Journal of Clinical Epidemiology 133 (2021) 130−139

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

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

Anna Noel-Storr[a,b,*], Gordon Dooley[c], Julian Elliott[d,e], Emily Steele[b], Ian Shemilt[f], Chris Mavergames[g], Susanna Wisniewski[a], Steven McDonald[d], Melissa Murano[d], Julie Glanville[h], Ruth Foxlee[i], Deirdre Beecher[g], Jennifer Ware[a], James Thomas[f]

[a]*Radcliffe Department of Medicine, University of Oxford, Level 4, Academic Block, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK*
[b]*People Services Department, Cochrane, St Albans House, 57-59 Haymarket, London SW1Y 4QX, UK*
[c]*Metaxis Ltd, Elmbank Offices, Main Road Curbridge, Witney, Oxfordshire OX29 7NT, UK*
[d]*Cochrane Australia, School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia*
[e]*Department of Infectious Diseases, Monash University and Alfred Hospital, 55 Commercial Rd, Melbourne, Victoria 3004, Australia*
[f]*EPPI-Centre, Department of Social Science, University College London, 18 Woburn Square, London, WC1H 0NR, UK*
[g]*Informatics and Technology Systems, Cochrane, St Albans House, 57-59 Haymarket, London SW1Y 4QX, UK*
[h]*York Health Economics Consortium, University of York, Enterprise House, York YO10 5NQ, UK*
[i]*Editorial and Methods Department, Cochrane, St Albans House, 57-59 Haymarket, London SW1Y 4QX, UK*

Accepted 13 January 2021; Published online 18 January 2021

Abstract

Background and Objectives: 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 health care. Accompanying each task are brief, interactive training modules, and agreement algorithms that help ensure accurate collective decision-making.The aims of the study were to evaluate the performance of Cochrane Crowd in terms of its accuracy, capacity, and autonomy and to examine contributor engagement across three tasks aimed at identifying randomized trials.

Study Design and Setting: 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 randomized controlled trials (RCTs) for Cochrane's Central Register of Controlled Trials. The sensitivity for each task was 99.1% for the RCT identification task (RCT ID), 99.7% for the RCT identification task of trials

Authors' contributions: A.N.-S. contributed to conceptualization, methodology, investigation, data curation, visualization, supervision, and writing the original article. G.D. contributed to conceptualization, methodology, resources, data curation, and reviewing and editing the article. J.E. contributed to conceptualization, methodology, and reviewing and editing the article. E.S., I.S., C.M., S.M., M.M., J.G., and R.F. contributed to conceptualization and reviewing and editing the article. S.W. contributed to conceptualization, data curation, and reviewing and editing the article. D.B. and J.W. contributed to data curation and reviewing and editing the article. J.T. contributed to conceptualization, methodology, and reviewing and editing the article.

Availability of data and materials: The data sets used and/or analyzed during the present study are available from the corresponding author on request.

Declarations: The following authors on this article are current or former members of the Cochrane Crowd team who helped to develop and maintain the Cochrane Crowd platform: Anna Noel-Storr, Gordon Dooley, Julian Elliott, Emily Steele, Susanna Wisniewski, and James Thomas.

Ethics approval and consent to participate: Not applicable.

Funding: The Cochrane Collaboration funded the development of the Cochrane Crowd Platform and continues to fund its ongoing development and maintenance. Evaluations of Crowd performance and data output form a part of that ongoing funding.

[1] In Ovid MEDLINE: 2019*.ed. = 1,041,651; In Ovid Embase: 2019*.dc. NOT MEDLINE = 1,416,448; In ClinicalTrials.gov: First posted from January 1, 2019, to January 1, 2020 = 32,524; In ICTRP = trials added from January 1, 2019, to January 1, 2020 = 62,738. Deduct 32,524 = 30,214. Total number: 2,520,837. For weekly average: 2,520,837/52 = 48,478.

* Corresponding author. Tel.: +44(0)1865 234 306.
E-mail address: anna.noel-storr@rdm.ox.ac.uk (A. Noel-Storr).

https://doi.org/10.1016/j.jclinepi.2021.01.006

from ClinicalTrials.gov (CT ID), and 97.7% for the identification of RCTs 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 CT ICTRP ID. The capacity of the combined Crowd and machine learning workflow has increased fivefold in 6 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.

Keywords: Randomized controlled trial; Screening; Cochrane; Crowdsourcing; Citizen science; Machine learning; Human intelligence tasking; Systematic review; Evidence production; Information management

1. Background

Over the last two decades, published health research output has more than doubled [1,2]. In 2019, just more than 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[1]. This equates to an average of 48,000 unique biomedical-related and health care−related research artifacts 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 organization that produces high-quality systematic reviews about the effectiveness of health care interventions [6,7]. In Cochrane systematic reviews alone, we estimate that reviewers assess in excess 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 while 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, 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 health care 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 [10−13].

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 [14−16]. Over the last decade, a range of crowdsourcing initiatives within health care have surfaced [17−19], 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 [24−26].

1.1. 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 organization is trying to solve: (1) *peer-vetted creative production* (sometimes termed "crowd creation") where the organization tasks the crowd with helping to generate new ideas, solutions, or designs; (2) *broadcast search*, which is a call to find a solution to an empirical (often scientific or technological) problem; (3) *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 (4) *distributed human intelligence tasking*, where the organization

What is new?

- Crowdsourcing the identification of RCTs via Cochrane Crowd is 99% accurate·
- Cochrane Crowd has contributors from over 160 countries

Key findings

- Crowdsourcing and machine learning working together produces accuarte results and ability to scale

What this adds to what is known

- Crowdsourcing has an important role to play in study identification for health evidence production

tasks the crowd with analyzing or categorizing 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 organizations, such as Cochrane, a new way of tackling the challenges described previously.

Additional File 1 provides a brief history of Cochrane Crowd.

1.2. The Cochrane Crowd platform

1.2.1. 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 (Fig. 1 provides an example). The focus of this article is on our evaluations of three microtasks to identify randomized controlled trials (RCTs) from bibliographic databases (task name: RCT ID); the US National Library of Medicine's ClinicalTrials.gov clinical trials registry (CT ID), and the World Health Organization's meta-registry of clinical trials, the International Clinical Trials Registry Platform (ICTRP ID).

1.2.1.1. RCT ID: Identifying randomized 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].

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

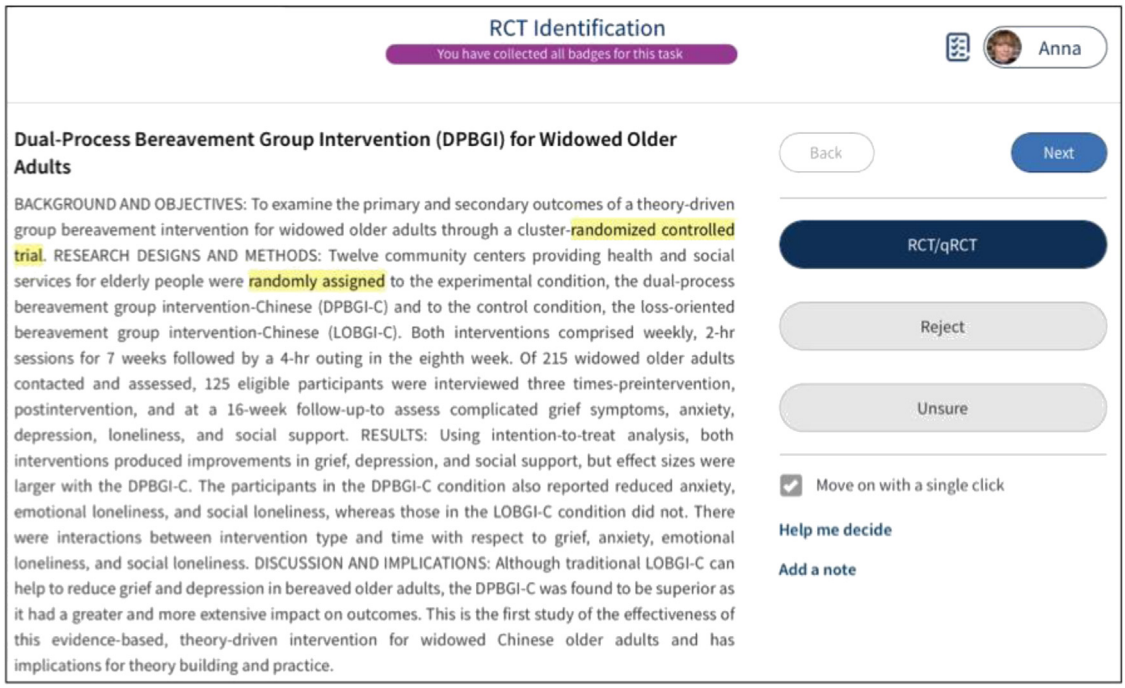

Fig. 1. Screenshot of the randomized controlled trials identification (RCT ID) task.

1.2.1.2. 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 randomized trials from the world's largest clinical trials registry, ClinicalTrials.gov (www.clinicaltrials.gov). The second, ICTRP ID, focuses on the identification of randomized trials from the World Health Organization's meta-registry of clinical trials, the International Clinical Trials Registry Platform (ICTRP; http://apps.who.int).

Although all three microtasks aim to identify randomized 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 categorization of trial registry records are potentially easier and more rewarding for beginners because (1) the information in these records is more structured compared with bibliographic records and (2) the prevalence of RCTs that can be correctly identified is higher, hopefully providing a higher level of satisfaction with the task.

1.2.2. 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].

1.2.3. Supporting crowd accuracy: guidance and training

From the outset, we wanted to avoid restrictions on who could contribute to the Cochrane Crowd. Recognizing 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 requires a pass mark, so on completion of these practice records, the contributor can progress straight to assessing "live" records.

As well as supporting contributors through task-focused training, we recognized the need to enable contributors to track their own progress with each task. Timely, accurate, individualized 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.

1.2.4. 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 minimize the effects of errors made at an individual level while 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 (Fig. 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*.

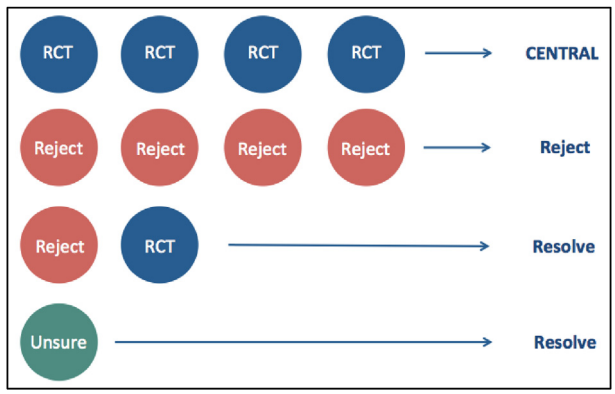

Fig. 2. The Cochrane Crowd agreement algorithm in place for standard screeners.

2. Methods

2.1. 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 the highest educational attainment, age at sign-up, country of residence, and level of experience with health research.

2.2. 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" [TP] count), the number of irrelevant items correctly identified as such (the "true negative" [TN] count), the number of relevant items incorrectly classified as irrelevant (the "false negative" [FN] count), and the number of irrelevant items, incorrectly classified as relevant (the "false positive" [FP] count). 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 follows:

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

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

2.3. Crowd autonomy

Crowd autonomy 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, because 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.

2.4. 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 centralized search and screen model. This is an appropriate baseline, as the Cochrane study identification workflow aims to prospectively identify all randomized 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).

3. Results

3.1. Crowd characteristics

Fig. 3 shows the steady rate of growth in the number of registered Cochrane Crowd contributors since the

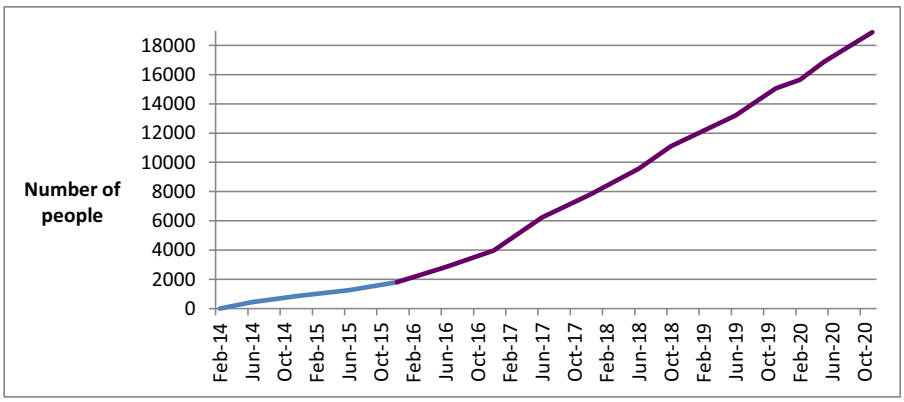

Fig. 3. Cochrane Crowd sign-up with the blue line representing the pilot Embase project phase. Data as of November 10, 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

platform's launch. Approximately 18,800 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.

Cochrane Crowd contributors are resident in 163 countries, of which 96 are low- and middle-income countries. The top five countries are the United Kingdom (17% of contributors), the 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. More than 2,800 new contributors have completed these questions, providing us with additional insight into our Crowd. Although 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 and 24 years at sign-up. Perhaps unsurprisingly, a large proportion of new contributors are students in a health-related area (42.4%).

3.2. Crowd accuracy

Table 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 randomization method used in the study) as a highlighted phrase (in Crowd, we have highlighted more than 80 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 FPs, several were records in which participants had been randomly selected rather than randomly assigned to groups. Another common error occurred with records that provide an overview of a topic, with a brief mention of a specific randomized trial. Other FPs 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 randomized trial identification tasks, similarly high accuracy was achieved, as shown in Table 1. For CT ID, almost all the 17 FNs (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 because of the lack of study design information shown in the record as a result of a display problem. This was because of the API not receiving the study design information for trial registry records from one of the main registries in ICTRP. Although the link to the full record with more information was available, contributors were not expected to access this link.

3.3. 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, shows 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%. Fig. 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 centralized 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, whereas the RCT machine learning classifier, calibrated to achieve a recall of 99%, processed a further 243,996 records for this

Table 1. Accuracy data for the three study identification microtasks

Microtask	Number of crowd participants	Number of records (number of RCTs)	TP	TN	FP	FN	Sensitivity (%), (95% CI)	Specificity (%), (95% CI)	Accuracy (%)
RCT ID	94	6,041 (461)	457	5,522	58	4	99.1 (97.79–99.76)	99.0 (98.66–99.21)	99.0
CT ID	179	11,040 (5,613)	5,596	5,350	77	17	99.7 (99.52–99.82)	98.58 (98.23–98.88)	99.1
ICTRP ID	109	5,000 (1,036)	1,012	3,941	23	24	97.7 (96.57–98.51)	99.1 (99.13–99.63)	99.1

Abbreviations: CI, confidence interval; FN: false negative; FP, false positive; RCT, randomized controlled trials; TN, true negative; TP, true positive.

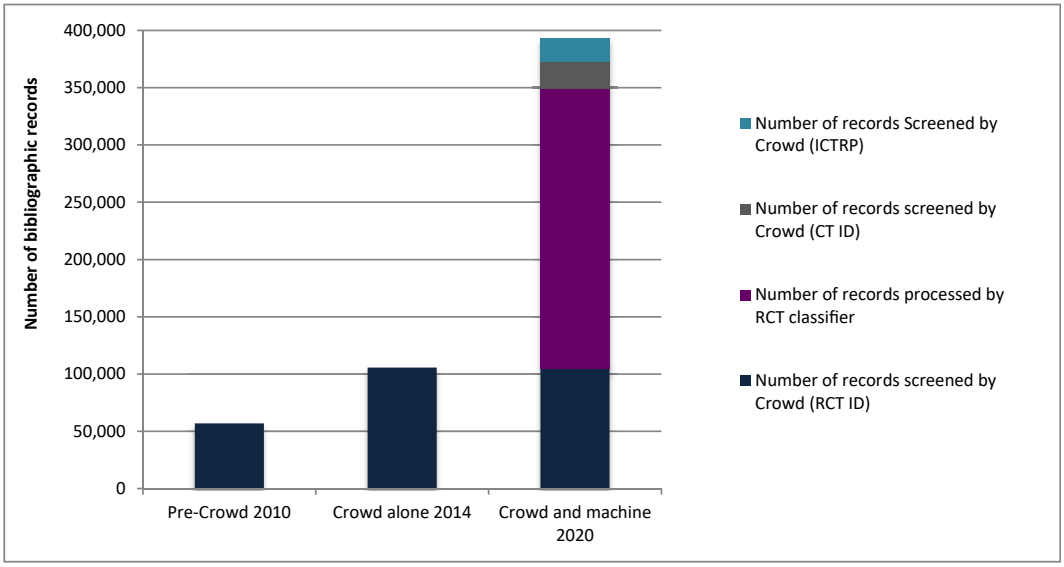

Fig. 4. Cochrane's capacity for identifying RCTs (2010–2020).

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 microtasks available in Cochrane Crowd as well as work on a range of other tasks now available on the platform.

As of November 2020, the 18,900 registered contributors have collectively identified more than 175,000 reports of randomized trials for inclusion in CENTRAL. Table 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 to find one relevant record.

4. Discussion

Cochrane established its crowdsourcing initiative primarily in response to the challenge posed by the rapid increase in global research output. Over the last 5 years, Cochrane Crowd has evolved to become an essential part of Cochrane's ongoing efforts to identify randomized trials for inclusion in its reviews. The Crowd now has approximately 18,900 contributors from 163 countries and has collectively processed over one million records, helping to identify more than 175,000 reports of randomized 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 randomized trial identification microtasks, with fewer than 20% of records needing resolution and a greater than fivefold 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 randomized controlled trial*

Table 2. The three study identification microtask metrics

Microtask	Date task went live	Number of records processed	Number of classifications	Number of RCTs (percentage of total identified for that microtask)	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

Data accurate as of November 10, 2020.

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 randomized 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 a work by Nama et al., 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 also in making accurate individual classifications. Although 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 health care−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. Although 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. Although 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.

4.1. Ongoing challenges

Although accuracy measures from our evaluations are very high, they are not 100%. As we have shown, FNs (missed studies) and FPs 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, while 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 to the Cochrane Crowd. 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 continue to grow and diversify.

5. Conclusions

To date, the Cochrane Crowd community has classified more than 1,021,227 records (756,916 from bibliographic databases and around 264,311 from trial registries). From this, more than 175,000 reports of randomized 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 randomized 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% and 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

while at the same time offer willing contributors a way to get involved and play a crucial role in health evidence production.

Acknowledgments

The authors would like to acknowledge the efforts of all those who have signed up and contributed to the Cochrane Crowd initiative. It is greatly appreciated. The authors would like to acknowledge the support of present and past members of the Cochrane Dementia and Cognitive Improvement Group. The authors would also like to thank the Cochrane Collaboration for funding the project. Support for Project Transform was provided by Cochrane and the National Health and Medical Research Council of Australia (APP1114605). The contents of the published material are solely the responsibility of the Administering Institution, a Participating Institution or individual authors and do not reflect the views of the NHMRC.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2021.01.006.

References

[1] Van Noorden R. Global scientific output doubles every nine years. Nature News Blog. 2014. Available at http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html#. Accessed November 20, 2020.

[2] 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;66:2215−22.

[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;147:224−33.

[5] Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, et al. 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. Eval Health Prof 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. J Am Med Assoc 1995;274:1935−8.

[8] Morley RF, Norman G, Golder S, Griffith P, 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, Synnot A, Nunn J, Hill S, 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, et al. Development of the ACTIVE framework to describe stakeholder involvement in systematic reviews show less. J Health Serv Res Policy 2019;24:245−55.

[11] Kreis J, Puhan MA, Schunemann HJ, Dickersin K, et al. Consumer involvement in systematic reviews of comparative effectiveness research. Health Expect 2013;16:323−37.

[12] Brett J, Staniszewska S, Mockford C, Herron-Marx S, Hughes J, Tysall C, et al. A systematic review of the impact of patient and public involvement on service users, researchers and communities. Patient 2014;7:387−95.

[13] Involve. Available at https://www.involve.org.uk. Accessed November 20, 2020.

[14] Muller CL, Chapman L, Johnston S, Kidd C, Illingworth S, Foody G, et al. Crowdsourcing for climate and atmospheric sciences: current status and future potential. Int J Climatol 2015;35:3185−203.

[15] Zhao Y, Zhu Q. Evaluation on crowdsourcing research: current status and future direction. Inf Syst Front 2014;16:417−34.

[16] Von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: human-based character recognition via web security measures. Science 2008;321:1465−8.

[17] Tucker JD, Day S, Tang W, Bayus B. Crowdsourcing in medical research: concepts and applications. PeerJ 2019;7:e6762.

[18] Wang C, Han L, Stein G, Day S, Bien-Gund C, Mathews A, et al. Crowdsourcing in health and medical research: a systematic review. Infect Dis Poverty 2020;9(1):8.

[19] Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing − harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2013;29:187−203.

[20] Brown AW, Allison DB. Using crowdsourcing to evaluate published scientific literature: methods and example. PLoS One 2014;9: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, et al. 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. An exploration of crowdsourcing citation screening for systematic reviews. Res Synth Methods 2016;8(3):366−86.

[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;122:160−2.

[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, Sandarage R, Menon K, Macartney G, et al. Crowdsourcing the citation screening process for systematic reviews: validation study. J Med Internet Res 2019;21(4):e12953.

[27] Brabham DC. Crowdsourcing. Cambridge, MA: The MIT Press; 2008.

[28] Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med 2014;46(2):179−87.

[29] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane. .; 2019. Available at www.training.cochrane.org/handbook. Accessed January 31, 2021.

[30] Should I publish this record to CENTRAL?. Available at https://community.cochrane.org/sites/default/files/uploads/Should%20I%20publish%20this%20record%20to%20CENTRAL.pdf. Accessed November 20, 2020.

[31] Noel-Storr AH, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, et al. An evaluation of Cochrane Crowd finds that

crowdsourcing produces accurate results in identifying randomised trials. J Clin Epidemiol 2020;130:23−31.

[32] Thomas J, McDonald S, Noel-Storr AH, Shemilt I, Elliott J, Mavergames C, et al. 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. https://doi.org/10.1016/j.jclinepi.2020.11.003. [Epub ahead of print].

[33] Krivosheev E, Casati F, Benatallah B. Crowd-based multi-predicate screening of papers in literature reviews. WWW. The 2018 Web Conference, April 23-27, 2018, Lyon, France 2018.

[34] Ashkanase J, Nama N, Sandarage RV, Penslar J, Gupta R, Ly S, et al. Identification and evaluation of controlled trials in pediatric cardiology: crowdsourced scoping review and creation of accessible. Can J Cardiol 2020;36(11):1795−804.