from the previous six chapters is applied and evaluated during a public health emergency context. New knowledge is also generated with the introduction of multi-question crowd tasks (Chapter 7). Chapter 8 evaluates a crowd tasked with identifying studies across a range of review question types and under very tight time constraints; Chapter 9, adopting a similar methodology described in Chapter 3, describes the development and evaluation of a machine learning classifier designed to identify COVID-19 related primary research. Below is a synopsis of each chapter and a description of how each part of the investigation connects and contributes to the overall research project.

1.8.2 Chapter 2

An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials[52]

This chapter introduces Cochrane Crowd[53], the web application that provides the functionalities and crowd management underpinning the methodological work presented in this thesis. I have led the development of the Cochrane Crowd platform since inception building on an earlier initiative in which I led the crowdsourcing component[47]. Cochrane Crowd was launched in May 2016 and in line with the principals of open innovation and Health 2.0 we were keen to have as few barriers to entry as possible. Anyone with an internet connection can join the initiative and start contributing without having had any prior experience or knowledge. At the time of writing (January 2022) Cochrane Crowd has attracted over 23,000 contributors from 170 countries. The platform began with a single microtask but expanded quickly and, to date, has hosted over fifty.

In order to make Cochrane Crowd as accessible as possible whilst also ensuring high quality data output by the crowd, both task training and the method of decision aggregation are vital. Every microtask is supported by a brief, interactive training module which is mandatory for potential

crowd contributors to complete. The training, usually made up of a qualification set of practice records is designed to introduce people to the task and help to ensure accurate decision-making by individuals. However, it would not be safe to rely on a single classification as the ground truth. Behind each microtask sits an algorithm that automatically aggregates the individual classifications made by contributors into a final classification. There are multiple ways to aggregate crowd responses. The primary aim is to reliably eliminate or reduce the need for further manual assessment. Majority voting is one popular approach. This rule is relatively straightforward to implement but can require a high number of individual classifications which has implications on crowd capacity (a small crowd will not produce enough unique classifications). The agreement algorithm we developed and implemented is similar to majority voting: each record assessed needs a certain number agreeing classifications made consecutively for a final classification to be generated. A record achieving the required number of agreeing classifications requires no further manual scrutiny. If a break in the consecutive chain occurs (i.e., a crowd member makes a conflicting classification in comparison to an already made classification) the record will enter a new workflow involving further manual assessment by a 'resolver' crowd member.

With data quality being our initial primary concern, Chapter 2 reports evaluations of the first three microtasks developed for, and hosted on, the Cochrane Crowd platform. The aim of each task was the identification of randomised controlled trials (RCTs) from three external sources. For each evaluation, a gold standard data set was used to compare the collective crowd decisions. The agreement algorithm developed for each task is described and crowd sensitivity, crowd specificity and crowd consensus were calculated for each.

In addition to these individual evaluations for the three microtasks, this work also includes wider analysis of the crowd's capacity to keep up with the flow of records from external sources such as Embase and ClinicalTrials.gov. Crowd capacity is a critical consideration; high quality data output is of little value if it proves difficult to recruit a large enough crowd to perform the task on an ongoing basis. Creating a flexible model of contribution where no minimum commitment is required brings with it the risk that crowd effort will not be continuously sustained. We demonstrate that compared to a previous model of study identification, the crowd, and subsequently the crowd plus machine learning capability (described below) was not only able to keep pace with the ever-increasing number of records retrieved by the searches, but to significantly outpace the previous approach, thereby enabling further expansion in terms of the number of external sources searched and assessed for RCTs in this way.

As well as helping to identify reports of randomised and quasi-randomised trials in a highly accurate and efficient way, thereby enabling a constant flow of current RCTs to be submitted to Cochrane's central repository of trials, the crowd also produced a valuable by-product: a large quantity of high-quality training data. These data were used in the development, and subsequent implementation, of machine learning classifiers.

1.8.3 Chapter 3

Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane reviews[55]

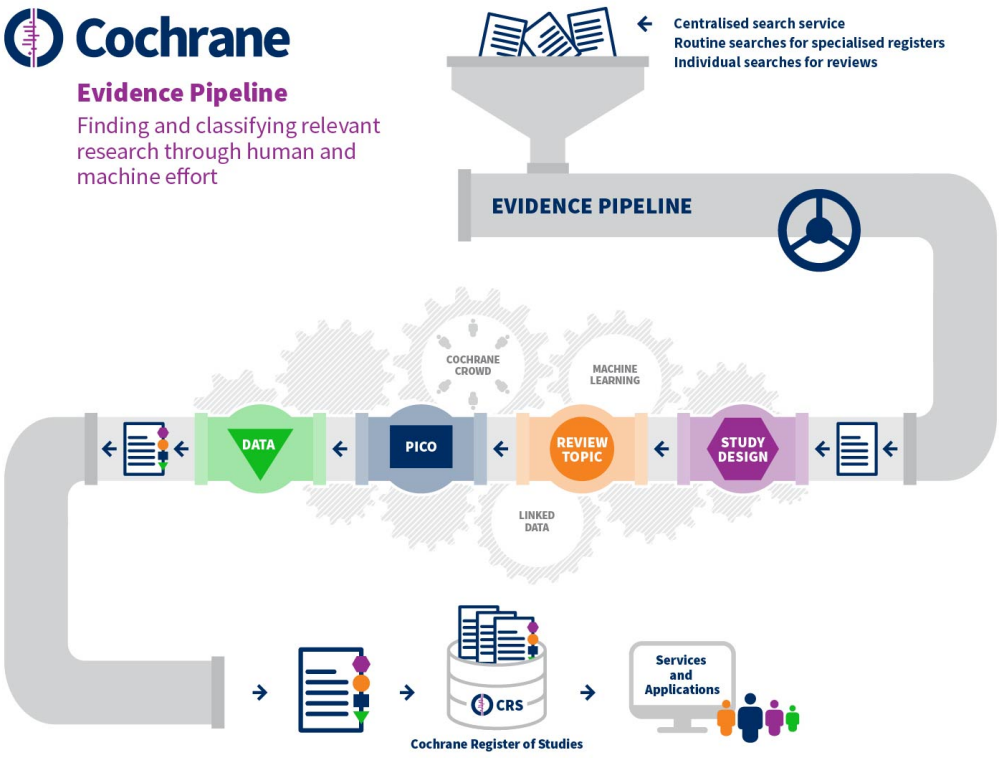
In this chapter, the development of the 'RCT Classifier' is described. One important potential use of high-quality crowd-generated data is as training data for machine learning. As described in Chapter 2, crowd accuracy in terms of both crowd sensitivity and crowd specificity was very high. This, together with the size of the data set and the representation of both positive and negative classes (i.e., RCTs and non-RCTs), made it highly suitable for training a machine learning classifier.

Machine learning in this context comprises a group of algorithms that 'learn' to perform a task via exposure to representative data sets. In this study we used supervised machine learning (training the algorithm on records for which the true label is known) and built an ensemble classifier made up of two support vector machine (SVM) models. With each SVM, the bag-of-words approach was used where each record is represented as a vector of 0's and 1's, depending on the presence or absence of each unique word from the article set vocabulary[56]. As well as the creation of a machine learning classifier that would output likelihood scores for records in terms of the probability that the record is describing an RCT, we were keen to determine a cut-point or threshold between RCT and non-RCT classifications. To do this we used an independent (i.e., one not generated by Cochrane Crowd), yet representative, data set. The data set we used is known as the Clinical Hedges data set. It was built for the purpose of testing and validating methodological search filters[57]. Using bootstrap sampling, we used this data to identify the threshold that would achieve a 99% recall (a threshold that would capture at least 99% of the RCTs in the set). Finally, we validated the ensemble classifier and its cut-point on a third, independent, yet highly representative data set: the included studies from Cochrane intervention reviews. This third data set was made up of 58,283 studies from 4,296 Cochrane reviews. The ensemble classifier correctly identified 99.5% of studies. This work describes in detail the training, calibration and validation of a machine learning classifier. The Cochrane RCT

Classifier has now been deployed in Cochrane. Its development was made possible by the collective efforts of the Cochrane Crowd, who produced the valuable training data.

Chapters 2 and 3 therefore describe the development and deployment of two technological enablers, crowdsourcing and machine learning. Together, these enablers form a core part of a larger vision called the Cochrane Evidence Pipeline. The Evidence Pipeline seeks to transform study identification for Cochrane and other evidence synthesis producers (see Figure 1.1). Research enters the Evidence Pipeline and goes through tailored workflows involving crowdsourcing and machine learning, working together to produce accurate, reliable metadata about studies. The implementation of this ensemble classifier into the Evidence Pipeline has brought significant efficiency to the process of identifying RCTs, with approximately 30-40% of records that enter The Pipeline being handled by machine alone (through being rejected by the machine as non-RCTs). This has created a virtuous cycle, as machine-input frees up human resource for the parts of the task that still require human input or indeed for other human computation tasks. Chapter 4 goes on to describe a retrospective evaluation of this implemented workflow that incorporates both the RCT Classifier and the Cochrane Crowd, working together in partnership to identify RCTs.

Figure 1.1 The Cochrane Evidence Pipeline



1.8.4 Chapter 4

Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: a retrospective analysis[58]

Cochrane's Centralised Search Service (CSS) forms a core part of the Evidence Pipeline. It encompasses the retrieval of reports of RCTs from external sources such as PubMed and Embase and the process involved in their subsequent assessment and publication in Cochrane's Central Register of Controlled Trials (CENTRAL). CENTRAL is a bibliographic database accessible via the Cochrane Library[59]. It is a valuable resource for healthcare researchers and professionals, and it is mandatory for Cochrane systematic reviewers to search CENTRAL for Cochrane intervention reviews. CENTRAL is populated with reports of randomised and quasi-randomised controlled trials that have been submitted to CENTRAL in one of two ways: (1) via Cochrane Information Specialists manually adding trial records via Cochrane's reference management software, called the Cochrane Register of Studies (CRS), and (2) via the Centralised Search Service. The CSS uses four main approaches: 'direct feeds' of records already indexed as RCTs in the external sources; sensitive search strategies to retrieve records from the source databases that might be RCTs but have not been indexed as such; machine learning using the Cochrane RCT Classifier described in Chapter 3, which primarily models decisions about what to ignore using a calibrated cut-point; and finally, crowdsourcing via Cochrane Crowd, as described in Chapter 2, who assess the remaining records.

This chapter describes a retrospective analysis conducted to assess the effectiveness of this CSS workflow and each of its component parts. We used a convenience sample of 650 references to RCTs that had been included in Cochrane reviews. We performed an audit trail on each record to determine if it had been identified by the CSS, and if so, how (i.e., through which component). We also performed an analysis on any references to RCTs that had been missed by the CSS workflow. The results showed that 97.5% of RCTs in our sample had been identified by the CSS. Some studies, however, were missed: four by the sensitive search filters, three were collectively mis-classified by the crowd, one was incorrectly rejected by the RCT Classifier. This analysis helped us to better understand weak points in our workflow but primarily indicated the effectiveness of this approach.

The implications of this analysis are far-reaching. As CENTRAL becomes ever-more comprehensive in terms of RCT coverage, the need for multi-source searching in the way it is currently done, is significantly lessened. In 2021, over 95% of reports of RCTs submitted to CENTRAL were identified by the CSS via the Evidence Pipeline. This therefore marks a potential step-change in the study

identification process for health evidence reliant on randomised trials. It will bring significant efficiencies in the identification of RCTs for systematic reviews and other evidence outputs.

In summary, Chapters 2, 3 and 4 in this collection describe the development and deployment of a human-machine workflow geared towards the identification of RCTs for populating a critical repository of randomised trials. The RCT use case is an important one due to over 90% of Cochrane intervention systematic reviews relying on inclusion of evidence from randomised trials only. However, many reviews seek to incorporate a range of evidence not encapsulated by an RCT. Chapters 5 and 6 describe two pilot studies that explore the feasibility of crowdsourcing microtasks based on topic assessment for both an RCT-based systematic review and for a complex mixed studies review.

1.8.5 Chapter 5

Citation screening using crowdsourcing and machine learning produced accurate results: evaluation of Cochrane's modified Screen4Me service[60]

This chapter formally introduces the Screen4Me workflow (S4M) and presents an evaluation of a modified S4M workflow that enabled us to test the crowd's ability to perform a citation screening task based on topic relevance, rather than just study design. The Screen4Me workflow was deployed in April 2019 with the aim of enabling systematic review author teams access to both the RCT Classifier (described in Chapter 3) and the Cochrane Crowd (Chapter 2) in assessing the search results for their systematic review. Screen4Me is therefore about offering reviewers a way to lessen the screening burden *within* the current review production paradigm[61].

The workflow starts with the de-duplicated set of search results, against which two components of the S4M workflow are run simultaneously: (1) the RCT Classifier, and (2) a component we have termed Known Assessments. One particularly inefficient aspect of the current production model for systematic reviews is the reuse (or rather *lack of* reuse) of data. The Known Assessments component of Screen4Me aims to make better use of already known metadata about records. Every year millions of records are screened for potential eligibility for reviews (an estimated four million records are assessed annually for new Cochrane reviews alone). The majority of records (over 90%) are rejected on grounds of being ineligible[18]. A sub-set of these rejections will be based on the record reporting an ineligible study design. The Screen4Me workflow is currently only suitable for reviews that seek to include reports of randomised or quasi-randomised trials. The Known Assessments

component of the S4M workflow therefore indicates which records in the search results set have already been assessed by Cochrane Crowd via the centralised workflow, described in Chapter 4, as either describing an RCT or as not describing an RCT. Since launch, the Screen4Me workflow has been used in the development of 109 new Cochrane intervention systematic reviews. The mean reduction in the number of search results for author teams to assess is 63% (inter-quartile range of 28%-86%, based on an evaluation conducted in 2021).

We now wanted to evaluate the crowd when reframing the question from: *Is the record describing or reporting an RCT?* to: *Does this record look potentially relevant to the review?* The current requirement in the production of systematic reviews to run highly sensitive searches across multiple databases means that many of the search results retrieved will not be relevant. We therefore wanted to test whether a crowd could accurately remove the not relevant records and retain potentially relevant records having been trained on a test set of 15 records.

In this pilot study, the crowd achieved 100% sensitivity (collectively classifying all the included studies as potentially relevant). Overall, this modified workflow achieved an 81% workload reduction in terms of the number of records left for the core author team to assess. However, the topic of the review was not complex and therefore not far-removed from the RCT identification crowd tasks described in Chapter 2. Chapter 6 describes a study that sought to assess crowd performance for a more complex, mixed-studies systematic review.

1.8.6 Chapter 6

Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study[62]

Here we examined the crowd's performance in assessing the search results for a complex, mixed studies systematic review on the topic of training for healthcare professionals in intrapartum electronic fetal heart rate monitoring with cardiotocography[63]. All primary empirical research studies evaluating cardiotocography training were eligible for inclusion within the review. As in previous studies we assessed crowd accuracy in terms of sensitivity and specificity, and crowd consensus - the proportion of records not requiring resolution by a crowd resolver. However, in this study we also measured time: the overall time to task completion by the crowd, as well as the mean time taken per record in comparison to the core author team who performed the same task in parallel. Additionally, we sought to better understand crowd contributor motivations for taking part, as well as their views about the task's difficulty and their enjoyment of it.

Crowd performance for this task was good, but not perfect. In the initial running of this task, the crowd did not collectively reject any of the included studies. However, several of the included studies had needed resolving by a resolver crowd contributor. The resolver was a highly experienced crowd screener but unfamiliar with the topic area. The resolver incorrectly rejected eight studies, bringing overall crowd sensitivity down to 84%. This was an unexpected outcome. It led us to try an alternative approach to the record resolution component of the crowd process. We ran the task again, replicating it in all aspects but with a modification to the record resolution part of the agreement algorithm. In the replicated task, instead of using a single person to make the final decision on records that needed resolving, we engaged two crowd resolvers, each assessing all records that needed resolving. They did this task independently of each other with any conflicting classifications between them resulting in an automatic final classification of *Possibly relevant*.

Re-running the task not only enabled us to test a new algorithm related to resolving conflicting crowd classifications, it provided us with useful study replication data. One valid concern regarding using a crowd to assess search results is how replicable the results are. It was encouraging to see that with the replication task, strikingly similar metrics were achieved across all outcome measures despite using a completely new crowd.

With regards to time, individual crowd contributors took on average twice as long as individual members of the author team to screen a record yet in terms of overall time to task completion, the crowd's performance was impressive. It took the crowd 33 hours to complete the task (assessing around 10,000 records), whilst taking the review author team 410 hours to complete the same task.

For this study we also included a qualitative component using a questionnaire sent out to all participants once the task was completed. The response rate was excellent for both the original running of the task (81% responded) and for the replicated task (75%). Feedback about the task itself was positive, with many comments reflecting that it was both a doable task and provided the contributor with a way to be usefully involved in a worthwhile activity, for example one contributor wrote: "It was good to have a smaller task on offer as it felt more 'doable' and that my contribution would really make a difference"; another: "I think it is a very useful way to spend half an hour when I have the spare time; it made me feel connected, and it seemed to achieve a lot for the review".

Despite the crowd not achieving 100% accuracy in terms of crowd sensitivity, this study compounded our understanding of people's desire to be involved and to help in a flexible and easy way. It also emphasised the advantage of a crowd model in terms of overall time to task completion.

1.8.7 Chapter 7

Crowdsourcing and COVID-19: a case study of Cochrane Crowd[64]

The COVID-19 pandemic unleashed a corresponding 'infodemic' defined by the World Health Organization as "too much information including false or misleading information in digital and physical environments during a disease outbreak…An infodemic can intensify or lengthen outbreaks when people are unsure about what they need to do to protect their health"[65]. This brief case study describes four ways in which crowd effort was harnessed to help tackle the infodemic during the first twelve months of the pandemic.

We developed COVID Quest, a new crowd task hosted on Cochrane Crowd. This task marked a departure from previous tasks. Contributors were tasked with identifying COVID-related studies eligible for Cochrane's COVID-19 Study Register (CCSR)[66], and then to tag those studies with additional metadata regarding the study's design characteristics and aims. We were able to develop, test and deploy this task, despite its increased complexity, within weeks of the first UK national lockdown. This was largely helped by already having a solid technical infrastructure, as well as a willing and able crowd.

As with other tasks, we included elements of gamification as additional incentives: digital badges could be earned as contributors progressed in the task, and every week we conducted 'weekly challenges'. These challenges were three-hour blocks of time where we encouraged the community to work specifically on that task to see how many COVID studies could be collectively identified and tag.

The final two use-cases described concern crowd input into Cochrane rapid reviews related to COVID-19, and the development of a machine learning classifier for helping to identify COVID-19 related studies, trained, in part, on data generated by the crowd from the already described COVID Quest crowd task. Chapters 8 and 9 describe each of these latter two use cases in detail.

1.8.8 Chapter 8

Crowdsourcing the identification of studies for COVID-19 related Cochrane rapid reviews[67]

Rapid reviews are a form of evidence synthesis that aim to provide more timely information for decision making compared with standard reviews. The methods involved in producing rapid reviews are evolving and can vary substantially between producers[67]. However, one part of the rapid review often pared back due to time constraints, is the search for studies. As the infodemic produced by the pandemic gained momentum, identifying relevant and emerging evidence related to the virus in a timely manner became increasingly challenging. These circumstances presented us with an opportunity to further evaluate crowd capability, specifically with a focus on potential crowd input into study identification for rapid reviews related to COVID-19. Here we tasked the crowd with assessing sets of search results within a 48-hour time period. Our previous work, described in Chapter 6, had indicated that a crowd could collectively assess a set of records much more quickly than a 'traditional' author team. We also had encouraging evidence from the Screen4Me crowd tasks, where the crowd are given two weeks to complete the screening task (over 95% of Screen4Me tasks complete comfortably within that two-week time frame).

In this study, we created four crowd tasks, each one based on a different Cochrane COVID-19 rapid review. The crowd performed achieved accuracy measures ranging from 94% to 100% sensitivity across the four reviews, and completed three of the four tasks within the 48 hours, and one in 52 hours. As well as measuring crowd accuracy for time-sensitive tasks, we also assessed (a) whether the conclusions of the reviews would have been altered by the missed studies, and (b) whether any of the missed studies would have been identified by the core author teams performing citation tracking on the studies that had been correctly identified.

Overall, this methodological work done within a COVID-19 context, provided us with further evidence of effectively utilizing a crowd to help in the rapid identification of evidence needed for rapid reviews. Acknowledging that no system will be 100% accurate at all times, we also discuss how the generation of crowd data is only one part of the equation; it is how that data is then used that is critical, and provide three possible configurations of how author teams could interact with crowd-generated data based on what their priorities are.

1.8.9 Chapter 9

Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier[69]

Chapter 9 describes the development of the Cochrane COVID-19 machine learning classifier. Using the same methodology described in Chapter 3, we trained a support vector model to identify potentially relevant studies for the Cochrane COVID-19 Study Register (CCSR) with the aim of reducing manual screening burden. Both crowd-generated screening data produced via COVID Quest (described in Chapter 7) as well as 'in-house' data generated by Cochrane information and data curation specialists were used. As with the RCT Classifier, a calibration stage was performed to enable us to use this classifier in a binary fashion within our COVID-19 study identification workflow. Given the importance of not missing eligible studies for the CCSR, we determined a cut-point that would help to remove ineligible records (rather than identify eligible records). Records therefore scoring below the determined cut-point would be discarded, thereby reducing manual screening effort.

This COVID-19 machine learning classifier is now fully implemented into the COVID study identification workflow. Analogous to the RCT Classifier's deployment, as described in Chapter 4, this classifier forms a core part of a study identification workflow that harnesses both machine and human effort. Having been trained, calibrated and validated on high quality data, the implemented classifier reduces the number of records for manual screening by around 25% overall, and approximately halves the number of ineligible records for manual assessment.

As the deluge of information being produced during this pandemic, of both highly variable quality and structure, shows no sign of abating, adapting existing infrastructure, systems, workflows and processes to operate within a COVID-19 context has been invaluable. It has also helped to further test both crowd and machine capability on 'messy' literature and moved us convincingly beyond just the identification of RCTs.

1.9 Summary

In his chapter entitled *Human Computation in the Wild*, from the *Handbook of Human Computation*, Haym Hirch reminds us that "one of the backbones of human society has been finding ways to organise human labor to achieve desired outcomes…The advent of computing has allowed to bring

to bear the ideas and tools of computing to this task, giving rise to what we are now calling 'human computation'"[40].

This work as a whole represents significant progress regarding our growing understanding of the huge potential of crowdsourcing human computation tasks relevant to health evidence synthesis. It has the advantage of being based on robust methodological work conducted 'in the wild', as part of an evidence ecosystem that cannot be paused whilst we run 'lab-based' experiments. This brings it huge external validity: Cochrane Crowd is a real crowd, a diverse, global community of people brought together to improve health. Within this context, we have tested a crowd model under various conditions, with a range of different tasks, and perhaps even more importantly, we have deployed both crowd, and machine learning models (trained on crowd-generated labels) into production workflows. This has had a direct impact on study identification within the *current* information retrieval paradigm via Screen4Me, but also in plotting a future course into new territory of upstream, ongoing metadata creation and curation via the Evidence Pipeline.

1.10 References

1. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009 Jul 21;339:b2700. doi: 10.1136/bmj.b2700. PMID: 19622552; PMCID: PMC2714672.
2. Roberts D, Brown J, Medley N, Dalziel SR. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database of Systematic Reviews 2017, Issue 3. Art. No.: CD004454. DOI: 10.1002/14651858.CD004454.pub3.
3. British Thoracic Society. Guidelines for management of asthma in adults: I. Chronic persistent asthma. BMJ 1990;301:651-3.
4. Majeed A, Ferguson J, Field J. Prescribing of beta-2 agonists and inhaled steroids in England: trends between 1992 and 1998, and association with material deprivation, 42 chronic illness and asthma mortality rates. J Pub Health Med 1999;21:395-400.
5. Kelly MP, Capewell S. Relative contributions of changes in risk factors and treatment to 43 the reduction in coronary heart disease mortality. Health Development Agency, 2004.
6. Lau BD, Haut ER. Practices to prevent venous thromboembolism: a brief review. BMJ 44 Qual Safety 2014;23:187-95.
7. Felson DT. Bias in meta-analytic research. J Clin Epidemiol. 1992 Aug;45(8):885-92. doi: 10.1016/0895-4356(92)90072-u. PMID: 1624971.
8. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021). Cochrane, 2021. Available from www.training.cochrane.org/handbook.
9. Turner T, Green S, Tovey D, McDonald S, Soares-Weiser K, Pestridge C, Elliott J; Project Transform Team; IKMD developers. Producing Cochrane systematic reviews-a qualitative study of current approaches and opportunities for innovation and improvement. Syst Rev. 2017 Aug 1;6(1):147. doi: 10.1186/s13643-017-0542-3. PMID: 28760162; PMCID: PMC5537977.
10. Ioannidis JPA. Meta-research: Why research on research matters. PLoS Biol 2018; 16(3): e2005468. https://doi.org/10.1371/journal.pbio.2005468.
11. Greenhalgh T, Howick J, Maskrey N; Evidence Based Medicine Renaissance Group. Evidence based medicine: a movement in crisis? BMJ. 2014 Jun 13;348:g3725. doi: 10.1136/bmj.g3725. PMID: 24927763; PMCID: PMC4056639.
12. Allen D, Harkins K. Too much guidance? Lancet 2005;365:1768.

13. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 2015;66(11):2215–2222.
14. Burns CS, Nix T, Shapiro RM 2nd, Huber JT. MEDLINE search retrieval issues: A longitudinal query analysis of five vendor platforms. PLoS One. 2021 May 6;16(5):e0234221. doi: 10.1371/journal.pone.0234221. PMID: 33956834; PMCID: PMC8101950.
15. Gurung P, Makineli S, Spijker R, Leeflang MMG. The Emtree term "diagnostic test accuracy study" retrieved less than half of the diagnostic accuracy studies in Embase. Journal of Clinical Epidemiology 2020;126:116-121.
16. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, Catalá-López F, Li L, Reid EK, Sarkis-Onofre R, Moher D. Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. PLoS Med. 2016 May 24;13(5):e1002028. doi: 10.1371/journal.pmed.1002028. PMID: 27218655; PMCID: PMC4878797.
17. Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, Perrodeau E, Altman DG, Ravaud P. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ. 2017 Jun 8;357:j2490. doi: 10.1136/bmj.j2490. Erratum in: BMJ. 2017 Aug 8;358:j3806. PMID: 28596181.
18. Manca A, Cugusi L, Cortegiani A, Ingoglia G, Moher D, Deriu F. Predatory journals enter biomedical databases through public funding. BMJ. 2020;8;371:m4265. doi: 10.1136/bmj.m4265. PMID: 33293265.
19. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 27;7(2):e012545. doi: 10.1136/bmjopen-2016-012545. PMID: 28242767; PMCID: PMC5337708.
20. Thomas J, Noel-Storr AH, McDonald S. Evidence surveillance to keep up to date with new research. Chapter 9 In Levay P (ed.), Craven J (ed.). Systematic searching: practical ideas for improving results. Facet Publishing 2019.
21. Glanville J, Kotas E, Featherstone R, Dooley G. Which are the most sensitive search filters to identify randomized controlled trials in MEDLINE? J Med Libr Assoc. 2020 Oct 1;108(4):556-563. doi: 10.5195/jmla.2020.912. PMID: 33013212; PMCID: PMC7524635.
22. Cohen JF, Korevaar DA, Bossuyt PM. Diagnostic accuracy studies need more informative abstracts. European Journal of Clinical Microbiology and Infectious Diseases 2019;38:1383-1385.
23. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, König F. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized

controlled trial. J Clin Epidemiol. 2020 May;121:20-28. doi: 10.1016/j.jclinepi.2020.01.005. Epub 2020 Jan 21. PMID: 31972274.

24. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007 Aug 21;147(4):224-33. doi: 10.7326/0003-4819-147-4-200708210-00179. Epub 2007 Jul 16. PMID: 17638714.
25. Howe J. The Rise of Crowdsourcing. Wired 2006. https://www.wired.com/2006/06/crowds/ [Accessed 27 December 2021].
26. Longitude Rewards: https://en.wikipedia.org/wiki/Longitude_rewards [Accessed 27 December 2021].
27. Estellés-Arolas E., Navarro-Giner R., González-Ladrón-de-Guevara F. (2015) Crowdsourcing Fundamentals: Definition and Typology. In: Garrigos-Simon F., Gil-Pechuán I., Estelles-Miguel S. (eds) Advances in Crowdsourcing. Springer, Cham. https://doi.org/10.1007/978-3-319-18341-1_3.
28. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med. 2014 Feb;46(2):179-87. doi: 10.1016/j.amepre.2013.10.016. PMID: 24439353.
29. Ghezzi A, Gabelloni D, Martini A, Natalicchio A. Crowdsourcing: a review and suggestions for future research. International Journal of Management Reviews 2017; https://doi.org/10.1111/ijmr.12135.
30. Threadless: https://www.threadless.com [Accessed 27 December 2021].
31. HeroX: https://www.herox.com [Accessed 27 December 2021].
32. The Longitude Prize: https://longitudeprize.org [Accessed 27 December 2021].
33. Crowdmap: https://www.crowdmap.com [Accessed 27 December 2021].
34. OpenStreetMap: https://www.openstreetmap.org/#map=5/54.910/-3.432 [Accessed 27 December 2021].
35. Amazon Mechanical Turk: https://www.mturk.com [Accessed 27 December 2021].
36. Clickworker: https://www.clickworker.com [Accessed 27 December 2021].
37. Minijobz: http://www6.minijobz.com [Accessed 27 December 2021].
38. Haklay M, Dörler D, Heigl F, Manzoni M, Hecker S, Vohland K. What Is Citizen Science? The Challenges of Definition In The Science of Citizen Science eds. Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, Samson R, Wagenknecht K. Springer 2021.https://doi.org/10.1007/978-3-030-58278-4 [Accessed 27 December 2021].
39. eBird Project: https://ebird.org/home [Accessed 27 December 2021].
40. Michelucci P (ed) Handbook of human computation. Springer, New York, pp 561–572.

41. Raddick JM, Bracey G, Gay PL, Lintott CJ, Cardamone C, Murray P, Schawinski K, Szalay AS, Vandenberg J. Galaxy Zoo: Motivations of Citizen Scientists. Astronomy Education Review 2013; arXiv:1303.6886v1.
42. Von Ahn L, Dabbish L. Data generated as a side effect of game play also solves computational problems and trains AI algorithms. Communications of the ACM 2008;51(8):58-67.
43. StallCatchers: https://stallcatchers.com/main [Accessed 27 December 2021].
44. Estellés-Arolas E, González-Ladrón-De-Guevara F. Towards an integrated crowdsourcing definition. J Inf Sci. 2012;38(2):189-200.
45. Geiger D, Seedorf S, Schader M. Managing the crowd: towards a taxonomy of crowdsourcing processes. In: Proceeding of the seventeeth Americas conference on information systems, Detroit, Michigan, 4-7 August 2011.
46. Follet R, Strezov V. An analysis of citizen science based research: usage and publications patterns. Plos One 2015;10(11):e0143687
47. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, Seymour AK, Merchant RM. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med. 2014;29(1):187-203.
48. von Ahn L. Human computation. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering.
49. Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of Crowdsourcing in Health: Systematic Review. J Med Internet Res. 2018;15;20(5):e187.
50. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, Hinder S, Fahy N, Procter R, Shaw S. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res. 2017;1;19(11):e367.
51. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Shemilt I, Thomas J, Glasziou P, Wolfe MS. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). Syst Rev. 2019;20;8(1):57.
52. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 May;133:130-139. doi: 10.1016/j.jclinepi.2021.01.006. Epub 2021 Jan 18. PMID: 33476769.
53. Cochrane Crowd: https://crowd.cochrane.org [Accessed 27 December 2021].

54. Noel-Storr A, Dooley G, Glanville J, Foxlee R. The Embase project: crowdsourcing citation screening. 23rd Cochrane Colloquium, Vienna, Austria; 3rd – 7th October 2015.
55. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane reviews. J Clin Epidemiol. 2021 May;133:140-151.
56. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. Res Synth Methods. 2018;9(4):602-614.
57. Wilczynski NL, Morgan D, Haynes RB; Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak. 2005 Jun 21;5:20. doi: 10.1186/1472-6947-5-20. PMID: 15969765; PMCID: PMC1183213.
58. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020;127:142-150. doi: 10.1016/j.jclinepi.2020.08.008. Epub 2020 Aug 13. PMID: 32798713.
59. The Cochrane Library: https://www.cochranelibrary.com [Accessed 27 December 2021].
60. Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. J Clin Epidemiol. 2021;130:23-31.
61. Noel-Storr AH, Thomas J, McDonald S, Dooley G. 'Screen For Me': harnessing the efficiencies of machine learning and Cochrane Crowd to identify randomized trials for Cochrane reviews. 25th Cochrane Colloquium, Edinburgh, UK; 16th – 18th September 2018.
62. Noel-Storr AH, Redmond P, Lamé G, Liberati E, Kelly S, Miller L, Dooley G, Paterson A, Burt J. Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study. BMC Med Res Methodol. 2021;26;21(1):88.
63. Kelly S, Redmond P, King S, Oliver-Williams C, Lamé G, Liberati E, Kuhn I, Winter C, Draycott T, Dixon-Woods M, Burt J. Training in the use of intrapartum electronic fetal monitoring with cardiotocography: systematic review and meta-analysis. BJOG 2021; https://doi.org/10.1111/1471-0528.16619.
64. Noel-Storr A, Dooley G, Featherstone R, Wisniewski S, Shemilt I, Thomas J, Gartlehner G, Nußbaumer-Steit B, Mavergames C. Crowdsourcing and COVID-19: a case study of Cochrane Crowd. JEAHIL [Internet] 2021;17(2):27-1.

65. World Health Organization: https://www.who.int/health-topics/infodemic#tab=tab_1 [Accessed 27 December 2021].
66. Cochrane COVID-19 Study Register (CCSR): https://covid-19.cochrane.org [Accessed 27 December 2021].
67. Noel-Storr A, Gartlehner G, Dooley G, Persad E, Nussbaumer-Streit B. Crowdsourcing the identification of studies for COVID-19 related Cochrane rapid reviews. Research Synthesis Methods 2022;13(5):585-594.
68. Hamel C, Michaud A, Thuku M, Skidmore B, Stevens A, Nussbaumer-Streit B, Garritty C. Defining rapid reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. J Clin Epidemiol. 2021;129:74-85.
69. Shemilt I, Noel-Storr A, Thomas J, Featherstone R, Mavergames C. Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier. Systematic Reviews 2022;11(1):15.

Chapter 2

An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials

This original manuscript was published in *Journal of Clinical Epidemiology*

Citation: Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 May;133:130-139.

DOI: 10.1016/j.jclinepi.2021.01.006

URL: https://www.jclinepi.com/article/S0895-4356(21)00008-1/fulltext

2.1 Abstract

Background

Filtering the deluge of new research to facilitate evidence synthesis has proven to be unmanageable using current paradigms of search and retrieval. Crowdsourcing, a way of harnessing the collective effort of a 'crowd' of people, has the potential to support evidence synthesis by addressing this information overload created by the exponential growth in primary research outputs. Cochrane Crowd, Cochrane's citizen science platform, offers a range of tasks aimed at identifying studies related to healthcare. Accompanying each task are brief, interactive training modules and agreement algorithms that help ensure accurate collective decision-making. Our objectives were: (1) to evaluate the performance of Cochrane Crowd in terms of its accuracy, capacity and autonomy; and (2) to examine contributor engagement across three tasks aimed at identifying randomised trials.

Study design

Crowd accuracy was evaluated by measuring the sensitivity and specificity of crowd screening decisions on a sample of titles and abstracts, compared with 'quasi gold-standard' decisions about the same records using the conventional methods of dual screening. Crowd capacity, in the form of output volume, was evaluated by measuring the number of records processed by the crowd, compared with baseline. Crowd autonomy, the capability of the crowd to produce accurate collectively-derived decisions without the need for expert resolution, was measured by the proportion of records that needed resolving by an expert.

Results

The Cochrane Crowd community currently has 18,897 contributors from 163 countries. Collectively, the crowd has processed 1,021,227 records, helping to identify 178,437 reports of randomised trials (RCTs) for Cochrane's Central Register of Controlled Trials. The sensitivity for each task was 99.1% for the randomised controlled trial identification task (RCT ID), 99.7% for the randomised controlled trial identification task of trial from ClinicalTrials.gov (CT ID) and 97.7% for identification of randomised controlled trials from the International Clinical Trials Registry Platform (ICTRP ID). The specificity for each task was 99% for RCT ID, 98.6% for CT ID and 99.1% for ICTRP ID. The capacity of the combined crowd and machine learning workflow has increased five-fold in six years, compared with baseline. The proportion of records requiring expert resolution across the tasks ranged from 16.6% to 19.7%.

Conclusion

Cochrane Crowd is sufficiently accurate and scalable to keep pace with the current rate of publication (and registration) of new primary studies. It has also proved to be a popular, efficient and accurate way for a large number of people to play an important voluntary role in health evidence production. Cochrane Crowd is now an established part of Cochrane's effort to manage the deluge of primary research being produced.

2.2 Background

Over the last two decades, published health research output has more than doubled[1,2]. In 2019, just over one million records were added to PubMed, a further 1.4 million unique records to Embase, and approximately 60,000 clinical trials were registered around the world[*]. This equates to an average of 48,000 unique biomedical- and healthcare-related research artefacts published every week. This information deluge is putting health evidence production systems under strain, as systematic reviewers often need to sift through large numbers of records, identified from sensitive searches performed across these and other databases, in search of eligible studies[3]. This bottleneck in the evidence production process can cause delay and contributes to often lengthy production times for systematic reviews and other evidence syntheses such as guidelines and technology assessments; leaving important clinical questions unanswered, and possibly resulting in reliance on out-of-date, and potentially inaccurate, evidence for clinical and policy decision-making[4,5].

Cochrane is an international organisation that produces high-quality systematic reviews about the effectiveness of healthcare interventions[6,7]. In Cochrane systematic reviews alone, we estimate that reviewers assess in excess of four million records annually (based on dual screening) in search of a relatively small number of relevant studies; this also means that large numbers of irrelevant records are being assessed by more than one editorial or review team. We therefore continue to face the major, ongoing challenge of keeping pace with the sheer quantity of information being produced that is potentially relevant for consideration in reviews, whilst also avoiding unnecessary effort and duplication of effort.

It has also been challenging to offer prospective contributors to Cochrane meaningful ways to get involved with producing Cochrane systematic reviews; particularly those with little or no experience of health research[8,9]. Many willing potential contributors are understandably unable, or do not want,

[*] In Ovid MEDLINE: 2019*.ed. = 1041651; In Ovid Embase: 2019*.dc. NOT MEDLINE = 1416448; In ClinicalTrials.gov: First posted from 01/01/2019 to 01/01/2020 = 32524; In ICTRP = Trials added 01/01/2019 to 01/01/2020 = 62738. Deduct 32524 = 30214. Total number: 2520837. For weekly average: 2520837/52=48,478

to take on the full workload and responsibility of authoring a Cochrane review. Yet wider patient and public involvement in health research can bring important benefits to the contributor, to the research process and its outputs, and to the healthcare community at large. This involvement can be at the primary research level, such as helping to design and be involved in a clinical trial, or at the secondary research level, such as evidence synthesis[10,11,12,13].

New approaches are needed to meet these challenges. Specifically, more efficient applications of human effort and better systems for managing information could: (1) significantly reduce current bottlenecks in health evidence synthesis production; and (2) provide people with further opportunities to get involved in the evidence production process. One such approach is crowdsourcing. Other applied fields, such as environmental science and ecology, have successfully incorporated crowdsourcing into their research processes[14,15,16]. Over the last decade a range of crowdsourcing initiatives within healthcare have surfaced[17,18,19], including a number of pilot studies and evaluations focusing specifically on the potential role of crowdsourcing within health evidence synthesis. These studies have largely been exploratory, seeking to test and evaluate different aspects of crowd involvement, including general feasibility[20,21], individual accuracy[22], performance based on different agreement algorithms[23,24], and crowd involvement in other task types beyond study selection[24,25,26].

What is crowdsourcing?

Crowdsourcing is the practice of engaging a large group of people in performing tasks or helping to generate ideas, usually via the internet. There are several different types of crowdsourcing[19]. One commonly used typology[27,28] comprises four main types based on the nature of the 'problem' the host organisation is trying to solve: (i) *peer-vetted creative production* (sometimes termed 'crowd creation') where the organisation tasks the crowd with helping to generate new ideas, solutions or designs; (ii) *broadcast search*, which is a call to find a solution to an empirical (often scientific or technological) problem; (iii) *knowledge discovery and management*, where the crowd is tasked with finding or reporting information, such as gathering data on the use of public spaces; and (iv) *distributed human intelligence tasking*, where the organisation tasks the crowd with analysing or categorising large amounts of information.

Distributed human intelligence tasking is the type most identifiable with the 'wisdom of crowds' concept, because it leverages the collective decision-making abilities of the group over its individual members. Multiple classifications or decisions are required to be submitted by different crowd members, so that an aggregate or collective answer can be reached using an agreement algorithm.

The possible classifications or decisions that can be made must therefore be prospectively well defined. It is this type of crowdsourcing that has been successfully used in many citizen science initiatives that involve processing, filtering or classifying large data sets; and also the type that offers organisations like Cochrane, a new way of tackling the challenges described above.

The Cochrane Crowd platform

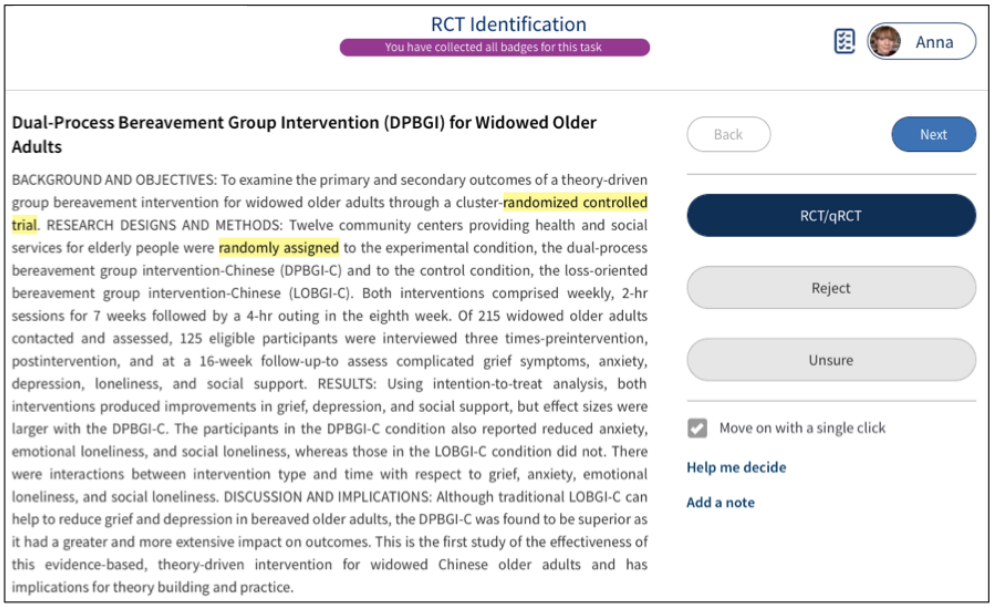
Microtasks

Cochrane Crowd is a web-based application designed to host microtasks. These are small, discrete tasks that require the contributor to perform a classification task, for example, reading a short piece of text and choosing between two (or more) ways that it should be classified (see Figure 2.1 for an example). The focus of this article is on our evaluations of three micro-tasks to identify randomised controlled trials (RCTs) from: bibliographic databases (task name: RCT ID); the U.S. National Library of Medicine's ClinicalTrials.gov clinical trials registry (task name: CT ID); and the World Health Organization's meta-registry of clinical trials, the International Clinical Trials Registry Platform (task name: ICTRP ID).

RCT ID: identifying randomised trials from bibliographic databases

The RCT ID task involves the identification of RCTs and quasi-RCTs from bibliographic sources such as Embase. The definitions of RCT and quasi-RCT are based on the definitions provided in the Cochrane Handbook and the Cochrane Central Register of Controlled Trials (CENTRAL) eligibility record type criteria[29,30].

Figure 2.1 Screen shot of the randomised controlled trials identification (RCT ID) task in Cochrane Crowd



For each record, a contributor must make one of three decisions: *RCT/qRCT*, *Reject*, or *Unsure*, before being able to move on to the next record.

CT ID and ICTRP ID

In September 2017 and September 2018, two new microtasks were launched on Cochrane Crowd. The first, CT ID, aims to identify randomised trials from the world's largest clinical trials registry, ClinicalTrials.gov (www.clinicaltrials.gov). The second, ICTRP ID, focuses on the identification of randomised trials from the World Health Organization's meta-registry of clinical trials, the International Clinical Trials Registry Platform (ICTRP) (http://apps.who.int).

Whilst all three microtasks aim to identify randomised trials, we created a separate task for each source for two reasons. The first was that the record format varies between the sources. RCT ID is based on bibliographic records – such as journal articles and conference publications. For these records we display the titles and abstracts, whereas for the trial registry records, a different set of fields is displayed. The second was that we wanted to create microtasks more suitable for beginners. Microtasks involving categorisation of trial registry records are potentially easier and more rewarding for beginners, because (a) the information in these records is more structured compared with bibliographic records and (b) the prevalence of RCTs that can be correctly identified is higher, hopefully providing a higher level of satisfaction with the task.

The processes and workflows

For each study identification microtask on Cochrane Crowd, a bespoke workflow has been developed to make efficient use of human effort and ensure a steady intake of records from the source databases. These workflows, many of which use a combination of human and machine effort, have been described in detail elsewhere[31,32].

Supporting crowd accuracy: guidance and training

From the outset we wanted to avoid restrictions on who could contribute to Cochrane Crowd. Recognising that people might want to contribute without having much experience with health research, we developed brief training modules for each microtask. The format of the training modules for all the study identification microtasks is the same: between 10 and 20 interactive practice records, selected to reflect the range of records that contributors are likely to encounter in the 'live' task, guide the contributor through the basics of what each specific task is about and how it

should be completed. None of the training modules require a pass mark, so upon completion of these practice records, the contributor can progress straight to assessing 'live' records.

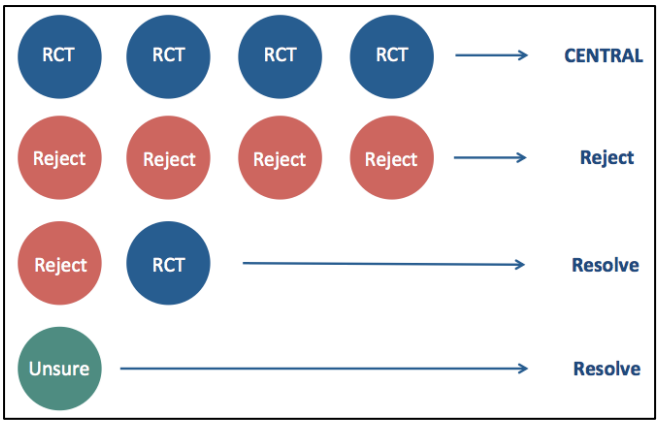
As well as supporting contributors through task-focused training, we recognised the need to enable contributors to track their own progress with each task. Timely, accurate, individualised feedback can be challenging to provide in a live environment where the 'answers' are not yet known. However, it is possible to show each contributor a comparison of *their* decisions against the *final* crowd decisions (based on the task's agreement algorithm - see below); and contributors are encouraged to review their *History* tab and can seek further clarification on final decisions. However, for such feedback to be of value, the agreement algorithm itself has to be robust.

Supporting crowd accuracy: the agreement algorithm

In a crowdsourced model such as ours, an 'agreement algorithm' is used to ensure, at a collective level, that classifications are accurate. All contributors, even experienced screeners can make mistakes. The agreement algorithm is designed to minimise the effects of errors made at an individual level whilst maintaining as much efficiency as possible.

Currently, for the RCT identification microtask in Cochrane Crowd, four consecutive, identical classifications are needed to positively identify a record as an RCT/qRCT (see Figure 2.2), which is then submitted to CENTRAL. If four contributors classify a record as *Reject*, that record will not be submitted to CENTRAL. Classifications by individual contributors are made blinded to any previous classifications. Where classifications disagree, the consecutive chain is broken and the records are automatically sent to be resolved by a subgroup of Crowd contributors known as 'resolvers'. Any *Unsure* classifications are also sent to 'resolvers'. In Cochrane Crowd, contributors can progress from standard contributors, to 'experts', and finally to 'resolvers'. An 'expert' carries the weight of two standard contributors in the decision-making for the task at which they have become an 'expert' (i.e., instead of four classifications needed, only two are needed if both are made by contributors with 'expert' status). To gain expert status, a contributor must have completed 1,000 classifications and achieved 90% or above on both sensitivity and specificity metrics. 'Resolvers' make final classification decisions about records that have either not received the required number of consecutive agreement decisions, or that have been classified as *Unsure*.

Figure 2.2 The Cochrane Crowd agreement algorithm in place for standard screeners



2.3 Methods

Crowd characteristics

We describe the rate of sign-up and the characteristics of the Cochrane Crowd based on information collected from contributors the first time they log-in. This includes information regarding highest educational attainment, age at sign-up, country of residence and level of experience with health research.

Crowd accuracy

We compared the crowd's collective decisions against a gold/reference standard for each of the three microtasks. For RCT ID the evaluation set was a single month of Embase records requiring screening, as described earlier. For CT ID the evaluation set were records screened in the first month after going live with the task, and for ICTRP ID we evaluated the first 5,000 records processed by the crowd. In each of these evaluations, the reference standard data sets were produced by two experts (three different pairs across the three evaluations) who were highly experienced information or data curation specialists with extensive experience of screening, independently classifying the same sets of records as the crowd. For each evaluation, a third screener resolved disagreements between the expert screeners.

In all data sets we counted the number of relevant items identified correctly (the 'true positive' count (TP)); the number of irrelevant items correctly identified as such (the 'true negative' count (TN)); the number of relevant items incorrectly classified as irrelevant (the 'false negative' count (FN)); and the number of irrelevant items, incorrectly classified as relevant (the 'false positive' count (FP)). We then calculated the crowd's collective accuracy in terms of sensitivity (the crowd's ability

to classify relevant records correctly) and specificity (the crowd's ability to exclude irrelevant records correctly) as:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Crowd autonomy

Crowd autonomy (synonymous with crowd consensus) is defined here as the proportion of records that Crowd contributors can process without requiring action by 'resolver' crowd members. The more records that can be dealt with by non-resolvers the better, since resolvers are more experienced members of the crowd, are fewer in number, and are therefore a scarce resource. If a high proportion of records need to be resolved collective accuracy may still be high but the system becomes less autonomous and less efficient, because more time is needed from contributors overall to achieve the same level of output.

Crowd capacity

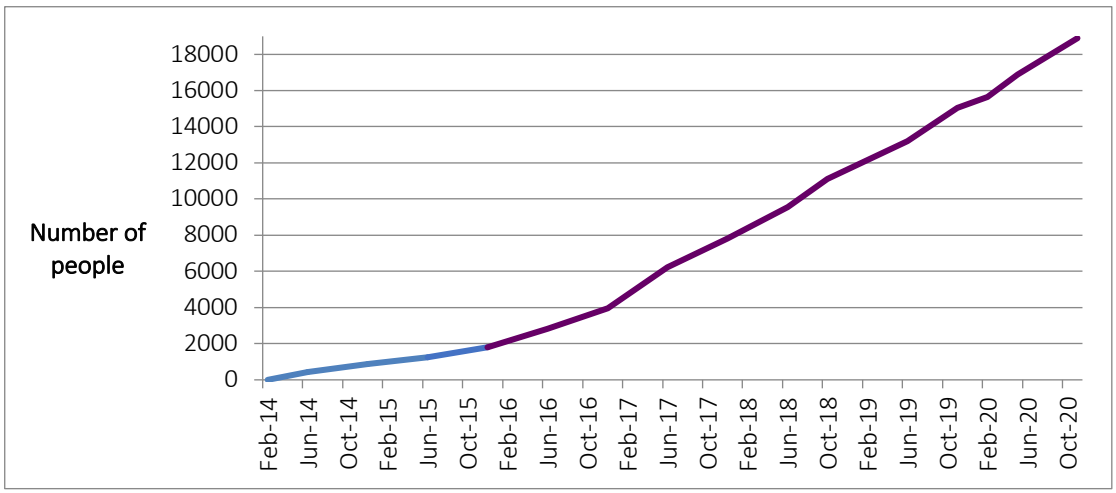
Crowd capacity is defined as the number of records that the crowd workflow can process annually, compared with the baseline. The baseline is the number of records processed by the previous centralised search and screen model. This is an appropriate baseline, as the Cochrane study identification workflow aims to prospectively identify all randomised trials. We compared the number of records handled by the previous method (2010) with the number assessed by crowd alone during the first year of the crowd model being in place (2014), as well as the number assessed by crowd enhanced with machine learning (2020).

2.4 Results

Crowd characteristics

Figure 2.3 shows the steady rate of growth in the number of registered Cochrane Crowd contributors since the platform's launch. Approximately 19,800 people had signed-up to contribute by November 2020, with the average number of active 'sessions' per month (where contributors log in and screen at least one record) being 3,482 since the start of 2020.

Figure 2.3 Cochrane Crowd sign-up with the blue line representing the pilot Embase project phase. Data as of 10th November 2020



Cochrane Crowd contributors are resident in 163 countries, of which 96 are low- and middle-income countries. The top five countries are: United Kingdom (17% of contributors), United States (15%), India (8%), Canada (6%), and Australia (5%). In March 2020 we introduced some optional questions for new contributors regarding educational attainment and experience with health research. Over 2,800 new contributors have completed these questions, providing us with additional insight into our crowd. Whilst many new contributors are already familiar with what a systematic review is, 11% stated that they did not know what a systematic review was and a further 21% only have some sense of what a systematic review was. Twenty seven percent answered that they were completely new to health research. Cochrane Crowd also appears to attract young people with 33% aged between 17-24 years at sign-up. Perhaps unsurprisingly, a large proportion of new contributors are students in a health-related area (42.4%).

Crowd accuracy

Table 2.1 details the results of our evaluation of the accuracy of the crowd across the three study identification microtasks. For the RCT ID evaluation, the data set comprised 6,041 records. The crowd correctly identified 457 RCTs but missed four RCTs, resulting in 99.1% sensitivity. Three missed studies were rejected by the crowd outright (i.e., the records had received the requisite number of consecutive *Reject* classifications). One of the four had gone to resolution but had then been misclassified by the crowd resolver. Of the four missed reports of RCTs, one was an RCT but perhaps confusingly the methods section of the abstract was at the end of the abstract. Another was also clearly an RCT, but at the time we did not have the phrase “random number table” (the randomisation method used in the study) as a highlighted phrase (in Cochrane Crowd we have highlighted over eighty words and phrases to help direct the contributor to the parts of the record

that might describe the study design). The third and fourth missed RCTs were more obvious 'edge cases', in which it was not clear whether the study participants were randomly allocated. Records such as this should be classified as *Unsure* so that the corresponding full-text publication can be checked to see whether random allocation was used. The crowd also correctly rejected 5,522 records out of 5,580 non-RCT records resulting in 99.0% specificity. Among the 58 false positives, several were records in which participants had been randomly selected rather than randomly assigned to groups. Another common error occurred with records that provided an overview of a topic, with a brief mention of a specific randomised trial. Other false positives included five RCTs on animals and one cadaveric study (i.e., records that should be rejected because they do not involve live human participants).

For the other two randomised trial identification tasks, similarly high accuracy was achieved, as shown in Table 2.1. For CT ID, almost all of the 17 false negatives (i.e. relevant records incorrectly classified as irrelevant) contained conflicting information within them. This included records describing the study as a "single-arm" trial in their study design field, but also describing a method of random allocation of participants in their study description field. In ICTRP ID, the majority of the 24 missed RCTs appear to be due to the lack of study design information shown in the record as a result of a display problem. This was due to the API not receiving the study design information for trial registry records from one of the main registries in ICTRP. While the link to the full record with more information was available, contributors were not expected to access this link.

Table 2.1 Accuracy data for the three study identification microtasks

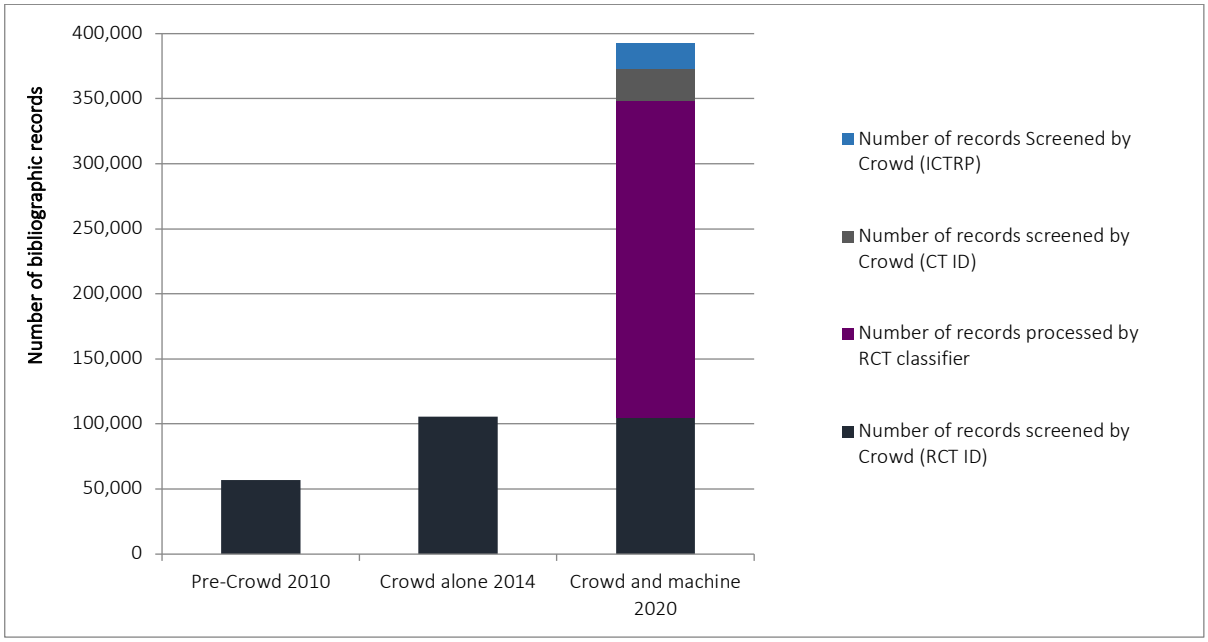
Micro-task	No. of Crowd participants	No. of records (no. of RCTs)	TP	TN	FP	FN	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]	Accuracy (%)
RCT ID	94	6041 (461)	457	5522	58	4	99.1 [97.79 – 99.76]	99.0 [98.66 – 99.21]	99.0
CT ID	179	11,040 (5613)	5596	5350	77	17	99.7 [99.52 – 99.82]	98.58 [98.23 – 98.88]	99.1
ICTRP ID	109	5,000 (1036)	1012	3941	23	24	97.7 [96.57 – 98.51]	99.1 [99.13 – 99.63]	99.1

Crowd autonomy and crowd capacity

An analysis of crowd autonomy, as measured by the proportion of records that need resolving for the three microtasks in Cochrane Crowd show that across each task the proportion of records needing to be resolved is very similar: RCT ID: 16.6%, CT ID: 19.7%, ICTRP ID: 14.9%. Figure 2.4 presents data on Crowd capacity (the number of records that can be processed by the crowd). The 2010 'standard practice' baseline showed that the original centralised search and screen workflow

(staffed by a small team of information specialists) assessed 57,034 records in 2010. During its first year of operation in 2014, Cochrane Crowd assessed 105,747. During 2020 the Cochrane Crowd assessed around the same number of records for the RCT ID task, while the RCT machine learning classifier, calibrated to achieve a recall of 99%, processed a further 243,996 records for this task. The introduction of the RCT Classifier into the workflow in 2016 has significantly increased the number of records that can now be processed. This has freed up the crowd to perform the other two RCT identification micro-tasks available in Cochrane Crowd as well as work on a range of other tasks now available on the platform.

Figure 2.4 Cochrane's capacity for identifying RCTs (2010-2020)



As of November 2020, the 18,900 registered contributors have collectively identified over 175,000 reports of randomised trials for inclusion in CENTRAL. Table 2.2 shows further output metrics for each of the three RCT study identification tasks, including the total number of records screened by the crowd to date and the number of RCTs identified. The relative prevalence of RCTs is indicated in the 'number needed to screen', which is the average number of records that a crowd contributor screens in order to find one relevant record.

Table 2.2 The three study identification microtask metrics. Data accurate as of 10th November 2020

Micro-task	Date task went live	Number of records processed	Number of classifications	Number of RCTs (% of total identified for that micro-task)	Number needed to screen
RCT ID	February 2014	756,916	2,639,800	68,936 (9.1)	11.0
CT ID	September 2017	178,855	507,814	98,269 (54.9)	1.8
ICTRP ID	September 2018	85,456	310,573	11,232 (13.1)	7.6

2.5 Discussion

Cochrane established its crowdsourcing initiative primarily in response to the challenge posed by the rapid increase in global research output. Cochrane Crowd has evolved to become an essential part of Cochrane's ongoing efforts to identify randomised trials for inclusion in its reviews. The crowd now has approximately 18,900 contributors from 163 countries and has collectively processed over 1 million records, helping to identify over 175,000 reports of randomised trials for inclusion in CENTRAL. Each month the platform logs around 3,500 unique sessions from contributors. Our evaluations demonstrate very high levels of accuracy for the three randomised trial identification microtasks, with fewer than 20% of records needing resolution, and a greater than five-fold increase in the number of records processed each year. Cochrane Crowd can now comfortably keep pace with the rate of publication of new studies.

There are several factors that contribute to the success of this crowd model. First, the nature of the tasks themselves plays a key role. Several studies report on the feasibility of using a crowd to assess the search results for systematic reviews[20,21] but do not contain evaluations of accuracy. Those that do report on accuracy measures often report lower accuracy measures[22,23]. However, in contrast to these studies, we are not asking contributors to assess whether a record is relevant to a particular review against all relevant PICO elements – a complex task that typically comprises several judgments relating to different elements of the review's eligibility criteria. Our approach has been to break this complex task down to a simpler binary question: *is this record describing a randomised controlled trial or not?* This makes the task easier to communicate and support with brief, yet targeted training. It also has the advantage of high applicability to the Cochrane use case, given that 90% of published Cochrane reviews use only randomised trial evidence.

Second, and potentially most critical to achieving collective accuracy, is the robustness of the agreement algorithm. This algorithm helps to create an environment where errors made by individuals do not impact on the final decisions. Our current accuracy levels indicate that the crowd misses fewer than one in every hundred trials and incorrectly classifies one in every hundred records submitted to CENTRAL as an RCT. An analysis of records incorrectly classified as trials from the three evaluations showed that common errors included studies where participants had been randomly selected rather than randomly assigned, crossover studies and long-term follow-up studies of RCTs. This is an issue we have now addressed in the support materials for these microtasks. The critical importance of the agreement algorithm has also been shown in other studies, notably in the work by Nama and colleagues who report comparable levels of crowd accuracy in their evaluations[24,26].

Third, the individual contributors that make up the crowd itself clearly play a critically important role; not only in being able to keep up with the constant flow of records fed into the system, but in making accurate individual classifications. Whilst our recruitment is open and, we hope, attracts contributors from a wide variety of backgrounds, it is clear that we appeal largely to those who either work or study in a healthcare-related field. This potentially quite 'expert' crowd implies that even without such a robust agreement algorithm, we could expect higher accuracy than is obtained in other crowdsourcing initiatives. More work is needed to access the impact of prior knowledge and experience on performance measures, as well as the role of the task training and feedback mechanisms on individual accuracy measures over time.

In November 2020 we exceeded 4.5 million classifications. While to our knowledge Cochrane Crowd is the largest crowdsourcing initiative linked to evidence synthesis, several smaller research studies have also evaluated crowdsourcing for study identification[20,22,23,26,33] plus other review production tasks, such as critical appraisal[21,25,34]. These studies all show the potential of crowdsourcing to support these tasks. One notable difference, however, is that Cochrane Crowd is already a fully implemented system that forms part of an important 'end-to-end' process in Cochrane. Whilst crowd accuracy is of critical import, we have also sought to create an efficient, operational workflow that makes the best use of human and machine effort.

Ongoing challenges

Whilst accuracy measures from our evaluations are very high, they are not 100%. As we have shown, false negatives (missed studies) and false positives can arise from consecutive crowd errors as well as from mistakes made by resolver level screeners. In addition, the introduction of machine learning

into the workflow, whilst bringing undoubted gains in the number of records that we are able to handle, has also introduced an interesting challenge for us. With the RCT Classifier now handling a large proportion of the 'easy to reject' records, this has subtly changed the nature of the task itself. In short, the task has potentially become less accessible to beginners. Related to this point, another ongoing challenge is around attracting non-health professionals to contribute. Expanding opportunities for contributors who are new to health research could become increasingly challenging as the machine handles most of the 'easier' records; but on the other hand, new opportunities may arise for those new to health research as the range and content of available crowd tasks continues to grow and diversify.

2.6 Conclusions

To date, the Cochrane Crowd community has classified over 1,021,227 records (756,916 from bibliographic databases and around 264,311 from trial registries). From this, over 175,000 reports of randomised trials have been identified. These reports have been submitted to CENTRAL, helping to enrich that important resource with reports that might not otherwise have been identified.

Identifying reports for CENTRAL or other repositories in this way contributes to the production of Cochrane evidence, but also moves us closer to a more dynamic, upstream model of study identification by identifying accurately *all* reports of RCTs as they are published, indexed or registered so that the evidence for specific reviews can be identified more quickly, with far greater specificity, and without compromising sensitivity.

In addition to populating CENTRAL with reports of randomised trials, this substantial crowd effort has helped to create high-quality data sets for machine learning. Across the current RCT identification tasks the machine classifiers now handle between 50-75% of the records, significantly helping to scale our efforts. This virtuous cycle, where crowd and machine play to their strengths of accuracy and speed respectively, has become the standard model for all future crowd tasks.

We have found that crowdsourcing can be a valuable way of reimagining the research curation work needed to support the timely production and updating of systematic reviews at scale. Cochrane Crowd is now an established and important system within Cochrane's transforming technological landscape. The crowd has proved highly effective, both in terms of accuracy and efficiency, when provided with small tasks supported by brief training and robust agreement algorithms. In short, Cochrane Crowd is transforming the way we identify and curate health evidence; helping us to keep

up with the information overload whilst at the same time offer willing contributors a way to get involved and play a crucial role in health evidence production.

2.7 Author contributions

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Gordon Dooley: conceptualisation, methodology, resources, data curation, writing – reviewing and editing

Julian Elliott: conceptualisation, methodology, writing – reviewing and editing

Emily Steele: conceptualisation, writing – reviewing and editing

Ian Shemilt: data curation, writing – reviewing and editing

Chris Mavergames: conceptualisation, writing – reviewing and editing

Susanna Wisniewski: conceptualisation, data curation, writing – reviewing and editing

Steve McDonald: conceptualisation, writing – reviewing and editing

Melissa Murano: conceptualisation, writing – reviewing and editing

Julie Glanville: conceptualisation, writing – reviewing and editing

Ruth Foxlee: conceptualisation, writing – reviewing and editing

Deirdre Beecher: data curation, writing – reviewing and editing

Jennifer Ware: data curation, writing – reviewing and editing

James Thomas: conceptualisation, methodology, writing - reviewing and editing

2.8 Abbreviations

RCT – randomised controlled trial

CENTRAL – Cochrane Central Register of Controlled Trials

2.9 References

1. Van Noorden R. Global scientific output doubles every nine years. Nature News Blog 2014: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html# [Accessed 28 December 2021].
2. Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 2015;66:2215-2222.
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545.
4. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007; 147: 224-233
5. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, Gruen RL. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. PLoS Med 2014;11(2):e1001603.
6. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. Evaluation & the health professionals 2002;25(1):12-37.
7. Bero L, Rennie D. The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. Journal of the American Medical Association 1995;274:1935-1938.
8. Morley RF, Norman G, Golder S et al. A systematic scoping review of the evidence for consumer involvement in organisations undertaking systematic reviews: focus on Cochrane. Res Involv Engagem 2016;2:36.
9. Pollock A, Campbell P, Struthers C, et al. Stakeholder involvement in systematic reviews: a scoping review. Syst Rev. 2018;7:208.
10. Pollock A, Campbell P, Struthers C, Synnot A, Nunn J, Hill S, Goodare H, Morris J, Watts C, Morley R. Development of the ACTIVE framework to describe stakeholder involvement in systematic reviews show less. J Health Serv Res Policy. 2019: https://doi.org/10.1177/1355819619841647.
11. Kreis J, Puhan MA, Schunemann HJ, et al. Consumer involvement in systematic reviews of comparative effectiveness research. Health Expect 2013;16:323-337.
12. Brett J, Staniszewska S, Mockford C, et al. A systematic review of the impact of patient and public involvement on service users, researchers and communities. Patient 2014;7:387-395
13. INVOLVE. https://www.involve.org.uk [Accessed 28 December 2021].

14. Muller CL, Chapman L, Johnston S, Kidd C, Illingworth S, Foody G, Overeem A, Leigh RR. Crowdsourcing for climate and atmospheric sciences: current status and future potential. International Journal of Climatology 2015;35:3185-3203.
15. Zhao Y, Zhu Q. Evaluation on crowdsourcing research: current status and future direction. Information Systems Frontier 2014;16:417-434.
16. Von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: human-based character recognition via web security measures. Science 2008;321:1465-1468.
17. Tucker JD, Day S, Tang W, Bayus B. Crowdsourcing in medical research: concepts and applications. PeerJ. 2019;7:e6762. doi: 10.7717/peerj.6762.
18. Wang C, Han L, Stein G, Day S, Bien-Gund C, Mathews A, Ong JJ, Zhao PZ, Wei SF, Walker J, Chou R, Lee A, Chen A, Bayus B, Tucker JD. Crowdsourcing in health and medical research: a systematic review. Infect Dis Poverty 2020;9(1):8. doi: 10.1186/s40249-020-0622-9.
19. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, Seymour AK, Merchant RM. Crowdsourcing – harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med. 2013;29(1):187-203.
20. Brown AW, Allison DB. Using crowdsourcing to evaluate published scientific literature: methods and example. PLoS One 2014;9(7):e100647.
21. Bujold M, Granikov V, Sherif RE, Pluye P. Crowdsourcing a mixed systematic review on a complex topic and a heterogeneous population: Lessons learned. Educ. Inf. 2018;34:293-300.
22. Ng L, Pitt V, Huckvale K, Clavisi O, Turner T, Gruen R, Elliott JH. Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): a pilot randomised controlled trial of title and abstract screening by medical students. Syst Rev. 2014;3(121).
23. Mortensen JM, Adam GP, Trikalinos TA, Kraska T, Wallace BC. Res Synth Methods. Research Synthesis Methods 2016;8(3):366-386.
24. Nama N, Barrowman N, O'Hearn K, Sampson M, Zemek R, McNally JD. Quality control for crowdsourcing citation screening: the importance of assessment number and qualification set size. J Clin Epidemiol. 2020 Jun;122:160-162. doi: 10.1016/j.jclinepi.2020.02.009.
25. Pianta MJ, Makrai E, Verspoor KM, Cohn TA, Downie LE. Crowdsourcing critical appraisal of research evidence (CrowdCARE) was found to be a valid approach to assessing clinical research quality. J Clin Epidemiol. 2018;104:8-14.
26. Nama N, Sampson M, Barrowman N, et al. Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study. J Med Internet Res. 2019;21(4):e12953.
27. Brabham DC. Crowdsourcing. The MIT Press, Cambridge, Massachusetts 2008

28. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med 2014;46(2):179-187.
29. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook
30. Should I publish this record to CENTRAL? https://community.cochrane.org/sites/default/files/uploads/Should%20I%20publish%20this%20record%20to%20CENTRAL.pdf [Accessed 28 December 2021].
31. Noel-Storr AH, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd finds that crowdsourcing produces accurate results in identifying randomised trials. Journal of Clinical Epidemiology 2020;130:23-31.
32. Thomas J, McDonald S; Noel-Storr AH, Shemilt I, Elliott J, Mavergames C, Marshall I. Machine learning reduces workload with minimal risk of missing studies: development and evaluation of an RCT Classifier for Cochrane reviews. J Clin Epidemiol. 2021 May;133:140-151. doi: 10.1016/j.jclinepi.2020.11.003.
33. Krivosheev E, Casati F, Benatallah B. Crowd-based multi-predicate screening of papers in literature reviews. WWW 2018: The 2018 Web Conference, April 23-27, 2018, Lyon, France.
34. Ashkanase J, Nama N, Sandarage RV, et al. Identification and Evaluation of Controlled Trials in Pediatric Cardiology: Crowdsourced Scoping Review and Creation of Accessible Searchable Database [published online ahead of print, 2020 Feb 15]. Can J Cardiol. 2020;S0828-282X(20)30174-4.

Chapter 3

Machine learning reduces workload with minimal risk of missing studies: development and evaluation of an RCT classifier for Cochrane reviews

This original manuscript was published in *Journal of Clinical Epidemiology*

Citation: Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane reviews. J Clin Epidemiol. 2021 May;133:140-151.

DOI: doi: 10.1016/j.jclinepi.2020.11.003

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8168828/

3.1 Abstract

Background

To describe the development, calibration and evaluation of a machine learning classifier designed to reduce study identification workload in Cochrane for producing systematic reviews.

Methods

A machine learning classifier for retrieving RCTs was developed (the 'Cochrane RCT Classifier'), with the algorithm trained using a data set of title-abstract records from Embase, manually labelled by the Cochrane Crowd. The classifier was then calibrated using a further data set of similar records manually labelled by the Clinical Hedges team, aiming for 99% recall. Finally, the recall of the calibrated classifier was evaluated using records of RCTs included in Cochrane reviews that had abstracts of sufficient length to allow machine classification.

Results

The Cochrane RCT Classifier was trained using 280,620 records (20,454 of which reported RCTs). A classification threshold was set using 49,025 calibration records (1,587 of which reported RCTs) and our bootstrap validation found the classifier had recall of 0.99 (95% CI 0.98 to 0.99) and precision of 0.08 (95% CI 0.06 to 0.12) in this data set. The final, calibrated RCT classifier correctly retrieved 43,783 (99.5%) of 44,007 RCTs included in Cochrane reviews but missed 224 (0.5%). Older records were more likely to be missed than those more recently published.

Conclusions

The Cochrane RCT Classifier can reduce manual study identification workload for Cochrane reviews, with a very low and acceptable risk of missing eligible RCTs. This classifier now forms part of the Evidence Pipeline, an integrated workflow deployed within Cochrane to help improve the efficiency of the study identification processes that support systematic review production.

3.2 Background

Cochrane is a leading producer of systematic reviews, with more than 8,000 currently published in the Cochrane Library[1]. These reviews incorporate the results of tens of thousands of randomised controlled trials (RCTs) and other primary studies. The manual effort invested in identifying primary studies eligible for inclusion in these and other systematic reviews is vast. Author teams and information specialists typically search a large number of bibliographic databases to find the comparatively small number of studies eligible to be included[2]. These searches are sensitive, to

identify as many relevant studies as possible, but therefore yield large numbers of irrelevant records which are then screened manually by author teams. This is a time-consuming and therefore costly process, especially when all records are checked by at least two people to aid reliability. With the rapidly increasing volume of research being conducted and published[3], systematic reviews tend to be resource-intensive projects, which can take years to complete. As a consequence, many important research questions are not covered by systematic reviews, and it is increasingly difficult to maintain an up-to-date synthesised evidence base[4]. This is a waste of global investment in research, leading to suboptimal decision-making and poorer health outcomes[5].

Automation has been proposed as one possible solution to reduce the manual burden of many systematic review tasks[6]. For example, machine learning classification algorithms ('classifiers') can 'learn' the eligibility criteria of a review through exposure to a manually classified set of documents, thus reducing the human effort required to find relevant studies[7].

To date, most automation approaches operate at the level of individual reviews[8,9], rather than addressing structural deficiencies in research curation[10]. This paper describes an important component in an alternative approach which aims to improve the efficiency of study identification across multiple systematic reviews of RCTs. The system comprises: 1) database searching; 2) machine learning; and 3) crowdsourcing (via the Cochrane Crowd citizen science project) to populate an existing database of RCTs (CENTRAL)[11]. The interlinked system or 'workflow' is known as the Cochrane 'Evidence Pipeline'. Here we describe the machine learning component of the Pipeline workflow; the other components (the Cochrane Crowd and a Centralised Search Service) are detailed elsewhere[12,13] (chapters one and three of this thesis). The reason that this is so beneficial for Cochrane reviews is twofold. First, on the basis that RCT study designs can be ethically implemented to produce results capable of supporting causal claims about the beneficial effects of the large majority of healthcare interventions evaluated in Cochrane reviews, approximately 90% of Cochrane reviews aim to include only RCTs. Thus, the capability to efficiently identify studies with designs at scale will generate large corollary cost savings and efficiency gains in review production and updating systems across thousands of Cochrane reviews, reducing research waste. Second, searches conducted for Cochrane and non-Cochrane health and non-health systematic reviews of RCT evidence also retrieve many records of studies that are not RCTs (often well over 50%). Thus, the capability to automatically exclude non-RCTs from manual checking in such reviews will reduce manual workload (since, even if they are about the right topic, the fact that they are not RCTs means that they are ineligible for inclusion), with corollary cost savings and efficiency gains.

We have previously described methods for automatically identifying RCTs from research databases[8]. In that evaluation we found machine learning classification systems are more accurate than manually crafted Boolean string searches of databases (the current standard practice). Yet showing higher accuracy in a validation study is not sufficient to ensure new technologies are adopted in practice. We have engaged with the Cochrane Information Retrieval Methods Group (IRMG) with whom we agreed additional requirements for this technology to be adopted by Cochrane. First, the classifier must recall at least 99% of RCTs (a more stringent threshold than we had applied in our previous work)[8]. Second, the classifier should provide an indicative probability score to users. Third, an additional assessment should be done of whether the classifier would be at risk of missing any of the studies included in existing Cochrane reviews. In this article, we describe the development, calibration and evaluation of a machine learning classifier designed to meet these requirements, which has subsequently been adopted by and deployed within Cochrane.

3.3 Methods

Cochrane Evidence Pipeline workflow

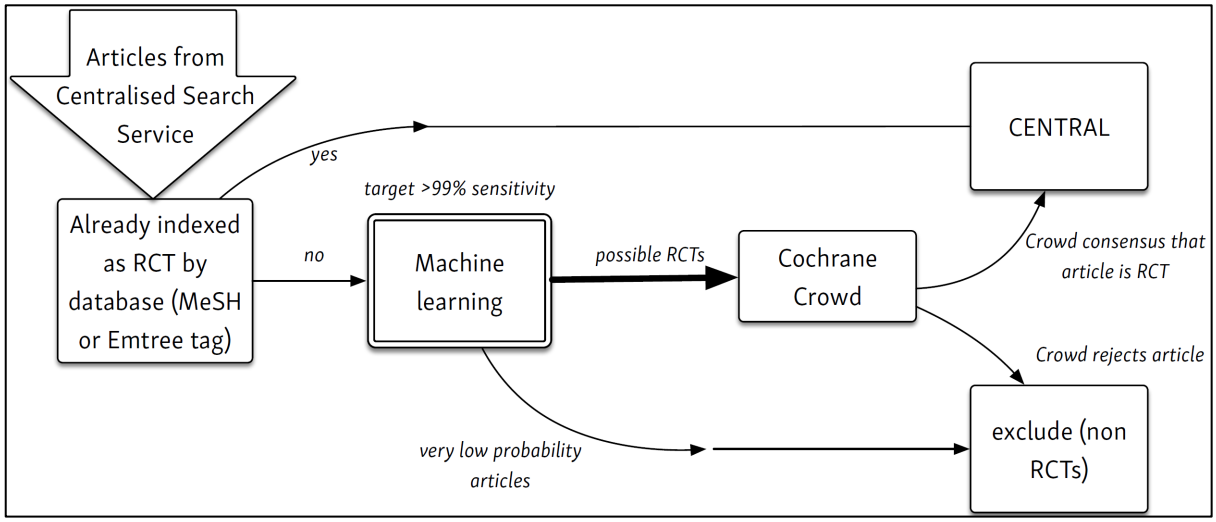
Cochrane publishes a database of randomised controlled trials (RCTs) that are relevant for current or potential future reviews (CENTRAL), with an administrative interface for Cochrane users, known as the Cochrane Register of Studies (CRS)[11]. Although a rapidly increasing minority of reviews synthesise non-randomised research designs (including qualitative and quasi-experimental studies), CENTRAL focuses on RCTs which currently remain the basis of the large majority of published Cochrane reviews. We likewise focus our efforts on the discovery of RCTs.

We seek to benefit from efficiencies in two ways. First, current practice is to identify RCTs through searches of bibliographic databases using highly sensitive RCT filters. Such filters have *low precision*, retrieving as many as 20 non-RCTs for every true RCT[12]. These irrelevant articles then need to be manually screened and removed. Second, the same studies are retrieved and assessed multiple times by different people across the global systematic review workforce. The Pipeline therefore aims to avoid this duplication of effort, by facilitating the reuse of previous assessments as to whether a given report describes, or does not describe, an RCT.

Figure 3.1 depicts the role of machine learning within the Pipeline. To populate CENTRAL, Cochrane regularly searches a range of online resources (e.g., biomedical literature databases) through the 'Centralised Search Service', which is described elsewhere[12] (Chapter 4). Abstracts of these

candidate articles (of which the majority are not RCTs) are 'fed' into the top of the Pipeline.* The machine learning classifier (described in this work) is used to filter out records that are highly unlikely to be an RCT study report. The remaining articles are then handed over to the Cochrane Crowd, which filters out all further records that do not report an RCT[13]. Finally, the remaining articles (which should all describe RCTs) are stored in CENTRAL. Crowd 'labels' are also used to update the machine learning algorithm, so that it becomes more accurate at distinguishing between relevant and irrelevant records (see 'Machine learning for RCT identification', below).

Figure 3.1 The Cochrane Evidence Pipeline workflow, depicting the flow of records from the centralised search service, through machine and crowd classification services to the CENTRAL database



Data sets and their role in this study

High-quality data sets are vital for the development, calibration and evaluation of reliable machine learning classifiers. Most evaluations of such classifiers utilise a single data set, which is split at random between 'training' and 'test' data (for example, with 70% of the data reserved for training). The training data are used to estimate the model parameters, and the test data to evaluate its performance. However, while a single data set evaluation can provide estimates of classifier performance that have strong internal validity, it cannot tell us how well a classifier will perform in the real world, where data may come from sources that differ in important ways from those used to produce this data set. As outlined by Nevin, it is important to consider the external validity of machine learning models before deployment[14]. Here, we examined external validity in terms of whether the use of our machine learning model would risk missing RCTs included in Cochrane reviews.

* The scope and detail of these searches is described here: https://www.cochranelibrary.com/central/central-creation.

We therefore utilised three distinct data sets in the current study:

- *Training* data, from which the machine learning models are built;
- *Calibration* data, on which the threshold for determining the cut-off between 'RCT' and 'non-RCT' classifications was based; and
- *Validation* data, on which the calibrated classifier was evaluated.

Figure 3.2 summarises the contribution made by each data set that we now describe in detail.

Training data

The data set used to train the classifier comprises a corpus of 280,620 title-abstract records retrieved from Embase using a highly sensitive search for RCTs[†]. This search has been carried out each month since January 2014, for the purpose of identifying relevant studies for inclusion in CENTRAL (see 'Materials and Methods'). In this study we used records retrieved between January 2014 and July 2016 inclusive. During this period, any records indexed with the Emtree headings 'Randomized controlled trial' or 'Controlled clinical study' were automatically marked for inclusion in CENTRAL, without any manual checking, on the basis that this rule produced a false positive rate for identifying reports of RCTs that was judged sufficiently low. To account for the historical use of this rule, records with these specific Emtree headings were also excluded from our training data set. Because obvious RCTs and obvious non-RCTs had already been filtered out of this data set (using Emtree headings and the sensitive search filter for RCTs respectively) before we used it to train the classifier, the data set therefore comprises records that are, on average, more difficult to classify according to whether or not they report an RCT, compared with an unfiltered sample from the raw database.

Next, each record in the training data set was labelled by Cochrane Crowd members according to whether it reported an RCT (n = 20,454) or not (n = 260,166). Each record was labelled by multiple crowd members, with the final crowd decision being determined by an agreement algorithm; Noel-Storr and colleagues report that the crowd recall and precision for identifying RCTs both exceeded 99%[13].

This data set ('Cochrane Crowd data' in Figure 3.2) has characteristics that make it highly suitable for training a machine learning classifier: it is both large – so represents a wide range of instances of both the positive and negative classes (i.e., RCTs and non-RCTs) – and also very accurately labelled.

[†] https://www.cochranelibrary.com/central/central-creation

However, it is also comprised of records added to Embase during a relatively short (<3 years) period, which could limit the generalisability of the resulting machine learning classifier.

Calibration data

When machine learning classifiers are used for prediction, they output a score (often scaled to be bounded by 0 and 1) that gets assigned to each title-abstract record, with a higher value representing an increased likelihood that the record reports an RCT. However, to use the classifier to reduce manual screening workload, we also needed to set a threshold score below which records (unlikely to be RCTs) are discarded, and conversely above which records (possible RCTs) are retained for manual screening. Higher score thresholds can be expected to lead to a higher prevalence of reports of RCTs (true positives) among fewer retained records (i.e., higher precision), but at the expense of having discarded some reports of RCTs (false negatives) with scores below the threshold (i.e., lower recall). Conversely, a lower threshold can be expected to lead to lower precision but higher recall.

We sought advice from the Cochrane Information Retrieval Methods Group (IRMG) and were advised that the classifier would need to have a threshold score calibrated to retrieve at least 99% of relevant RCT study reports in order to be adopted for use in Cochrane; and also that achieving this high level of recall should be prioritised over any reductions in manual screening workload. These specifications reflect the strong aversion that we have, when conducting systematic reviews, to inadvertently failing to identify studies that should be included.

The Clinical Hedges data set (Figure 3.2) was built during 2000 and 2001 for the purposes of testing and validating sensitive search filters for RCTs[15]. It contains 49,028 title-abstract records manually identified and selected by information specialists using a combination of hand- and electronic search methods. Corresponding full-text reports of all records were manually checked in order to ascertain with confidence whether or not each reported an RCT, making this a highly accurate data set for our current purpose. Three records from this data set were no longer available in PubMed, so our final calibration data set comprised 49,025 PubMed title-abstract records, of which 1,587 reported an RCT (and the remaining 47,438 did not report an RCT).

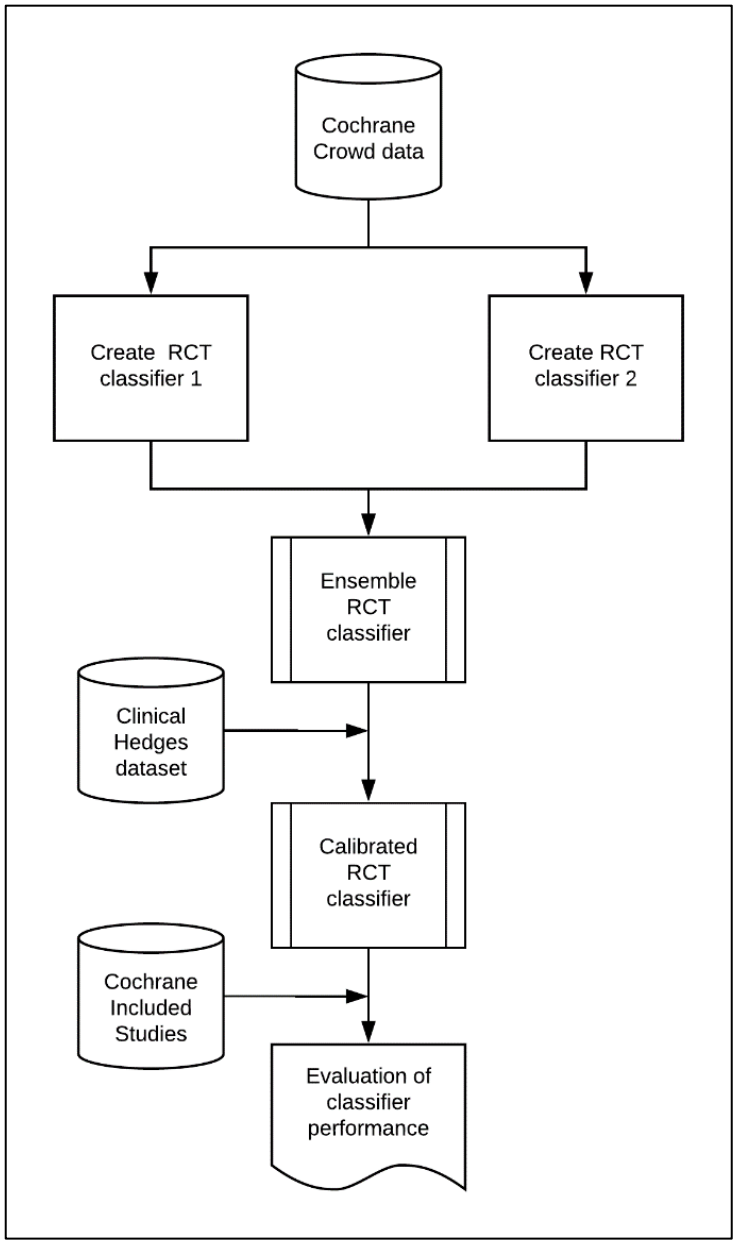
It was more demanding to calibrate our RCT classifier on this data set (compared with using a proportion of records held back from the Cochrane Crowd data set) because: a) the records are older and are less likely to have a consistent reporting structure for RCTs, because the study reports

were published only a few years after the CONSORT statement (and before the latter became widely used)[16]; and b) the Clinical Hedges Team's assessments were based on full-text reports, and there is no indication in some of the corresponding titles and abstracts that they actually report an RCT. We used this data set to identify the threshold for achieving 99% recall and thereby calibrate our classifier, and we also present results concerning the precision with which these records can be identified.

Validation data

As described above, this machine learning classifier was primarily designed to identify records of study reports potentially eligible for inclusion in Cochrane reviews. We therefore validated the classifier using a third data set, to determine whether the desired level of 99% recall (calibrated using the Clinical Hedges data set), could be achieved in practice. This validation data set ('Cochrane Included Studies' in Figure 3.2) comprises title and abstract records of all study reports included in Cochrane reviews in which eligible study designs are restricted to 'RCTs only', published up to April 2017. The data set comprises 94,305 records of 58,283 included studies, across 4,296 Cochrane reviews. Although it could be assumed that the vast majority of these records report an RCT, in practice we found that some records of included study reports did not report an RCT (for example, they reported a meta-analysis of RCTs; a related editorial; or personal correspondence). These records were retained in the validation set, as removing them all would have required the manual screening of all records.

Figure 3.2 Development and evaluation of the classifier, showing where the various data sets were used in the classifier development process



Excluded data

Articles without an abstract (i.e., title-only records) may contain insufficient information for accurate machine (or human) classification. However, title-only records that include the words 'randomised controlled trial' (as per CONSORT guidance) should be labelled correctly by a classifier. In consultation with the IRMG, and in the light of manual assessment of records with some content in their abstract field (but not a full abstract), we determined that pragmatic cut-offs for including a record in the training, calibration, or validation data sets would be set at: 400 characters as a minimum abstract length; and 15 characters as a minimum title length. These cut-offs aimed to

balance the need for sufficient text to be present in the abstract field for the machine learning to operate, while not referring too many records for manual assessment. It is important to note that, in the Cochrane Evidence Pipeline workflow (see above), all records with title and/or abstract fields that have fewer characters than the minimum cut-off are referred for manual screening by members of the Cochrane Crowd. When the minimum character cut-off is applied to the Cochrane Included Studies data set, the final number of studies in the evaluation falls to 44,007.

Machine learning methods for RCT identification

Machine learning describes a group of algorithms which seek to 'learn' to do a task by exposure to (typically large amounts of) data. The approach we used here can be described as *supervised machine learning:* meaning that the algorithm is 'trained' on articles for which the true label is already known. Although the current state-of-the-art approach for text classification is the use of neural network models, we have previously found that support vector machine (SVM) models (and specifically *ensembles*[‡] of multiple of SVM models) were similarly accurate for high sensitivity document recall[8]. SVMs are less computationally intensive than neural models, and therefore can run quickly and without the need for any special computer hardware. The final Cochrane RCT Classifier model also needed to be deployed in a live web service that might need to cope with heavy user demand. For these reasons we chose SVMs for the current work. We refer the interested reader to a detailed description of machine learning methods as applied to abstract classification[8]. In our previous work, we incorporated metadata from the database describing study design into our models (the Publication Type tag in MEDLINE, which is manually added by MEDLINE staff, often several months after publication). However, as the Evidence Pipeline retrieves mainly very new records, which usually lack this metadata, we used a model which utilises titles and abstract text without additional metadata.

We used the *bag-of-words* approach, in which each title-abstract record is represented as a vector of 0s and 1s, depending on the presence or absence of each unique word from the article-set vocabulary[8]. These vector representations are then used to 'train' (i.e., find optimal parameters for) an SVM model. We pre-processed the records to remove commonly used words (e.g., 'and', 'the') that appear on the PubMed "stopword" list[8]. During our initial development phase, we found that an ensemble of two SVM models with minor differences resulted in greatest accuracy when evaluated on the training data. We therefore selected an ensemble of two SVM variants for use in the study

[‡] Ensembling describes a strategy of using multiple machine learning models together, with the aim of improving performance compared with any model individually.

(see Figure 3.2). The first classifier (SVM1) represented the texts as individual words, pairs of words, and triplets of words *(uni-, bi-, and trigrams)*. This accounts for situations in which adjacent words affect document class (e.g. the text 'randomized controlled trial' might be more strongly indicative of an RCT than any word individually). The second classifier (SVM2) used a *unigram* model (i.e. each word is considered individually), and used a strategy of *oversampling*. This strategy aims to reduce the likelihood of missing a 'rare class' (here the 'rare class' is RCTs, which account for ~5% of the data set) by artificially increasing the number of RCTs in the training data set by random sampling with replacement (this process is not repeated with the calibration or validation data sets, which are left in their original state).

The source code for building SVM1 is available here: https://github.com/alan-turing-institute/DSSG19-Cochrane/blob/dev/analyses/partner_baseline/create_model.py
The source code for building SVM2 is available here: https://github.com/ijmarshall/robotsearch

Generating calibrated probability estimates

SVMs estimate the distance between a given record and a 'hyperplane'[17] which, in the current use, separates RCTs (the positive class) from non-RCTs (the negative class). The hyperplane distance metric is not readily interpretable (in our data set this metric had a numeric value approximately between –1 and +8), and we therefore sought to add probability estimates to meet the needs of Cochrane users, who have found this feature to be particularly useful in understanding the output. To achieve this, we calibrated the ensemble SVM scores on the Clinical Hedges data set using a logistic regression model (known as Platt scaling)[18]. This generated a score for each 'unseen' record in the calibration or validation data sets that is bounded by 0 and 1. These scores are closer to representing the true probability that a given record reports an RCT; as such they are readily interpretable, with higher scores representing a higher likelihood that the record reports an RCT (and vice-versa). When viewed graphically, the distribution of scores is often U-shaped, with the majority of records being assigned either a high (close to 1) or low (close to 0) probability score (see Figure 3.4), and a smaller number of records in the middle that are more ambiguous in terms of class membership.

Evaluation metrics

In this paper we use the conventional information retrieval terminology *recall* and *precision*, which are synonymous with sensitivity and positive predictive value respectively. As outlined above, the recall statistic is of primary concern in the current use scenario – i.e., that eligible study reports are

not incorrectly discarded from the Evidence Pipeline workflow. Cochrane required the system to have at least 99% recall. Recall is calculated as the proportion of relevant items (i.e., records describing an RCT) that are correctly identified by the Evidence Pipeline workflow compared with the total number of records genuinely reporting an RCT that should have been identified.

To evaluate the discriminative performance and quality of calibration of our machine learning strategy on the Clinical Hedges data, we used bootstrap sampling as described by Steyerberg and colleagues (19). In short, a series of artificial new data sets were 'bootstrapped' by random sampling with replacement from the Clinical Hedges data set. Logistic regression models (which served the dual purposes of ensembling the individual SVM models, and producing calibrated probability outputs) were trained on each sampled data set, and evaluated on the original data set. This process was repeated 5000 times and used to estimate performance metrics with 95% confidence intervals. Although the primary use of the system is for binary classification, a key secondary use is providing indicative probability scores to users. We evaluate the quality of the probabilities via a calibration plot, and by calculation of the Brier score and C statistics[20].

Common practice in Cochrane reviews is to find, use and cite all published (and unpublished) reports of each included study ('study reports'). Many studies included in Cochrane reviews are comprised of multiple study reports. This means that if the classifier 'misses' one of several study reports of the same RCT, this does not necessarily mean the RCT study has been 'lost'. We therefore adopted the following approach. We first classified all study reports in Cochrane reviews of RCTs using the machine learning classifier and then we considered a study to be 'lost' only if *all* reports of that study fell below the threshold. As such, the 'study' is our unit of analysis rather than the 'study report'. We made this decision since we found many secondary citations in reviews referred to indirectly related non-RCT studies, and also since we would expect the retrieval of a single article would alert the review team to the existence of the trial.

Precision is also a metric of interest because it can be used to compute the number of articles requiring manual screening by Cochrane Crowd in the Evidence Pipeline workflow. Here we were concerned with the number of irrelevant records (i.e., records not reporting an RCT) that are incorrectly classified by machine learning as relevant (i.e., records with an assigned probability score above the identified threshold score), which must then be filtered out manually by the Cochrane Crowd. Precision is calculated as the proportion of retrieved records which genuinely report an RCT.

Recall and precision were calculated from a 2 x 2 table representing positive/ negative (relevant/ irrelevant) classes and whether they were correctly or incorrectly classified (Table 3.1).

Table 3.1 2x2 table from which precision and recall are calculated

	RCTs (gold standard)	non-RCTs (gold standard)
Machine learning classed RCTs	True positives	False positives
Machine learning classed non-RCTs	False negatives	True negatives

Formulae used to calculate precision and recall are as follows:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

We computed statistics for precision at 99% recall against the Clinical Hedges calibration data set. As specified above, recall was set at 99% by the IRMG. Since the Cochrane reviews we examined contain only RCTs (and the non-RCTs excluded during searches are not usually comprehensively recorded) we were not able to calculate precision on the Cochrane reviews data set, and report recall only.

For the primary analysis, the denominator was all articles in Cochrane reviews meeting the minimum character length criteria described above (i.e., very short titles and abstracts were excluded). We assume that manual assessment will yield 100% recall of these records. We also report results on the full data set, without removing articles with small or non-existent abstracts, as a secondary analysis. The first figure can be interpreted as the recall of the overall workflow, because it takes account of our decision to remove records with insufficient information for machine classification from the workflow.

We present absolute values of the total number of eligible studies 'lost' to Cochrane reviews. Finally, we also present the distribution of 'lost' study reports according to the year of publication, since we hypothesise that the classifier may perform less well on older study reports because: a) it has been trained on newer reports; and b) trial reporting may have improved as a result of the CONSORT statement.

3.4 Results

The machine learning classifier for identifying reports of randomised trials (Cochrane RCT Classifier) was built as per the above methods from the screening of 280,620 Embase records (January 2014 to July 2016) by Cochrane Crowd. Of these, 20,454 (7.3%) were deemed to be RCTs.

Threshold setting, and binary classification performance

The 49,025 records from the Clinical Hedges data set were scored by the machine learning classifier. The records were ordered according to classifier score, and precision and recall statistics were calculated for every record in sequence. The classifier probability, which corresponded with 99% recall, was recorded and used as the classification threshold for the later validation (and the deployed system). The discriminative and calibrative performance of this strategy, estimated using bootstrap sampling is presented in Table 3.2. We estimate that precision was 8.3%, meaning that one in every 12 records retrieved described an RCT. Setting the classifier at this level of recall resulted in 58% of records in this data set being automatically discarded as highly unlikely to be reporting a randomised trial.

Estimates of the C statistic and Brier score were 0.978 and 0.048 respectively, indicating excellent discriminative performance. We present a calibration plot showing point estimates from the bootstrap evaluation and the final model (trained on the whole data set) in Figure 3.3. We show how the predicted scores are distributed for RCTs and non-RCTs in Figure 3.4.

Figure 3.3 Calibration plot showing bootstrap estimates of predicted *vs* observed probabilities of an article being an RCT in Clinical Hedges data set (each blue point represents an estimate of a model generated from one bootstrap sample), and the performance of the final model (orange)

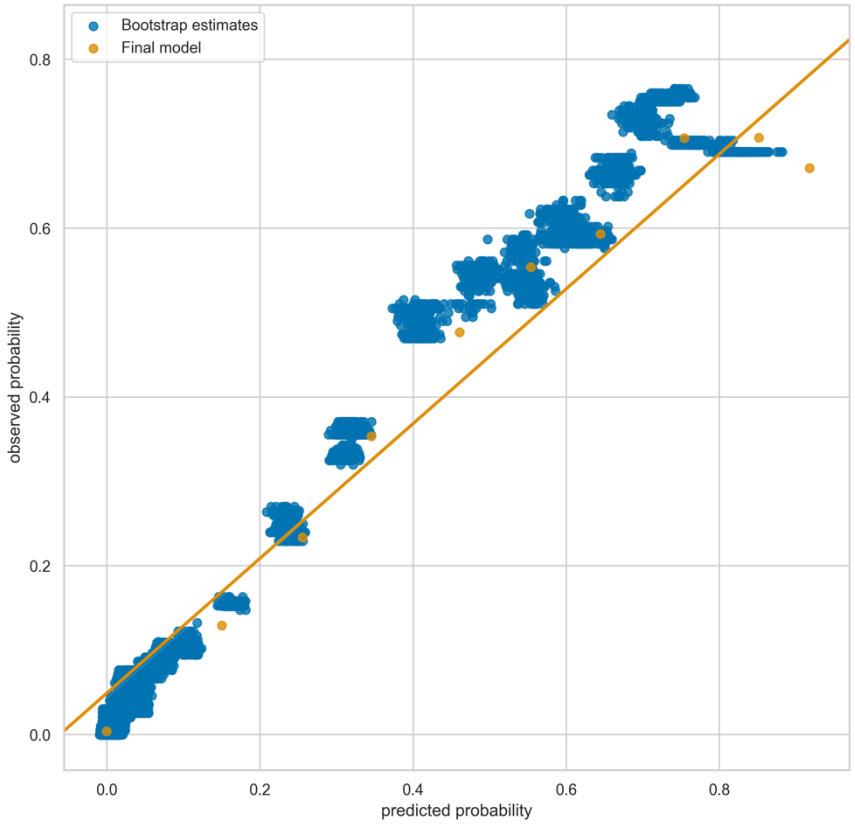
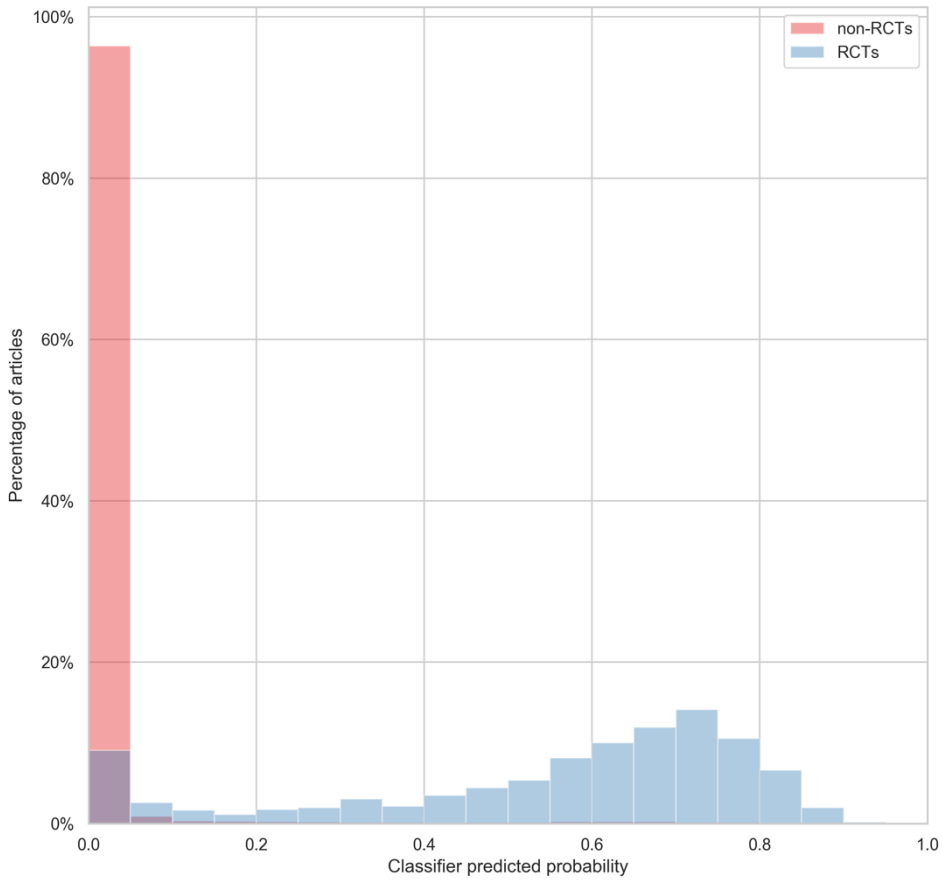


Table 3.2 Bootstrap estimates of model performance on Clinical Hedges data set, with 95% confidence intervals

Validation precision	Validation recall	Validation specificity	C statistic	Brier score
0.08 (0.06, 0.12)	0.99 (0.98, 0.99)	0.63 (0.48, 0.76)	0.98 (0.98, 0.98)	0.05 (0.05, 0.05)

Figure 3.4 Distribution of classification scores for RCTs and non-RCTs in Clinical Hedges data set



Validating the classifier recall on studies included in Cochrane reviews

The title and abstract records of 58,283 studies included in 4,296 Cochrane reviews were fed through the classifier. Records with a score equal to or above the threshold identified in the previous step were automatically classified as potentially reporting an RCT; those scoring below this threshold were automatically classified as not reporting an RCT.

Table 3.3 summarises the number of eligible studies that are 'lost' to reviews as a result of all of their corresponding study reports scoring lower than the threshold. When records that contain insufficient information for machine classification are excluded from machine classification, and assumed to be manually assessed (see Methods), the classifier correctly identifies 99.5% (43,783 out of 44,007) of studies.

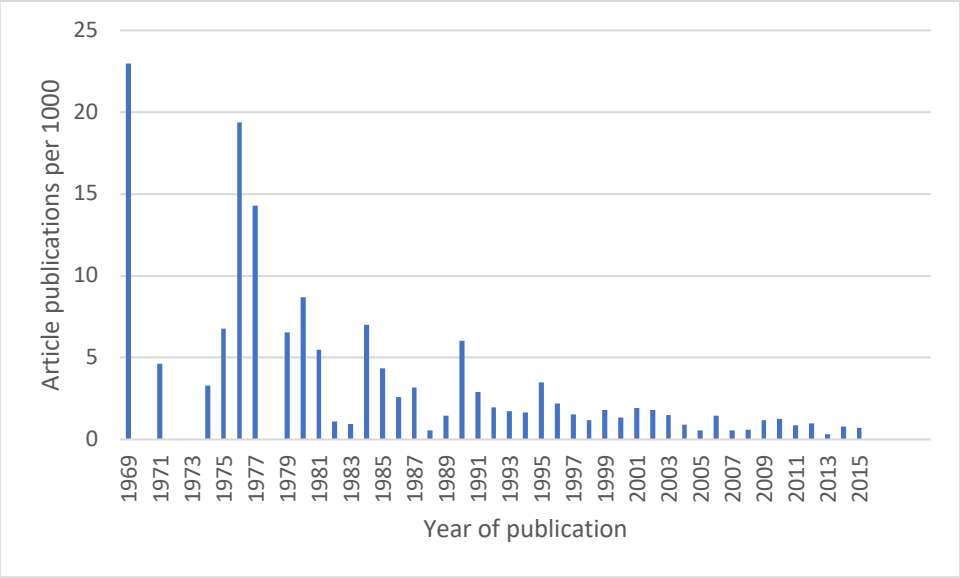
In our secondary analysis, when we include for machine classification data for all studies (including the subset of studies which contain insufficient information for accurate machine classification (see 'Methods', above)), we find that 3,396 studies would be potentially 'lost' to reviews (compared to 224 studies when only those with sufficient information are included).

Table 3.3: Number of included studies in Cochrane reviews classified as RCTs

	RCTs correctly identified by the classifier (recall)	RCTs not identified by the classifier
All studies (n = 58,283)	54,683 (93.8%)	3,600 (6.2%)
Studies with sufficient information for machine classification (n = 44,007 studies)	43,783 (99.5%)	224 (0.5%)

Figure 3.5 shows the 224 randomised trials 'lost' by the classifier per 1000 published, by year of publication, for all but one of the publications (the age of one publication could not be ascertained). These results show that older reports are much more likely to be misclassified by the machine learning classifier.

Figure 3.5 RCTs 'lost' by the classifier per 1000 published, by year of publication, showing that the risk of 'losing' a publication decreases over time



3.5 Discussion

Summary of findings

We conducted a three-stage study that involved training, calibrating and evaluating a machine learning classifier designed to distinguish between bibliographic title-abstract records that report an RCT and those that do not. Recall falls to an unacceptably low level (94%) if records with limited information in their titles and/or abstracts are submitted for machine classification. However, when these records are excluded, the classifier exceeds the standard required by Cochrane with recall at

99.5% of all those records scored. It should be noted that this means that some records are unsuitable for machine learning, and so must necessarily be checked manually; however, this mirrors current practice, whereby records with limited information in their titles and abstracts are retained for further assessment on the basis of their corresponding full text reports.

We deem the recall level as 'acceptable' for use in 'live' reviews on the basis that: a) this exceeds the recall of validated RCT search filters that have been used in systematic review production for many years; and b) this threshold was agreed by methodologists in Cochrane for use in Cochrane reviews.

While the precision of 8% estimated against the Clinical Hedges data set appears low, this is partly because of the age of that data set and relatively low prevalence of RCTs. In the Cochrane Evidence Pipeline workflow (Figure 3.1), the classifier saved Cochrane Crowd from needing to check 185,039 records manually (out of a total of 449,480) during the 2018 calendar year; a very large saving in manual workload[13].

Systematic reviews are frequently used to support decision-making processes, for both policymakers and practitioners, and are also key sources of evidence in drug licensing regulation. Reviews need to be accurate representations of the state of current knowledge, as decisions that are based on their findings can affect people's lives. Reviews also need to be demonstrably correct, as the way in which evidence is synthesised can have implications, for example, for drug licensing, and can therefore be open to legal challenge. These joint imperatives – for systematic reviews to be correct, and to be seen to be correct – generate the normative expectation that they should contain all relevant research evidence and the corollary concern that review findings based on bodies of evidence that inadvertently exclude some eligible studies are potentially unreliable. To this end, our study provides data demonstrating the reliability of implementing what could be seen as a major innovation in study identification methods for systematic reviews, the automatic eligibility assessment of study reports, and the exclusion of a portion without any manual checking by humans, rolled out at scale across Cochrane: the largest producer of systematic reviews globally and an organisation committed to minimising risk of bias in review production through methodological and editorial rigour. We note that the recall threshold set by Cochrane (99%) exceeds the performance of conventional search methods (for example, the Cochrane Highly Sensitive Search Strategy was found to have recall of 98.5% by the Clinical Hedges team)[21], and we have demonstrated in previous work that our machine learning approach can exceed the precision achieved by conventional search filters[8].

Although our results indicated that 0.5% of studies could have been 'lost' to Cochrane reviews if authors had used this classifier (affecting 178 reviews, leaving 4,118 reviews unaffected), this is almost certainly an overestimate when considering the prospective use of this classifier to support identification of newly published RCTs for new, updated and/or living systematic reviews. First, other means of finding studies are routinely employed in Cochrane reviews alongside conventional electronic searching – such as checking reference lists or contacting researchers who are active in the topic area – so some of these 'lost' studies would likely be found using these complementary search methods. Second, studies that are potentially lost are overwhelmingly older reports. While we do not dismiss the potential importance of identifying older trials for consideration in systematic reviews, it is reassuring that more recent studies (relevant especially for newer treatments and review updates) are far less likely to be missed. One reason the classifier performs better for more recent studies could be improvement in the reporting of RCTs over time, for example, in response to the CONSORT statement[16,22]. Trialists are now widely expected to detail trial methodology in the report's abstract and to include the fact that they are reporting an RCT in its title.

Strengths and weaknesses of this evaluation

We have described a robust evaluation of the performance of an RCT classifier in a large data set of systematic reviews. We were fortunate in having three large, independently generated, high-quality data sets available to train, calibrate and validate the classifier. This is an unusual position to be in and there are probably few study designs other than RCTs with comparably high-quality data sets available. We note that this may limit the potential to evaluate the performance of similar workflows, created to identify other types of study design, using the same three-stage process.

The current classifier has been trained almost exclusively on records published in English, so it does not necessarily generalise to other languages, However, this important limitation is, in principle, surmountable, as machine learning technology is language-agnostic and would therefore be capable of modelling any language, so long as sufficient training data were available.

The focus of this work has been to build a machine learning classifier for deployment in a specific workflow. The machine learning classifier we have developed meets required levels of recall, but inevitably results in some studies being 'lost' to reviews. This study does not attempt to ascertain the impact of these losses on the affected reviews' statistical and narrative results and findings, and a future extension of this study will investigate this important question. We also note that only 178 out of 4,296 reviews were affected, leaving results unchanged in at least 96% of the reviews.

3.6 Next steps: the Screen4Me service

We are currently piloting an extension to the Evidence Pipeline for use with individual Cochrane reviews. Authors using this service will compile their set of potentially eligible records from searches of multiple databases (including CENTRAL) as is typical for any systematic review. Given that the RCT Classifier and Cochrane Crowd have already classified more than 800,000 study records (and increasing by > 10,000 per month), it is likely that a proportion of the records retrieved and uploaded to the Classifier by authors have already been classified according to whether they report an RCT or not. Where this is the case, the records which are already known not to describe RCTs will be removed from the workflow. The remaining studies will then be sent to the RCT Classifier, and those records classified as not reporting an RCT (i.e., that fall below the 99% recall threshold) will be discarded. Finally, the records classified as potentially reporting an RCT will be screened by Cochrane Crowd. The review team is then left with a much smaller pool of records to examine, containing only RCTs. In early pilots, this new workflow reduced manual screening workload by between 40 and 70%, depending on the prevalence of RCTs in the search results of individual reviews.

3.7 Conclusions

The Cochrane RCT Classifier is now deployed by Cochrane for reducing screening workload in review production. As part of a wider workflow that includes prospective database searches and crowdsourcing to build a comprehensive database of RCTs, machine learning can reduce the manual screening burden associated with research synthesis, while ensuring a very high level of recall that is acceptable for an organisation which depends on having comprehensive access to the published research that falls within healthcare topics relevant to its scope.

3.8 Author contributions

James Thomas, **Anna Noel-Storr**, **Steve McDonald** and **Iain Marshall** designed the study.
James Thomas and **Iain Marshall** built the classifiers and calibration models.
Anna Noel-Storr and **Steve McDonald** worked on evaluation data sets.
Chris Mavergames provided overall Cochrane direction and governance.
Ian Shemilt and **Julian Elliott** provided methodological input throughout.
All authors read and approved the final article.

3.9 Abbreviations

CENTRAL Cochrane Central Register of Controlled Trials

CRS	Cochrane Register of Studies
IRMG	Cochrane Information Retrieval Methods Group
RCT	Randomised controlled trial
SVM	Support vector machine

3.10 References

1. Cochrane. Cochrane Library [Internet]. 2019 [cited 2019 Oct 26]. Available from: https://www.cochranelibrary.com [Accessed 27 December 2021].
2. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf M, et al. Chapter 4: Searching for and selecting studies. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions 2nd Edition. Chichester (UK): John Wiley & Sons; 2019. p. 67-99.
3. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med [Internet]. 2010 Sep 21;7(9):e1000326. Available from: http://dx.plos.org/10.1371/journal.pmed.1000326.
4. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S MD. How Quickly Do Systematic Reviews Go Out of Date ? A Survival Analysis. Ann Intern Med. 2007;147:224–33.
5. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: Increasing value, reducing waste. Lancet. 2014;383(9912):101-4.
6. Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. BMJ [Internet]. 2013;346(jan10 1):f139–f139. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.f139.
7. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev [Internet]. 2015;4(1):5. Available from: http://www.systematicreviewsjournal.com/content/4/1/5.
8. Marshall I, Noel-Storr AH, Kuiper J, Thomas J, Wallace BC. Machine Learning for Identifying Randomized Controlled Trials: an evaluation and practitioner's guide. Res Synth Methods [Internet]. 2018;(December 2017):1–13. Available from: http://doi.wiley.com/10.1002/jrsm.1287.
9. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Informatics Assoc [Internet]. 2017;0(0):1-4. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocx053.
10. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol [Internet]. 2017;91:31–7. Available from: http://www.sciencedirect.com/science/article/pii/S0895435617306042.

11. Cochrane. About the CRS (Cochrane Register of Studies) [Internet]. Cochrane Community. 2019 [cited 2019 Oct 25]. Available from: https://community.cochrane.org/help/tools-and-software/crs-cochrane-register-studies/about-crs.
12. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020 Nov;127:142-150. doi: 10.1016/j.jclinepi.2020.08.008. Epub 2020 Aug 13. PMID: 32798713.
13. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, et al. An evaluation of Cochrane Crowd finds that crowdsourcing can help to address the challenge of information overload in evidence synthesis. J Clin Epidemiol. 2021 May;133:130-139. doi: 10.1016/j.jclinepi.2021.01.006.
14. Nevin L. Advancing the beneficial use of machine learning in health care and medicine: Toward a community understanding. PLoS Med. 2018;15(11):4-7.
15. Wilczynski NL, Douglas Morgan, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak. 2005;5(20):1-15.
16. Schulz KF, Altman DC, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMC Med. 2010;8(18):1-9.
17. Sain SR, Vapnik VN. The Nature of Statistical Learning Theory. Technometrics. 2006;
18. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv large margin Classif. 1999.
19. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001 Aug;54(8):774-81. doi: 10.1016/s0895-4356(01)00341-9.
20. Brier G. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;
21. McKibbon KA, Wilczynski N Lou, Haynes RB. Retrieving randomized controlled trials from medline: A comparison of 38 published search filters. Health Info Libr J. 2009;26(3):187-202.
22. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane reviewa. Syst Rev. 2012;29(1).

Chapter 4

A retrospective analysis of Cochrane reviews showed high sensitivity for the identification of randomised controlled trials by the Centralised Search Service

This original manuscript was published in *Journal of Clinical Epidemiology*

Citation: Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020 Nov;127:142-150.

DOI: 10.1016/j.jclinepi.2020.08.008

URL: https://www.jclinepi.com/article/S0895-4356(20)30116-5/fulltext

4.1 Abstract

Background

The Cochrane Central Register of Controlled Trials (CENTRAL) is compiled from a number of sources, including PubMed and Embase. Since 2017, we have increased the number of sources feeding into CENTRAL and improved the efficiency of our processes through the use of APIs, machine learning and crowdsourcing.

Objectives

Our objectives were twofold:

(1) Assess the effectiveness of Cochrane's centralised search and screening processes to correctly identify references to published reports which are eligible for inclusion in Cochrane systematic reviews of randomised controlled trials (RCTs).

(2) Identify opportunities to improve the performance of Cochrane's centralised search and screening processes to identify references to eligible trials.

Methods

We identified all references to RCTs (either published journal articles or trial registration records) with a publication or registration date between 1st January 2017 and 31st December 2018 that had been included in a Cochrane intervention review. We then viewed an audit trail for each included reference to determine if it had been identified by our centralised search process and subsequently added to CENTRAL.

Results

We identified 650 references to included studies with a publication year of 2017 or 2018. Of those, 634 (97.5%) had been captured by Cochrane's Centralised Search Service (CSS). Sixteen references had been missed by the CSS: six had PubMed-not-MEDLINE status, four were missed by the centralised Embase search, three had been misclassified by Cochrane Crowd, one was from a journal not indexed in MEDLINE or Embase, one had only been added to Embase in 2019, and one reference had been rejected by the automated RCT machine learning classifier. Of the sixteen missed references, eight were the main or only publication to the trial in the review in which it had been included.

Conclusions

This analysis has shown that Cochrane's centralised search and screening processes are highly sensitive. It has also helped us to understand better why some references to eligible RCTs have been missed. The CSS is playing a critical role in helping to populate CENTRAL and is moving us towards making CENTRAL a comprehensive repository of RCTs.

4.2 Background

The Cochrane Central Register of Controlled Trials (CENTRAL) is a bibliographic database populated with reports of randomised and quasi-randomised controlled trials (RCTs and q-RCTs)[1,2]. CENTRAL is available through the Cochrane Library. Most review teams, whether they are producing Cochrane or non-Cochrane reviews, can access CENTRAL for free, either through national licenses or institutional subscriptions. Reports of RCTs are added to CENTRAL through two main routes: (1) via Cochrane Information Specialists identifying and manually adding trial records, and (2) by a centralised search initiative, called the Centralised Search Service (CSS), managed by Cochrane's Editorial and Methods Department.

Five sources are searched centrally: PubMed and ClinicalTrials.gov, both produced by the US National Library of Medicine (NLM); Embase.com produced by Elsevier; the World Health Organization's International Clinical Trials Registry Platform (ICTRP); and KoreaMed produced by the Korean Association of Medical Journal Editors. The service is also adding a sixth source: CINAHL hosted by EBSCOhost. CINAHL records are expected to appear in CENTRAL at the end of the first quarter of 2020. For each source we have developed bespoke workflows with the aim of capturing all possible reports of RCTs and q-RCTs.

The CSS uses four main approaches to identify relevant records. Not all are used for each of the sources covered (a summary of the overall workflow is given in Table 4.1). The four approaches are:

1. Direct feed
2. Sensitive search
3. Machine learning
4. Crowdsourcing

Direct feed

The first approach is the 'direct feed' which consists of records that have been indexed in the source databases as RCTs. Wherever possible, we aim to identify potential 'direct feeds' of records reporting an RCT into CENTRAL. This route is the most efficient approach because it does not require any manual assessment/screening of records. We currently have 'direct feeds' in place for four of the five sources: PubMed, Embase, ClinicalTrials.gov and ICTRP. However, the 'direct feeds' only capture a proportion of the eligible records from each source. Other approaches are therefore needed to identify the remaining RCTs.

Sensitive search

The second approach is to use a sensitive search. Records from all sources which cannot be identified through the 'direct feed' (i.e., they do not have the required index terms) are identified through a search which has been developed for each source[3,4]. As the results from a sensitive search for RCTs inevitably contain many non-relevant records, additional checks are then required to ensure that only randomised study reports are retained. These additional checks are in two phases: first, records are passed through a machine learning classifier to eliminate clearly irrelevant records[5]; and second, the remainder, are checked by Cochrane Crowd[6].

Machine Learning

The third approach, which supplements the sensitive search, uses machine learning. The automated machine learning classifiers provide likelihood scores as to whether the record is describing an RCT. The CSS uses two machine learning classifiers, one developed for the bibliographic records such as those identified from Embase, and one developed for trial registry records from ClinicalTrials.gov. For more detail on the training, calibration and validation of the bibliographic RCT classifier, see Thomas[5] (and Chapter 3). The RCT machine learning classifiers are currently used to remove 'noise' (non-relevant records) from large record sets. In other words, we are not using the RCT classifiers to identify RCTs with high precision; we are using them to remove the obvious non-RCT records, thereby reducing the amount of manual screening required.

Crowdsourcing

The final component in the workflow is 'the crowd'. Records that have not been accounted for in either the direct feeds or excluded by the RCT classifiers need to be manually screened. These records are sent to Cochrane Crowd (https://crowd.cochrane.org), Cochrane's citizen science platform that hosts tasks aimed at identifying particular types of health research. Cochrane Crowd is

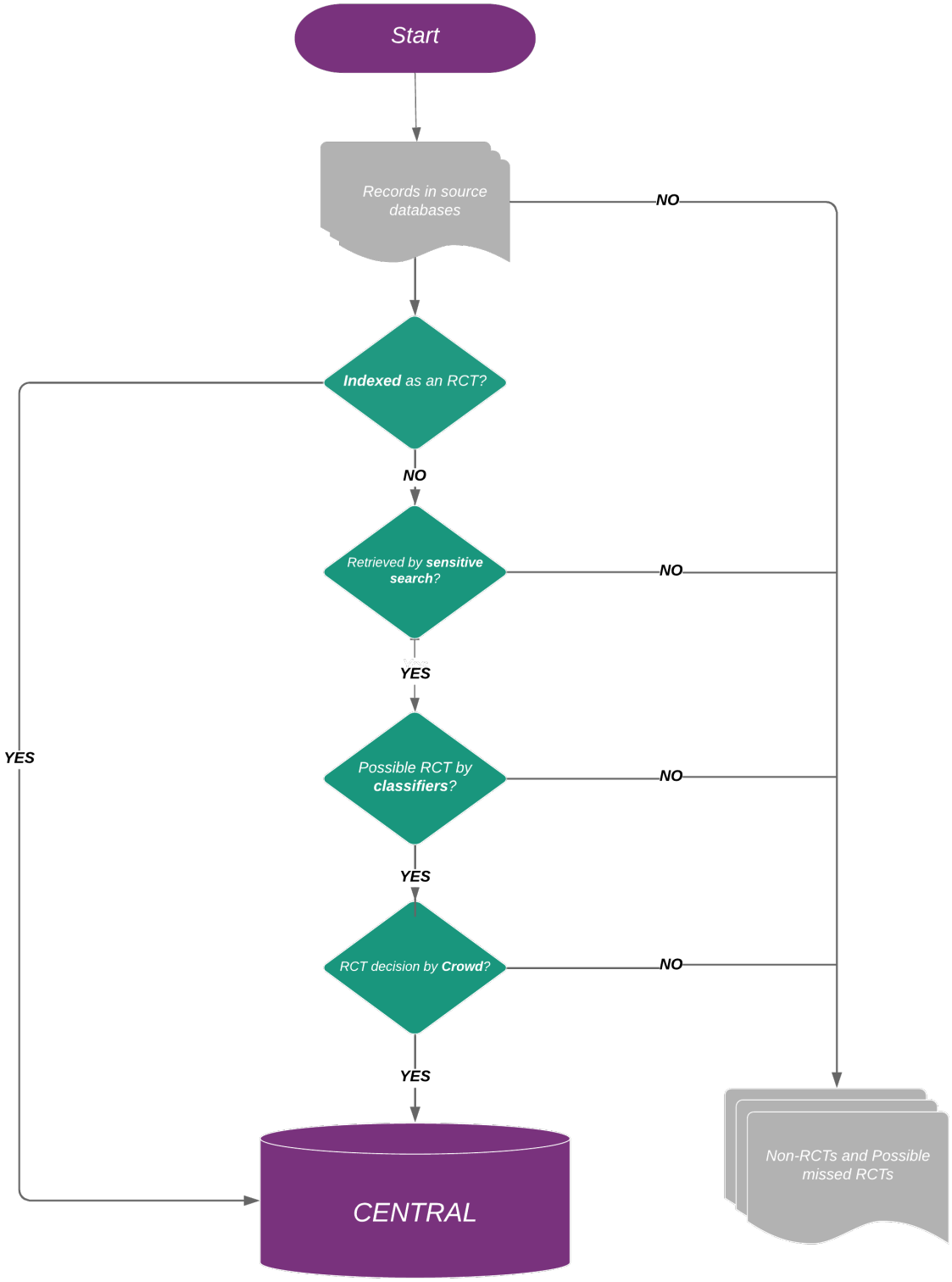
open to all but contributors must first complete a brief training module before being able to screen live records. In addition to training, the crowd approach uses an agreement algorithm to help ensure collective accuracy of the data output. The current algorithm in place for the identification of RCTs from bibliographic sources requires that each record needs four consecutive agreement classifications for that record to be deemed either an RCT or not an RCT. The current agreement algorithm in place for the identification of RCTs from trials registries (i.e., the records from ClinicalTrials.gov and ICTRP) is that each requires three consecutive agreement classifications for that record to be deemed either an RCT or not an RCT. Disagreeing classifications or records that receive an *Unsure* classification go to 'resolver' screeners in the Cochrane Crowd. Resolvers are highly experienced screeners who are tasked with making a final decision on records that have not received the required consecutive agreement classifications. For more detailed information regarding the agreement algorithms used and the accuracy of this crowdsourced approach, see Noel-Storr [6](and as described in Chapter 2). The Cochrane Crowd community stands at over 17,000 people from over 150 countries.

Table 4.1 Individual workflows for centrally searched sources as of December 2019. More detailed information on the current workflows in place can be found at http://www.cochranelibrary.com/help/central-creation-details.html

Source (provider)	Workflow description	Harvested from external source
PubMed (National Library of Medicine)	Direct feed of records into CENTRAL based on index terms: *"randomized controlled trial"[Publication Type] OR "controlled clinical trial"[Publication Type]*	Monthly API call on 16th of each month
Embase (Elsevier)	Direct feed of records into CENTRAL based on Emtree term: *Randomised controlled trial* Sensitive search of Embase.com via the Embase.com API; results sent to RCT Classifier and remaining records sent to Cochrane Crowd for manual screening	Monthly API call on 15th of each month
ClinicalTrials.gov (National Library of Medicine)	Direct feed of records into CENTRAL of all records with randomised controlled trial in study design field Download all other records; results sent to classifier. Those meeting threshold criteria are then sent to Cochrane Crowd for manual screening	Daily API call
ICTRP (World Health Organisation)	Download all records; remove CT.gov records. Remaining records sent to Cochrane Crowd for manual screening.	Monthly API call on 15th of each month
KoreaMed (Korean Association of Medical Journal Editors)	Download all records; records are sent to Cochrane Crowd for manual screening.	Monthly API call on 15th of each month
CINAHL (EBSCOhost)	Sensitive search of CINAHL via API; results sent to RCT Classifier and remaining records sent to Cochrane Crowd for manual screening	Daily API call from August 2020

By combining these approaches – API direct feeds, sensitive searches, machine learning classifiers, and crowdsourcing manual screening via Cochrane Crowd – the CSS has established an effective process for identifying RCTs for CENTRAL. This RCT identification workflow required evaluation to ensure and improve efficiency and accuracy.

Figure 4.1 Study identification workflow



4.3 Aims and objectives

Our aim was to evaluate Cochrane's CSS to assess its effectiveness at capturing reports of randomised trials for Cochrane intervention reviews. We sought to determine overall comprehensiveness as well as to assess the performance of each component of the workflow. We were concerned specifically with *sensitivity*: i.e., establishing whether our processes identify all the studies that they were designed to, rather than evaluating their efficiency in terms of their *specificity*.

We also sought to analyse in detail the reasons why references to studies were not captured by the CSS, and to recommend any improvements to our workflows and processes that we could identify.

4.4 Methods

We conducted a retrospective analysis of Cochrane intervention reviews available in March 2019 and downloaded references to their included studies that had a publication (or trial registration) date of either 2017 or 2018. We chose these two years because they are the two most recent years where we have had the CSS operating. We are able to use this data set because at present, in the vast majority of cases, studies included in Cochrane reviews are not identified from a single search of CENTRAL but through extensive and sensitive searches conducted across multiple sources in accordance with Methodological Expectations for Cochrane Intervention Reviews (MECIR)[7] and the Cochrane Handbook[8]. If studies had been identified through searches of CENTRAL only, we would not have been able to ascertain the comprehensiveness of CSS processes.

After downloading all the 2017 and 2018 references to included studies, we removed duplicate references and references to non-randomised studies. We identified these studies by examining the inclusion criteria for each review. If the review stated that it had included study designs other than randomised controlled trials, we then checked the Characteristics of Included Studies table within the review to discern whether the included studies were RCTs or not. Two assessors working independently then categorised each reference according to the following: *1) journal article* (including letters, errata etc.), *2) conference publication*, *3) trial registry record*, and *4) other* for record types not covered by the CSS, for example, clinical study reports and email correspondence.

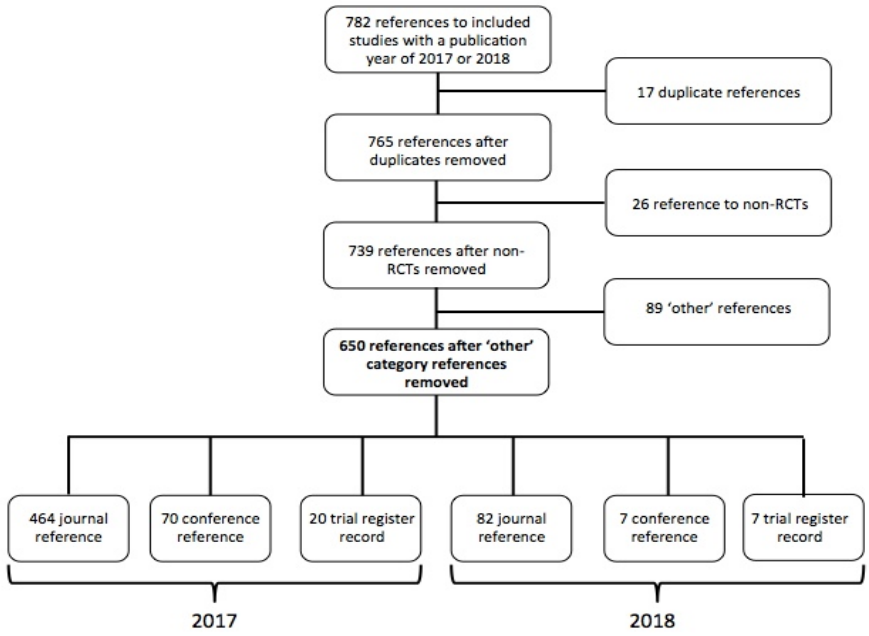
We also noted whether the reference had been flagged as the primary reference to an RCT or a secondary publication by the individual review author teams, because trialists very often produce more than one publication or research output for a single trial[9].

With this categorised data set we then constructed an audit trail for each record to ascertain whether it had been identified by the CSS and, if so, through which approach. For example, whether the reference had been picked up by the CSS via a direct feed or via a sensitive search and crowdsourcing. We did not assess whether the references to included studies were retrieved by the actual searches performed in CENTRAL for the reviews. Whilst this is an important question, it goes beyond the scope of this evaluation which sought to assess recall in terms of whether the centralised processes identified the RCTs included in reviews. We used a relative recall approach often used in studies evaluating the performance of methodological search filters[10]. This approach uses a set of known relevant records (the included studies) as its denominator, rather than a handsearched gold standard data set.

4.5 Results

We retrieved 782 references to included studies from 274 reviews with a publication year of 2017 or 2018. After removing the duplicates and the non-RCT records, we were left with 739 records. Figure 4.2 shows the flow of references used in this analysis, and the breakdown of record type based on the categories we used.

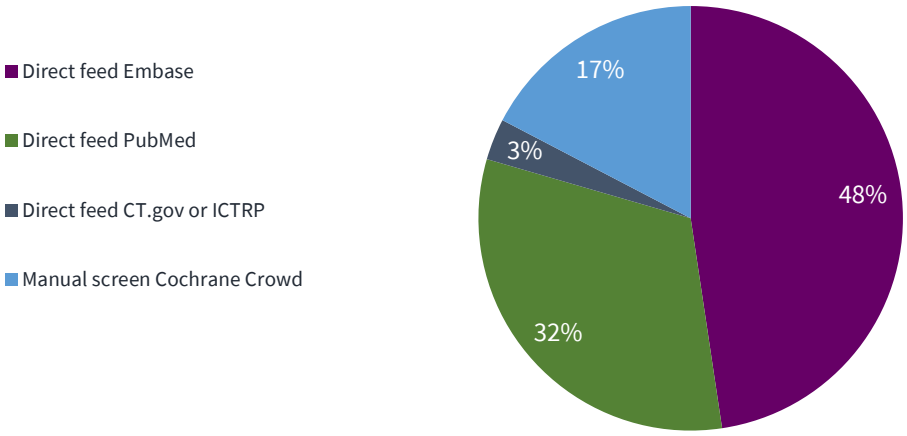
Figure 4.2 Flow diagram of references to studies included in this retrospective analysis



The 650 references to included studies from 262 Cochrane reviews were record types covered by the CSS. We reviewed the methods section of a random sample of 25% (65) of the 262 reviews to check that multiple sources (i.e., not just CENTRAL) had been searched. Within this sample two search approaches were described: 1. searches developed and carried out across multiple sources specifically for the review, and 2. searches carried out in specialised registers of the review group responsible for the review. For those reviews reliant on register-only searches we checked that the searches run for the maintenance of the register was across multiple sources. 51 of the 65 reviews checked reported carrying out bespoke searches across multiple sources specifically for the review; 14 reported using their specialised register as the main source searched. None of the review methods checked reported searching only CENTRAL.

Of the 650 references to RCTs included in Cochrane reviews, 97.5% (634) had been captured by the centralised study identification processes. The majority of these had been identified by the PubMed and Embase direct feeds of records 32% (202) and 48% (302) respectively. A further 110 (17%) references to included studies had been identified by Cochrane Crowd and 20 (3%) had come in through the direct feeds of trial registry records (see Figure 4.3).

Figure 4.3 Breakdown of RCTs identified by CSS approach



Sixteen references to included studies were not identified through the CSS. Of these, six (38%) were references in PubMed but they did not have the *randomised controlled trial* or *controlled clinical trial* publication type index term and so were not picked up by the PubMed direct feed or by the Embase direct feed or sensitive search. While the vast majority of PubMed records are in MEDLINE and therefore identifiable from Embase (which subsumed MEDLINE content in 2011), some records with

PubMed-not-MEDLINE status remain outside of the main data set. The two main reasons for records acquiring a PubMed-not-MEDLINE status are (1) they have yet to be indexed for MEDLINE, or (2) they are records to articles in journals not covered by MEDLINE or are in the National Library of Medicine's PMC (PubMed Central) open archive of full text journal articles. Some of these PubMed-not-MEDLINE status records will become retrievable from Embase over time; however, others may not. All six missed publications were the primary, and only, study records listed in the reviews for those trials.

Of the remaining ten missed references, three had been identified by the sensitive Embase search but had then been incorrectly rejected as non-RCTs by the Cochrane Crowd. The three references were: a long-term follow-up report of an RCT, a letter, published in a journal, about an RCT, and an analysis of a secondary outcome of an RCT. All three were secondary publications to the trial they were describing.

A further four references had been missed by the sensitive Embase search. Of these, one was the only reference for that included study; the other three were secondary publications (i.e., there were other references to those trials included in the review). For each of the missed references, the titles, abstracts and index terms contained no explicit description to indicate they were reporting or describing an RCT. One was a sub-group analysis to an RCT where the name of a trial was provided but there were no other descriptors that indicated that the trial was an RCT. Another was a letter published in a journal about a trial. This Embase record contained only the title and none of the index terms related to study design. The third missed secondary publication was a long-term follow-up of an RCT. The final missed reference described a controlled study, but did not provide details on how the participants had been allocated to each arm of the trial. The abstract described the trial's aim as examining the "comparative efficacy" of a twelve-week treatment programme versus a "treatment as usual" group, and was indexed with the Emtree headings *controlled study/* and *comparative effectiveness/*. We currently use the narrower Emtree term *controlled clinical study/* in the Cochrane sensitive Embase search, rather than *controlled study/;* therefore, despite the sensitivity of the Cochrane Embase centralised search, this reference was not captured.

The machine learning classifier was identified as the cause of one missing reference. The RCT Classifier works to remove the records with a very low probability of describing an RCT. In other words, it handles many of the clear-cut non-RCTs, thereby freeing up human effort (the crowd) to manually screen the records that would challenge the machine classifier. The expected recall rate of

the classifier is around 99.5% on studies included in Cochrane reviews[5]. Therefore, only missing 1 study is exceeding this expected performance. The missed study was a secondary publication of a randomised controlled trial that assessed biomarker data available from a subset of the original trial's participants. With the exception of the word 'randomised' being used once in the abstract, there were no other indications that this report was related to a randomised trial.

The final two missed references were a conference publication to a 2017 study that was not added to Embase until week 34 of 2019 and not retrieved by the feeds during the period of interest, and a reference in a journal not indexed by any of the sources covered by the CSS: *Modern Approaches in Drug Designing*. The former missed reference was a secondary one and the latter was flagged as the primary reference to the included study.

4.6 Discussion

This analysis found that the CSS – a Cochrane initiative that aims to identify as many reports of RCTs as efficiently and as accurately as possible through a combination of direct feeds, sensitive searching, crowdsourcing and machine learning – is achieving high sensitivity. While some studies were missed through these approaches, the number missed was small and comparable to the expected recall of traditional methodological filters and the screening of abstracts by review author teams[11]. Only 2.5% of references to included studies in our test set were not picked up by the CSS, and of those 16 missed references, only eight (50%) were flagged as the primary paper to the RCT in the Cochrane review. In addition, this analysis has shown the valuable role of Cochrane Crowd, which identified 17% of the references.

In terms of the range of sources we currently search as part of the CSS initiative, this analysis also indicates that Cochrane's coverage of English-language journal articles, conference publications and trial register records is highly sensitive. Importantly, only one reference to a journal article included in a Cochrane review was missed because it was not indexed in any of the bibliographic databases covered by the CSS. This is helpful information in terms of prioritising which sources should be the next focus for any centralised searching efforts. However, the fact there was only one missed study could indicate that searches for Cochrane reviews are potentially not broad enough in terms of less mainstream databases and non-English language sources. Options for future objectives of the CSS could be to target non-English language material and other record types such as clinical study reports[12,13].

This analysis also provided us with a better understanding of what we can do to improve our current processes, and some of that work has already begun. For example, we have revised the Cochrane Crowd Quick Reference Guide to make clear that follow-up studies to RCTs are to be selected for CENTRAL. We also now require that crowd contributors repeat the training module every 6 months to remind them of the inclusion criteria for CENTRAL. We are reviewing the Embase sensitive search to see whether it should be amended slightly in light of the few missed references, and we are currently evaluating the existing PubMed RCT filters to capture those references that are in PubMed but have not been indexed with the RCT publication type term. We have also recently updated the RCT Classifier and tested whether references rejected by the old version would now be included.

One question frequently posed to the CSS team is whether searches for randomised evidence can now be limited to searching only CENTRAL. Several research papers have sought to evaluate the comprehensiveness of other major bibliographic databases[14,15,16,17] or trial registries[18]. This analysis indicates that the vast majority of published articles, conference proceedings and trial registry records, are being successfully identified by the centralised searching and screening processes. However, there are a number of factors that should be taken into consideration when deciding which sources to search and, specifically, whether there is still a need to search the source databases currently covered by the Cochrane CSS. We will start first with specific limitations of this analysis before describing a number of more general factors that could help inform decisions about which sources to search.

Limitations of this analysis

This analysis has focused on a very specific time frame: studies with a publication year of 2017 or 2018, therefore our results are limited to more recent reports of randomised trials. The reporting of randomised trials has likely improved over time due to the CONSORT initiative[19,20]; this may have made identifying randomised trials easier. Most new reviews would normally plan to search for trials across all years and not just those published more recently. This analysis does not help to answer the question of whether someone looking for trials across all dates by just searching CENTRAL would be likely to find them all (or even 97.5% of them). Another limitation is that our sample size for trial registry records is small; therefore the findings of this study should be viewed with caution in relation to this record type.

There are other, broader factors to take into consideration when deciding which sources to search, specifically with regard to limiting a search to CENTRAL.

Time-lag from source database to CENTRAL

Currently, the shortest time possible for a record in a source database to appear in CENTRAL following identification from a source database, is between three to four weeks. This is because the source databases are currently queried once every month and CENTRAL is updated once every month. However, some records can take much longer to reach CENTRAL. These are records that need to be resolved in Cochrane Crowd (an average of 11.3% of records across the three Crowd tasks that feed CENTRAL need resolving), either because the crowd has disagreed in their classifications of a record or has classified a record with an *Unsure* classification. These records can take time to receive a final classification. Resolver screeners, members of the Cochrane Crowd community tasked with making a final decision on records that need resolving, are few in number due to the expertise level required and often have to obtain the full text to make a final decision. This issue however does raise the question around what would be considered acceptable levels of ineligible records being submitted to CENTRAL. This analysis has focused entirely on the sensitivity of current processes. However, we do know that some ineligible records reach CENTRAL via the direct feeds and via Cochrane Crowd. A further small drop in specificity may be acceptable if it enabled faster delivery of these RCTs into CENTRAL.

Different versions of centralised searches and processes

This analysis has focused on records retrieved by the most recent version of the searches and processes in place for the CSS. However, these searches and processes have evolved over time. For example, records identified for CENTRAL from Embase were identified on the basis of a different search strategy pre-2014. The latest search strategy in use by the CSS is considered to be more sensitive than previous iterations. It is therefore feasible that a higher proportion of RCTs may have been missed by older, less sensitive searches. Similarly, when new sources are added to the CSS process there is often a large initial set of records for all years to process, after which monthly processing is quicker. Strategies to deal with large backlogs often differ slightly from the process used to deal with the prospective data feed for the same source. In taking a pragmatic approach to managing backlogs – which we must do because of resource constraints – it is possible that, despite our best efforts, some eligible records may have been lost.

The search interface

Another consideration for those interested in restricting their searching to CENTRAL is the difference in the search interface and search capabilities in CENTRAL compared with those of the source

databases. Only a sub-set of metadata is harvested for CENTRAL, so supplementary searches of source databases may yield additional records. This is particularly relevant to the trial registry records where often much more information is available to search within the regional and international registries[21]. We hope to conduct a further analysis using the CENTRAL search strategies reported in the Cochrane reviews we used for this study, and to assess whether the trials were successfully captured by those strategies instead or as well as by the searches run directly in the source databases.

Inclusion criteria for CENTRAL

Study designs that are eligible for CENTRAL have not changed for many years, and remain: randomised or quasi-randomised controlled trials, controlled before-and-after studies and interrupted time series. The centralised search processes were designed to capture randomised or quasi-randomised controlled trials. We do not currently have any centralised processes in place to identify controlled before-and-after studies or interrupted time series. In addition, while criteria in terms of study design have been stable for some time, over the last few years (since 2014) the types of reports eligible for CENTRAL has broadened. For example, post-hoc and secondary analyses of RCTs, are now included in CENTRAL. The expanded eligibility criteria have implications particularly for those seeking *all* publications relevant to a single randomised trial, rather than just the main or primary publication.

Limitations of search strategies

The effectiveness of database retrieval is also impacted by the quality of the search strategies used to search the database. If the only database to be searched is CENTRAL then the quality of the single search strategy becomes crucial to the success of the review. A more conservative approach of searching a range of databases with different search translations may increase the chances to retrieve relevant records.

Searching for what purpose?

The final factor to consider, and perhaps the most obvious one, concerns the objective of the search itself. For example, rapid reviews or scoping searches may accept lower sensitivity in favour of precision, while searches for Cochrane intervention reviews and living systematic reviews[22] will be primarily concerned with maximising sensitivity. Searchers conducting rapid reviews or scoping reviews may be content to use only CENTRAL with a highly sensitive strategy, whereas searchers

populating full systematic reviews may wish to search beyond CENTRAL for the reasons discussed above. Context will therefore always be an important consideration.

There are numerous factors that information specialists should consider when deciding which sources to include in their searches. To help inform this decision-making, we present our methods for identifying trial records for CENTRAL transparently and completely.

5.7 Conclusions

The Centralised Search Service has established processes for identifying RCTs for inclusion in CENTRAL by using a combination of API direct feeds, sensitive searches, machine learning classifiers and crowdsourced manual screening via Cochrane Crowd. Our evaluation has found that the workflow achieves a very high level of sensitivity. We have also identified ways to improve the CSS. We present our process and the results of this evaluation in an effort to support the decision-making of information specialists seeking the best source databases for their work. Although highly sensitive in its coverage, CENTRAL may not yet be seen as a comprehensive source of all relevant trials for systematic reviews for all purposes, it may however be comprehensive enough for some use cases such as searchers undertaking rapid or scoping reviews. In these cases it will be important that the quality of the search itself is high and takes into account the limitations discussed. Our processes for identifying RCTs for CENTRAL will continue to evolve through the use of machine learning, and the contribution of the Cochrane Crowd community. During these transformations, we will continue to share the results of our process evaluations and our methods for identifying RCTs for CENTRAL.

4.8 Abbreviations

CENTRAL	Cochrane Central Register of Controlled Trials
CRS	Cochrane Register of Studies
CSS	Cochrane's Centralised Search Service
q-RCT	Quasi-randomised controlled trial
RCT	Randomised controlled trial

4.9 Author contributions

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Gordon Dooley: methodology, resources, data curation, writing – reviewing and editing

Susi Wisniewski: methodology, data curation, validation, writing - reviewing and editing

Julie Glanville: methodology, visualisation, writing – reviewing and editing

James Thomas: conceptualisation, methodology, visualisation, writing - reviewing and editing

Sam Cox: methodology, writing - reviewing and editing

Robin Featherstone: methodology, visualisation, writing – reviewing and editing

Ruth Foxlee: conceptualisation, methodology, writing - reviewing and editing

4.10 References

1. McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
2. The Cochrane Library. CENTRAL Creation: How CENTRAL is created (https://www.cochranelibrary.com/central/central-creation) [Accessed 28 December 2021].
3. Glanville J, Dooley G, Wisniewski S, Foxlee R, Noel-Storr A. Development of a search filter to identify reports of controlled clinical trials within CINAHL Plus. Health Information and Libraries Journal 2019;36(1):73-90.
4. Glanville J, Foxlee R, Wisniewski S, Noel-Storr A, Edwards M, Dooley G. Translating the Cochrane EMBASE RCT filter from the Ovid interface to Embase.com: a case study. Health Information and Libraries Journal 2019;36(3):264-277.
5. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021 May;133:140-151. doi: 10.1016/j.jclinepi.2020.11.003.
6. Noel-Storr AH, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 May;133:130-139. doi: 10.1016/j.jclinepi.2021.01.006.
7. Higgins JPT, Lasserson T, Chandler J, Tovey D, Thomas, J, Flemyng E, Churchill R. Methodological Expectations of Cochrane Intervention Reviews. Cochrane: London, Version October 2019.
8. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
9. Ebrahim S, Montoya L, Kamal el Din M et al. Randomized trials are frequently fragmented in multiple secondary publications. J Clin Epidemiol 2016;79:130-39.
10. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, Klassen TP, Moher D. An alternative to the hand searching gold standard: validating methodological search filters using relative recall. BMC Medical Research Methodology 2006;6:33. doi: 10.1186/1471-2288-6-33.

11. McKibbon KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials from Medline: a comparison of 38 published search filters. Health Information and Libraries Journal. 2009;26(3):187-202.
12. Doshi P, Jefferson T. Clinical study reports of randomised controlled trials: an exploratory review of previously confidential industry reports. EMJ Open 2013;3:e002496.
13. Jefferson T, Doshi P, Boutron I, Golder S, Heneghan C, Hodkinson A, Jones M, Lefebvre C, Stewart LA. When to include clinical study reports and regulatory documents in systematic reviews. BMJ Evidence-Based Medicine 2018; 23: 210-217.
14. Halladay CW, Trikalinos TA, Schmid IT, Schmid CH, Dahabreh IJ. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. Journal of Clinical Epidemiology 2015; 68: 1076-1084.
15. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. The contribution of databases to the results of systematic reviews: a cross-sectional study. BMC Medical Research Methodology 2016;16:127.
16. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, St John PD, Viola R, Raina P. Should meta-analysts search Embase in addition to Medline? Journal of Clinical Epidemiology 2003;56:943-955.
17. Sampson M, de Bruijn B, Urquhart C, Shojania K. Complementary approaches to searching MEDLINE may be sufficient for updating systematic reviews. Journal of Clinical Epidemiology 2016;78:108-115.
18. Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ 2017;356:j448.
19. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T for the CONSORT Group. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. Ann Intern Med. 2001 Apr 17;134(8):663-94. doi: 10.7326/0003-4819-134-8-200104170-00012.
20. Hopewell S, Ravaud P, Baron G, Boutron I. Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. BMJ 2012;344:e4178.
21. Isojarvi J, Wood H, Lefebvre C, Glanville J. Challenges of identifying unpublished data from clinical trials: getting the best out of clinical trials registers and other novel sources. Research Synthesis Methods 2018:561-578.

22. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, Salanti G, Meerpohl J, MacLehose H, Hilton J, Tovey D, Shemilt I, Thomas J; Living Systematic Review Network. Living systematic review: 1. Introduction-the why, what, when, and how. Journal of Clinical Epidemiology 2017;91:23-30.

Chapter 5

Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service

This original manuscript was published in *Journal of Clinical Epidemiology*

Citation: Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. J Clin Epidemiol. 2021 Feb;130:23-31.

DOI: 10.1016/j.jclinepi.2020.09.024

URL: https://www.jclinepi.com/article/S0895-4356(20)31110-0/fulltext

5.1 Abstract

Objective

To assess the feasibility of a modified workflow that uses machine learning and crowdsourcing to identify studies for potential inclusion in a systematic review.

Study design and setting

This was a sub-study to a larger randomised study; the main study sought to assess the performance of single screening search results versus dual screening. This sub-study assessed the performance in identifying relevant RCTs for a published Cochrane review of a modified version of Cochrane's Screen4Me workflow which uses crowdsourcing and machine learning. We included participants who had signed up for the main study but who were not eligible to be randomised to the two main arms of that study. The records were put through the modified workflow where a machine learning classifier divided the data set into "Not RCTs" and "Possible RCTs". The records deemed "Possible RCTs" were then loaded into a task created on the Cochrane Crowd platform and participants classified those records as either "Possibly relevant" or "Not relevant" to the review. Using a pre-specified agreement algorithm, we calculated the performance of the crowd in correctly identifying the studies that were included in the review (sensitivity) and correctly rejecting those that were not included (specificity).

Results

The RCT machine learning classifier did not reject any of the included studies. In terms of the crowd, 112 participants were included in this sub-study. Of these, 81 completed the training module and went on to screen records in the live task. Applying the Cochrane Crowd agreement algorithm, the crowd achieved 100% sensitivity and 80.71% specificity.

Conclusions

Using a crowd to screen search results for systematic reviews can be an accurate method as long as the agreement algorithm in place is robust.

Trial registration

Open Science Framework: https://osf.io/3jyqt

5.2 Background

The current process of identifying studies for inclusion in systematic reviews is hampered by the sheer volume of research produced and by a model of identification that is both inefficient and costly[1]. The task of assessing what are often thousands of search results, in duplicate by two reviewers independently, undoubtedly contributes to lengthy production times meaning that important questions about treatment effects remain unanswered[2,3].

Cochrane, an international non-profit organisation that produces systematic reviews, has been working on a number of solutions to help expedite the review production process. Much of this effort has focused on the study identification stage of review production through the use of two increasingly popular technologies: crowdsourcing and machine learning. These two approaches, working in partnership, have the potential to transform the traditional review production paradigm in terms of study identification.

5.3 Introduction

Machine learning

Machine learning in this context means supervised machine learning. This is where the machine-learning model has been trained on a large, already categorised data set. Once the model (or classifier) has been built using this training data, it is then able to provide a score on new data. This score reflects how likely the new data is describing what is being looked for (the class of interest).

RCT Classifier

Cochrane uses a supervised model known as the RCT Classifier. Its development, calibration, validation and implementation has been described in detail elsewhere[4,5,6] (see Chapter 3). In brief, the RCT Classifier was developed to distinguish between reports of randomised controlled trials (RCTs) and non-randomised controlled trials (non-RCTs). It was built and trained using a large high-quality training data set produced by Cochrane's crowdsourcing initiative, Cochrane Crowd (described below). It was then calibrated using an independent, already categorised data set. This stage further tested the classifier and enabled a cut-point to be established which would enable its use as a binary classifier with records scoring above or equal to the cut-point as being possible RCTs, and those scoring below it, as very likely non-RCTs. The final validation stage made use of a third, independent, data set (the already included studies in Cochrane intervention reviews). This stage tested the cut-point by assessing the classifier's recall – its ability to correctly classify RCTs included in Cochrane reviews, as RCTs. Cochrane now uses the RCT Classifier in its process to identify possible

reports of RCTs as part of its Centralised Search Service initiative[6] (as described in Chapter 4). It is also used as part of the study identification process for *individual* Cochrane reviews through a workflow called Screen4Me (described below).

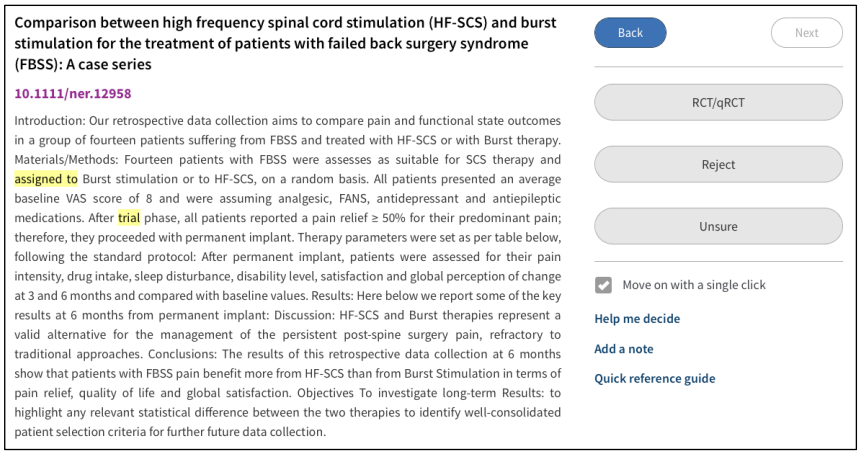
Crowdsourcing

Crowdsourcing is the outsourcing of tasks or activities to a large community, usually via the internet. The type of crowdsourced approach will depend on the nature of the problem that the host organisation is trying to solve. If an organisation is, for example, wanting to solve an empirical problem or generate innovative or creative ideas they will need to adopt different approaches than an organisation requiring help to collect, process and categorise large amounts of data[7]. Crowdsourcing in medical research has taken on a range of applications as summarised by Tucker and colleagues[8]. It has demonstrated huge potential in this domain and is now an established part of Cochrane's technological eco-system.

Cochrane Crowd

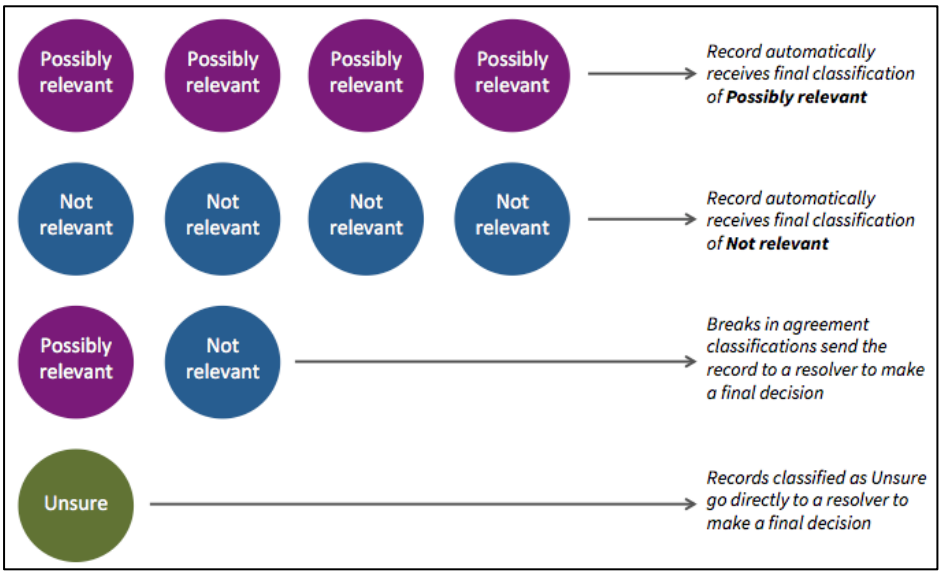
Cochrane Crowd[9] was launched in 2016. It is an online platform that hosts 'microtasks' – small classification tasks all of which are centred on identifying and describing health research in a consistent and standardised way (an example task is shown in Figure 5.1)[10]. To date, over 17,500 people based in 158 countries have signed up and helped to classify over one million records sourced from bibliographic databases and trial registries. Brief, interactive training modules, that are mandatory for contributors to complete, accompany each micro-task in Cochrane Crowd. Providing comprehensive training is an important aspect in ensuring accurate classifications are made by individuals. Cochrane Crowd is open to anyone to join regardless of their experience and prior knowledge of health research.

Figure 5.1 Screen shot of the Cochrane Crowd RCT identification task



In addition to providing training, each task is supported by an agreement algorithm. This agreement algorithm is critical in a crowdsourced model such as this as it helps to ensure accurate *collective* decision-making. In Cochrane Crowd the main current study identification tasks employ an agreement algorithm where each record requires four consecutive agreement classifications for it to be deemed either a randomised controlled trial (RCT) or not. If that consecutive chain is broken either through disagreeing classifications or by *Unsure* classifications, the record goes to a resolver screener in Cochrane Crowd to make a final decision. In Cochrane Crowd there are three levels of screeners: standard screeners, expert screeners and resolvers. Everyone begins as a standard screener. Contributors can then progress to become expert screeners within tasks. With expert screeners the algorithm is slightly altered with more weight given to the classifications made by expert screeners than standard screeners. Resolvers are screeners with an exceptional track record in Cochrane Crowd and are tasked with making the final decision on records where contributors have either disagreed in their classifications or classified a record as *Unsure* (see Figure 5.2).

Figure 5.2 Infographic of the Cochrane Crowd agreement algorithm



Screen4Me

In April 2019 Cochrane launched its Screen4Me (S4M) workflow. The S4M workflow is a search results screening service available for Cochrane systematic reviews. It is comprised of three components brought together into one workflow: crowdsourcing via Cochrane Crowd, a component called Known Assessments which indicates records, and their corresponding labels of *RCT* or *Not an RCT*, that have *already been through* Cochrane Crowd, and the RCT machine learning classifier. Screen4Me is available to Cochrane review author teams to help in the identification of randomised

trials from the search results retrieved for specific reviews. At the time of writing, S4M has been used by over 60 review teams across 15 different Cochrane review groups[11,12].

Currently, each component part of Screen4Me works to identify potential RCTs in the data set. None of the three components therefore involve an assessment of whether the RCTs identified are relevant to the review. This is a limitation of the workflow as it stands. For this pilot study we modified the crowd component of the S4M workflow to assess the performance of a crowd in undertaking a *topic-based* screening task based on the review's inclusion criteria.

5.4 Aims and objectives

Our primary aims were to assess the accuracy and autonomy of a crowd in *collectively* classifying studies as either potentially relevant or not relevant for a specific systematic review using the expert classifications from the main study as the ground truth.
Specifically, we sought to determine:

- Crowd sensitivity, determined by the crowd's ability to collectively correctly identify the records that were for inclusion within the review.
- Crowd specificity, determined by the crowd's ability to collectively correctly identify the records that were not relevant to the review.
- Crowd autonomy, determined by the proportion of records which were sent to the crowd resolver for a final decision.

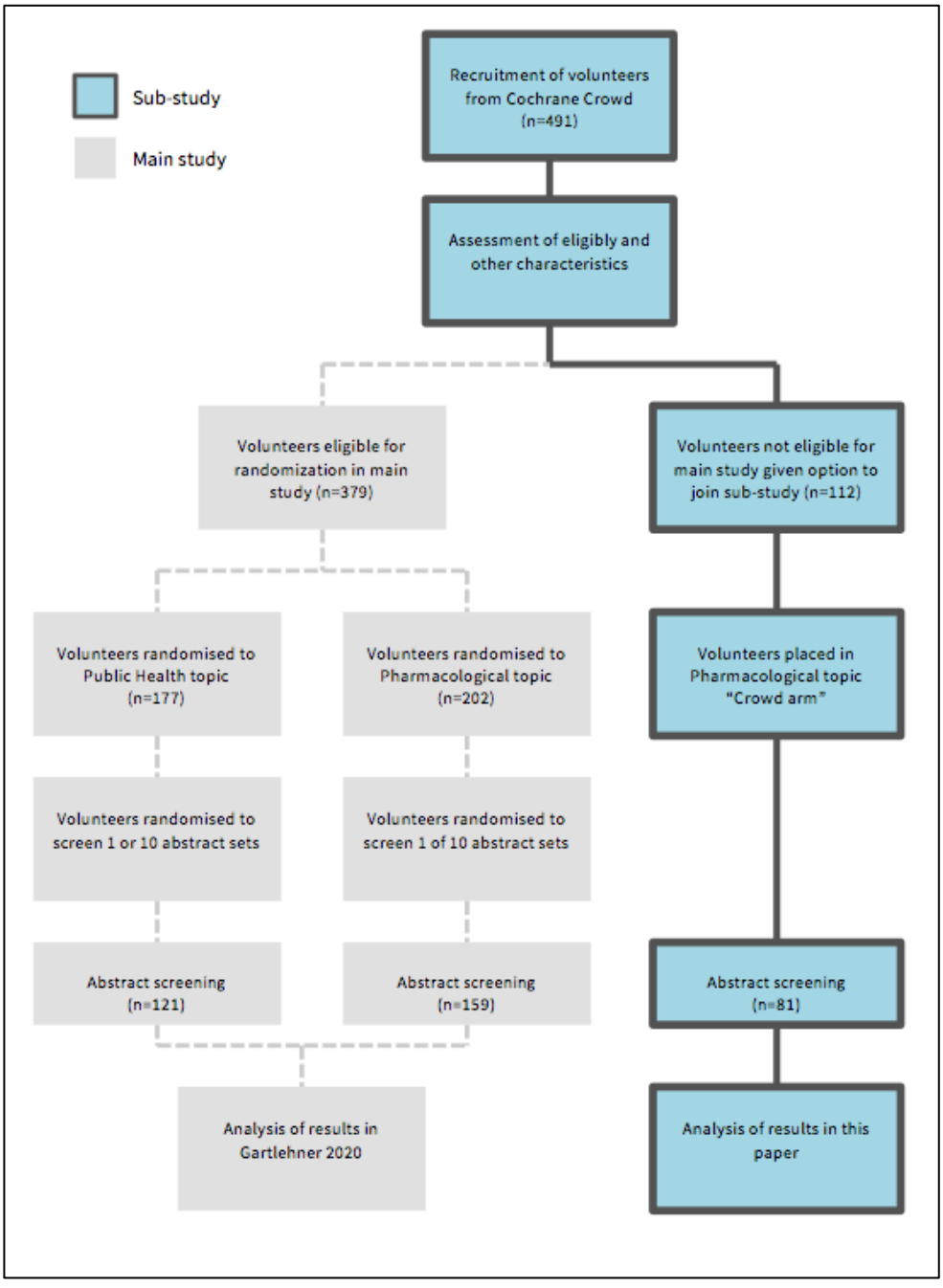
Our secondary aims were to assess the sensitivity of the other two unchanged components in the Screen4Me workflow: the Known Assessments and the RCT Classifier in terms of their ability to not reject any included RCTs.

5.5 Methods

This exploratory study was run as an additional sub-study to a larger main study[13]. In brief, the main study sought to assess single versus dual screening of abstracts for two systematic reviews: one pharmacological review and one public health review. The main study randomised eligible participants to one of the two candidate reviews. To be eligible for randomisation, participants had to have experience of screening search results for a specific review. They were recruited from the existing Cochrane Crowd community and through a number of student networks. Participants who did not meet the inclusion criteria for the main study (due to not having the pre-requisite experience

of having worked on a specific review before) were offered the option to join the modified Screen4Me sub-study. The protocol of the main study was registered in the Open Science Framework (https://osf.io/3jyqt) and the results have been described by Gartlehner and colleagues[13]. See Figure 5.3 for the flow of participants for the main and sub-study.

Figure 5.3 Participant flow for main and sub-study



Data source for abstract screening

For the modified Screen4Me arm, we used the pharmacological review as the review against which to compare crowd classifications against a gold standard[14]. As previously described by Gartlehner

and colleagues, the review chosen as the data source for this evaluation required the following characteristics: 1) it had to be a review focused on efficacy or effectiveness of an intervention or interventions; 2) the search needed to have yielded at least 2,000 results and contain a fairly high prevalence of included studies (at least 40); 3) the decisions regarding inclusions and exclusions of abstracts needed to have been based on current gold standard methods (i.e. performed by two screeners independently) as recommended by Cochrane's MECIR standards[15] and the Cochrane Handbook[16]. We did not use the other review used in the main study as it was a non-RCT based review. We would therefore not have been able to use the RCT Classifier or Known Assessments components of S4M.

The process and study participants

The records were first loaded into the Cochrane Register of Studies (CRS). The CRS is Cochrane's reference and study management software used by Cochrane Information Specialists to manage the search results for Cochrane systematic reviews. Once loaded, the records were sent through the *current* Screen4Me Known Assessment component. Here, previously flagged *Not RCT* records were put into a sub-folder. Any previously flagged RCT records were left to flow through to the other components. Next, the RCT Classifier component scored and divided the remaining records into *Not RCT* records, which were placed into a sub-folder, and *Possible RCT* records which were left to flow through to the modified crowd component of the workflow.

We recruited the study participants through emailing the existing Cochrane Crowd community and via a number of professional and student networks. Interested potential participants were then invited to complete a questionnaire. This was designed to assess their eligibility to join the main study. The participants who were not eligible to join the main study were allocated to this modified Screen4Me sub-study.

Each participant in this sub-study went through an interactive training exercise comprised of 15 example records. This training module was a bespoke module developed specifically to train contributors in the inclusion and exclusion criteria for the review. In order to be able to proceed to the task, participants had to achieve 80% (12) or more on the training module. Participants could repeat the training module as often as they liked. Once they had successfully completed the training exercise participants proceeded to abstract screening. We put no limit on the number of abstracts a participant could screen. We applied the same agreement algorithm used for most tasks on Cochrane Crowd, which meant that abstracts required four consecutive agreeing classifications for

the record to be deemed either possibly relevant or not relevant. Disagreeing classifications would break the consecutive chain. Where this happened, the records would be sent to a crowd resolver. The resolver would make the final decision on the record (as shown in Figure 5.2). The resolver screener used for this study was the main Cochrane Crowd resolver for the standard RCT Identification task on Cochrane Crowd. The resolver was required to do the same compulsory training module as the other participants. She was given no additional training or guidance. We did not provide contributors with an *Unsure* classification option.

We anticipated that the crowd would likely screen all the records in the data set before the end of the main study period. In order to enable potential participants to continue to enrol and join the modified Screen4Me sub-study (if not eligible for the main study), and to provide us with some replication data, we decided that should the crowd finish screening the batch of records, we would send the 'finished' records back through the crowd again to see whether we achieved the same final classification on the record after the first run through. In doing so, we enabled it so that completed records that went back to the crowd could not be re-screened by the same crowd contributors who had screened them the first time around.

Data collection and statistical analysis

We counted the number of relevant items (included studies) identified correctly (the 'true positive' count (TP)); the number of irrelevant items (not included studies) correctly identified as such (the 'true negative' count (TN)); the number of relevant items incorrectly classified as irrelevant (the 'false negative' count (FN)); and the number of irrelevant items, incorrectly classified as relevant (the 'false positive' count (FP)). We then calculated the crowd's collective performance in terms of sensitivity (the crowd's ability to classify relevant records correctly), and specificity (the crowd's ability to exclude irrelevant records correctly) as:

Crowd sensitivity:

$$\frac{TP}{TP + FN}$$

Crowd specificity:

$$\frac{TN}{TN + FP}$$

Crowd autonomy is the proportion of records that the crowd can process without requiring resolution by 'resolver' crowd members:

$$\frac{No.\,of\ records\ not\ requiring\ resolution}{Total\ number\ of\ records\ in\ dataset}$$

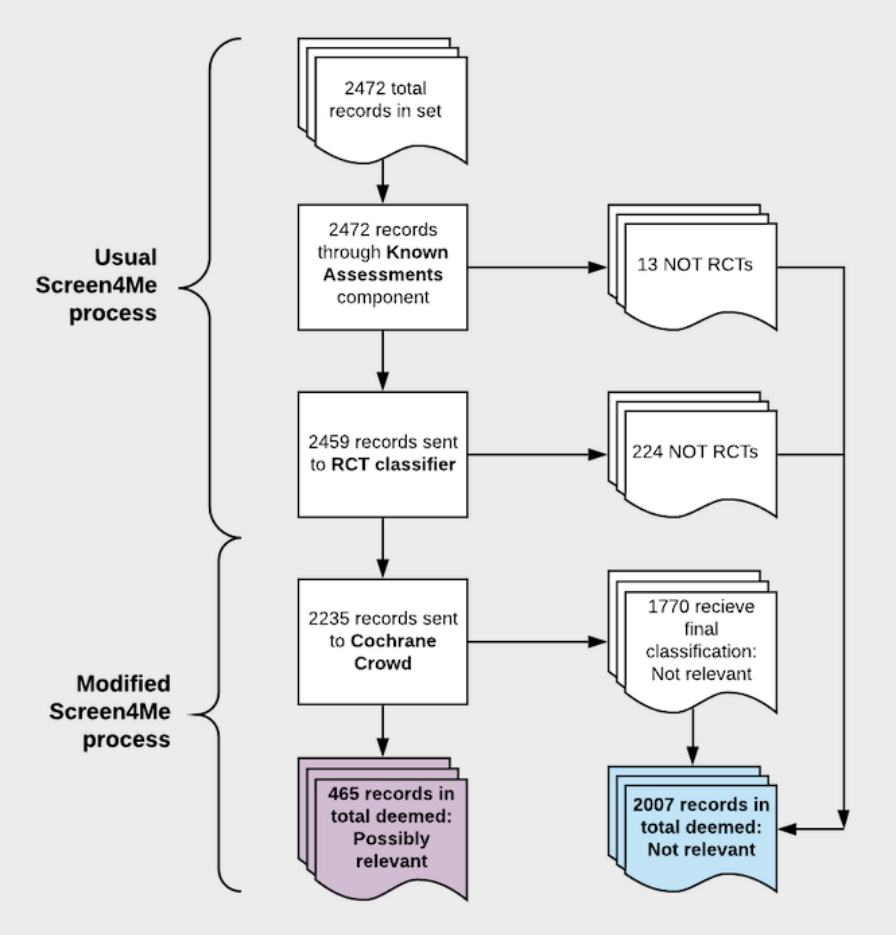
We conducted all statistical analyses in SPSS v26.

5.6 Results

Data flow through each of the Screen4Me components

The total record set was made up of 2472 records. The Known Assessments component of Screen4Me identified 13 records that had already been through Cochrane Crowd and received a final classification of *Not-RCT*. The RCT machine learning classifier then identified a further 224 records as being highly likely *Not-RCTs*. The remaining records, 2235 were then sent to Cochrane Crowd where the crowd collectively rejected 1770 as being *Not Relevant* and collectively classified 465 as being *Possibly Relevant* (see Figure 5.4).

Figure 5.4 Data flow diagram showing each Screen4Me component. Note: this figure shows the flow of records from the main data set. It does not include the replication data set



Crowd characteristics

In total 112 people were assigned to this sub-study (see Table 5.1 for crowd characteristics). Of these, 81 (72%) completed the training and went on to screen records in the live task.

Table 5.1 Characteristics of the participants assigned to the modified Screen4Me arm

Participant characteristics	n=112
Age (mean and IQR, years)	37.7 (26-45)
Gender (proportion, female, n=36*)	61.1%
English native speaker	15.5%
Number of countries of residence (n=112)	43

*Because of a technical problem, data were not recorded for all participants

Crowd performance

In total, contributors in this sub-study made 11,789 classifications (8,657 for the original batch and 1281 for the replication batch, further 1,851 classifications were made on records being re-run through the crowd task, the replication set, but did not receive the required number of classifications for a final decision to be assigned). Contributors screened an average of 149 records (range: 5-701).

Using the agreement algorithm illustrated in Figure 5.2 whereby four consecutive agreeing classifications, made independently by four contributors, are required for the record to be deemed either possibly relevant or not relevant, the crowd's collective sensitivity on the original batch (n=2235) was 100%. In other words, all references to included studies were correctly identified as *possibly relevant* by the crowd. In terms of specificity, the crowd incorrectly classified 423 references as possibly relevant resulting in a specificity of 80.71%. On the small replication batch (n=361), crowd sensitivity was again 100% and crowd specificity was 62.43%

Crowd autonomy

The proportion of records that needed to be resolved (records for which individual contributors had made conflicting classifications) on the original batch (n=2235) was: 24.6% (551 records needed resolving). Of the records resolved, 13 were for included studies and 538 were for not relevant records. On the replication batch the proportion that needed resolving was much higher at 52.9%. However, this figure should be treated with caution as the replication batch was made up of records that had 'completed' screening by the end of the study period. Records sent to be resolved would likely have completed sooner than records awaiting final classifications by normal contributors. If the study had continued for longer it is likely the proportion needing resolution would have decreased.

5.7 Discussion

In terms of performance these results are very encouraging. In both the original and replication data sets no included studies were missed. Overall, the modified Screen4Me workflow delivered a significant 81.2% workload reduction in terms of number of records deemed potentially relevant at the end of the process. With the Known Assessments component of this study, only a very small number of records matched up with records that had already been through Cochrane Crowd as part of the RCT identification task. This is due to the fact that currently the Known Assessments in Screen4Me use Embase accession numbers to identify records that have already been through Cochrane Crowd. In the data set used for this study few records had Embase accession numbers. In addition, the RCT Classifier did not contribute substantially to the overall workload reduction (9.1%). This is likely because within the set of search results used for the study, the prevalence of randomised and quasi-randomised trials was high due to there being a high number of RCTs in this domain area and the use of a methodological search filter applied to the search.

There are a number of factors that likely contributed to the success of the crowd component for this sub-study, all of which have implications for being able to scale this modified Screen4Me workflow. The first factor is the training module, which we developed specifically for this task. Developing a bespoke training module is a resource-intensive task. In the current Screen4Me workflow contributors are tasked with only identifying certain study designs and so tailored training is not required. One possible future approach that would negate the need to develop a new training module for each and every review would be to task the crowd with identifying certain components of the review's PICO (Population, Intervention, Comparator and Outcomes) rather than requiring a full eligibility assessment. An example might be to task the contributor with identifying all RCTs and all pharmacological studies. This would also have the added advantage of collecting useful additional metadata (i.e., that a record is either describing a pharmacological study or not) that could then be re-used when the same record enters the system for another review. This approach would require some training but the training would likely be applicable to multiple review questions, rather than to just one review.

Another factor is the crowd itself. Though we were not primarily interested in individual accuracy measures, mean individual sensitivity on those who screened 100 or more records was high (92%), especially given that the contributors assigned to this sub-study were those that did not meet the systematic review experience required to join the main study. This high accuracy could in part be explained by the fact that while participants in this sub-study may have lacked experience of working

on specific reviews, some had experience screening citations in Cochrane Crowd (18 had screened 100 or more records in the main RCT ID task, with eight of those having screened over 1000 records). Those 18 who had done 100 or more records screened 3,676 out of 11,789 (31%) of records in this study. However, another interesting finding is that only 15% of contributors in this arm reported that English was their first language. This makes their accuracy all the more impressive. One limitation of this study was that due to a technical error we did not collect much data on the demographics of the participants assigned to this arm. This is an area we hope to address in future studies as it could have significant implications on crowd performance and choice of appropriate agreement algorithms.

The agreement algorithm clearly played an important role in achieving 100% sensitivity. This finding is supported by previous studies indicating that with the use of an appropriate algorithm, high accuracy can be achieved[10,17,18]. We opted to use the same algorithm that is in place for the RCT identification task in Cochrane Crowd. However, with the classifications generated by the crowd for this task we will now be able to run some simulations to assess whether, for example, three consecutive agreements instead of four would have been adequate.

In addition, critical to the success of this crowd task was the involvement of the crowd resolver. As with the rest of the participants in this arm, the crowd resolver was new to this task and exposed only to the training module; she was given no additional guidance. This was deliberate as an important part of this pilot study was to understand better what guidance a crowd resolver might need for a topic-based assessment task. Having suitable resolver capacity is important. Without it author teams could expect a lower workload reduction. For example, if we had not used a resolver for this study (i.e., if we had simply assigned all records with disagreeing classifications to the final *Possibly relevant* pot) workload reduction would have fallen from 81.2% to 55.4%. That said, a 55% reduction is still substantial indicating that one viable approach could be to not use a resolver at all and have the author team screen both the records that needed resolving as well as those that get a final decision of “possibly relevant”.

The final factor is the review question itself. This was perhaps a relatively straightforward question. It would be interesting to see how well the crowd performed with identifying relevant RCTs for a complex intervention. This then raises a further question around acceptable levels of sensitivity for a crowdsourced approach. No method is infallible; current methods involved in the study identification process do not achieve 100% sensitivity. For example, the Cochrane highly sensitive

RCT methodological filter has been reported achieving 98.4%[19]. In addition, recommended activities such as citation checking and peer review are designed to act as safety nets to capture studies missed by earlier processes.

Our study has two main limitations: the first is that the size of the original data set and replication set were both relatively small. More work is needed with larger data sets and covering a variety of healthcare domains and questions, as well as review type outputs. The second limitation concerns the set of records used for training the crowd. The training batch was developed retrospectively by those who were already familiar with the studies that had been included in the review. This may have influenced the range of examples used within the training. Developing training prospectively i.e., before the included studies are known (as would be the case for real use) might produce different and potentially less effective training.

5.8 Conclusions

The modified Screen4Me approach used for this pilot study produced highly encouraging results. The crowd achieved 100% sensitivity and 80.71% specificity. Steps should now be taken to explore a range of use cases to identify those where a crowd approach of this nature could make a significant difference to workload reduction in study identification. As Cochrane and other review producers explore the role of other review types, including living systematic reviews[20] and rapid reviews[21], harnessing the potential of machine learning and crowdsourcing could bring significant efficiencies with limited impact on quality. Wide-scale adoption of these approaches will have operational and ethical implications including co-ordination of crowd effort and how the crowd should be rewarded and acknowledged for their increasing contribution in review production.

5.9 Author contributions

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Gordon Dooley: conceptualisation, methodology, resources, data curation, writing – reviewing and editing

Lisa Affengruber: conceptualization, methodology, writing - review & editing

Gerald Gartlehner: conceptualization, methodology, writing - review & editing

5.10 Abbreviations

CENTRAL Cochrane Central Register of Controlled Trials

CRS	Cochrane Register of Studies
MECIR	Methodological Expectations of Cochrane Intervention Reviews
q-RCT	Quasi-randomised controlled trial
RCT	Randomised controlled trial
S4M	Screen4Me

5.11 References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545.
2. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, Gruen RL. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. PLOS Medicine 2014;11(2):e1001603.
3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Annals of Internal Medicine 2007;147:224-233.
4. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021;133:140-151.
5. Marshall IJ, Noel-Storr AH, Kuiper J, Thomas J, Wallace B. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Research Synthesis Methods 2018;9(4):602-614.
6. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020;127:142-150.
7. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. American Journal of Preventive Medicine 2014;46(2):179-187.
8. Tucker JD, Day S, Tang W, Bayus B. Crowdsourcing in medical research: concepts and applications. PeerJ 2019;7:e6762.
9. Cochrane Crowd: https://crowd.cochrane.org [Accessed 05 January 2022].
10. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021;133:130-139.
11. Noel-Storr A. Screen4Me: using machine learning and crowdsourcing for evidence identification. Cochrane Tech Symposium, 21 October 2019, Santiago, Chile.
12. Thomas J, Noel-Storr A, McDonald S. Data reuse, machine learning, and crowdsourcing in Screen4Me: https://www.youtube.com/watch?v=WGeHo9dWS1k Cochrane Tech Symposium, 21 October 2019, Santiago, Chile [Accessed 28 December 2021].

13. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, König F. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. 2020;121:20-28.
14. Gartlehner, G., Gaynes, B.N., Forneris, C., and Lohr, K.N. Comparative benefits and harms of antidepressant, psychological, complementary, and exercise treatments for major depression. Ann Intern Med. 2016;165:454.
15. Higgins JPT, Lasserson T, Chandler J, Tovey D, Thomas J, Flemyng E, et al. Methodological Expectations of Cochrane Intervention Reviews. London: Cochrane; 2019. Available at https://community.cochrane.org/sites/default/files/uploads/inlinefiles/MECIR%20October%2019%20Final%20Online%20version.pdf
16. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook
17. Nama N, Sampson M, Barrowman N, et al. Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study. Journal of Medical Internet Research 2019; 21(4): e12953
18. Mortensen JM, Adam GP, Trikalinos TA, Kraska T, Wallace BC. An exploration of crowdsourcing citation screening for systematic reviews. Research Synthesis Methods 2016;8(3):366-386.
19. McKibbon KA, Wilczynski NL, Haynes RB; Hedges Team. Retrieving randomized controlled trials from Medline: a comparison of 38 published search filters. Health Information and Libraries Journal 2009;26(3):187-202.
20. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. 2014;18;11(2):e1001603.
21. Khangura, S., Konnyu, K., Cushman, R. et al. Evidence summaries: the evolution of a rapid review approach. Syst Rev 1, 10 (2012). https://doi.org/10.1186/2046-4053-1-10.

Chapter 6

Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study

This original manuscript was published in *BMC Medical Research Methodology*

Citation: Noel-Storr AH, Redmond P, Lamé G, Liberati E, Kelly S, Miller L, Dooley G, Paterson A, Burt J. Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study. BMC Med Res Methodol. 2021 Apr 26;21(1):88.

DOI: doi: 10.1186/s12874-021-01271-4

URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01271-4

6.1 Abstract

Background

Crowdsourcing engages the help of large numbers of people in tasks, activities or projects, usually via the internet. One application of crowdsourcing is the screening of citations for inclusion in a systematic review. There is evidence that a 'crowd' of non-specialists can reliably identify quantitative studies, such as randomised controlled trials, through the assessment of study titles and abstracts. In this feasibility study, we investigated crowd performance of an online, topic-based citation-screening task, assessing titles and abstracts for inclusion in a single mixed-studies systematic review.

Methods

This study was embedded within a mixed studies systematic review of maternity care, exploring the effects of training healthcare professionals in intrapartum cardiotocography. Citation-screening was undertaken via Cochrane Crowd, an online citizen science platform enabling volunteers to contribute to a range of tasks identifying evidence in health and healthcare. Contributors were recruited from users registered with Cochrane Crowd. Following completion of task-specific online training, the crowd and the review team independently screened 9,546 titles and abstracts. The screening task was subsequently repeated with a new crowd following minor changes to the crowd agreement algorithm based on findings from the first screening task. We assessed the crowd decisions against the review team categorizations (the 'gold standard'), measuring sensitivity, specificity, time and task engagement.

Results

78 crowd contributors completed the first screening task. Sensitivity (the crowd's ability to correctly identify studies included within the review) was 84% (N=42/50), and specificity (the crowd's ability to correctly identify excluded studies) was 99% (N=9373/9493). Task completion was 33 hours for the crowd and 410 hours for the review team; mean time to classify each record was 6.06 seconds for each crowd participant and 3.96 seconds for review team members. Replicating this task with 85 new contributors and an altered agreement algorithm found 94% sensitivity (N=48/50) and 98% specificity (N=9348/9493). Contributors reported positive experiences of the task.

Conclusion

It is feasible to recruit and train a crowd to accurately perform topic-based citation-screening for mixed studies systematic reviews, though resource expended on the necessary customised training

required should be factored in. In the face of long review production times, crowd screening may enable a more time-efficient conduct of reviews, with minimal reduction of citation-screening accuracy, but further research is needed.

6.2 Background

Systematic reviews are essential to locate, appraise and synthesize the available evidence for healthcare interventions[1]. Citation-screening is a key step in the review process whereby the search results identified from searches often performed across multiple databases, are assessed based on strict inclusion and exclusion criteria. The task is performed through an assessment of a record's title and abstract (what we term 'citation'). The aim is to remove records that are not relevant and determine those for which the full-text paper should be obtained for further scrutiny. This is no easy task. One study found a mean of 1781 citations were retrieved in systematic review searches (ranging from 27 to 92,020 hits retrieved from searches), from which a mean of 15 studies were ultimately included in each review: an overall yield rate of only 2.94%[2]. In part driven by the resources required to undertake citation-screening, reviews are typically time and labour intensive, taking an average of 67.3 weeks to complete[2]. Moreover, the challenge of locating relevant evidence for reviews is becoming ever greater: over the last decade, research output has more than doubled, and approximately 4000 health-related articles are now published every week[3,4,5]. New approaches are needed to support systematic review teams to manage the screening of increasing numbers of citations.

One possible solution is crowdsourcing[6]. Crowdsourcing engages the help of large numbers of people in tasks, activities or projects, usually via the internet. Such approaches have been trialled in a number of health research areas, using volunteers to process, filter, classify or categorise large amounts of research data[7,8]. More recently, the role of crowdsourcing in systematic reviews has been explored, with citation-screening proving a feasible task for such a crowdsourced approach[9,10,11,12]. Cochrane, an international not-for-profit organization and one of the most well-known producers of systematic reviews of RCTs, is an early adopter of the use of crowdsourcing in the review process. Since the launch of their Cochrane Crowd citizen science platform in May 2016, over 18,000 people from 158 countries have contributed to the classification of over 4.5 million records[13].

To date, crowdsourcing experiments in citation-screening have often focussed on identifying studies for intervention reviews, with included studies often limited to randomised or quasi-randomised

controlled trials[9,10,11]. Whilst this supports traditional systematic reviews concerned with evidence of effectiveness, an increasing number of reviews in health and healthcare now address research questions requiring the identification and synthesis of both quantitative and qualitative evidence[14,15,16]. Much less has been done to explore the effectiveness of using a crowd to screen citations for complex, mixed studies reviews. One study by Bujold and colleagues used a small crowd (n=15) to help screen the search results for a review on patients with complex care needs. The study was not a validation study and so does not provide crowd accuracy measures; however, the authors concluded that crowdsourcing may have a role to play in this stage of the review production process, bringing benefit to the author team and crowd contributor alike[17].

6.3 Aims and objectives

The aim of this feasibility study was to investigate whether a crowd could accurately and efficiently undertake citation-screening for a mixed studies systematic review. Our objectives were to assess:

(1) Crowd sensitivity, determined by the crowd's ability to correctly identify the records that were subsequently included within the review by the research team

(2) Crowd specificity, determined by the crowd's ability to correctly identify the records that were subsequently rejected by the research team

(3) Crowd efficiency, determined by the speed of the crowd in undertaking the task and the proportion of records which were sent to crowd resolvers for a final decision

(4) Crowd engagement, determined by qualitative assessment of their satisfaction with the citation-screening task and readiness to participate

6.4 Methods

The systematic review

This study was embedded within a mixed studies systematic review exploring training for healthcare professionals in intrapartum electronic fetal heart rate monitoring with cardiotocography[18]. Cardiotocography is widely used in high-risk labours to detect heart rate abnormalities which may indicate fetal distress, in order to intervene or expedite birth as required. The aim of the review was to examine the effects of training for healthcare professionals in intrapartum cardiotocography and to assess evidence for optimal methods of training. All primary empirical research studies that evaluated cardiotocography training for healthcare professionals were eligible for inclusion in the review, irrespective of study design.

Crowdsourcing platform

The citation-screening task was hosted on the Cochrane Crowd platform[19]. This citizen science platform offers contributors a range of small, discrete tasks aimed at identifying and describing evidence in health and healthcare. There is no requirement for contributors to have any relevant background in research or healthcare: anyone with an interest in helping may volunteer to do so. The main activities available to contributors to Cochrane Crowd are tasks aimed at identifying or describing reports of RCTs. In these, contributors are asked to look at a series of citations (titles and abstracts of journal articles or trial registry records) and classify them as either reporting an RCT or not reporting an RCT.

Cochrane Crowd employs two strategies to ensure accuracy of contributor screening decisions. Firstly, each contributor is required to complete an interactive, customised training module prior to commencing each task, designed to improve their likelihood of making the correct classification for each record ('individual accuracy'). Secondly, each record is reviewed and classified by multiple contributors, with an agreement algorithm used to improve the crowd's likelihood of making the correct classification for each record ('collective accuracy'). Typically, either three or four (depending on the task and experience of the individual screeners) consecutive identical classifications are required for a record to be labelled as either an RCT or not an RCT and removed from the screening task. Breaks in the consecutive chain, or 'unsure' classifications, are reviewed by crowd 'resolvers', highly experienced crowd contributors who make a final classification decision. Contributors are also supported in the screening task by the use of pre-specified highlighted words and phrases added automatically to each record they assess. These highlights flag notable parts of a title or abstract, and are used to direct a screener's attention to key phrases or words which may help them make a classification decision (Figure 6.1). On Cochrane Crowd, red highlights are used to flag words that may appear on citations that are unlikely to be relevant, whilst yellow highlights indicate particularly relevant keywords (such as 'randomly assigned', in an RCT classification task).

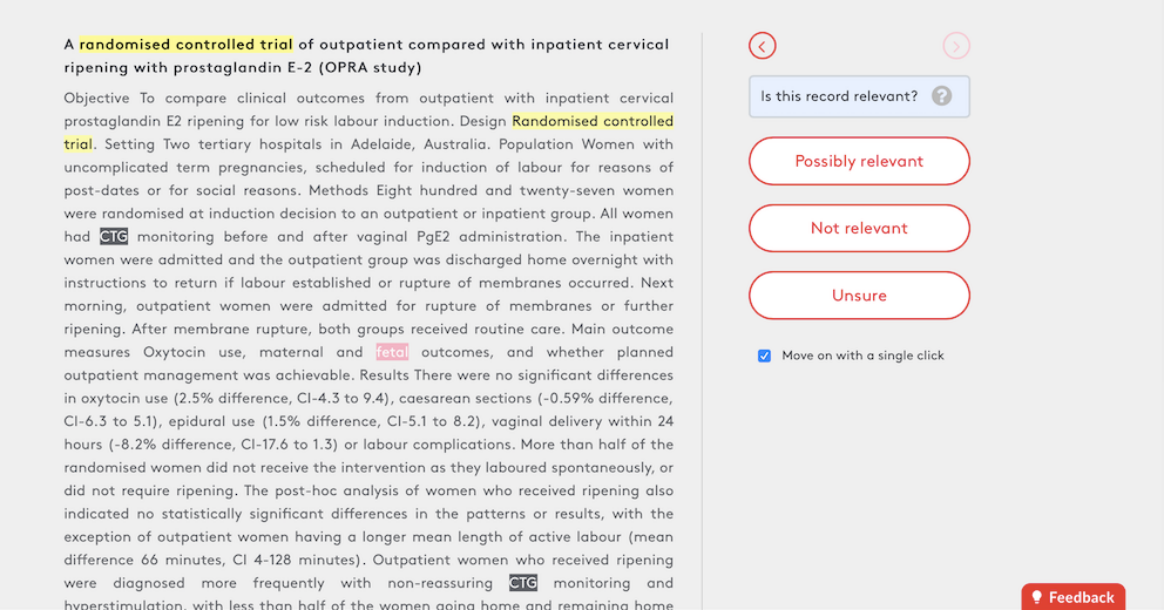
Citation screening task

Citations for screening were generated through searches conducted according to the review protocol[18]. The initial search identified a total of 10,223 records; after the removal of duplicates, 9,546 records remained for screening.

We created two identical citation-screening tasks on the Cochrane Crowd platform: one task for the systematic review author team (n = 4), comprising experienced researchers undertaking the

systematic review (hereafter the 'review team'), and the second task for existing registered users on the Cochrane Crowd platform (hereafter 'the crowd' – see details below). The screening task presented all contributors with a series of journal titles and abstracts identified from the review searches and asked them to determine whether each record may be relevant to the topic of the review (see Figure 6.1). Unlike previous tasks hosted on Cochrane Crowd, this task was a 'topic-based' assessment task whereby the crowd was tasked with determining the potential topic relevance of each citation rather than assessing it based on study design. Three possible classification choices were available: *Possibly relevant*, *Not relevant*, and *Unsure* (for more information on what these classification terms meant see the *Task Training* section below*)*. Terms for highlighting were pre-specified and based on the review search terms used (e.g., cardiotocography; training; course) and added as yellow highlights to records where they appeared. Red warning highlights were not used for this task.

Figure 6.1 Screenshot from the task hosted on the Cochrane Crowd platform



The crowd

Cochrane Crowd contributors were approached via email invitation, giving details of the review and the citation-screening task. All crowd contributors who had screened 100 or more records in Cochrane Crowd's RCT identification task were invited to participate: this is a standard entry criterion for all more complex tasks on the platform. A 'frequently asked questions' document gave more detail about the study. We offered a certificate of participation to all contributors, and acknowledgement in the published systematic review for those who screened 250 or more records. The Cochrane Crowd community is open to anyone with an interest in healthcare including

healthcare professionals and students, researchers, patients, carers and members of the general public.

Task training

We developed a task-specific online training module, hosted on Cochrane Crowd. This consisted of an introduction to the review topic and 20 practice records. It also included a description of the classification options available: *Possibly relevant*, *Not relevant*, and *Unsure* and guidance on when to use which option. In brief, *Possibly relevant* was to be used when records described or reported on both healthcare professional training and cardiotocography; *Not relevant* was to be used for records that were clearly not about both of one of those elements; *Unsure* was to be used if a participant was not sure either because the record contained very little information (for example a title-only record) or because the available information was simply not clear. There was no pass mark for the training module and contributors could repeat the training as often as they wished. Both the crowd and the author team completed the same set of training records.

Agreement algorithm

We used different agreement algorithms for the review team and the crowd. For the review team, we used the standard recommended algorithm for citation-screening for systematic reviews as recommended by the Cochrane Handbook[20]. Two independent contributors assessed each record and made a judgement as to whether the record was potentially relevant, not relevant or that they were unsure. Records that received discordant assessments had a final decision determined by a third member of the review team.

The agreement algorithm for the crowd required each record to be assessed by three independent contributors. Records that received discordant assessments (e.g., two *Possibly relevan*t and one *Not relevant*) were decided by a separate 'crowd resolver', in this case a highly experienced crowd contributor and data curation specialist selected by Cochrane Crowd (Table 6.1). Crowd resolvers are Cochrane Crowd contributors who have achieved exceptional accuracy on specific crowd tasks or who have extensive experience screening citations for Cochrane systematic reviews.

Table 6.1 The agreement algorithm used for the crowd task. Breaks in the consecutive chain or any 'unsure' classification sends the records to resolvers to make the final decision

Decision 1	Decision 2	Decision 3	Final decision
Possibly relevant	Possibly relevant	Possibly relevant	Possibly relevant
Not relevant	Not relevant	Not relevant	Not relevant
Possibly relevant	Possibly relevant	Not relevant	Resolver decision
Possibly relevant	Not relevant	NA	Resolver decision
Not relevant	Not relevant	Possibly relevant	Resolver decision
Not relevant	Possibly relevant	NA	Resolver decision
Unsure	NA	NA	Resolver decision

Calculating crowd sensitivity, specificity and efficiency

We calculated crowd sensitivity by identifying the proportion of citations which were subsequently included within the review by the author team and which were also correctly identified as *Possibly relevant* by the crowd (Table 6.2). We calculated crowd specificity by identifying the proportion of citations which were subsequently rejected from inclusion within the review by the research team and which were also rejected from inclusion by the crowd (*Not relevant*). We additionally considered crowd efficiency in terms of the speed at which the crowd completed the citation-screening task derived from the time taken for each screening classification (automatically logged by the Cochrane Crowd platform) as well as the proportion of records which were sent to crowd resolvers for a final decision.

Table 6.2 Outcome variables assessed

Outcome variable	Definition
Final sensitivity	The number of citations deemed relevant by the research team (included in the final set of studies for the review after both screening and full-text review) that were correctly identified by the crowd (true positives), divided by the number of true positives plus the number of citations included in the final set of studies by the research team that were not included by the crowd (false negative)
Screening specificity	The number of citations excluded by the crowd that were also excluded from the final set of studies by the research team (true negative), divided by the number of true negatives plus the number of citations included by the crowd that were not deemed relevant by the research team after both screening and full-text review (false positive).
Efficiency	Total time taken for the crowd versus the research team to complete the screening task.

Replication of citation-screening task

Following completion of the citation-screening task by the crowd and assessment of the initial findings, we amended the crowd agreement algorithm to include two resolvers acting independently (rather than one resolver, as used in the first round). In this replication exercise, the two resolvers each screened all records that needed resolving. Where there was a disagreement between resolver classifications (i.e., one *Possibly relevant* resolver classification and one *Not relevant* resolver

classification, since resolvers could not classify a citation as *Unsure*), the citation was to be kept in. We then repeated the citation-screening task for the crowd, using the adjusted resolver algorithm. As before, we invited all registered users of Cochrane Crowd who had screened 100 or more records in the RCT identification task. Those who had already taken part in the first round were excluded.

Evaluation questionnaire

In order to evaluate crowd motivations and engagement, all crowd contributors (in both the original and replication task) were asked to complete a brief online survey at the end of the task. The questionnaire covered areas including motivation to participate; experience of citation-screening; and brief socio-demographic details. Most questions were picklist-type questions but with many providing a free text option in addition. All questions were optional.

6.5 Results

Contributors

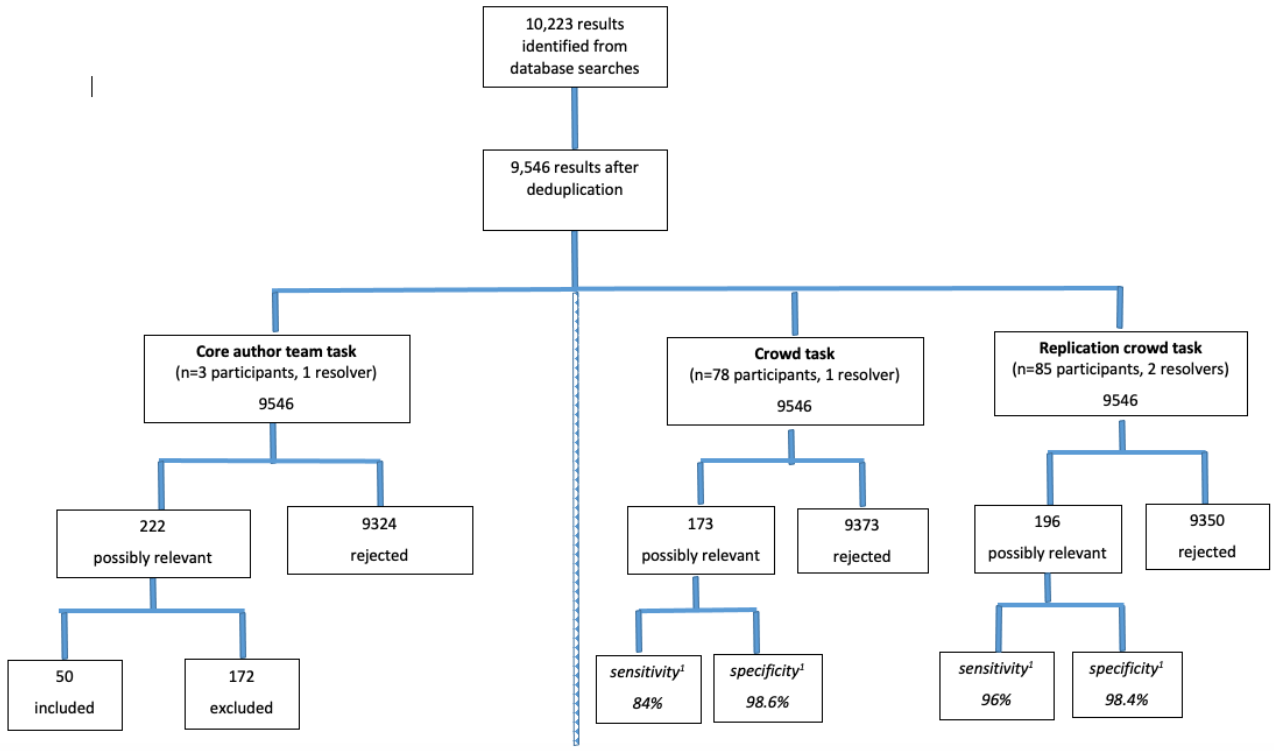
Within the review team, three researchers undertook the screening task, and one researcher acted as resolver. We invited 903 Cochrane Crowd contributors to take part in citation-screening; of these, 78 (9%) participated, with 48 (62%) screening over 250 records each. An additional contributor acted as the resolver. The response rate to the post-task survey was 63/78 (81%). Fifty-one percent of respondents worked in a health-related area, 10% were patients, and 5% were carers.

Crowd sensitivity

Following citation-screening, the review team classified 222 of the 9,546 records as *Possibly relevant* to the review, whilst the crowd classified 173 records as *Possibly relevant* (Figure 6.2). Following full-text assessment of the 222 *Possibly relevant* citations, the review team identified 50 studies for inclusion within the review. All 50 studies had been classified by individual crowd contributors as either *Possibly relevant* or *Unsure*. However, eight of these studies which received at least one *Unsure* classification or conflicting classifications by crowd contributors were subsequently rejected by the crowd resolver. This reduced overall crowd sensitivity to 84%.

Figure 6.2 Citation screening decisions made by the review team and the crowd

[1] Sensitivity and specificity compared to core author team as reference standard



Crowd specificity

The review team rejected 9,324 records at the citation-screening stage, and a further 172 at the full-text stage, bringing the total rejected by the author team to 9,493. The crowd rejected 9,373 records at the citation-screening stage, leading to a crowd specificity of 98.6%.

Crowd efficiency

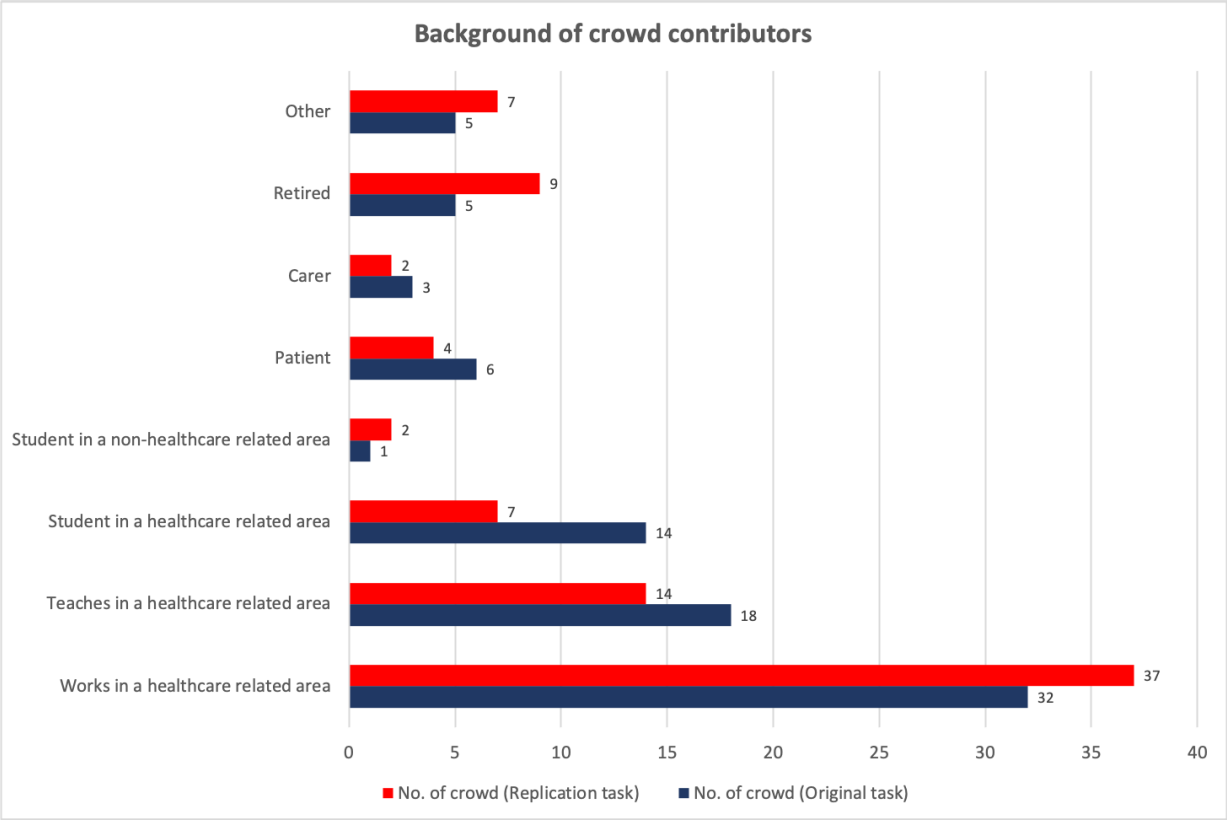
The crowd took a total of 33 hours from when the task went live on Cochrane Crowd to complete screening of the 9,546 records. This included the resolution of records where the crowd had discordant classifications or had classified a record as *Unsure*. The review team took a total of 410 hours to complete screening. However, crowd contributors took longer on average to screen an individual record compared to a member of the core author team (mean of 6.06 seconds per record for the crowd compared to 3.97 seconds per record for the core author team). For the crowd task, 677 (7.09%) records needed resolving; in the review team task 420 (4.39%) records needed to be resolved.

Replication of citation-screening task

The citation-screening task was replicated to assess consistency of crowd performance, and to evaluate a different agreement algorithm whereby two resolvers rather than one resolver assessed all records that needed resolving. This was because in the original task, resolver error had led to a reduction in crowd sensitivity.

Eighty-five participants contributed to the replication task. None of the 85 contributors for this task had taken part in the original task. There was little variation in the background of contributors in the replication study compared to the original contributors (Figure 6.3). The response rate to the post-task survey for the replication task was 64/85 (75%). Of those responders, 58% of respondents worked in a health-related area, 6% were patients, and 3% percent were carers.

Figure 6.3 Clustered bar chart showing crowd contributor backgrounds for original and replication tasks. 63 out of 78 (81%) participants completed the survey for the original task; 64 out of 85 (75%) participants completed the survey for the replication task



The crowd took 48 hours to complete the second citation-screening task. 889 (9.3%) records of the 9,546 screened were referred to the crowd resolvers, either due to discordant or *Unsure* classifications. No included study referred to the resolvers was subsequently rejected. Two previously included studies were however rejected by the crowd during the replicated task. Crowd

sensitivity improved from 84% in the original study to 96% (48 out of the 50 studies correctly classified as *Possibly relevant*) in the replication study. The crowd's specificity in the replication task was similar to the original value, at 98.4% in the repeated study compared to 98.6% in the original study.

Crowd engagement

The primary motivation to participate amongst respondents to the questionnaire was 'to do something for a well-respected organisation'. This was followed by the chance to get acknowledgement on a review. 97% of respondents reported enjoying the task in both the original and replication task surveys, and a similar proportion, again for both surveys, said that they found the task either easy or very easy (84% for the original task and 90% for the replication task). None of the respondents reported that they found the task difficult.

When asked whether they preferred the usual RCT identification task available in Cochrane Crowd or the new topic-focused task, the responses were evenly split between preferring the new topic-focused task or liking both tasks. Only 10% reported that they definitely preferred the RCT identification task (for the replication task, only 6% preferred the RCT task).

When asked about the use of highlighted words and phrases, 92% of respondents felt that they had been useful. This was also the case for the replicated task where 89% of respondents felt they had found them useful. At the end of the survey people had the chance to make any further comments. Thirty-six people made comments, with the vast majority reflecting their enjoyment of the task or the satisfaction of being involved in the feasibility study:

> *"I thought this was an excellent pilot project. If these were offered more frequently, I would assign my students to participate"*
>
> *"I think it is a very useful way to spend half an hour when I have the spare time, it made me feel connected, and it seemed to achieve a lot for the review"*
>
> *"Please do more. Please keep doing this. I feel much more connected to things, being offered a role however small in somebody's research, I value this immensely."*
>
> *"Was good to have a smaller task on offer as it felt more 'doable' and that my contribution would really make a difference"*

6.6 Discussion

In this feasibility study examining the potential for crowdsourcing citation-screening in a mixed studies systematic review, crowd contributors correctly identified between 84% and 96% of citations included in the completed review, and 99% of citations which were not included. On both occasions, citation-screening was completed by the crowd in two days or less. These results compare very favourably to other studies exploring topic-based crowd citation-screening and time outcomes[9,11,21] though direct comparisons are difficult due to the variation in review types and tasks being evaluated.

Whilst the sensitivity and specificity of the crowd appear high, there were misclassifications made by both contributors and resolvers when compared to the decisions of the review team. In the first citation-screening task, eight studies identified as potentially relevant by the crowd, and included in the review by the review team, were later rejected by the crowd resolver. In the replication task, the crowd collectively rejected two studies included in the review by the review team. With a strengthening of the resolver function in the replication task (with two resolvers working independently, rather than one resolver), no included studies that needed resolving were rejected, suggesting crowd sensitivity was boosted by the use of a more robust agreement algorithm. In part, this may be due to a decreased risk of screening fatigue with more than one resolver available to adjudicate screening disagreements or uncertainties[22] but also, as two recent studies have confirmed, a single screener (versus dual screening) is likely to miss includable studies[23,24].

Both of the studies rejected by the crowd in the replication task[25,26] were amongst the eight rejected by the resolver in the first round of the study. Rejection of these papers by the crowd may have been influenced by the highlighting of words and phrases in the records. The record for Blomberg 2016, contained only one highlighted word ('training'), whilst the Byford 2014 paper contained no highlighted words. In comparison, the records for other studies identified through screening tended to contain a larger number of highlighted words.

The relative speed of the crowd in completing the citation-screening was similar to previous tasks undertaken by Cochrane Crowd. In a series of pilot studies run in July 2017, contributors were tasked with screening search results for four Cochrane reviews. The number of results to screen within each review ranged from approximately 1000 to 6000: the mean time taken by the crowd to complete citation-screening was 24 hours[27]. Such speed enables a review team to move rapidly from search

results to full-text screening: an advantage when the time taken to complete many reviews means searches have to be updated and further screening and data extraction undertaken prior to publication. Whilst the increased speed of screening raises the potential for time and cost savings through using crowdsourcing, this does not account for the time taken to design, build and pilot the training and instructions for each review. This training was customised to the review, marking a departure from the RCT identification tasks normally hosted on Cochrane Crowd, and thus more resource intensive to develop. Therefore, the trade-off between speed of crowd screening and resources to enable crowd screening needs further scrutiny. With searches for mixed studies reviews often generating very high numbers of search results to assess, the time spent creating customised topic-based training modules might be well justified.

An alternative approach to crowd citation-screening might be to reframe the nature of the crowd screening task itself. This study, like others before it, asked the crowd to screen search results against the same criteria used by the core author team. This provides good comparative data for crowd performance calculations. However, a more effective approach may be to ask crowd screeners to focus on the identification of very off-topic citations, changing the overarching question from "*Does the record look potentially relevant?"* to *"Is the record obviously not relevant?"* Approaching crowd tasks for complex reviews in this way might make the compilation of the training module less time and resource intensive, as well as improving crowd sensitivity. The obvious detrimental impact would be on specificity, as a greater proportion of irrelevant records would be kept in. However, following 'first pass screening' from crowd contributors, author teams would be able to undertake title-abstract screening of a substantially smaller number of remaining records with a higher prevalence of potentially relevant records. To our knowledge, this approach has not been explored.

Another approach may be to explore the role of machine learning in combination with crowd effort. Machine learning classifiers are being used increasingly to help identify RCTs and other study designs[28,29,30,31]. Within a mixed studies context, the main challenge would be generating enough high-quality training data for the machine. However, for searches that retrieve a high volume of hits it may be feasible to build a machine learning model from a portion of crowd- or author-screened records, that could then either help to prioritise/rank remaining records by likelihood of relevance or be calibrated at a safe cut-off point to automatically remove the very low-scoring records.

Our findings are inevitably limited by their dependence on citation-screening of search results from only one systematic review. Different search results from different reviews may generate different sensitivity and specificity estimates in crowdsourced citation-screening. However, it is notable that there is little published evidence on how screening decisions vary: different expert review teams could also be anticipated to make different screening decisions when presented with the same set of search results. With a complex mixed studies review, the likelihood of human error, whether from the crowd or the 'expert' review team, is further increased: there is often limited information in abstracts to judge topical relevance. It is not clear what an acceptable level of crowd accuracy is, to be able to confidently use crowdsourcing without comparing crowd decisions to those of an expert review team. For reviews of evidence of effectiveness, there may be very little tolerance for divergence of decisions. For other reviews – such as the current example on training in the use of cardiotocography – overall review results may be little influenced by the inclusion or exclusion of a few studies of marginal quality and depth of information. The level of error deemed to be acceptable in relation to a specific degree of time saving may depend on both the type of review being conducted and the breadth and volume of potentially includable studies. These are factors that require determining if crowdsourcing in this way is to become an acceptable model of research contribution.

In terms of the generalisability of these results we should address the characteristics of the crowd participants. Whilst we recruited a non-selective crowd (contributors did not need to have any topic knowledge or expertise to be able to participate) we can see from the survey responses that many participants did have a healthcare background which may have made the task easier. In addition, in order to be able to participate in this study, potential participants had to have completed 100 assessments in another Cochrane Crowd task, RCT Identification. The RCT Identification task on Cochrane Crowd requires contributors to complete a brief training module made up of twenty practice records. While this task is different from the study task, it does mean that the participants were already familiar with screening citations within Cochrane Crowd. We therefore must exercise caution in generalising that a crowd consisting of either fewer healthcare professionals or those without any experience of screening citations, would perform as successfully.

Finally, the very positive responses from this study's participants were highly encouraging. However, successful, widespread implementation of crowdsourcing in this way brings with it a number of important ethical considerations. Providing meaningful opportunities for people to get involved with the research process must be matched by appropriate measures of acknowledgement and reward.

In this study named acknowledgement in the review proved a suitable reward but as crowdsourced tasks become more involving or challenging, as they no doubt will, it stands that the requisite reward should be greater. This then potentially presents a conflict with current academic publishing guidance, with criteria for authorship often requiring full involvement of all authors across all or many parts of the study. In some circumstances, payment might be appropriate, yet micro-payment or piece-rate models such as those used by Amazon Mechanical Turk have come under fire in recent years with studies revealing poor working conditions of an "unrecognised labour"[32,33]. As crowdsourcing in this way becomes more accepted as an accurate and efficient method of study identification, these ethical factors will need to be understood and addressed in parallel, for the benefit of both contributor and task proposer alike.

6.7 Conclusions

In support of a complex mixed-studies systematic review, a non-specialist crowd tasked with undertaking citation-screening performed well in terms of both accuracy and efficiency measures. Importantly, crowd members reported that they enjoyed being part of the review production process.

Further research is required to develop effective approaches to pre-task training for contributors to crowdsourced citation-screening projects, the refinement of agreement algorithms, and establishing 'acceptable' levels of performance (for example, by investigating the variation in performance by both crowd and 'expert' screening teams, such as clinicians).

Review teams, particularly those engaged in locating a broad range of evidence types, face significant challenges from information overload and long production times. With further refinements in its approach, crowdsourcing may offer significant advantages in terms of time-saving, building capacity, engagement with the wider evidence community and beyond, with a minimal loss to quality.

6.8 Author contributions

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Patrick Redmond: conceptualisation, methodology, resources, data curation, writing – reviewing and editing

Guillaume Lamé: conceptualisation, methodology, data curation, writing – reviewing and editing

Elisa Liberati: conceptualisation, data curation, writing – reviewing and editing

Sarah Kelly: conceptualisation, methodology, data curation, visualisation, writing – reviewing and editing

Lucy Miller: conceptualisation, data curation, writing – reviewing and editing

Gordon Dooley: conceptualisation, data curation, writing – reviewing and editing

Andy Paterson: conceptualisation, methodology, writing – reviewing and editing

Jenni Burt: conceptualisation, methodology, investigation, data curation, supervision, writing - reviewing and editing

6.9 Abbreviations

RCT	Randomised controlled trial
THIS Institute	The Healthcare Improvement Studies Institute

6.10 References

1. Mulrow CD. Rationale for systematic reviews. BMJ. 1994 Sep 3;309(6954):597-9. doi: 10.1136/bmj.309.6954.597. PMID: 8086953; PMCID: PMC2541393.
2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 27;7(2):e012545. doi: 10.1136/bmjopen-2016-012545. PMID: 28242767; PMCID: PMC5337708.
3. Bastian H, Glaszio P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010;79:1-6.
4. Van Noorden R. Global scientific output doubles every nine years. Nature News Blog 2014: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html#
5. Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol. 2015;66: 2215-2222. doi:10.1002/asi.23329.
6. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med. 2014;46(2):179-187. doi:10.1016/j.amepre.2013.10.016.
7. Lee YJ, Arida JA, Donovan HS. The application of crowdsourcing approaches to cancer research: a systematic review. Cancer Med. 2017;6:2595–605. doi:10.1002/cam4.1165.
8. Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of Crowdsourcing in Health: Systematic Review. J Med Internet Res. 2018 May 15;20(5):e187. doi: 10.2196/jmir.9330. PMID: 29764795; PMCID: PMC5974463.
9. Mortensen ML, Adam GP, Trikalinos TA, Kraska T, Wallace BC. An exploration of crowdsourcing citation screening for systematic reviews. Res Synth Methods. 2017 Sep;8(3):366-386. doi: 10.1002/jrsm.1252. Epub 2017 Jul 4. PMID: 28677322; PMCID: PMC5589498.
10. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Inform Assoc. 2017 Nov 1;24(6):1165-1168. doi: 10.1093/jamia/ocx053. PMID: 28541493; PMCID: PMC5975623.
11. Nama N, Sampson M, Barrowman N, Sandarage R, Menon K, Macartney G, Murto K, Vaccani JP, Katz S, Zemek R, Nasr A, McNally JD. Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study. J Med Internet Res. 2019 Apr 29;21(4):e12953. doi: 10.2196/12953. PMID: 31033444; PMCID: PMC6658317.
12. Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me

service. J Clin Epidemiol. 2020 Sep 30;130:23-31. doi: 10.1016/j.jclinepi.2020.09.024. Epub ahead of print. PMID: 33007457.

13. Noel-Storr AH, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd finds that crowdsourcing produces accurate results in identifying randomised trials. J Clin Epidemiol. 2021 [article in press]
14. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J Health Serv Res Policy. 2005 Jul;10 Suppl 1:6-20. doi: 10.1258/1355819054308576. PMID: 16053580.
15. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. J Health Serv Res Policy. 2005 Jan;10(1):45-53. doi: 10.1177/135581960501000110. PMID: 15667704.
16. Pluye P, Hong QN. Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. Annu Rev Public Health. 2014;35:29-45. doi: 10.1146/annurev-publhealth-032013-182440. Epub 2013 Oct 30. PMID: 24188053.
17. Bujold M, Granikov V, Sherif RE, Pluye P. Crowdsourcing a mixed systematic review on a complex topic and a heterogeneous population: lessons learned. Education for Information 2018;34(4):293-300.
18. Kelly S, Redmond P, King S, Oliver-Williams C, Lamé G, Liberati E, Kuhn I, Winter C, Draycott T, Dixon-Woods M, Burt J. Training in the use of intrapartum electronic fetal monitoring with cardiotocography: systematic review and meta-analysis. BJOG 2021; https://doi.org/10.1111/1471-0528.16619
19. Cochrane Crowd: https://crowd.cochrane.org [last accessed 4 November 2020]
20. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook
21. Brown AW, Allison DB. Using crowdsourcing to evaluate published scientific literature: methods and example. PLoS One. 2014 Jul 2;9(7):e100647. doi: 10.1371/journal.pone.0100647. PMID: 24988466; PMCID: PMC4079692.
22. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med. 2002 Nov 30;21(22):3431-46. doi: 10.1002/sim.1253. PMID: 12407682.
23. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological

systematic review. BMC Med Res Methodol. 2019 Jun 28;19(1):132. doi: 10.1186/s12874-019-0782-0. PMID: 31253092; PMCID: PMC6599339.

24. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, König F. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. 2020 May;121:20-28. doi: 10.1016/j.jclinepi.2020.01.005. Epub 2020 Jan 21. PMID: 31972274.
25. Blomberg M. Avoiding the first cesarean section--results of structured organizational and cultural changes. Acta Obstet Gynecol Scand. 2016 May;95(5):580-6. doi: 10.1111/aogs.12872. Epub 2016 Mar 15. PMID: 26870916.
26. Byford S, Weaver E, Anstey C. Has the incidence of hypoxic ischaemic encephalopathy in Queensland been reduced with improved education in fetal surveillance monitoring? Aust N Z J Obstet Gynaecol. 2014 Aug;54(4):348-53. doi: 10.1111/ajo.12200. Epub 2014 Mar 6. PMID: 24597944.
27. McDonald S, Noel-Storr AH, Thomas J. Harnessing the efficiencies of machine learning and Cochrane Crowd to identify randomised trials for individual Cochrane reviews. Global Evidence Summit, Cape Town, South Africa; 13th – 16th September 2017
28. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. Res Synth Methods. 2018;9(4):602-614. doi: 10.1002/jrsm.1287. Epub 2018 Feb 7. PMID: 29314757; PMCID: PMC6030513.
29. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2020;7:S0895-4356(20)31172-0. doi: 10.1016/j.jclinepi.2020.11.003. Epub ahead of print. PMID: 33171275.
30. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020;127:142-150.
31. Bannach-Brown A, Przybyła P, Thomas J, Rice ASC, Ananiadou S, Liao J, Macleod MR. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst Rev. 2019;15;8(1):23. doi: 10.1186/s13643-019-0942-7. PMID: 30646959; PMCID: PMC6334440.
32. Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. Proceedings of the 2018 CHI Conference on

Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Paper 449, 1–14; https://doi.org/10.1145/3173574.3174023

33. Gray M, Suri S. Ghost work: how to stop Silicon Valley from building a new global underclass. Boston, Houghton Miffin Harcourt; 2019

Chapter 7

Crowdsourcing and COVID-19: a case study of Cochrane Crowd

This original manuscript was published in *Journal of EAHIL*

Citation: Noel-Storr A, Dooley G, Featherstone R, Wisniewski S, Shemilt I, Thomas J, Gartlehner G, Nußbaumer-Steit B, Mavergames C. Crowdsourcing and COVID-19: a case study of Cochrane Crowd. JEAHIL [Internet]. 24Jun.2021 [cited 27Nov.2021];17(2):27-1.

DOI: https://doi.org/10.32384/jeahil17467

URL: http://ojs.eahil.eu/ojs/index.php/JEAHIL/article/view/467

7.1 Abstract

Cochrane has used crowdsourcing effectively to identify health evidence since 2014. To date, over 175,000 trials have been identified for Cochrane's Central Register of Controlled Trials via Cochrane Crowd (https://crowd.cochrane.org), Cochrane's citizen science platform, engaging a crowd of over 20,000 people from 166 countries. The COVID-19 pandemic presented the evidence synthesis community with the enormous challenge of keeping up with the exponential output of COVID-19 research. This case study will detail the new tasks we developed to aid the production of COVID-19 rapid reviews and supply the Cochrane COVID-19 Study Register. The pandemic initially looked set to disrupt the crowd team's plans for 2020 but has in fact served to further our understanding of the potential role crowdsourcing can play in the health evidence ecosystem.

7.2 Introduction

Crowdsourcing in health research has become increasingly popular over the last decade[1]. Cochrane, an international network that produces systematic reviews, has been harnessing a type of crowdsourcing called 'human intelligence tasking' since 2014[2,3]. Human intelligence tasking involves filtering or classifying large amounts of data or information via an online community. In May 2016, Cochrane launched Cochrane Crowd (https://crowd.cochrane.org), its citizen science platform, with its first crowdsourcing task: the identification of reports of randomised controlled trials (RCTs) from Embase. Other tasks followed soon after and new tasks are in development and being rolled out on an ongoing basis. Our evaluations of the crowd's performance in terms of accuracy demonstrated that a crowdsourcing approach to identifying RCTs was both robust and efficient[2]. By early 2020, over 20,000 contributors had signed up to Cochrane Crowd from 166 countries and generated over 5 million individual classifications, helping to identify around 175,000 reports of randomised trials.

2020 looked to be a busy year, but we did not anticipate how large an impact the COVID-19 pandemic would have on Cochrane Crowd. We had launched a new version of the crowd platform in early March 2020 and work was about to begin on a new PICO extraction task as part of Cochrane's trial surveillance initiative. Initially, the pandemic was hugely disruptive to the latter planned work, with our efforts immediately re-focussed to help.

One of the main challenges presented by the pandemic was the corresponding infodemic. According to the World Health Organization,

> *[A]n infodemic is too much information including false or misleading information in digital and physical environments during a disease outbreak. It causes confusion and risk-taking behaviors that can harm health. It also leads to mistrust in health authorities and undermines the public health response. An infodemic can intensify or lengthen outbreaks when people are unsure about what they need to do to protect their health and the health of people around them.*[4]

The dramatic increase in COVID-19 research production and publication throughout 2020 and 2021 has created significant information retrieval challenges, both from the sheer volume of research and in the nature of the research output. One example was the so-called "preprint rush," with both demand for, and availability of, preprints soaring during 2020[5,6]. Cochrane was able to adapt existing skills and systems for the organisation of COVID-19 research to assist with review production.

Cochrane prioritised resources and developed initiatives to respond to the pandemic, including a programme of work to produce rapid reviews and the production of special collections of existing relevant health evidence on topics such as infection control and prevention measures and remote care through telehealth[7].

Another major undertaking within the network was the development of a curated register of COVID-19 studies, the Cochrane COVID-19 Study Register (CCSR) (https://covid-19.cochrane.org)[8] . The CCSR is a continuously updated open access repository of COVID-19 human studies that have been identified from a range of sources and tagged by study type, study design and study aim. Related reports about the same study are linked together to create a 'study based' register. The register went live in April 2020 and within twelve months over 57,000 COVID-19 studies had been identified and described.

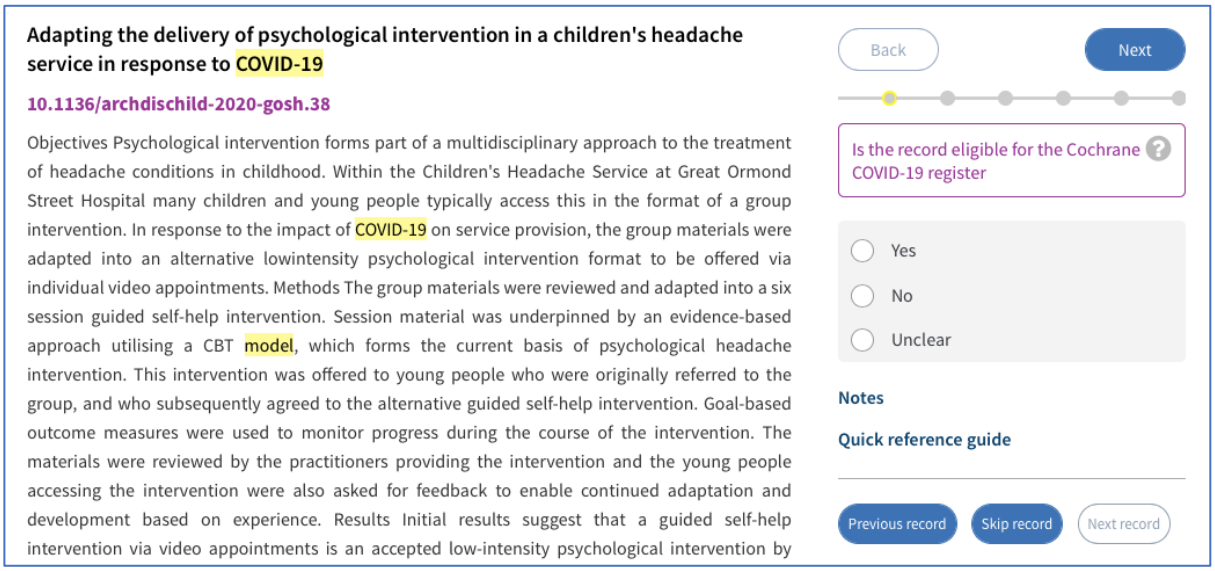
Cochrane Crowd was uniquely placed to help as our thriving community of contributors were eager to support Cochrane's response to the pandemic. This case study details four main areas of work undertaken by Cochrane Crowd during the first twelve months of the pandemic: (1) COVID Quest – a new Cochrane Crowd task; (2) direct review input and methodological research; (3) weekly screening challenges; (4) a COVID-19 machine learning classifier.

7.3 COVID Quest

We developed a new crowdsourced task: COVID Quest. In COVID Quest the crowd identify COVID-related studies based on assessing title-abstract records (Figure 7.1). Unlike most Cochrane Crowd tasks, it is a 'multi-question' task – made up of a series of questions about the record.

COVID Quest tasks contributors with identifying a range of different study types and study designs, which is another key difference with this task compared to other mainstream tasks on Cochrane Crowd, which relate to identification or description of randomised controlled trials. This is crucial because in a pandemic, a range of study types are needed to answer urgent questions regarding treatment, diagnostics, health services, mental health and the larger societal impact. Controlled vocabularies are used for each question within the task. Anyone can join, though completion of a brief training module is mandatory.

Figure 7.1 Screen capture of Cochrane Crowd's COVID-19 task: COVID Quest



We launched the task in June 2020 after a rapid development and testing phase, and to date (June 2021) the crowd have amassed around 60,000 assessments helping to identify and describe thousands of studies for the CCSR. We have evaluated crowd accuracy against a gold standard data set made up of 2000 records assessed by Cochrane information specialists working on the register. Within this set, 566 records were eligible for the CCSR. The crowd correctly identified 558 as eligible giving a crowd sensitivity of 98.5%. The crowd achieved similarly high levels of sensitivity across the study type (whether the study described was an observational, interventional, qualitative, or mathematical modelling study) and the specific study design used (RCT, cohort study/case control, case report, cross-section etc.) components of COVID Quest: 98.2% and 97.6% respectively. In

addition, around 85% of records assessed had matching classifications under our agreement algorithm, with only 15% requiring resolution by an "expert" after discordant classifications between crowd contributors.

COVID Quest forms part of a study identification workflow that is largely based on processes that Cochrane's Centralised Search Service already had in place for identifying studies for the Cochrane Central Register of Controlled Trials (CENTRAL)[9] (Figure 7.2). Having some of the foundations and technical infrastructure in place facilitated rapid implementation of this end-to-end process.

7.4 Review input

As already described, Cochrane undertook a programme of COVID-related, rapidly produced reviews. This work presented an opportunity to test the crowd's ability to identify studies for reviews in a time-sensitive context. Four reviews were used in this methodological work: Quarantine alone or in combination with other public health measures to control COVID-19[10]; Barriers and facilitators to healthcare workers' adherence with infection prevention and control (IPC) guidelines for respiratory infectious diseases[11]; Universal screening for Severe Acute Respiratory Syndrome Coronavirus 2[12]; and Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19[13]. We created a corresponding crowdsourced task for each of these reviews in Cochrane Crowd. Crowd contributors were tasked with assessing the search results and making one of two possible classifications on each title-abstract record: *Possibly relevant* or *Not relevant*.

As with COVID Quest, these new crowd tasks marked a departure from crowd tasks focussed on identifying RCTs. This collection of rapidly produced reviews covered a wide range of eligible study types and designs including mathematical modelling studies, observational studies, interventional studies, and qualitative and mixed study designs. The crowd had to become familiar with both the topic of the review and study types eligible for the review. They were also only given 48 hours to complete each task. The crowd performed well, comfortably completing the screening task for three of the four reviews within 48 hours (one review took just over 48 hours to complete). Crowd accuracy levels were high, ranging from 90%-100% recall across the four reviews. This methodological work furthered our understanding of crowdsourcing capabilities in topic-based screening tasks under tight time constraints. The crowd also inputted directly into the update of the rapid review on quarantine measures, where 65 crowd contributors screened the 5000 results retrieved from the update search in 22 hours (https://www.cochrane.org/news/cochrane-crowd-does-it-again-rapid-study-identification-cochrane-rapid-review).

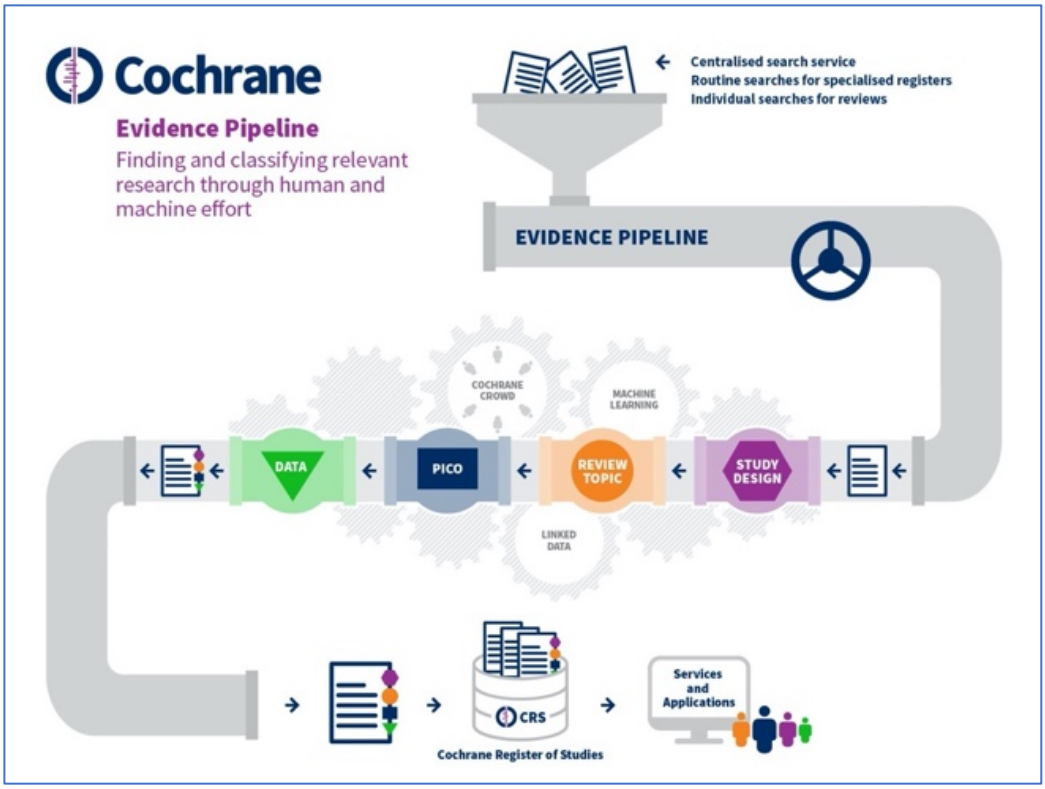
7.5 Weekly screening challenges

From April 2020, we started a series of weekly 3-hour crowd 'challenges'. Each week we select a task and encourage as many as possible to get online and join in. During the early days of the pandemic, when most of us were in strict lockdown with many not able to work, this felt like a suitable community engagement activity that enabled us to keep some of our 'business as usual' tasks going. We have now completed over 50 weekly challenges and in that time, screened approximately 100,000 records mostly from the RCT Identification task.

7.6 COVID-19 machine learning classifier

The final area of crowd input is related to the development of a machine learning classifier for COVID-19 studies. In July 2020 members of the CCSR team and the COVID EPPI-Centre Map team, based at University College London, set up a series of meetings with the aim of sharing best practice and reducing duplication of effort across the two initiatives. One area of focus was on strategies to reduce study identification screening burden. The EPPI-Centre Map team had already developed a binary machine learning classifier that worked to reduce screening workload as well as to help prioritise screening. Given the differing scope regarding studies eligible for the CCSR and the EPPI-Centre COVID Map, we decided that a new binary machine learning classifier should be developed specifically for the CCSR workflow. We therefore used high quality data generated by both the core Cochrane register team and Cochrane Crowd to train, calibrate and evaluate a COVID-19 study classifier. We followed the same stages of training, calibration and validation as we had done for the development of the Cochrane RCT classifier[14]. The result is a classifier that helps to accurately identify records that are not eligible for the CCSR. We have been using this classifier since February 2021, reducing screening burden by between 20-25%.

Figure 7.2 Cochrane's Evidence Pipeline vision



7.7 Conclusions

COVID-19 presented us with major information retrieval challenges, but also provided important opportunities for research and development on methods, processes, and tools. Our experiences have highlighted the benefit of focussed and collaborative working. Development, testing and full implementation of Cochrane Crowd's most complex task to date took eight weeks instead of the more usual 12-24 months. We were able to use and adapt existing systems (such as the Cochrane Crowd platform), processes, for example Cochrane's Centralised Search Service, and expertise across information and data science disciplines. The Cochrane Crowd community itself played an invaluable role in enabling us to keep-up, advancing our expectations of crowdsourced capability in evidence synthesis. We are now working on extending the crowd's role to include PICO extraction of both COVID-19 studies as well as studies in other healthcare areas. This will, we hope, significantly improve search precision, and support accurate surveillance of the evidence as it emerges.

In its early days, the pandemic appeared to be highly disruptive to 'business as usual', but in hindsight it has accelerated our work and our understanding of the value of human and machine input in the production of health research. Sharing an overarching mission to help during a global health crisis, organisations at different levels of the evidence ecosystem pulled together to make the emerging evidence base FAIR (findable, accessible, interoperable, and reusable). Duplication of

effort still occurred and enormous challenges remain as the deluge of information around COVID-19 shows little sign of abating, but for the Cochrane Crowd team, the experience and the learning of the last twelve months has been important and lasting.

7.8 Abbreviations

CCSR	Cochrane COVID-19 Study Register
CENTRAL	Cochrane Central Register of Controlled Trials
FAIR	Findable, Accessible, Interoperable, Reusable
IPC	Infection, prevention and control
PICO	Population, Intervention, Comparator, Outcome
RCT	Randomised controlled trial

7.9 Author contributions

Anna Noel-Storr: conceptualisation, methodology, writing – original draft preparation

Gordon Dooley: conceptualisation, methodology, writing – reviewing and editing

Robin Featherstone: conceptualisation, methodology, writing – reviewing and editing

Susanna Wisniewski: conceptualisation, methodology, writing – reviewing and editing

Ian Shemilt: conceptualisation, methodology, writing – reviewing and editing

James Thomas: conceptualisation, methodology, writing – reviewing and editing

Gerald Gartlehner: conceptualisation, methodology, writing – reviewing and editing

Barbara Nußbaumer-Steit: conceptualisation, methodology, writing – reviewing and editing

Chris Mavergames: conceptualisation, methodology, writing – reviewing and editing

7.10 References

1. Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of Crowdsourcing in Health: Systematic Review. J Med Internet Res. 2018 May 15;20(5):e187. doi: 10.2196/jmir.9330. PMID: 29764795; PMCID: PMC5974463.
2. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 Jan 18:S0895-4356(21)00008-1. doi: 10.1016/j.jclinepi.2021.01.006. Epub ahead of print. PMID: 33476769.
3. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med. 2014 Feb;46(2):179-87. doi: 10.1016/j.amepre.2013.10.016. PMID: 24439353.
4. World Health Organization. https://www.who.int/health-topics/infodemic#tab=tab_1 [last accessed 30 April 2021].
5. Odone A, Salvati S, Bellini L, Bucci D, Capraro M, Gaetti G, Amerio A, Signorelli C. The runaway science: a bibliometric analysis of the COVID-19 scientific literature. Acta Biomed. 2020 Jul 20;91(9-S):34-39. doi: 10.23750/abm.v91i9-S.10121. PMID: 32701915; PMCID: PMC8023084
6. Else H. How a torrent of COVID science changed research publishing - in seven charts. Nature. 2020 Dec;588(7839):553. doi: 10.1038/d41586-020-03564-y. PMID: 33328621.
7. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD202002/full
8. Featherstone R, Last A, Becker L, Mavergames C. Rapid development of the Cochrane COVID-19 Study Register to support review production. In: Collaborating in response to COVID-19: editorial and methods initiatives across Cochrane. Cochrane Database Sys Rev. 2020;(12 Suppl 1):37–40. doi: 10.1002/14651858.CD202002
9. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020 Nov;127:142-150. doi: 10.1016/j.jclinepi.2020.08.008. Epub 2020 Aug 13. PMID: 32798713.
10. Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, Wagner G, Siebert U, Christof C, Zachariah C, Gartlehner G. Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. Cochrane Database Syst Rev. 2020 Apr 8;4(4):CD013574. doi: 10.1002/14651858.CD013574. PMID: 32267544; PMCID: PMC7141753.
11. Houghton C, Meskell P, Delaney H, Smalle M, Glenton C, Booth A, Chan XHS, Devane D, Biesty LM. Barriers and facilitators to healthcare workers' adherence with infection prevention and control (IPC) guidelines for respiratory infectious diseases: a rapid qualitative evidence synthesis.

Cochrane Database Syst Rev. 2020 Apr 21;4(4):CD013582. doi: 10.1002/14651858.CD013582. PMID: 32315451; PMCID: PMC7173761.

12. Viswanathan M, Kahwati L, Jahn B, Giger K, Dobrescu AI, Hill C, Klerings I, Meixner J, Persad E, Teufer B, Gartlehner G. Universal screening for SARS-CoV-2 infection: a rapid review. Cochrane Database Syst Rev. 2020 Sep 15;9:CD013718. doi: 10.1002/14651858.CD013718. PMID: 33502003.
13. Valk SJ, Piechotta V, Chai KL, Doree C, Monsef I, Wood EM, Lamikanra A, Kimber C, McQuilten Z, So-Osman C, Estcourt LJ, Skoetz N. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a rapid review. Cochrane Database Syst Rev. 2020 May 14;5(5):CD013600. doi: 10.1002/14651858.CD013600. Update in: Cochrane Database Syst Rev. 2020 Jul 10;7:CD013600. PMID: 32406927; PMCID: PMC7271896.
14. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2020 Nov 7:S0895-4356(20)31172-0. doi: 10.1016/j.jclinepi.2020.11.003. Epub ahead of print. PMID: 33171275.

Chapter 8

Crowdsourcing the identification of studies for COVID-19 related Cochrane rapid reviews

This original manuscript was published in the journal *Research Synthesis Methods*

Citation: Noel-Storr A, Gartlehner G, Dooley G, Persad E, Nussbaumer-Streit B. Crowdsourcing the identification of studies for COVID-19 related Cochrane rapid reviews. Research Synthesis Methods 2022;13(5):585-594.
DOI: 10.1002/jrsm.1559

8.1 Abstract

Background

Utilization of crowdsourcing within evidence synthesis has increased over the last decade. Crowdsourcing platform Cochrane Crowd has engaged a global community of 22,000 people from 170 countries. The COVID-19 pandemic presented an opportunity to engage the community and keep up with the exponential output of COVID-19 research.

Aims

To test whether a crowd could accurately assess study eligibility for reviews under time constraints. Outcome measures: time taken to complete each task, time to produce required training modules, crowd sensitivity, specificity, and crowd consensus.

Methods

We created four crowd tasks, corresponding to four Cochrane COVID-19 rapid reviews. The search results of each were uploaded and an interactive training module was developed for each task. Contributors who had participated in another COVID-19 task were invited to participate. Each task was live for 48-hours. The final inclusion and exclusion decisions made by the core author team were used as the reference standard.

Results

Across all four reviews 14,299 records were screened by 101 crowd contributors. The crowd completed each screening task within 48-hours for three reviews and in 52 hours for one. Sensitivity ranged from 94% to 100%. Four studies, out of a total of 109, were incorrectly rejected by the crowd. However, their absence ultimately would not have altered the conclusions of the reviews. Crowd consensus ranged from 71% to 92% across the four reviews.

Conclusion

Crowdsourcing can play a valuable role in study identification and offers willing contributors the opportunity to help identify COVID-19 research for rapid evidence syntheses.

8.2 Background

The COVID-19 pandemic highlighted the need to produce reliable syntheses of health evidence as quickly as possible. An unprecedented volume of research has been undertaken resulting in a 'tidal wave' of trials and research publications[1]. This infodemic makes the production of reliable health

evidence synthesis especially challenging when it is needed most. Timely dissemination of accurate information is critical in the fight against both COVID-19 and the harmful spread of mis-information[2]. Many questions have arisen regarding mechanism, transmission, diagnosis, prognosis, treatment and management of COVID-19. In response to this global crisis, Cochrane launched a rapid review initiative (https://www.cochrane.org/cochranes-work-rapid-reviews-response-covid-19). Rapid reviews are needed urgently to assess and appraise both existing actionable literature (on areas such as transmission mitigation, oxygen therapy, respiratory failure, and others) and to assess and appraise the exponentially growing corpus of research being produced as a direct result of COVID-19[3].

Crowdsourcing may help solve this data deluge challenge. Crowdsourcing is the outsourcing of needed tasks or activities to a large community of people, usually via the internet. Many domains and disciplines have implemented a range of crowdsourcing models to solve organisational or research problems. In psychology for example, crowdsourced research methods have been applied to overcome challenges of small sample sizes and enable research replication[4,5]. Crowds have also been engaged in helping to classify or categorise large amounts of data, from assessing underwater images from the Great Barrier Reef to helping to classify galactic data as part of the Galaxy Zoo citizen science project[6].

Cochrane has used crowdsourcing as a means of effectively identifying health evidence since 2014. To date, over 200,000 trials have been identified for Cochrane's Central Register of Controlled Trials via Cochrane Crowd (https://crowd.cochrane.org), Cochrane's citizen science platform. Cochrane Crowd has attracted over 22,000 contributors from 170 countries. Accuracy evaluations have shown that the crowd, when performing a task with an appropriate agreement algorithm, can achieve 99% accuracy in terms of the crowd's ability to correctly identify studies of interest (for example, randomised trials) and the crowd's collective ability to reject the records that should be rejected[7].

In April 2019, Cochrane launched a workflow called Screen4Me. This workflow enables Cochrane review author teams to send search results to Cochrane Crowd. Prior to this the crowd had focused on identifying studies for central repositories, such as Cochrane's Central Register of Controlled Trials. The Screen4Me workflow requires the crowd to work to a given deadline, assessing search results for a specific review, in return for named acknowledgement in the review when it is published[8,9].

Rapid reviews on COVID-19 present us with two specific new challenges with regards to the feasibility of recruiting and using a crowd effectively. The first is that it is likely that many rapid reviews undertaken will not be reliant on evidence from randomised controlled trials (RCTs) due either to the research or clinical question not being appropriate for RCTs or to the current lack of completed RCTs in this area. Therefore, the crowd will need to be able to identify and assess a range of different study types and designs. They will also be required to perform a more topic-based assessment of the search results for rapid reviews. This has been shown to be feasible in two recent pilot studies performed with the Cochrane Crowd community. In the first pilot, the crowd were tasked with performing a topic-based assessment for potentially relevant studies for an RCT-based systematic review and, in the second, to perform a topic-based assessment for a review that sought to include a range of different study types, including qualitative and mixed studies. In both pilot studies the crowd performed with a very high degree of accuracy: 100% and 96% sensitivity respectively [10,11]. Beyond Cochrane Crowd, other feasibility studies exploring the role of crowdsourcing in study identification have produced similar results[12,13]. Mortensen and colleagues tasked a crowd, via Amazon Mechanical Turk, with assessing the search results for four systematic reviews. The reviews included a range of study types and designs including randomised controlled trials and diagnostic studies. The crowd was able to achieve high sensitivity (ranging from 96% to 99%) and moderate specificity (68% to 81%)[12]. Nama and colleagues' validation study used data from six systematic reviews across a wide range of healthcare areas and similarly demonstrated the feasibility of engaging a crowd to perform citation screening to high degree of accuracy[13].

Our second challenge relates to time-to-task-completion. Rapid reviews aim to be produced within a few weeks, with the results screening stage needing to be completed within 24 to 48 hours. Cochrane's current Screen4Me workflow allows the crowd two weeks to complete the results screening task. This deadline is met for 95% of Screen4Me tasks[14]. This is encouraging, but two weeks is a substantial increase on the hoped for 24 to 48 hours for task completion for rapid reviews. The shorter timeframe therefore needs to be tested within the context of rapid reviews for COVID-19, especially given that the task itself is different (as described above). In addition, time and accuracy are not mutually exclusive; one may adversely impact the other. Time pressure may increase crowd inaccuracy or reduce consensus (the proportion of records that do not require arbitration to reach a final decision) or both. We need to explore these factors in order to be able to better understand the role the crowd could play in the production of rapid reviews in this area.

8.3 Aims and objectives

Our aim was to test whether a crowd could accurately assess the eligibility of search results for a range of rapid reviews when given a short deadline to do so. Our main outcome measures were time taken, in hours, to complete each of the screening tasks and time taken to prepare the customized training modules and other guidance materials required for each task. Additionally, we sought to measure crowd accuracy in terms of crowd sensitivity, specificity and crowd consensus.

8.4 Methods

The data sets

We conducted a crowdsourced screening exercise using the sets of search results identified from a convenience sample of four Cochrane rapid reviews produced in response to the COVID-19 pandemic. The four reviews were:

- Quarantine alone or in combination with other public health measures to control COVID-19 (hereafter shortened to: *Review 1: Quarantine*)*[15]
- Barriers and facilitators to healthcare workers' adherence with infection prevention and control (IPC) guidelines for respiratory infectious diseases (*Review 2: IPC Adherence*)[16]
- Universal screening for Severe Acute Respiratory Syndrome Coronavirus 2 (*Review 3: Universal Screening*)[17]
- Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19 (*Review 4: Convalescent Plasma*)[18]

The size of the search results sets varied with the smallest being the set for Review 4: Convalescent Plasma (948 records) to the largest set for Review 1: Quarantine (5606). The inclusion criteria in terms of eligible study designs also varied across the four reviews. Review 1: Quarantine, included mathematical modelling studies, as well as interventional and observational study types. Review 2: IPC Adherence, included qualitative and mixed methods studies. Review 3: Universal Screening, included diagnostic test accuracy designs as well interventional studies as it considered both the accuracy and effectiveness of universal screening approaches, and Review 4: Convalescent Plasma, included both observational and interventional designs (see Table 8.1 for review characteristics). The final inclusion and exclusion decisions of studies made by the core author team for each of the four reviews was used as the reference standard. The screening process in place for rapid reviews differs slightly from the process for mainstream Cochrane systematic reviews in that records need only one

assessment from a member of the core author team unless the record is rejected; rejected records are dual-screened[3].

Table 8.1 Key task characteristics

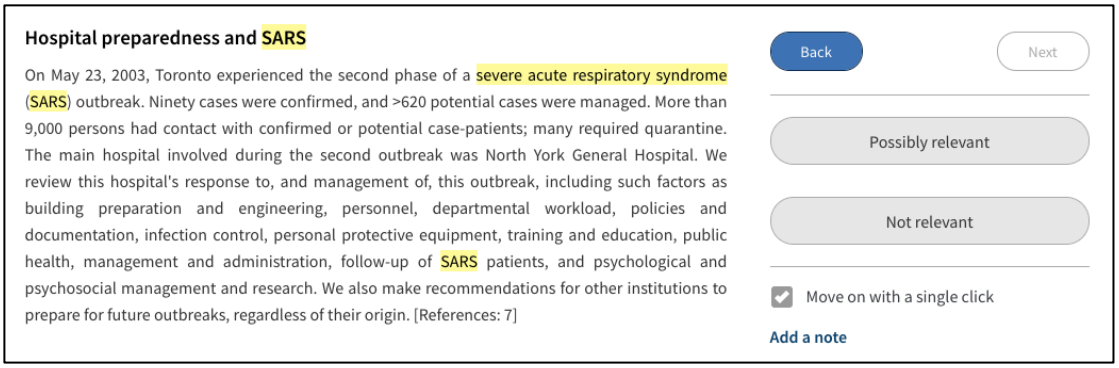
Review	Eligible study types	Size of set	No. of included studies*	No. of people invited	No. of people contributed
Review 1: Quarantine	Observational Modelling Interventional	5606	47	123	65
Review 2: IPC Adherence	Qualitative Observational Interventional	3367	32	85	36
Review 3: Universal Screening	Observational (Diagnostic) Interventional	4378	18	104	38
Review 4: Convalescent Plasma	Observational Interventional	948	12	122	12
Total		14,299	109	287**	101**

*no. of included studies used in the evaluation data sets (some includes studies were used in the training modules so were not then included in the evaluation data sets)
**unique contributors

The process

We created four separate tasks in Cochrane Crowd. With each, the crowd was tasked with classifying the search results based on an assessment of title-abstract records (see Figure 8.1). We created a brief training module to accompany each of the four crowd tasks. Each module was composed of a series of introductory screens describing the topic of the review and the types of eligible studies followed by an assessment made up of sixteen practice records. We included two title-only records within the training module for each review to help contributors know how to assess records that did not have abstracts. Crowd contributors needed to pass the assessment with a score of 80% or more to be able to progress to the live task. This pass mark is the standard pass mark used for other citation screening tasks in Cochrane Crowd. In addition to the training module, we employed an agreement algorithm which required three consecutive agreement classifications on a record for that record to be deemed either *Not relevant* (in the case for three independently made *Not relevant* classifications) or *Possibly relevant* (three consecutively made *Possibly relevant* classifications). We set each task to run initially for 48 hours, with the option to extend the time if needed.

Figure 8.1 Screen shot of Review 1: Quarantine



The crowd

Eligible crowd contributors were those who had completed and passed the training module for another task available in Cochrane Crowd: *COVID Quest*. *COVID Quest* was launched in May 2020[19]. The task was built to help feed the Cochrane COVID-19 Study Register (https://covid-19.cochrane.org). For this task, contributors need to be able to identify COVID-19 related research as described by a title and abstract, and to then tag that research by study type and design, as well as assign study aims (e.g., treatment and management, or diagnostic, etc.). They must pass the *COVID Quest* training module by 80% or more to gain access to the live task[20]. Once each rapid review crowd task had been built, contributors who had assessed at least one record in *COVID Quest* within the last month were contacted by email to inform them that they were eligible to participate in these rapid review tasks.

Data collection and statistical analysis

Crowd sensitivity was measured as the proportion of records correctly and collectively identified as *Possibly relevant* and crowd specificity, the proportion of records correctly and collectively identified as *Not relevant* to the review. We used the final set of studies included/not included in the review as the reference standard.

Crowd sensitivity:

$$\frac{TP}{TP + FN}$$

Crowd specificity:

$$\frac{TN}{TN + FP}$$

In terms of accuracy, we are primarily interested in crowd sensitivity rather than crowd specificity. The crowd missing or rejecting studies that should have been included is of more significance than the crowd mistakenly classifying irrelevant records as possibly relevant.

Crowd consensus is the proportion of records that the crowd assesses that do not require arbitration due to disagreeing classifications.

$$\frac{No.\,of\ records\ not\ requiring\ resolution}{Total\ number\ of\ records\ in\ dataset}$$

We conducted all statistical analyses in Microsoft Excel v16.50 and SPSS v26.

8.5 Results

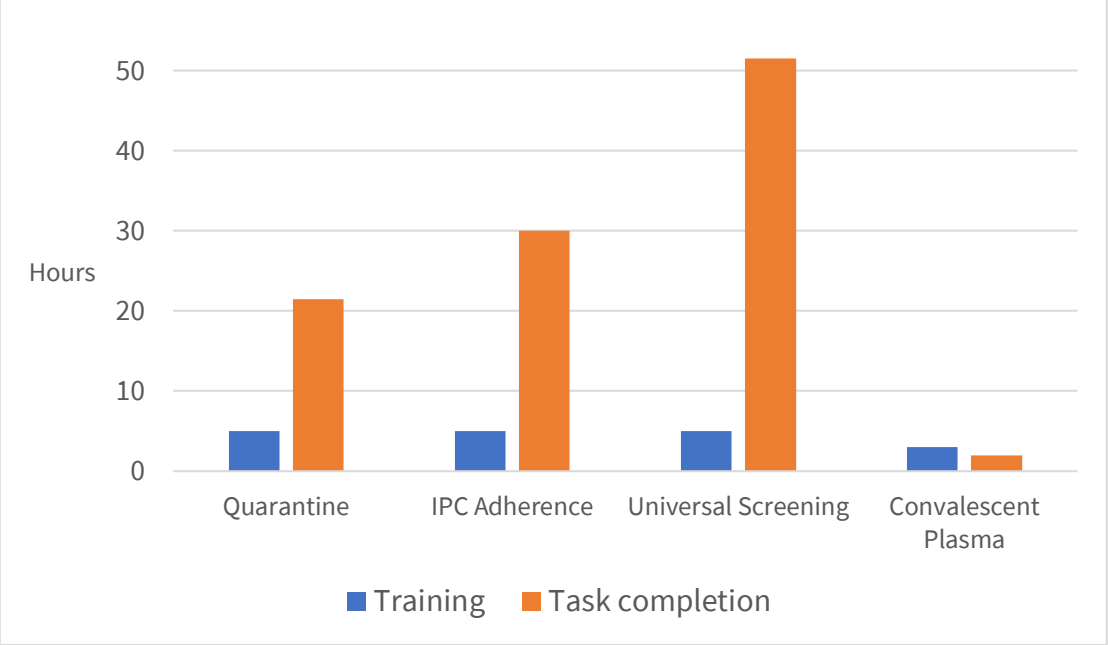
Crowd characteristics

We created and ran four Cochrane Crowd tasks, one for each of the Cochrane rapid reviews used for this pilot study[15,16,17,18]. Table 8.1 shows, for each of the tasks, the number of contributors invited to take part, the number that took part, the size of each data set and the time taken to complete the task. Eligible Crowd contributors were those who had taken part in the Cochrane Crowd task, *COVID Quest* within the last month prior to the date the rapid review task went live on the platform. For the Review 1: Quarantine, 65 crowd participants took part; Review 2: IPC Adherence, 36; Review 3: Universal Screening, 38; Review 4: Convalescent Plasma, 12. Of those who took part, 65% took part in only one of the tasks; the remainder (35%) took part in more than one. Crowd contributors screened on average 268 records (ranging from 4-1201) for Review 1, 274 (range 2-1500) for Review 2, 333 (range 10-3168) for Review 3, and 248 (range 1-711) for Review 4.

Time

Our main outcome measure was time, both in terms of time taken to produce the bespoke training modules and time to task completion by the crowd. Figure 8.2 shows the time taken to develop each training module, which ranged from 3 to 5 hours, and the time-to-task-completion which ranged from 2 hours to 51.5 hours). Time per 100 records for each of the reviews was therefore 22 minutes for Review 1: Quarantine, 53 minutes for Review 2: IPC Adherence, 74 minutes for Review 3: Universal Screening, and 13 minutes for Review 4: Convalescent Plasma.

Figure 8.2 Outcome measure: Time



Crowd accuracy: sensitivity and specificity

In terms of crowd accuracy, sensitivity (i.e., the crowd's collective ability to correctly identify the included studies) ranged from 94% to 100% (see Table 8.2). In Review 1: Quarantine, two included studies were missed by the crowd. In Review 2: IPC Adherence and Review 3: Universal Screening, one included study was incorrectly rejected. In Review 4: Convalescent Plasma, no included studies were missed.

Crowd specificity (i.e., the crowd's collective ability to correctly reject ineligible references to studies) for each of the four reviews was: Review 1: Quarantine 71%, Review 2: IPC Adherence 73%, Review 3: Universal Screening 71%, and Review 4: Convalescent Plasma 89%.

Table 8.2 Crowd accuracy

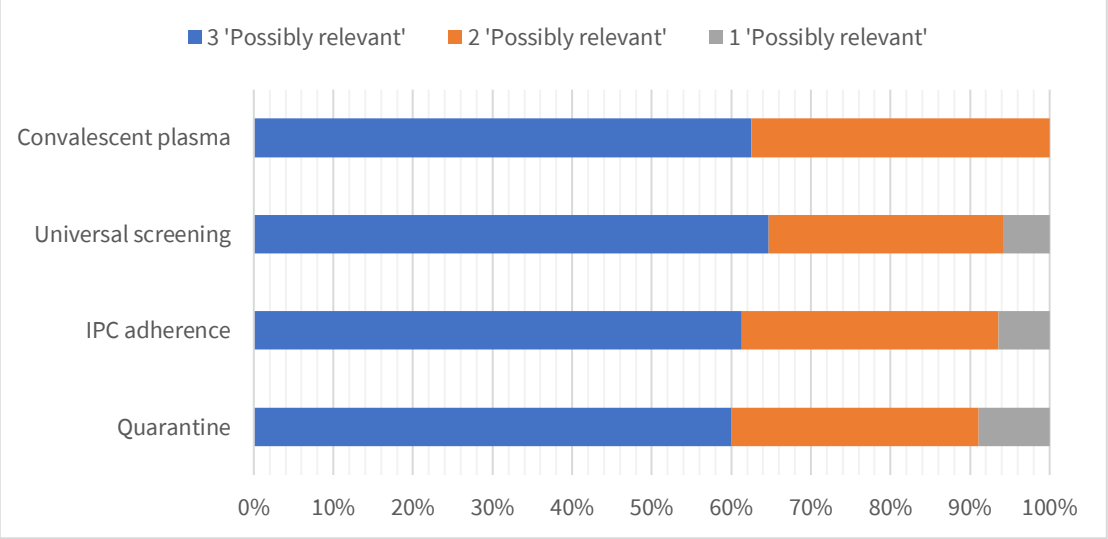
Review	N	TP	TN	FP	FN	Sensitivity	Specificity	Consensus
Review 1: Quarantine	5606	45	3942	1617	2	95.7	70.9	72.02
Review 2: IPC Adherence	3367	31	2437	897	1	96.9	73.0	74.96
Review 3: Universal Screening	4378	17	3075	1285	1	94.4	70.5	71.34
Review 4: Convalescent Plasma	948	12	827	109	0	100.0	88.7	92.19

TP = True Positive; the number of records correctly classified as possibly relevant
TN = True Negative; the number of records correctly classified as not relevant
FP = False Positive; the number of records incorrectly classified as possibly relevant
FN = False Negative; the number of records incorrectly classified as not relevant

Crowd consensus

The level of crowd consensus (i.e., the proportion of records receiving three consecutive agreeing classifications) was 72% for Review 1: Quarantine, 75% for Review 2: IPC Adherence, 71% for Review 3: Universal Screening, and 92% for Review 4: Convalescent Plasma. As well as evaluating crowd consensus for each data set as described above, we also calculated crowd consensus for just the eligible studies for each review. The proportion of included studies that received the required three *Possibly relevant* classifications was similar across all four reviews: Review 1: Quarantine 60%, Review 2: IPC Adherence 61%, Review 3: Universal Screening 65% and Review 4: Convalescent Plasma 63% (See Figure 8.3).

Figure 8.3 Crowd consensus for included studies



8.6 Discussion

The crowd performed three of the review tasks comfortably within the 48-hour time limit, and one (Review 3: Universal Screening) in just over the time limit. This is an encouraging result. The development of each training module took on average four hours. We had hoped to run the tasks either concurrently or in very quick succession to gauge the capacity of the crowd to handle multiple tasks simultaneously or continuously. However, we were unable to do that due to the availability of the data sets and the prioritization of other COVID-19 related activities. However, one advantage of having the tasks run approximately 4 weeks apart, meant that we were more likely to attract different crowd contributors for each task, giving us a better sense of generalizable crowd performance.

Analysis of missed studies

The crowd performed well across all reviews in terms of accuracy measures. Overall, out of a total of 109 included studies, the crowd incorrectly rejected four studies (3.7%). The titles of the four missed studies were:

1. Factors that make an infectious disease outbreak controllable[21] (Review 1: Quarantine)
2. Severe Acute Respiratory Syndrome Coronavirus 2 Infection among Returnees to Japan from Wuhan, China[22] (Review 1: Quarantine)
3. SARS: key factors in crisis management[23] (Review 2: IPC Adherence)
4. Suppression of COVID-19 outbreak in the Italian municipality of Vo, Italy[24] (Review3: Universal Screening)

Two of the missed studies were from the quarantine review. One was a small modelling study pre-dating the pandemic but deemed relevant in terms of modelling the effects of pre-symptomatic infections. However, it provided only indirect evidence on SARS, not specifically on SARS-CoV-2. The other, an observational study, reported on the screening and quarantining of a cohort of Japanese nationals repatriated to Japan from Wuhan, China in early 2020. It may have been mistakenly perceived as a diagnostic study rather than of relevance to the quarantine measures review. The missed study from the IPC Adherence review was a qualitative study. It had very broadly stated aims to: "identify the key factors enabling the hospital to survive SARS unscathed." The results described in the abstract make no direct mention of IPC Adherence but instead refer more broadly to good crisis management principles adopted by this specific hospital during the 2003 SARS epidemic. The final missed study was from the Universal Screening review (Review 3). It was not described explicitly as a screening study which may account for why it was missed.

Despite crowd sensitivity not achieving 100% for three of the four reviews used in this evaluation study, sensitivity was comparable to other similar studies run by this and other research teams[10,11,12,13] and potentially more accurate than having the search results screened by a single human assessor[25]. However, it is arguable that providing a measure of sensitivity where the prevalence of included studies within each of the review data sets was very low, should be considered with caution: Review 1 had a prevalence of 0.87%, Review 2: 1.07%, Review 3: 0.53%, Review 4: 2%.

What is perhaps a more meaningful measure of performance is whether the conclusions of each review would have been altered by the missed studies. We contacted the lead authors for each of the reviews to ascertain whether conclusions would have changed. For Review 1: Quarantine, the missed studies would not have altered the conclusions of the review. The missed modelling study by Fraser and colleagues[21] pre-dated COVID-19 and was based on SARS. This study therefore received less weight in the review's analysis than direct evidence based on SARS-CoV-2. The second missed study was deemed more important to the review. It was one of two observational studies on the quarantine of travelers. However, it would not have changed the direction of the finding nor the certainty of evidence grading (which was already very low). Therefore, missing this study would not have changed the review's conclusions. For Review 2: IPC Adherence, the missed study by Tseng and colleagues[23] contributed to nine findings in the review. However, given the high number of other studies additionally contributing and the moderate to high confidence in these findings, it is likely the review would have drawn the same conclusions had the study not been included. Finally, for Review 3: Universal Screening, the missed study by Lavezzo and colleagues[24] would also not have changed the conclusions nor the strength of the evidence for the findings it contributed to. The review author team noted within the review itself that the Lavezzo study did not contain specificity estimates and so had already analysed the effect of excluding this study, concluding that excluding it did not change the findings or range of estimates[17].

As well as assessing the impact of missed studies, we also performed forward citation tracking to ascertain whether any of the missed studies would potentially have been retrieved via this method. This involves assessing the reference lists of included studies as a way of identifying additional studies missed by the electronic database searches. Of the four studies collectively rejected by the crowd, two were cited by other included studies in the reviews: one[21] from Review 1: Quarantine, and the other[24] from Review 3: Universal screening.

Another area of consideration is around whether domain or topic area affected crowd performance. One strength of this study was the range of review question types included: Review 1 was largely focused on observational and modelling studies (interventional designs were includable but unlikely to be found). Review 2 sought mixed methods studies and qualitative studies, Review 3, diagnostic and screening studies, and Review 4, interventional study designs. Research has highlighted the challenge in assessing studies for diagnostic-related reviews[26,27], and this appears to have been borne out in this evaluation study. In addition, no studies were incorrectly rejected for Review 4. This review sought to include studies that assessed the effectiveness of a treatment, convalescent

plasma. This review was most alike other tasks hosted on Cochrane Crowd, namely the RCT identification task. This might account for the crowd's highly accurate and speedy performance.

We also explored whether records that did not have an abstract had an impact on accuracy or consensus measures. The proportion of title-only records for each of the reviews was low (Review 1: 5.7%, Review 2: 7.2%, Review 3: 6.8%, Review 4: 6.6%). However, all four of the missed studies did have abstracts so this was not a factor in terms of negatively impacting crowd sensitivity. Where it did potentially have an impact on crowd performance is in terms of crowd consensus. Overall consensus ranged from 71% to 92% across each of the data sets. However, it was lower across both the eligible studies (range 60% to 65%) and lower still across records that did not have an abstract (54% to 61%). Neither finding is surprizing but both have implications for future potential applications of a crowd model for citation screening. The higher the prevalence of includable studies and/or the higher the proportion of title-only records, the lower crowd consensus is likely to be.

Two other factors are also worth exploration in terms of possible impact on crowd accuracy: the agreement algorithm and the training materials. In terms of the agreement algorithm, we chose an algorithm (three consecutive agreements) that had produced high collective accuracy in other similar pilot projects[9,10]. Would altering the consecutive number of agreeing classifications have made a difference to collective accuracy? Starting with the accuracy of a single classification, the mean accuracy of individual contributors for each review was: 84.2% sensitivity and 82.2% specificity for Review 1: Quarantine; 86.6% sensitivity, 84.1% specificity for Review 2: IPC Adherence; 85.1% sensitivity, 89.9% specificity for Review 3: Universal Screening; and 89.3% sensitivity, 90.9% specificity for Review 4: Convalescent Plasma. Taking the first two consecutive classifications made on each record across the four data sets would have resulted in reduced crowd sensitivity with one additional study being missed per review. With regards to how an algorithm based on four consecutive agreeing classifications would have performed, we do not have the data to model this. However, interesting recent work by Nama and colleagues indicates that excellent sensitivity can be achieved with three assessments per record. In their analysis, increasing this number made little difference to sensitivity but decreased specificity[28].

With regards to the training provided, we were able to provide highly representative records for the test set. We used a set of 16 records for each training module. In the recent evaluation by Nama and colleagues described above, the optimal size for the qualification set was explored. Their analysis

indicated that the optimal size for a qualification set made up of true positives and true negatives was between 10-15 records[28].

Despite this study's focus being on rapid reviews in the context of COVID-19, the range of study types and designs eligible across the four reviews, and the correspondingly high levels of accurate screening by the crowd bode well for this approach being applied beyond a public health setting. Indeed, a recent overview by Burgard and colleagues describes initiatives underway to support 'community-augmented meta-analyses' in the field of psychology, leveraging distributed human effort to help curate the evidence base and produce 'living' or dynamic syntheses[29].

This study has focussed exclusively on the use of crowdsourcing as a means of reliably expediting parts of the study identification stages of evidence synthesis. However, there is a growing field of research exploring the potential of machine learning for citation screening, for example using support vector machine learning classifiers that assign likelihood scores to records. The chief advantage of machine learning over crowdsourcing is time. Records can be classified by a machine learning classifier within minutes, irrespective of the size of the search results set; conversely a crowd will take a variable amount of time (though often still significantly faster than a small review author team). The significant challenge however with applying machine learning alone relates to the high-quality training data required to build a reliable classifier. Also, for a machine learning classifier to operate as a binary classifier (replicating the human classification task), a calibration stage would be needed to ascertain the appropriate score threshold. Another approach, however, would be a hybrid machine-crowd model. This might work well where there is limited training data or where sensitivity is paramount. One possible hybrid configuration would be to employ the classifier to help remove the more obviously not relevant material whilst engaging human effort to assess the remainder. This approach has been used to good effect in Cochrane in both its Screen4Me workflow and within Cochrane's broader Centralised Search Service initiative[30] (as described in Chapters Three and Four).

Despite the safeguards described above, no system will be 100% accurate all the time. As well as quality control measures aimed at maximising crowd performance, review author teams also have a range of possible ways in which they can use the data generated by the crowd within their review production process. Table 8.3 presents three possible workflows regarding the use of the crowd's collective output, each dependent on the required outcome: sensitivity maximizing (i.e., using the crowd in a way that reduces the risk of missing includable studies as much as possible), speed

maximizing, where time is the most critical factor and author team capacity is limited, or specificity maximizing (reducing the number of false positives). The most appropriate approach will depend on the nature, complexity, and scope of the review itself, as well as the time and resources available to the author team.

Table 8.3 Crowdsourcing workflows

Sensitivity maximizing	Crowd assessment + author team dual assessment of conflicting crowd records + author team single assessment of *Possibly relevant* records only
Speed maximizing	Crowd assessment + author team single assessment of *Possibly relevant* records only
Specificity maximizing	Crowd assessment + crowd resolver* + author team single assessment of crowd identified *Possibly relevant* records only

*A crowd resolver is a crowd contributor assesses only records that have received discordant classifications, and makes a final crowd classification on the record.

8.7 Conclusions

This pilot study has demonstrated the feasibility of using a crowd in the study identification process for Cochrane rapid reviews. The crowd performed consistently well across each of the four evaluations in terms of time and accuracy measures. During a global health crisis, when time is of the essence and robust health evidence is critical, using crowdsourcing in this way offers a viable means to expedite the review process and offer willing contributors meaningful ways to get involved. The exact method of crowd application and use of crowd-generated data will depend on the nature of the review itself and the urgency at which the evidence is required.

8.8 Abbreviations

CCSR	Cochrane COVID-19 Study Register
IPC	Infection, prevention and control
RCT	Randomised controlled trial

8.9 Author contributions

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Gerald Gartlehner: conceptualisation, methodology, resources, data curation, writing – reviewing and editing

Gordon Dooley: conceptualisation, data curation, writing – reviewing and editing

Emma Persad: conceptualisation, data curation, writing – reviewing and editing

Barbara Nussbaumer-Streit: conceptualisation, methodology, data curation, visualisation, writing – reviewing and editing

8.10 References

1. Else H. How a torrent of COVID science changed research publishing - in seven charts. Nature. 2020 Dec;588(7839):553. doi: 10.1038/d41586-020-03564-y. PMID: 33328621.
2. Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D'Agostino M, Garcia-Saiso S, Landry M, Rashidian A, Hamilton C, AbdAllah A, Ghiga I, Hill A, Hougendobler D, van Andel J, Nunn M, Brooks I, Sacco PL, De Domenico M, Mai P, Gruzd A, Alaphilippe A, Briand S. Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation. J Med Internet Res. 2020 Jun 26;22(6):e19659. doi: 10.2196/19659. PMID: 32558655; PMCID: PMC7332158.
3. Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, Affengruber L, Stevens A. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol. 2021 Feb;130:13-22. doi: 10.1016/j.jclinepi.2020.10.007. Epub 2020 Oct 15. PMID: 33068715; PMCID: PMC7557165.
4. McCarthy 2017 and Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., etal. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology. 2016;67:68–82. DOI: https://doi.org/10.1016/j.jesp.2015.10.012.
5. Miller JD, Crowe M, Weiss B, Maples-Keller JL, Lynam DR. Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. Personal Disord. 2017;8(1):26-34. doi: 10.1037/per0000191. PMID: 28045305.
6. Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AZ, Vandenberg J. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. 2009 arXiv:0909.2925v1
7. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 May;133:130-139. doi: 10.1016/j.jclinepi.2021.01.006. Epub 2021 Jan 18. PMID: 33476769.
8. Thomas J, Noel-Storr A, McDonald S, Marshall I. Data reuse, machine learning, and crowdsourcing in Screen4Me: how screening burden can be reduced substantially and reliably. Cochrane Colloquium, 2-6 December 2019, Santiago, Chile.
9. Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. J Clin Epidemiol. 2021 Feb;130:23-31. doi: 10.1016/j.jclinepi.2020.09.024. Epub 2020 Sep 30. PMID: 33007457.

10. Noel-Storr AH, Redmond P, Lamé G, Liberati E, Kelly S, Miller L, Dooley G, Paterson A, Burt J. Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study. BMC Med Res Methodol. 2021 Apr 26;21(1):88. doi: 10.1186/s12874-021-01271-4. PMID: 33906604; PMCID: PMC8077753.
11. Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. J Clin Epidemiol. 2021 Feb;130:23-31. doi: 10.1016/j.jclinepi.2020.09.024. Epub 2020 Sep 30. PMID: 33007457.
12. Mortensen ML, Adam GP, Trikalinos TA, Kraska T, Wallace BC. An exploration of crowdsourcing citation screening for systematic reviews. Res Synth Methods. 2017 Sep;8(3):366-386. doi: 10.1002/jrsm.1252. Epub 2017 Jul 4. PMID: 28677322; PMCID: PMC5589498.
13. Nama N, Sampson M, Barrowman N, Sandarage R, Menon K, Macartney G, Murto K, Vaccani JP, Katz S, Zemek R, Nasr A, McNally JD. Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study. J Med Internet Res. 2019 Apr 29;21(4):e12953. doi: 10.2196/12953. PMID: 31033444; PMCID: PMC6658317.
14. Noel-Storr A, Thomas J, Dooley G. Using crowdsourcing and machine learning for study identification: a quantitative and qualitative evaluation of Cochrane's Screen4Me workflow. 27th Cochrane Colloquium.
15. Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, Wagner G, Siebert U, Christof C, Zachariah C, Gartlehner G. Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. Cochrane Database Syst Rev. 2020 Apr 8;4(4):CD013574. doi: 10.1002/14651858.CD013574. PMID: 32267544; PMCID: PMC7141753.
16. Houghton C, Meskell P, Delaney H, Smalle M, Glenton C, Booth A, Chan XHS, Devane D, Biesty LM. Barriers and facilitators to healthcare workers' adherence with infection prevention and control (IPC) guidelines for respiratory infectious diseases: a rapid qualitative evidence synthesis. Cochrane Database Syst Rev. 2020 Apr 21;4(4):CD013582. doi: 10.1002/14651858.CD013582. PMID: 32315451; PMCID: PMC7173761.
17. Viswanathan M, Kahwati L, Jahn B, Giger K, Dobrescu AI, Hill C, Klerings I, Meixner J, Persad E, Teufer B, Gartlehner G. Universal screening for SARS-CoV-2 infection: a rapid review. Cochrane Database Syst Rev. 2020 Sep 15;9:CD013718. doi: 10.1002/14651858.CD013718. PMID: 33502003.
18. Valk SJ, Piechotta V, Chai KL, Doree C, Monsef I, Wood EM, Lamikanra A, Kimber C, McQuilten Z, So-Osman C, Estcourt LJ, Skoetz N. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a rapid review. Cochrane Database Syst Rev. 2020 May

14;5(5):CD013600. doi: 10.1002/14651858.CD013600. Update in: Cochrane Database Syst Rev. 2020 Jul 10;7:CD013600. PMID: 32406927; PMCID: PMC7271896.

19. COVID Quest: https://www.cochrane.org/news/help-find-studies-about-covid-19-join-covid-quest [last accessed: 13 November 2021].
20. Noel-Storr A, Dooley G, Featherstone R, Wisniewski S, Shemilt I, Thomas J, Gartlehner G, Nußbaumer-Steit B, Mavergames C. Crowdsourcing and COVID-19: a case study of Cochrane Crowd. JEAHIL [Internet]. 24Jun.2021 [cited 6Nov.2021];17(2):27-1. Available from: http://ojs.eahil.eu/ojs/index.php/JEAHIL/article/view/467.
21. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci U S A. 2004 Apr 20;101(16):6146-51. doi: 10.1073/pnas.0307506101. Epub 2004 Apr 7. PMID: 15071187; PMCID: PMC395937.
22. Arima Y, Shimada T, Suzuki M, Suzuki T, Kobayashi Y, Tsuchihashi Y, Nakamura H, Matsumoto K, Takeda A, Kadokura K, Sato T, Yahata Y, Nakajima N, Tobiume M, Takayama I, Kageyama T, Saito S, Nao N, Matsui T, Sunagawa T, Hasegawa H, Ohnishi M, Wakita T. Severe Acute Respiratory Syndrome Coronavirus 2 Infection among Returnees to Japan from Wuhan, China, 2020. Emerg Infect Dis. 2020 Jul;26(7):1596–600. doi: 10.3201/eid2607.200994. Epub 2020 Jun 21. PMID: 32275498; PMCID: PMC7323539.
23. Tseng HC, Chen TF, Chou SM. SARS: Key factors in crisis management. J Nurs Res. 2005 Mar;13(1):58-65. PMID: 15977136.
24. Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, Rossi L, Manganelli R, Loregian A, Navarin N, Abate D, Sciro M, Merigliano S, De Canale E, Vanuzzo MC, Besutti V, Saluzzo F, Onelia F, Pacenti M, Parisi SG, Carretta G, Donato D, Flor L, Cocchio S, Masi G, Sperduti A, Cattarino L, Salvador R, Nicoletti M, Caldart F, Castelli G, Nieddu E, Labella B, Fava L, Drigo M, Gaythorpe KAM; Imperial College COVID-19 Response Team, Brazzale AR, Toppo S, Trevisan M, Baldo V, Donnelly CA, Ferguson NM, Dorigatti I, Crisanti A; Imperial College COVID-19 Response Team. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature. 2020;584(7821):425-429. Epub 2020 Jun 30. Erratum in: Nature. 2021 Feb;590(7844):E11. PMID: 32604404.
25. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, König F. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. 2020;121:20-28.
26. Cohen JF, Korevaar DA, Bossuyt PM. Diagnostic accuracy studies need more informative abstracts. Eur J Clin Microbiol Infect Dis. 2019;38(8):1383-1385.

27. Gurung P, Makineli S, Spijker R, Leeflang MMG. The Emtree term "diagnostic test accuracy study" retrieved less than half of the diagnostic accuracy studies in Embase. J Clin Epidemiol. 2020;126:116-121.
28. Nama N, Barrowman N, O'Hearn K, Sampson M, Zemek R, McNally JD. Quality control for crowdsourcing citation screening: the importance of assessment number and qualification set size. J Clin Epidemiol. 2020;122:160-162.
29. Burgard, T., Bošnjak, M., & Studtrucker, R. (2021). Community-augmented meta-analyses (CAMAs) in psychology: Potentials and current systems. Zeitschrift für Psychologie, 229(1):15-23.
30. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020;127:142-150.

Chapter 9

Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier

This original manuscript was published in the journal *BMC Systematic Reviews*

Citation: Shemilt I, Noel-Storr A, Thomas J, Featherstone R, Mavergames C. Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier. Systematic Reviews 2022;11(1):15.
DOI: 10.1186/s13643-021-01880-6

9.1 Abstract

Background

This study developed, calibrated, and evaluated a machine learning (ML) classifier designed to reduce study identification workload in maintaining the Cochrane COVID-19 Study Register (CCSR), a continuously updated register of COVID-19 research studies.

Methods

A ML classifier for retrieving COVID-19 research studies (the "Cochrane COVID-19 Study Classifier") was developed using a data set of title-abstract records 'included' in, or 'excluded' from, the CCSR up to 18th October 2020, manually labelled by information and data curation specialists or the Cochrane Crowd. The classifier was then calibrated using a second data set of similar records 'included' in, or 'excluded' from, the CCSR between 19th October and 2nd December 2020, aiming for 99% recall. Finally, the calibrated classifier was evaluated using a third data set of similar records 'included' in, or 'excluded' from, the CCSR between 4th and 19th January 2021.

Results

The Cochrane COVID-19 Study Classifier was trained using 59,513 records (20,878 of which were 'included' in the CCSR). A classification threshold was set using 16,123 calibration records (6,005 of which were 'included' in the CCSR) and the classifier had a precision of 0.52 in this data set at the target threshold recall >0.99. The final, calibrated COVID-19 classifier correctly retrieved 2,285 (98.9%) of 2,310 eligible records but missed 25 (1%), with a precision of 0.638 and a net screening workload reduction of 24.1% (1,113 records correctly excluded).

Conclusions

The Cochrane COVID-19 Study Classifier reduces manual screening workload for identifying COVID-19 research studies, with a very low and acceptable risk of missing eligible studies. It is now deployed in the live study identification workflow for the Cochrane COVID-19 Study Register.

9.2 Background

The COVID-19 pandemic has resulted in an unprecedented level of article publications[1,2] of which only a small percentage report study data or analytics[3]. This presented the systematic review community with significant challenges to identify and classify relevant study evidence reliably, accurately, and efficiently, to enable the rapid synthesis and use of cumulative bodies of evidence to inform international, national and local responses to the evolving global health crisis.

As the pandemic took hold, a number of initiatives were started with the aim of identifying and classifying COVID-19 research. Two such initiatives are the COVID-19 Open Research Dataset (CORD-19) developed by the Semantic Scholar Team at the Allen Institute[4] and COVID-19 L·OVE by Epistemonikos[5]. Each initiative had variable aims and different approaches to collating the required information; but, to our knowledge, the Cochrane COVID-19 Study Register (CCSR) was the only product designed to support rapid evidence synthesis through the identification and classification of ongoing and completed primary studies. Cochrane was able to utilise existing technical infrastructure, processes and human resource to create an open access register of COVID-19 studies. The Cochrane COVID-19 Study Register (CCSR)[6] includes primary, human studies across a broad range of areas relevant to COVID-19, including the treatment and management of the virus, diagnosis, prognosis, transmission and prevention, mechanism, epidemiology and the wider impact of the pandemic on populations and health services. The CCSR study records are validated and maintained by a team of Cochrane information and data curation specialists. Automated searches retrieve results via daily or weekly API calls across a range of sources. The results are then de-duplicated and screened. A sub-set of results (those retrieved from Embase) are sent to Cochrane Crowd, Cochrane's citizen science platform[7]; the rest are screened by the core register team[8,9]. The screening process involves an assessment of record eligibility based on titles and abstracts. For records without abstracts, more information is sought before a judgement is made. Eligible studies are then tagged by the team or by the crowd according to study type, study design, and study aims. Intervention studies are also annotated according to their PICO (population, intervention, comparator and outcome) components. These tagging and annotation activities, together with the largely manual process of linking related reports together, are resource intensive.

In July 2020, we convened a series of meetings between the CCSR team and the team from the EPPI Centre (UCL) and Centre for Reviews and Dissemination (University of York), which has been maintaining a living map of COVID-19 research evidence (the 'C-19 living map') commissioned by the UK Department of Health and Social Care. The purpose of these meetings was to share best practice and reduce duplication of effort between the respective workflows being used to keep these two overlapping resources up to date; and we have initially focused on strategies to reduce manual screening burden in the selection of eligible articles.

As the rate of COVID-19 publishing shows little sign of slowing, introducing machine learning (ML) into COVID-19 study identification workflows could offer important gains in terms of workload

reduction[10] so long as the corollary risk of 'missing' (or 'losing') relevant research studies is acceptably low. The C-19 living map team had recently developed and deployed a ML classifier for this purpose; and similar classifiers have previously been deployed in Cochrane's Centralised Search Service and Screen4Me workflows, to support efficient identification of randomised controlled trials (RCTs)[11].

For both the CCSR and the C-19 living map, we decided to deploy a ML classifier to discard records scoring below an identified threshold score, calibrated to minimise the risk of 'missing' eligible articles. However, given differences between the respective scopes and eligibility criteria of these two resources, we decided that a new binary ML classifier should be specifically developed for the CCSR workflow.

9.3 Methods

In this study, we aimed to train (Stage 1), calibrate (Stage 2) and evaluate (Stage 3) a binary ML classifier ('the classifier') designed to reduce study identification workload in maintaining the CCSR, with an acceptably low corollary risk of 'missing' records of 'included' (eligible) studies. We therefore needed to assemble three separate data sets from the CCSR screening workflows (see below and 'Availability of data and materials').

Training (Stage 1)

In Stage 1, we assembled a training data set containing bibliographic title-abstract records of all articles manually screened for eligibility for the CCSR from its first search date (20th March 2020) up until 18th October 2020. Embase.com records had only been recently added to the CCSR's sources by mid-October and a backlog of medRxiv preprints was still being processed. As the CCSR's other sources were trial registers (not bibliographic title-abstract records), most of the training set records were from PubMed. These records had originally been identified using conventional Boolean searches of selected electronic bibliographic databases and trials registries, before being manually screened and labelled as either 'included' (eligible for the CCSR) or 'excluded' (ineligible) by Cochrane information specialists or the Cochrane Crowd[7]. The search strategies used can be seen on the *About* page of the CCSR[6]. After removing trials registry records, we were left 59,513 records, of which 20,878 were labelled as 'included' in the CCSR, and 38,635 were 'excluded'. These records were imported into *EPPI-Reviewer*[12], assigned to code sets, and used to train a logistic regression classifier using tri-gram 'bag of words' features, implemented in the SciKit-Learn python library, with

'included' records designated as the positive class (class of interest) and 'excluded' records as the negative class.

Calibration (Stage 2)

In Stage 2, we assembled a calibration data set containing 16,123 similar records manually screened for eligibility for the CCSR between 19th October and 2nd December 2020, again labelled as 'included' (6,005 eligible records) or 'excluded' (10,118 ineligible records) by the same people and process, and with trials registry records having again been removed. The records were imported into *EPPI-Reviewer*, assigned to code sets, and used to calibrate the classifier developed in Stage 1. Specifically, we applied the classifier to 16,123 calibration records, which automatically assigned a score (0-100) to each record. We then computed the threshold score that captured >99% of the 'included' records in this data set (i.e., recall >0.99). 0.99 is the threshold level of recall that is currently required for ML classifiers to be deployed in Cochrane systems and workflows[13]. We also computed standard performance metrics, namely: (cumulative) recall, (cumulative) precision and net workload reduction.

Evaluation (Stage 3)

In Stage 3, we assembled an evaluation data set of similar records containing 4,722 records manually screened for eligibility for the CCSR between 4th and 19th January 2021, once again labelled as 'included' (2,310 eligible records) or 'excluded' (2,412 ineligible records), with trials registry records removed. The records were imported into *EPPI-Reviewer*, assigned to code sets, and used to evaluate the classifier developed in Stage 1. Specifically, we applied the classifier to 4,722 evaluation records, identified 'included' and 'excluded' records scoring above and below the threshold score we had computed in Stage 2; and then we computed (cumulative) recall, (cumulative) precision and net workload reduction. We also analysed characteristics of 'included' articles that would have been 'missed' by the workflow if the classifier had been implemented.

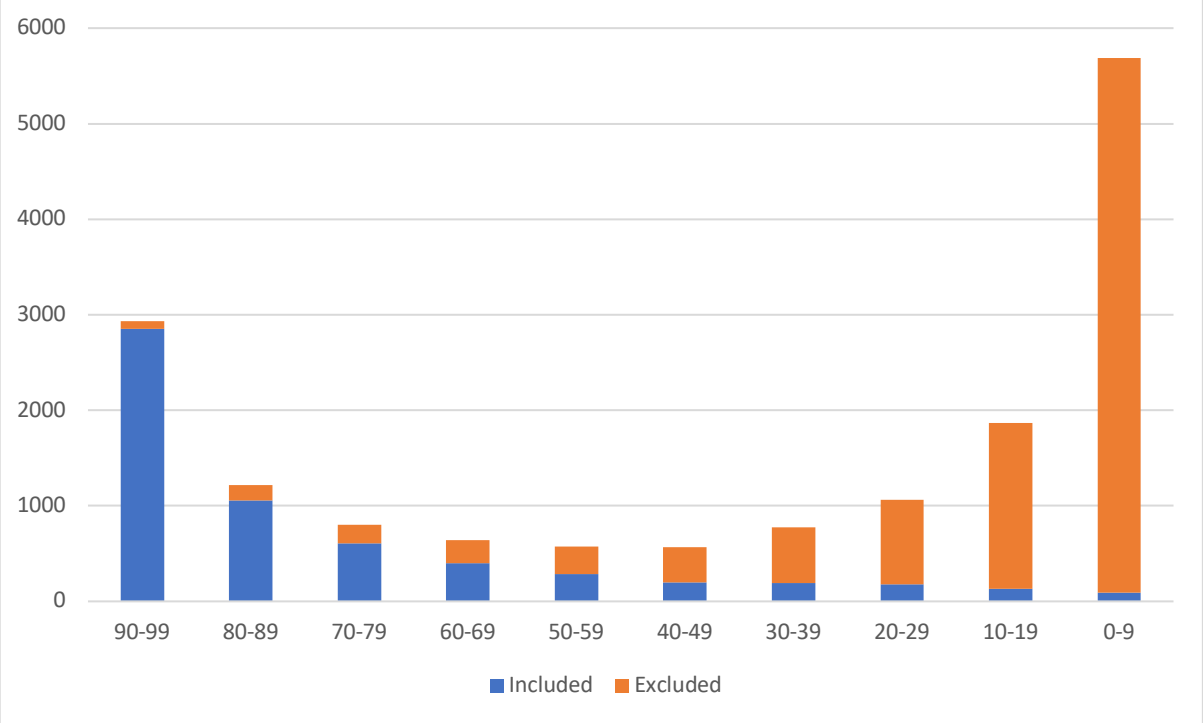
Finally, we compared key characteristics of articles between the three study data sets described above in this section (training, calibration, evaluation), to check post-hoc that they comprised similar enough sets of records to validate our results from calibrating and evaluating the classifier.

9.4 Results

Calibration

Results from calibrating the Cochrane COVID-19 Study Classifier (Stage 2) are shown in Figure 9.1 and Table 9.1. The threshold classifier score at target recall >0.99 was identified as 7 (Table 9.1), which means that >99% of 'included' records in the calibration set scored 7 or above. In this data set, retaining records scoring 7 or above, to achieve target recall >0.99 among 'included' records, would have resulted in an overall workflow precision of 0.52, with a corollary 29.1% reduction in manual screening workload.

Figure 9.1 Distribution of classifier scores among 'included' and 'excluded' calibration records (N=16,123) and related performance metrics



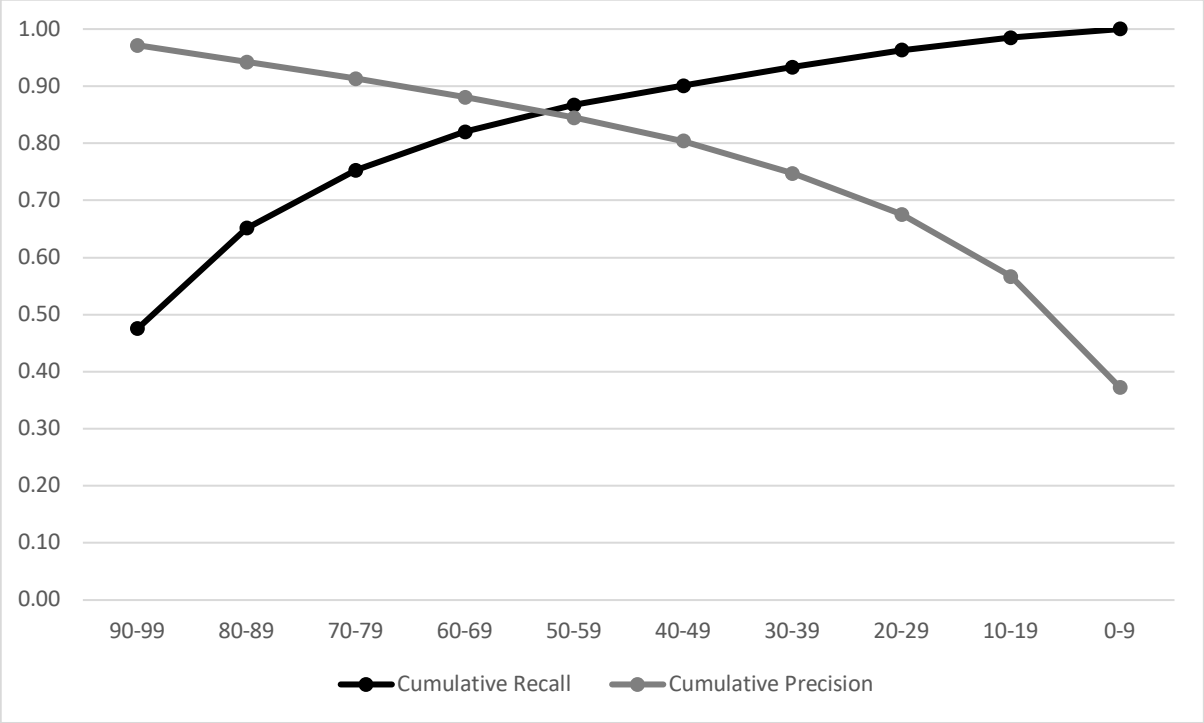


Table 9.1 Distribution of classifier scores among 'included' and 'excluded' calibration records and related performance metrics

Classifier Score	90-99	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19	0-9
Included N	2,853	1,059	610	402	284	202	195	180	129	91
Excluded N	83	156	190	237	290	364	578	885	1,736	5,599
Totals	2,936	1,215	800	639	574	566	773	1,065	1,865	5,690
Precision	0.97	0.87	0.76	0.63	0.49	0.36	0.25	0.17	0.07	0.02
Cumulative Recall	0.48	0.65	0.75	0.82	0.87	0.90	0.93	0.96	0.98	1.00
Cumulative Precision	0.97	0.94	0.91	0.88	0.84	0.80	0.75	0.68	0.57	0.37

Threshold Classifier Score (Recall >0.99)	7
Screened Included N*	5,950
Screened Excluded N*	5,487
Precision*	0.52
Discarded ('Lost') Included N*	55
Discarded Excluded N*	4,631
Net Workload Reduction N*	4,686
Net Workload Reduction %*	29.1%

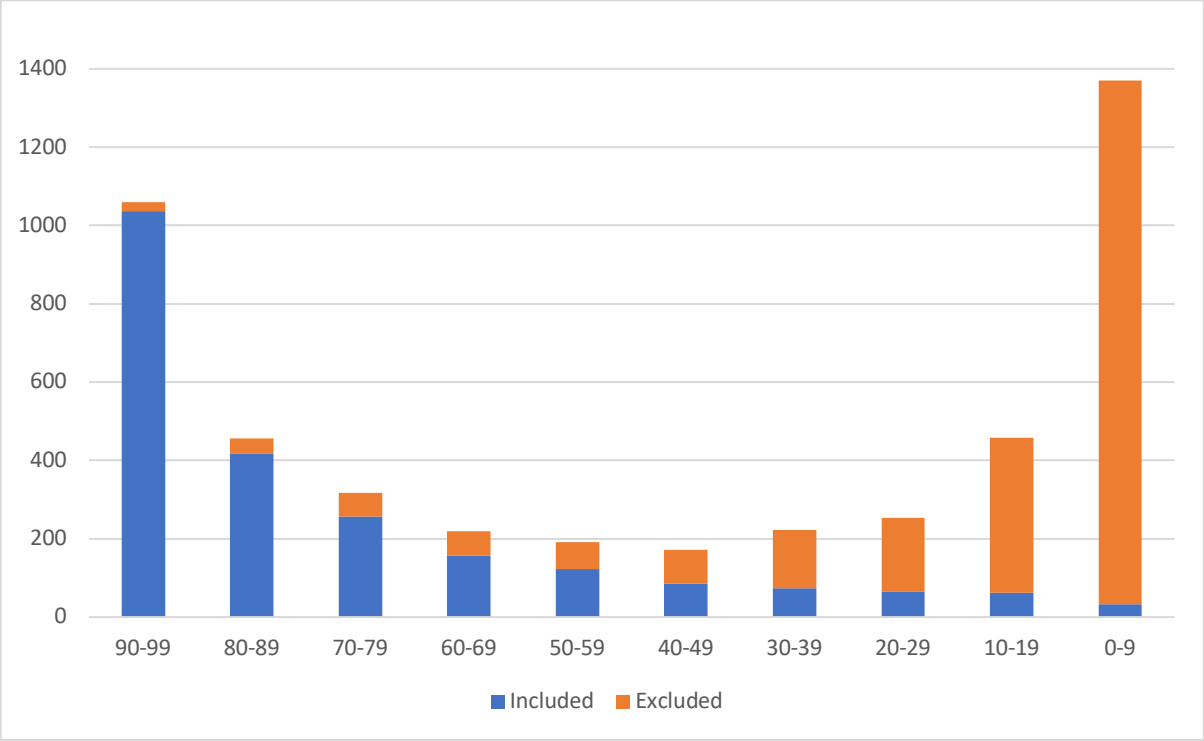
* At Threshold Score = 7 (Recall >0.99)

Evaluation

Evaluation results for the classifier are shown in Figure 9.2 and Table 9.2. In the evaluation data set, retaining records scoring at or above the calibrated threshold score would have resulted in 0.99

recall among ‘included’ records, with an overall workflow precision of 0.64 and a corollary 24.1% reduction in manual screening workload.

Figure 9.2 Distribution of classifier scores among ‘included’ and ‘excluded’ evaluation records (N=4,722) and related performance metrics



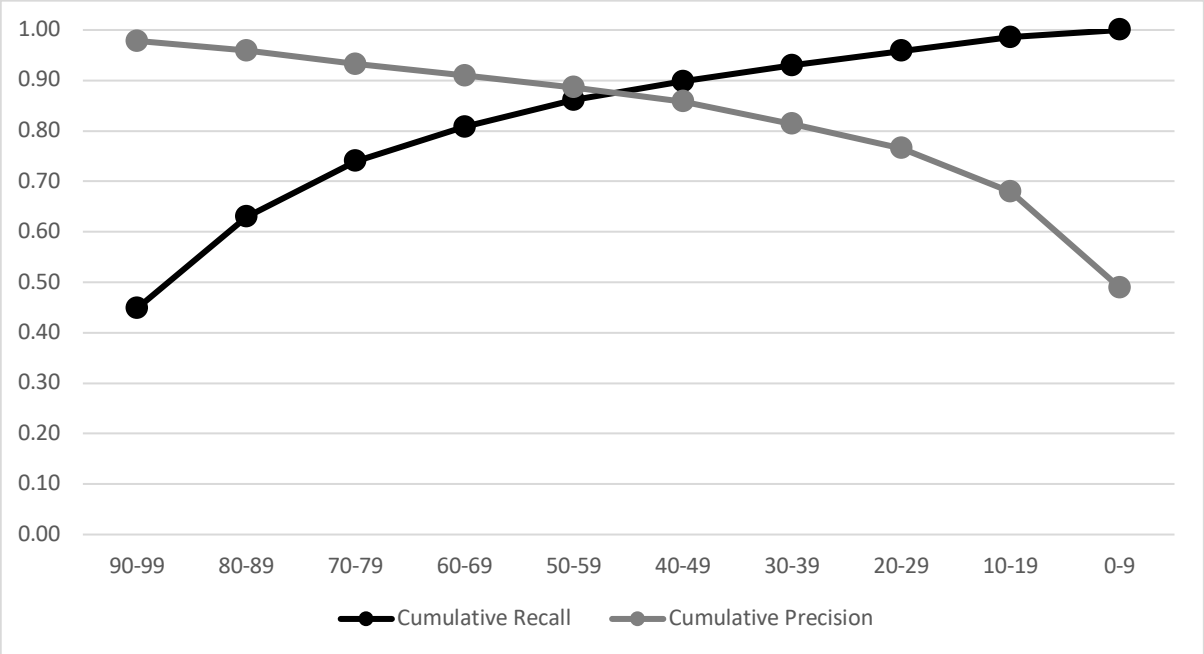


Table 9.2 Distribution of classifier scores among 'included' and 'excluded' evaluation records and related performance metrics

Classifier Score	90-99	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19	0-9
Included N	1037	417	256	157	122	85	74	66	63	33
Excluded N	23	39	62	62	69	87	149	188	395	1338
Totals	1060	456	318	219	191	172	223	254	458	1371
Precision	0.98	0.91	0.81	0.72	0.64	0.49	0.33	0.26	0.14	0.02
Cumulative Recall	0.45	0.63	0.74	0.81	0.86	0.90	0.93	0.96	0.99	1.00
Cumulative Precision	0.98	0.96	0.93	0.91	0.89	0.86	0.81	0.77	0.68	0.49

Threshold Classifier Score	7
Screened Included N*	2,285
Screened Excluded N*	1,299
Precision	0.64
Discarded ('Lost') Included N*	25
Discarded Excluded N*	1,113
Recall	0.99
Net Workload Reduction N*	1,138
Net Workload Reduction %*	24.1%

* At Threshold Score = 7

In our analysis of the 25 (1%) 'missed' (discarded) 'included' records, we found that 12 were title-only records. Of these, four were errata or replies to studies already included in the CCSR and were therefore not the primary reference to the study. All but one of the 'missed includes' had been sourced from PubMed. Only two were records of interventional studies, the rest were records of observational studies. One 'missed' interventional study was an RCT but it was not reporting the results of the RCT. The other one was a single arm study that was not about COVID-19 patients, but the broader impact of the pandemic on the mental health of students, and the effects of a mindfulness component of the intervention described. Of the remaining 'missed' observational studies, most were studies looking at the broader impact of the pandemic on health services or selected populations. Three were small case-control or cohort studies that were diagnostic or prognostic in their aims. The remining three 'missed' records were all studies concerned with virus mutations. It is likely that this kind of study was not part of our stage 1 (training) data set, which contains studies from the first few months of the pandemic.

Post hoc analysis of data set key characteristics

Results from comparing key characteristics between data sets used in the training, calibration, and evaluation of the COVID-19 Study Classifier are shown in Table 9.3. Stage 1 (training) and Stage 2 (calibration) data sets were very similar in terms of the proportion of 'included' records in each set

(35%, 37% respectively). The Stage 3 (evaluation) data set, compiled of records manually screened for the CCSR during January 2021, had a higher proportion of 'included' records, at almost 50%. Each data set included a substantial proportion of title-only records (i.e., records without abstracts). The Stage 1 data set had the largest proportion of such records: 18,669 records (31%), of which 4,495 were includes. Data sets 2 and 3 and a lower, but similar, proportion of title-only records: 23% and 19% respectively.

Table 9.3 Key characteristics of development, calibration and evaluation data sets

Data set (classifier development stage)	Size	Number of eligible records (%)	Number of title-only records (%)	Number of title-only records that were eligible (%)	Provenance of records
Data set 1 (Training)	59,513	20,878 (35.1%)	18,669 (31.4%)	4,495 (21.5%)	3229 (5.4%) – Embase 2083 (3.5%) – preprint 54201 (91.1%) - PubMed
Data set 2 (Calibration)	16,123	6,005 (37.2%)	3626 (22.5%)	821 (13.7%)	1994 (12.4%) – Embase 287 (1.8%) – pre-print 13842 (85.8%) - PubMed
Data set 3 (Evaluation)	4,722	2,310 (48.9%)	896 (19.0%)	285 (12.3%)	89 (1.9%) – Embase 202 (4.3%) – pre-print 4431 (93.8%) - PubMed

9.5 Discussion

We developed a binary ML classifier with the aim of reducing screening workload for the CCSR. Calibrated to achieve 99% recall, the classifier reduced screening workload by 24.1% in the evaluation data set. This finding was especially encouraging given the proportion of eligible records in this data set was close to 50%; and almost one in five of the records were 'title-only', with relatively few text features for classification, compared to records with accompanying abstracts. Title-only records in the context of the COVID pandemic can be resource- and time-intensive to manually assess. For many, more information will need to be found before a judgement on whether the record is eligible can be made. Having a classifier able to reliably reject ineligible title-only records is valuable and will free up human resource to assess the more unclear title-only records.

One of the main strengths of this study is the quality of the three data sets. We were able to use highly representative records for each stage, with a high level of confidence in the quality of each, derived as they were from the Cochrane Centralised Search Service team and Cochrane Crowd[7]. In addition, the training data set was fairly large (n=59,513), made up of both the class of interest ('included') and non-eligible records ('excluded'). Records within the class of interest set

encompassed all eligible study types (observational, interventional, qualitative, and modelling studies) and designs, and had good coverage across the range of possible study aims.

A potential limitation is that most records comprising each of the three study data sets were sourced from PubMed (of which a large proportion are also likely to have been indexed in Embase). This is unlikely to be an issue when applying the classifier to bibliographic records of journal articles identified from other database sources; but caution would be needed when applying the classifier to records with a different structure, for example, trial registry records. While many trial registry records contain similar information to a standard bibliographic record that could, in principle be parsed and added to the title-abstract records prior to their classification, it is important to be aware of which fields map well to each other across the different record types, and in some cases to exclude certain fields of information that might confuse the classifier – such as trial exclusion criteria. As such, further work would be needed to evaluate the performance of this classifier when applied to records incorporating selected text from trial registry records. We could also investigate the potential to incorporate such records into sets used to retrain and recalibrate periodically updated versions of this classifier.

In this paper we have focused on reporting the deployment of a machine learning classifier in a real-world scenario over a short period of time. The method employed, using train, test and calibration data sets and easily interpretable probabilities from a logistic regression classifier, provides a robust basis for future work, and has proved acceptable to Cochrane. A workload reduction of ~25% is substantial given the high recall that must be achieved. However, we do not rule out that deployment of more sophisticated machine learning classification algorithms may be able to push the reported savings in workload marginally higher.

Evolution in the scope, aims, and topics and text features of COVID-19 research over time suggest that ML classifiers which, like this one, that have been prospectively developed, are likely to need to be periodically retrained, recalibrated and re-evaluated, in order to minimise the risk of 'losing' (or 'missing') new bodies (or 'strands') of relevant research, with new 'previously unseen' text features, that are likely to emerge as the pandemic continues to unfold. Periodically updated training, calibration and evaluation data sets should be prospectively assembled to comprise records from three consecutive time periods, as we have done in the current study. This approach is robust in terms of its external validity, as it is consistent with the real-world use scenario in which such classifiers are deployed, where we do not know in advance how the research literature will evolve

following their (re-) deployment. (Re-)calibrating and (re-)evaluating the classifier using records from consecutive time periods immediately succeeding the one covered by records in the (re-)training data set therefore confers further confidence (alongside the size and breadth of our study data sets) that any subtle evolution or 'shifts' in the scope and text features of bibliographic records of published COVID-19 research over time are unlikely to adversely impact on the performance of the deployed classifier in the short-term.

In late January 2021, the classifier developed in this study was deployed in the Cochrane COVID-19 register workflow, with records retrieved from PubMed and Embase.com being run through it. Workload reduction in terms of screening effort has been reduced in practice by approximately 20%-25%, which is in line with the expected reduction based on this study. The classifier is also being used to help prioritise screening by ordering the records that score above the cut-point from highest to lowest score. Feedback from the screening team has indicated that records that receive high scores are almost always eligible studies, but they are often not the higher priority interventional studies. This is very likely due to the high prevalence of observational studies in the data sets used.

Next steps

The Cochrane COVID-19 Study Classifier reduces screening burden by cutting the number of excludes to assess by approximately half. This is a helpful start but with the proportion of records eligible being around 50% (as it has been for the last six months for the CCSR), an 'exclusion' classifier can only do so much. In addition, the rate of publication on COVID-19 shows no sign of slowing with the average number of new studies identified for the CCSR averaging 4600 per month over the last six months. Therefore, we are now developing additional automated approaches to maintain the CCSR. With over 60,000 COVID-19 related studies identified and tagged in the register, we are developing additional ML classifiers that will assign or suggest both study design characteristics and study aims to potentially eligible studies. We are also developing automated approaches to assigning PICO characteristics to interventional studies. Here we will use crowd and ML capabilities in a hybrid approach to keeping up with the deluge of publications on COVID-19.

9.6 Conclusions

The Cochrane COVID-19 Study Classifier can reduce manual screening workload for identifying COVID-19 research studies, with a very low and acceptable risk of missing eligible studies. This classifier is now deployed in the study identification workflow for the Cochrane COVID-19 Study Register.

9.7 Abbreviations

CCSR	Cochrane COVID-19 Study Register
ML	Machine learning
RCT	Randomised controlled trial

9.8 Author contributions

Ian Shemilt: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

Anna Noel-Storr: conceptualisation, methodology, investigation, data curation, visualisation, supervision, writing – original draft preparation

James Thomas: conceptualisation, methodology, data curation, writing - reviewing and editing

Robin Featherstone: conceptualisation, methodology, writing – reviewing and editing

Chris Mavergames: conceptualisation, methodology, writing – reviewing and editing

9.9 References

1. Odone A, Salvati S, Bellini L, Bucci D, Capraro M, Gaetti G, Amerio A, Signorelli C. The runaway science: a bibliometric analysis of the COVID-19 scientific literature. Acta Biomed. 2020 Jul 20;91(9-S):34-39. doi: 10.23750/abm.v91i9-S.10121. PMID: 32701915; PMCID: PMC8023084.
2. Else H. How a torrent of COVID science changed research publishing - in seven charts. Nature. 2020 Dec;588(7839):553. doi: 10.1038/d41586-020-03564-y. PMID: 33328621.
3. Raynaud, M., Zhang, H., Louis, K. et al. COVID-19-related medical research: a meta-research and critical appraisal. BMC Med Res Methodol 21, 1 (2021). https://doi.org/10.1186/s12874-020-01190-w.
4. COVID-19 Open Research Dataset (CORD-19). https://www.semanticscholar.org/cord19. [Accessed 04 January 2022].
5. COVID-19 L·OVE. https://app.iloveevidence.com. [Accessed 04 January 2022].
6. Cochrane COVID-19 Study Register. https://covid-19.cochrane.org. [Accessed 04 January 2022].
7. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, Wisniewski S, McDonald S, Murano M, Glanville J, Foxlee R, Beecher D, Ware J, Thomas J. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021 Jan 18:S0895-4356(21)00008-1. doi: 10.1016/j.jclinepi.2021.01.006. Epub ahead of print. PMID: 33476769.
8. Metzendorf MI, Featherstone RM. Evaluation of the comprehensiveness, accuracy and currency of the Cochrane COVID-19 Study Register for supporting rapid evidence synthesis production [published online ahead of print, 2021 Jun 5]. Res Synth Methods. 2021;10.1002/jrsm.1501. doi: https://doi.org/10.1002/jrsm.1501.
9. Featherstone R, Last A, Becker L, Mavergames C. Rapid development of the Cochrane COVID-19 Study Register to support review production. In: Collaborating in response to COVID-19: editorial and methods initiatives across Cochrane. Cochrane Database Sys Rev. 2020;(12 Suppl 1):37-40. doi: 10.1002/14651858.CD202002.
10. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S (2015) Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews 4:5. doi:10.1186/2046-4053-4-5.
11. Noel-Storr AH, Dooley G, Wisniewski S, Glanville J, Thomas J, Cox S, Featherstone R, Foxlee R. Cochrane Centralised Search Service showed high sensitivity identifying randomized controlled trials: A retrospective analysis. J Clin Epidemiol. 2020 Nov;127:142-150. doi: 10.1016/j.jclinepi.2020.08.008. Epub 2020 Aug 13. PMID: 32798713.

12. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M (2020) EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. EPPI-Centre, UCL Social Research Institute, University College London.
13. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021;133:140-151.

Chapter 10. Discussion

10.1 Introduction

A recent bibliometric analysis by Wang et al., demonstrates the wide adoption of crowdsourcing across many fields[1]. This expansion, coupled with diverse domain applications, initially made it challenging for the field to organise itself as a coherent academic discipline[2]. In addition, much research in this area is applied research, undertaken to address very domain-specific problems, and as such could be viewed as limited conceptually in terms of broader applicability to other domains. However, in the last decade crowdsourcing has emerged as an academic discipline in its own right. Recently formed international organisations have been established to set standards and provide researchers, academics, practitioners, and the public with forums for discussion and scholarly investigation[3,4]. This chapter aims to situate this research within the broader context of crowdsourcing as an emerging academic discipline.

10.2 The growth of crowdsourcing as an academic discipline

Theoretical foundations and reference frameworks are a critical part of any field's claim to scientific rationality. In the area of crowdsourcing, as more and more organisations undertake activities to engage dispersed populations (i.e., crowds), theoretical frameworks are emerging to support and develop our conceptual understanding of this activity. These frameworks build on established theories from existing, related research fields, including knowledge and information management, information systems, and artificial intelligence, to produce new frameworks with unique and innovative concepts of their own.

10.3 The Theory of Crowd Capital

One such framework, developed by John Prpić and Prashant Shukla, is the Theory of Crowd Capital[5,6]. The framework is a generalisable framework for studying IT-mediated crowd engagement. It takes as its starting point, the knowledge-based view of the organisation[7]. Here, knowledge is viewed as a difficult to replicate resource, and as such can account for the variation in organisational capabilities and performance. Crowd Capital is therefore a knowledge resource generated through the organisation's use of Crowd Capability. Crowd Capability is defined as the way in which an organisation engages with the dispersed knowledge of individuals, i.e., the crowd. The definition of a crowd in this conceptual model is broad: a crowd is any population of individuals who supply knowledge to the organisation. As such, a crowd can exist inside of an organisation, be external to it,

or a combination of both. The notion of the existence and value of dispersed knowledge is not new. Friedrich Hayek's 1945 text entitled *The Use of Knowledge in Society*[8], describes dispersed knowledge as a 'body of very important but inorganized knowledge'. The Theory of Crowd Capital centres on the view that organisations exist and compete in an environment of dispersed knowledge.

Traditionally, the production of scientific knowledge has been a predominantly top-down process, with researchers often portrayed in metaphorical ivory towers, somewhat separated from the rest of society until ready to communicate out the results of their research. Such an approach has been found wanting, and in health-related research has led to research being badly aligned with patient and societal needs. Advances in technology and cultural changes have enabled new models of knowledge production to emerge. Technology has advanced the way data is collected, processed, analysed and integrated with other systems and data. Culturally, Health 2.0 promotes individuals actively participating in their healthcare, including empowering patients to have greater control over their own healthcare needs, all in conjunction with Web 2.0 technology[9]. The huge rise in citizen science initiatives, across multiple domain areas, over the last decade demonstrates the increased recognition of the existence and value of dispersed knowledge.

In the Theory of Crowd Capital, the Crowd Capability concept is distinguished from the crowd. It is defined as an organisational level capability composed of three key parameters: structure, content, and process. The structure component, which is always an information systems-mediated phenomenon, refers to the organisation's technical infrastructure that is developed and deployed to engage the crowd. The content component connotes the data goals (i.e., the required knowledge or information) that the organisation seeks from the crowd. The final parameter, process, is defined as the procedures or workflows that the organisation will implement in order to organise, integrate and utilise the incoming knowledge, information or data. Through Crowd Capability therefore, the organisation puts in place the structure, content and processes to access and organise the dispersed knowledge from individuals. It is this capturing of dispersed knowledge through Crowd Capability efforts that endows organisations with information or knowledge previously unavailable to them. The internal processing of this in turn results in the generation of Crowd Capital within the organisation.

Crowd Capital is the knowledge resource. It is termed as such because of its ability to facilitate an organisation's productivity and potentially to generate economic benefit for the organisation. It also

requires investment as described above in the Crowd Capability tripartite parameters of structure, content, and process.

This conceptual framework is agnostic to organisational type (e.g., public, private etc.) and crowdsourcing type (e.g., paid microtasking, voluntary citizen science initiatives etc.), and modality (e.g., distributed human intelligence tasking, peer-vetted creative production etc.), and provides a holistic view of crowdsourcing from the organisational standpoint. This flexibility allows for, and appreciates, the diversity of possible crowdsourcing initiatives being both researched and practiced. The research undertaken as part of this thesis fits within this theoretical framework. Each research question detailed in the Introduction maps to part, or parts, of the model: Dispersed Knowledge is identified as existing and of being of potential value to the organisation (Chapters 2, 4, 5, 7 and 8); Crowd Capability is optimised through development of a state-of-the-art crowdsourcing platform, and accompanying workflows to turn dispersed knowledge into knowledge that is both useful and usable (Chapters 2, 3, and 4) and Crowd Capital – what is produced as a result of harnessing the crowd's capabilities – resulting in efficiencies in the review production process (Chapters 4 and 9).

10.4 The Four Pillars of Crowdsourcing

Hosseini and colleagues went on to develop a taxonomy for crowdsourcing which they divided into four parts, termed 'pillars', that together constitute the entire crowdsourcing operation[10]. Like Prpić's framework discussed above, Hosseini's taxonomy aims to accommodate both the diversity and commonality of crowdsourcing seen across multiple disciplines. The four pillars are:

(1) The Crowd: the people who take part in the crowdsourced activity
(2) The Crowdsourcer: the entity who seeks completion of a task via a crowd
(3) The Crowdsourcing task: the activity in which the crowd participates
(4) The Crowdsourcing Platform: the system with which the crowdsourced task is performed

Within each of the four pillars, Hosseini et al, through examination of 113 research papers on crowdsourcing, sought to identify and classify sub-features relevant to each pillar. This taxonomy highlights a number of key concepts relevant to this research and worthy of further discussion in this chapter.

10.4.1 Pillar One: The Crowd

Under Pillar One: The Crowd, five distinct features were identified and labelled as diversity, anonymity, largeness, undefined and suitability. Within each of these five features, further sub-features were identified. For example, diversity includes considerations of geographic diversity, gender diversity, age diversity, as well as expertise diversity. Expertise diversity has been considered in several of the evaluations included in this thesis (Chapter 2, 4 and 5). Largeness relates unsurprisingly to the size of the crowd, and the crowd's capacity therefore to complete the task or to keep up with the flow of data for ongoing, long-term tasks (as assessed in Chapters 2 and 4). Suitability refers to the fit of the crowd in performing the task, and particularly what motivates the crowd to take part. The following section looks in more detail at this important concept.

10.4.1.1 Incentivisation

Appropriate rewards and incentives are critical from a crowd engagement perspective. This involves an understanding of crowd contributor motivations to take part. Established theories of motivation are pertinent here, ranging from need-based theories such as Maslow's Hierarchy of Needs[11] and Herzberg's Two-Factor Theory[12], to process-based theories such as Equity Theory[13], Expectancy Theory[14] and Reinforcement Theory[15]. In addition, intrinsic and extrinsic motivation frameworks such as The Self Determination Theory[16], which make the distinction between intrinsic and extrinsic motivations, are relevant. Intrinsic motivation covers those types of tasks that are perceived as interesting, enjoyable or rewarding in, and of, themselves. In contrast, extrinsic motives are related to factors that are not related to the task but are appealing for some external reason, such as the possibility of improving social, professional or reputational status.

Chapter 6 highlights the range of both intrinsic and extrinsic motivations behind crowd participation in a study identification task for a complex, mixed studies systematic review. However, this was for a discrete task that had a clearly defined end, or finish, and a clearly defined reward of named acknowledgement in the published review. Rewards and incentives are more challenging for those tasks that are open-ended, for example, tasks that help to feed ongoing central repositories of trials, as described in Chapters 2 and 4. This is where alternative incentive mechanisms are especially important in order to help both recruit and retain crowd contributors. Two key areas are pertinent here: gamification and learning.

10.4.1.2 Gamification

One such approach, as described in Chapter 7, to encourage sustained participation involves the addition of game mechanics, or gamification, to microtasks. First explored by von Ahn and Dabbish in 2004[17], gamification refers to the use of game design elements in non-game contexts[18]. Common game elements include point systems, levels of progression, leader boards and badges. Many crowdsourcing initiatives have demonstrated substantial success in this area in terms of both crowd engagement and retention, as well as on measures of efficiency and productivity[19]. One study by Feyiseten at al., examined the potential of adding game elements to an image classification microtask[20]. Their results demonstrated clear evidence of the positive effects of game mechanics with up to five times more labels generated in comparison to the same task without any gamification added, while preserving a comparable level of crowd accuracy. However, gamification has also been met with criticism, with the assertion that it risks producing an effect termed overjustification – where the introduction of game elements into traditionally non-game contexts is seen as undermining potential intrinsic benefits and trivialising contributions into superficial goals[18].

The effects of gamification in the context of Cochrane Crowd are under researched and warrant further exploration. As briefly described in Chapter 7, a range of game features have been introduced into the platform, including a virtual badge system that rewards digital 'milestone' badges, a points system that rewards contributions with Cochrane membership, competition elements in the form of weekly 'screening challenges', progress bars and target settings enabling crowd contributors to set daily, weekly and monthly targets. As well as assessing the effect of gamification approaches on existing microtasks in Cochrane Crowd, consideration also needs to be given to incentivisation approaches for tasks that require a higher cognitive load, as it is likely that tasks will get more complex as automation capabilities improve. In a position paper by Ericson et al., the author posits that simple tasks with well-defined results lend themselves well to being embedded in games, whereas more complex tasks, in which individuals need to develop expertise may be more suited to alternative incentive mechanisms, such as those rooted in the social dynamics of communities[21]. This leads on to a consideration of learning as an incentive for participation.

10.4.2.3 Learning

The potential for learning was another key reason given by participants in the post-task questionnaire described in Chapter 7. Modern theories of learning recognise that science learning is complex and multi-faceted[22], influenced and affected by individual, social, cultural and institutional

factors. In addition to this, learning can occur in virtually any context and at any age. In understanding health research, scientific literacy is critical. The COVID-19 pandemic unleashed a corresponding infodemic, defined as too much information including false or misleading information in digital and physical environments. One factor underpinning the spread of misinformation can be attributed to low levels of health and scientific literacy. The role that citizen science could play therefore in improving scientific literacy should be explored in parallel to the benefits a crowd can bring in helping to generate scientific knowledge. There is evidence that citizen science can contribute to scientific literacy[23,24]. However, this is another under researched area in the context of Cochrane Crowd. Whilst mechanisms are in place to support learning (e.g., task-specific training materials, and an automated feedback mechanism that shows the contributor's classification compared to the final, deemed correct, classification), we have not formally evaluated whether learning occurs, and if so, the extent and nature of that learning.

In the recently published book, *The Science of Citizen Science*[25], the potential for learning is seen not only as a 'nice to have' but as an essential component or outcome that citizen science project and platform managers should ensure is delivered. However, it is recognised that this is not always straightforward. At times learning outcomes might not align with other desired outcomes, such as data generation. An example related to the Cochrane Crowd initiative concerns the contributor's valid desire to perform a task accurately, and to receive feedback related to their accuracy to be able to learn from mistakes, versus the organisation's prioritisation of outcomes related to recall or sensitivity rather than overall accuracy (as described in Chapters 2 and 5). Despite such challenges, opportunities for 'learning whilst doing' clearly exist and should be enhanced, not only for the benefit of the crowdsourcer but for the benefit of the crowd and their communities.

10.4.2 Pillar Two: The Crowdsourcer

Hossieni's second pillar of crowdsourcing is the crowdsourcer. This could be an individual, a company, a project or research team. In the research presented in this thesis, the crowdsourcer is a non-profit organisation, Cochrane[26]. Hosseini identified four distinct features of the crowdsourcer: incentive provision; open call; ethicality provision and privacy provision. Of these, ethical provision is worthy of more detailed consideration in the context of the Cochrane Crowd initiative and the research undertaken in this context.

10.4.2.1 Ethicality provision

In Hosseini's taxonomy, three acts are described in relation to ethical standards of conduct during a crowdsourcing activity. The first is that the crowd has a right to stop or opt-out of the activity at any time; the second, the crowdsourcer should provide feedback about the results of the crowdsourced activity; and third, the crowdsourcer should ensure that the crowd will not be harmed during the activity. Hosseini et al. have focussed on provisions applicable *during* a crowdsourced activity. These are clearly important and pertinent considerations, but it is worth looking more broadly at this important area as ethical considerations of crowdsourcing have generated much discussion in recent years.

Criticisms largely relate to the potential exploitation of crowd workers, with the perception that crowdsourcing often circumvents workplace practices, potentially leaving crowd worker rights, data and intellectual property unprotected and unregulated. Related to Hosseini's second act whereby the crowdsourcer should provide feedback on the results of the activity, there has been much discourse regarding the lack of acknowledgement for crowd effort across a wide gamut of crowdsourced instances. A study by Cooper et al.[27] examined the contribution of citizen science to a review paper by ornithologists in which they formulated ten central claims about the impact of climate change on avian migration. They found that many of the studies supporting the ten claims were based on crowdsourced generated data, but that despite the importance of citizen science in helping to substantiate claims, this crowd effort was rarely noted. Similarly, microtasking marketplaces such as Amazon Mechanical Turk have come under criticism in recent years as studies have revealed the often poor working conditions of an 'unrecognised labour'[28,29]. Some of these ethical issues have been touched upon in Chapter 6 but are worth more exploration here. The following sections describe the main ethical concerns emerging in crowdsourcing practice. In a recent paper by Tauginiene et al.[29], five key ethical areas are identified: exploitation, inclusiveness, research malpractice, collaboration with private partners, ownership and acknowledgement.

10.4.2.2 Exploitation

Tauginiere describes exploitation concerns in relation to data ownership[30]. This will ultimately relate to the nature of crowd contribution, which can vary hugely across crowdsourced and citizen science initiatives. In a paper by Scassa and colleagues, a typology of projects is presented from an intellectual property perspective[31]. Four broad categories of the nature of crowd contributions were identified: (1) classification or transcription data; (2) data gathering; (3) participation as a research subject; (4) the solving of problems, sharing of ideas or manipulation of data. The question therefore

of who owns the data produced by a model of crowd contribution ultimately depends on the nature of that crowd contribution, and should be addressed and defined when drafting the terms of participation for a project or task.

10.4.2.3 Inclusiveness

Inclusiveness concerns relate to the need to create an equitable model of participation. Depending, again, on the nature of the crowdsourced approach, a lack of inclusiveness includes both a lack of opportunity for certain of sections of society to be involved, and the potential loss of key viewpoints or perspectives. For crowd endeavours aimed at solving problems, or generating new ideas, input based on an unrepresentative sample of crowd participants might not result in an optimal solution or be generalisable or applicable to those it relates to or affects. For example, in the healthcare domain, crowd initiatives that aim to collect symptoms related to a particular condition or side effects related to an intervention or treatment could result in an unrepresentative data set (it might be that those with the most serious of side effects are too ill to report them).

Ensuring and supporting inclusiveness in the Cochrane Crowd microtasking initiative has brought with it its own set of challenges. Some are insurmountable, such as the 'digital divide' imposed by an online platform: it is impossible to contribute to the microtasks without access to a computer or smart device. However, we have enabled offline working and aim to make tasks work across a range of devices, understanding that not everyone has access to desktop or laptop computer. Another key challenge relates to language. As described in Chapters 2 and 6, Cochrane Crowd has attracted contributors based in 172 countries of the world. For many, English is likely not to be their first language, yet all current microtasks hosted on the platform are in English. The issue here is two-fold. Not providing tasks in other languages is a potential barrier to participation for huge sections of the global community. It is also an issue in terms of representativeness - providing an imbalanced view that the corpus of valuable health research is only in English. This issue was touched upon in Chapter 4 where the recommendation was made that non-English language sources of health research should be explored as potential additional sources to be added to Cochrane's Centralised Search Service.

10.4.2.4 Research malpractice

Ethical discourse on the use of crowdsourcing often focusses on ethical issues that impact the crowd contributor. However, one important area of ethical concern relates to the potential for research misconduct and malpractice in crowdsourced activities. Examples range from lay people destroying

archaeological sites[32] to disruptive online communities spamming projects with deliberate intent to undermine or influence the results[33].

Whilst research malpractice in the context of crowdsourcing activity has largely been associated within the context of disruption to the physical environment, a potential equivalent issue within the healthcare domain is around the outsourcing of tasks on sensitive topics. Such topic areas could be sensitive from a socio-political perspective, such as abortion or vaccination, or in a commercial or economic sense, for example tasks related to investigational drugs developed by 'Big Pharma'. Intentionally disruptive crowd contributors have not been an issue for Cochrane Crowd to date. Crowd activity on tasks is monitored carefully and unusual behaviour automatically flagged. In addition, the agreement algorithms in place are designed to absorb a degree of error, intentional or not, made by individual contributors.

10.4.2.5 Collaboration with private partners

Another area in which ethical concerns have been raised relates to crowd initiatives that are linked to commercial enterprise. The primary issue is the monetary value of the research which is based on crowd generated data. This might be an issue identified at the project initiation stage, in which case clear communication regarding the proposed use of the data is recommended. Potential complexities arise if new uses of the already generated data are developed that involve generating revenue for private partners.

10.4.2.6 Ownership and acknowledgement

Ensuring that the crowd are acknowledged and rewarded appropriately is another key ethical concern. For crowd contributions resulting in scientific publications, authorship or acknowledgement are two potential rewards. However, citizen science contributors are rarely included as authors on publications[34]. In medicine and healthcare-related research one reason for this is likely to be the current authorship criteria standards set by the International Committee of Medical Journal Editors (ICMJE)[35]. Many journals in this domain area follow the ICMJE standards which state that authors must have made a substantial contribution to any or all of the following: (a) the research design; (b) data acquisition; or (c) data analysis or interpretation. It also states that authors must read the submitted manuscript, agree with its conclusions, and take responsibility for their part of the research. Contributors who do not meet the criteria for authorship should be listed in an acknowledgements section at the end of the publication. Cochrane follows the ICMJE standards and offers crowd contributors named acknowledgement in the published reviews. To date, this has

proved an adequate reward for these types of tasks. We are however, left with two emerging concerns related to this particular ethical issue. The first is how best to appropriately acknowledge those who contribute to tasks that are not directly involved in a specific publication or review (as described in Chapters 2 and 4), and second, as tasks get more complex, while still falling short of the current requirements for authorship, named acknowledgement may not be enough.

10.4.3 Pillar Three: The Crowdsourced Task

A crowdsourced task can take many forms. A number of typologies exist that aim to define and categorise the nature and type of crowd tasks that are possible[36]. Hossieni et al. identified eight distinct features: traditional operation (a task normally done 'in-house' by employees of an organisation), outsourcing task – a task that would otherwise be outsourced; modularity – including complex tasks broken down into microtasks; complexity; solvability – how simple for a human participant to solve or answer; automation characteristics – the extent to which the task is difficult to automate; user-driven – where a crowd is used to generate ideas or create designs, or where the crowd participates in a production process in order to create a product; and finally, contribution type – how the contribution of the crowd is used.

This research has tested the modularity concept of microtasking within the context of health evidence synthesis production. Each empirical study has sought to evaluate measures of crowd performance across a range of experiments to identify studies of various designs (Chapters 2, 5, 6 and 7) and under tight time constraints (Chapter 8). The crowd's collective accuracy has proved consistently high across each study. This success is attributable to several key task characteristics related to their modularity, complexity and solvability. The microtasks developed and evaluated as part of this research were not designed to replicate the task undertaken traditionally by systematic reviewers, but rather the traditional task itself has been broken down into a micro format, with the aim of making it less complex and therefore more solvable. However, one danger of applied research of this nature is to assume that the findings are transferable.

10.4.3.1 Transferability

The concept of transferability, synonymous with replicability, refers to the extent to which an intervention's effectiveness could be achieved in another sample or setting. The setting in which this research has been conducted is highly specific, yet the problem it addresses is by no means unique to health evidence production. Global scientific output doubles every nine years[37]. Across multiple scientific domains, researchers are hindered by a deluge of data. It is becoming increasingly

challenging to identify what research has been done, is ongoing or has yet to be tackled. These unknowns hamper scientific progress and are detrimental to efforts to provide solutions to global societal challenges. In addition to this, research waste, defined as research with no societal benefits, is increasing. In medicine, the notion of research waste is well defined and has been estimated to cost US$85 billion annually[38]. While each domain area will bring with it its own set of specific challenges and requirements, the Evidence Pipeline approach, designed, evaluated and implemented as part of this research, provides a replicable model of research identification that could be adopted across any domain area where there is potential need for human computing power that crowdsourcing could fill in order to accelerate research, and reduce research waste. Indeed, we have had marked interest from a number of organisations across both the health sector and beyond, including areas such as education, agriculture, food safety and policing.

10.4.4 Pillar Four: The Crowdsourcing Platform

Four features are described in Hossieni's taxonomy related to the online environment: crowd-related interactions; crowdsourcer-related interactions; task-related facilities; and platform-related facilities. Of especial relevance here is the task-related facilities concept which relates to the interaction and aggregation of the knowledge from the crowd. The issue of the quality of crowd generated data has been the source of much research[39,40,41] and remains a challenge for many citizen science and crowdsourced initiatives[42]. This research developed, deployed, and evaluated an agreement algorithm designed to work across a range of microtasks with characteristics previously described.

A final key area worth discussion is the utility of the crowd-generated data itself. When issues regarding the quality of crowd generated data are raised, they should be done with a clear understanding of the potential use of the data, and what the end-users of the data's priorities are with regards to the data. Three questions should be asked: (1) Why do we need this data? (2) How will it, or how could it be used? (3) What are the priorities in terms of the data output (for example, timeliness, quality etc.)? These questions shape both the task and the task-related facilities. As shown by this research, we did not only create crowd tasks, we sought to evaluate their implementation in review production workflows to better understand how crowd generated data can be used, and explore the impact different variables have on the end-user or the end-product. Three main workflows were evaluated: (1) study identification at review-level (Chapters 5, 6 and 8); (2) study identification at repository level (Chapters 4 and 9); (3) study identification for machine

learning (Chapters 2 and 9). Across, and within, each of these workflows are variable requirements in terms of the composition and quality of the crowd data.

10.5 Future directions

Without a doubt, crowdsourcing activities, across multiple domains, will continue to grow. In terms of the science of crowdsourcing, there remain many unsolved issues that require more research. Some of these issues have been discussed above, such as ongoing ethical concerns and crowd incentivisation. With over five billion internet users around the world, crowdsourcing has the potential to be disruptive technology. In the context of this research, we have applied it within an existing scientific production and publishing context, accepting, perhaps even accommodating, weaknesses such as poor reporting of primary research, inconsistent and inaccurate record indexing by biomedical databases, lack of standardised record formatting and poor transparency of study audit trails. Yet crowds could not only help tackle challenges within the existing research paradigm but play a significant role in ushering in a new paradigm, one in which metadata is recognised as 'the liberator of knowledge, where everyone has a responsibility to improve it'[43]. As Chris Lintott wrote in his book, *The Crowd and the Cosmos*:

> There is a more interesting future in store – one in which the line between work done by amateurs and professionals, and between the amateurs and professionals themselves, blurs still further[44].

More specifically, future research will be needed to better understand which problems can be successfully solved by a crowd and how best to optimise the problem's solvability. This relates to the design and structure of the tasks themselves. Practitioners are now faced with multiple options and better understanding of which approach will be most effective is required. More research is also needed regarding how best to use and integrate crowd-generated data. This research has involved the development and evaluation of hybrid workflows incorporating human and artificial intelligence. We have demonstrated that such a hybrid intelligence system, which plays to the strengths of its component parts, can be highly effective. However, more research is needed to better understand the capabilities here and the conditions in which hybrid systems will flourish. In a recent essay by Ceccaroni et al., a comprehensive overview of current applications of AI in citizen science is provided[45]. The authors then discuss both future opportunities and potential risks, and posit that humans alone (both experts and citizen scientists alike) will be unlikely to be able to deliver the volumes of data needed to solve many of the global-scale challenges of today and the future. To that

end, the risks of both engaging AI in citizen science initiatives, as well as the risks of ignoring such capabilities, should be evaluated. Rafner and colleagues have produced a framework for types of human-AI interactions in citizen science based on established criteria of hybrid intelligence[46]. This framework forms a solid basis from which to identify the types of citizen science projects that may be supported by artificial intelligence. In the context of health evidence synthesis, we are now expanding our research to develop a range of crowd tasks and machine learning classifiers to generate high-quality, reusable metadata about primary studies according to the PICO (Population, Intervention, Comparator, Outcome) framework.

A plethora of research on crowdsourcing from multiple theoretical and domain perspectives now exists. However, many gaps in knowledge remain. Future work must continue to develop our understanding of crowdsourcing, particularly how we can best use crowd effort, leveraging the complementary functionalities of machine learning, while remaining attuned and attentive to crowd motivations and ethical requirements.

10.6 References

1. Wang L, Xia E, Li H, Wang W. A Bibliometric Analysis of Crowdsourcing in the Field of Public Health. Int J Environ Res Public Health. 2019;16(20):3825.
2. Ghezzi A, Gabelloni D, Martini A, Natalicchio A. Crowdsourcing: A Review and Suggestions for Future Research. International Journal of Management Reviews. 2017;20(2):343-363.
3. The European Citizen Science Association. https://ecsa.citizen-science.net [Last accessed 30 September 2022].
4. The Australian Citizen Science Association. https://citizenscience.org.au [Last accessed 30 September 2022].
5. Prpić J, Shukla P. The Theory of Crowd Capital. 46th Hawaii International Conference on System Sciences.
6. Prpić J, Shukla P. Crowd Science: Measurements, Models and Methods. 49th Hawaii International Conference on System Sciences.
7. Curado C, Bontis N. The knowledge-based view of the firm and its theoretical precursor. International Journal of Learning and Intellectual Capital 2006;3(4):367-381.
8. Hayek FA. The use of knowledge in society. The American Economic Review 1945;35(4):519-530.
9. Giustini D. How Web 2.0 is changing medicine. BMJ. 2006 Dec 23;333(7582):1283-4.

10. Hosseini M, Phalp K, Taylor J, Ali R. The four pillars of crowdsourcing: A reference model. *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 2014: 1-12, doi: 10.1109/RCIS.2014.6861072.
11. Maslow AH. A theory of human motivation. Psychological Review 1943;50(4):370-396.
12. Herzberg F. The motivation to work among Finnish supervisors. *Personnel Psychology 1965;18:*393-402.
13. Adams JS, Freedman S. Equity theory revisited: comments and annotated bibliography. Editor(s): Leonard Berkowitz, Elaine Walster, Advances in Experimental Social Psychology 1976;9:43-90.
14. Vroom VH. Work and motivation. 1964.
15. Skinner, BF. Science and human behavior. New York: Free Press 1953.
16. Deci EL, Ryan R. Intrinsic motivation and self-determination in human behavior. Perspectives in Social Psychology. Springer 1985.
17. Von Ahn L, Dabbish L. Labeling images with a computer game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04). Association for Computing Machinery, New York, NY, USA, 319-326.
18. Deterding S, Khaled R, Nacke L, Dixon D. Gamification: Toward a definition. In CHI 2011 Gamification Workshop Proceedings, 12-15, 2011.
19. Morschheuser B, Hamari J, Koivisto J. Gamification in crowdsourcing: a review. 49th Hawaii International Conference on System Sciences (HICSS), 2016;4375-4384.
20. Feyisetan O, Simperl E, Van Kleek M, Shadbolt N. Improving paid microtasks through gamification and adaptive furtherance incentives. In Proceedings of the 24th international conference on world wide web 2015;18:333-343.
21. Erickson T. Some thoughts on a framework for crowdsourcing. Position paper. CHI 2011 Workshop on Crowdsourcing and Human Computation. 2011.
22. Hodson D. Teaching and learning about science: language, theories, methods, history, traditions and values. 2009.
23. Cronje R, Rohlinger S, Crall A, Newman G. Does participation in citizen science improve scientific literacy? A study to compare assessment methods. Applied Environmental Education & Communication 2011;10(3):135-145.
24. Riesch H, Potter C. Citizen science as seen by scientists: methodological, epistemological and ethical dimensions. Public Understanding of Science. 2014;23(1):107–120.
25. Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, Samson R, Wagenknecht K (eds.). The Science of Citizen Science. 2021 Springer.
26. Cochrane. https://www.cochrane.org [Last accessed 30 September 2022].

27. Cooper CB, Shirk J, Zuckerberg B. The invisible prevalence of citizen science in global research: migratory birds and climate change. Plos One 2015;9(9):e106508.
28. Gray M, Suri S. Ghost work: how to stop Silicon Valley from building a new global underclass. Boston, Houghton Miffin Harcourt; 2019.
29. Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Paper 449, 1–14; https://doi.org/10.1145/3173574.3174023.
30. Tauginienė L, Hummer P, Albert A, Cigarini A, Vohland K. Ethical challenges and dynamic informed consent. Chapter 20 In: Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, Samson R, Wagenknecht K (eds.). The Science of Citizen Science. 2021 Springer.
31. Scassa T, Chung H. Typology of citizen science projects from an intellectual property perspective. 2015 https://www.wilsoncenter.org/publication/typology-citizen-science-projects-intellectual-property-perspective.
32. Davydov, D., Grünewald, C., Morscheiser, J., Tutlies, P., Vollmer-König, M., & Zeiler, M. (2017). Sondengänger und archäologie. LWL/LVR: Die rechtslage in NRW. http://www.roemisch-germanisches-museum.de/download/Sondengaenger_u_Arch.pdf.
33. Guerrini CJ, Majumder MA, Lewellyn MJ, McGuire AL. Citizen science, public policy. Science 2018;361(6398):134-136.
34. Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, Martin J, et al. The current state of citizen science as a tool for ecological research and public engagement. Frontiers in Ecology and the Environment 2012;10(6):291-297.
35. ICMJE. (2019). Defining the roles of authors and contributors. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html
36. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med. 2014 Feb;46(2):179-87.
37. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 2015;66(11):2215-2222.
38. Grainger MJ, Bolam FC, Stewart GB, Nilson EB. Evidence synthesis for tackling research waste. Nat Ecol Evol 2020;4:495-497.
39. Engel SR, Voshell JR Jr. Volunteer biological monitoring: can it accurately assess the ecological condition of stream? Am Entomol 2002;48:164-177.

40. Genet KS, Sargent LG. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildl Soc Bull 2003;31:703-714.
41. Sheng VS, Zhang J. Machine learning with crowdsourcing: a brief summary of the past research and future directions. In: Proceedings of the AAAI Conference on Artificial Intelligence 2019. https://doi.org/10.1609/aaai.v33i01.33019837.
42. Balázs B, Mooney P, Nováková E, Bastin L, Arsanjani JJ. Data quality in citizen science. Chapter 8 In: Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, Samson R, Wagenknecht K (eds.). The Science of Citizen Science. 2021 Springer. https://doi.org/10.1007/978-3-030-58278-4
43. Pentz E. Making the most of metadata. Research Information 2014: https://www.researchinformation.info/interview/making-most-metadata [Last accessed: 15 October 2022].
44. Lintott C. The crowd and the cosmos: adventures in the Zooniverse. OUP Oxford, 2019.
45. Ceccaroni L, Bibby J, Roger E, Flemons P, Michael K, Fagan L, et al. Opportunities and risks for citizen science in the age of artificial intelligence. Citizen Science: Theory and Practice 2019;4(1):29.
46. Rafner J, Gajdacz M, Kragh G, Hjorth A, Gander A, Palfi B, et al. Revisiting citizen science through the lens of hybrid intelligence. arXiv:2104.14961v1.

Chapter 11. Conclusion

11.1 Introduction

Evidence synthesis, usually in the form of a systematic review or a meta-analysis, aims to answer a pre-specified research question incorporating all relevant empirical evidence, and using explicit and replicable methods to minimise bias. Evidence synthesis is recognised as an essential component in bridging the gap between research findings (*what is known*) and health care practice (*what is done*): the 'know-do' gap. One of the most significant pressure points is in identifying the evidence as we find ourselves increasingly struggling to keep up with the amount of research produced. This information overload causes delays in the evidence synthesis process which ultimately means that important clinical questions about the effects of treatments remain unanswered, and clinicians and patients are left to make decisions based on a poor understanding of the available evidence. In other words, information overload is lengthening this 'know-do' gap. This research has sought to address this challenge. This collection of empirical studies, all of which have been published in academic, peer-reviewed journals, have had demonstrable impact both in terms of their academic significance and their practical application.

11.2 Academic significance

This work has contributed to the methods, theory, and application of crowdsourcing in the production of high-quality health evidence. From a methodological standpoint, this research has significantly contributed to our understanding of how to design, conduct and measure human computation microtasks to ensure high quality data output and task efficiency. More specifically, its theoretical contribution includes the introduction of a new method of crowd data aggregation, the application of the microtasking concept to a new domain area, and the methodological guidance for the development, calibration and validation of machine learning models trained on crowd-generated data. Each of these aspects are described in more detail below.

11.2.1 Agreement algorithms

Concerns regarding the quality of crowd generated data remain the predominant issue in the field of citizen science. Various approaches to aggregating crowd generated data have been developed to varying degrees of success. As part of this research, we developed a robust agreement algorithm, and an accompanying crowd contributor structure, that enables high quality data output (described

in detail in Chapter 2 and evaluated across a range of studies presented in Chapters 2, 4, 5 and 8). The algorithm and the structure are both simple to understand and implement, and constitute a valuable addition to a growing catalogue of approaches used to ensure high quality data output.

11.2.2 Human computation and microtasking

This work introduces the concept of microtasking to aspects health evidence production. Related to the above point regarding concerns about the quality of crowd work, one prior barrier to involving citizens and non-professionals in the research production process relates to questions regarding the crowd's ability to perform tasks traditionally undertaken by professional researchers. In the area of healthcare research, this issue has been arguably considered more important than in other domain areas. As part of this research, we applied the microtasking concept in a way that had not been applied before in this domain area (described in detail in Chapter 2). By decomposing larger, more complex tasks into a smaller microtasks, we created new tasks specifically designed to not require prior experience or expertise. Previous studies have struggled to obtain adequate levels of crowd accuracy in part because the tasks replicated the expert task rather than developing a more modular form of the task.

11.2.3 Machine learning and hybrid workflows

Many crowdsourcing initiatives fail to use the data generated. This work demonstrated the value of crowd-generated data in two main ways: (1) in the development of machine learning classifiers, and (2) in the development of hybrid workflows that utilise human and machine effort to their respective strengths. Chapters 3 and 9 describe in detail the methodology required to develop robust support vector models that can be implemented in a binary fashion. Chapters 4 and 5 evaluate hybrid workflows that combine these semi-automation approaches.

11.3 Practical impact

11.3.1 The Cochrane Crowd Platform

A major practical output of this research is the platform itself. Cochrane Crowd (https://crowd.cochrane.org) is a state-of-the-art crowdsourcing web application designed to create and host human computation microtasks. It was launched in 2016, hosting a single task. Since then, it has hosted dozens of microtasks, each one either directly contributing to the production of a systematic review, feeding central repositories of studies, or helping to create training data for machine learning models. Over two million records have been collectively assessed by the crowd,

equating to over eight million individual classifications. As described in Chapter 2, this has enabled us to keep pace with the quantity of research being produced.

11.3.2 The Cochrane Crowd community

Over 26,000 people have signed up to join the Cochrane Crowd community. This research has enabled many to get involved in the review production process, a significant proportion of whom had little or no previous experience with health research. Based on information we collect the first time someone logs in, almost a quarter are completely new to the area of health research, and a third of all contributors state they had either no idea or only a vague idea of what a systematic review is upon sign-up. The community is also geographically diverse, with contributors resident in 172 countries of which 52% are based in lower and middle income countries.

11.3.2 Machine Learning classifiers

This research includes the development of two support vector machine learning classifiers: (1) the RCT Classifier, and (2) the COVID Eligibility Classifier, described in detail in Chapters 3 and 9 respectively. These classifiers have been deployed to production, performing aspects of their respective tasks previously undertaken by the crowd, thereby freeing up human effort to focus on tasks not yet doable by the machine. Both classifiers have resulted in a significant workload reduction in terms of the number of records requiring manual assessment. The RCT Classifier has also been adopted by the wider research community having been incorporated into a number of non-Cochrane tools and workflows: Robot Reviewer (https://www.robotreviewer.net), Trialstreamer (https://www.sites.google.com/a/york.ac.uk/yhectrialsregisters/home/clinicaltrials/trialstreamer), Eppi Reviewer (https://eppi.ioe.ac.uk/CMS/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4&).

11.3.3 Impact on the current paradigm

My research supports the current search paradigm through the creation of crowdsourced capability that enables both rapid and accurate study identification. The identification of potentially relevant studies for reviews is a critical activity, applicable to all review types. The Screen4Me workflow (described in Chapter 5) which uses both Cochrane Crowd and the machine learning RCT Classifier, has to date been applied to 107 Cochrane intervention reviews, and resulted in author team workload reduction for study identification of between 52-87%. As well as workload reduction, time has been saved with the screen4me workflow designed to take a maximum of two weeks from start to finish in comparison to months spent on this task by review author teams.

11.3.4 Introduction of new paradigm

At a macro level, in terms of the development of the Evidence Pipeline, our human-machine workflows now contribute over 99% of reports of randomised and quasi-randomised trials (RCTs) to Cochrane's Central Register of Controlled Trials (CENTRAL). This constitutes 97.5% of RCTs that get included in Cochrane intervention reviews (as detailed in Chapter 4). This has made CENTRAL the most comprehensive repository of RCTs in the world.

11.4 Conclusion

Each study in this thesis has made a unique contribution to our understanding of the potential role of crowdsourcing in health evidence synthesis. Taken together, this body of work has unquestionably demonstrated the feasibility of crowdsourcing human computation tasks within the study identification stages of evidence synthesis. Crowdsourcing is now implemented into Cochrane review production processes both within the current information retrieval paradigm, in terms of assessing sets of search results retrieved for individual or suites of reviews, but also in terms of helping to produce and maintain highly curated repositories of studies as part of Cochrane's Evidence Pipeline. The crowd have proved capable with a multitude of varying tasks, and produced data that has enabled the development of machine learning models, creating virtuous cycles that have enabled us to create human-in-the-loop workflows that play to the strengths of crowd and machine alike.

As we move towards potentially tasking the crowd with perhaps more challenging tasks aimed at enriching repositories of studies with metadata about those studies, more thought must be given to rewarding and acknowledging the crowd effort (as discussed in Chapter 10). The evaluations in this thesis have largely focussed on outcomes regarding crowd performance. That has been an appropriate focus to date as ensuring (and proving it to the wider community) crowd accuracy was our primary concern. We do however need to recognise the vital importance of acknowledging contributions of this nature appropriately. Named acknowledgement for those who have helped to assess the search results of individual reviews has been well received by the Cochrane Crowd community, but how to best to reward and acknowledge those that play an arguably even more important role in helping to create and curate vital repositories of studies that will feed the reviews themselves? As the scientific publishing paradigm shifts towards enabling a more living evidence approach to evidence synthesis, we have an opportunity to make sure that crowd contributions are recognised in those evidence outputs. Other domains are ahead of us in addressing some of this. For example, it is not uncommon for physics paper to have dozens of authors, reflecting the large

numbers of contributors helping to maintain the data sets upon which the research is based. Cultural and organisational shifts are needed, as well as a technological one; the 'publish or perish' model encourages a competitive and siloed approach to science that is not always beneficial, undoubtably leads to duplication of effort and waste, and is not aligned with the notion of diverse communities of contributors continually curating the evidence base.

Cochrane Crowd was one of the first and is arguably the most successful implementation of crowdsourcing in evidence-based healthcare. This innovation has been pivotal for Cochrane in establishing a hybrid human-machine workflow now deployed at production level. By leveraging the speed and scale of automation via machine learning, working in partnership with the accuracy of manual verification by the crowd, we have been able to make substantial efficiency savings, without compromising on quality. The Cochrane Crowd initiative is a powerful example of generating economies of effort to produce results far greater than the sum of its parts. We will continue to develop this important human resource, understanding that people have always been our greatest asset. What might have started out as a novelty to some, or indeed something to be approached with caution, has now become a fundamental part of the evidence production ecosystem, and one that has the potential to transform the way we produce health evidence.

This work has demonstrated that crowdsourcing in this way not only expedites the study identification process for individual reviews, with significant workload reduction for author teams within the *current* production paradigm, but also enables a new production model to emerge; one that leverages economies of effort and scale to help create comprehensive, curated repositories of studies, provide ongoing high-quality training data for machine learning classifiers, and support the co-production of evidence by the very communities who need it.