Journal of the American Medical Informatics Association, 24(6), 2017, 1165–1168
doi: 10.1093/jamia/ocx053
Advance Access Publication Date: 25 May 2017
Brief Communication

Brief Communication

Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach

Byron C Wallace,[1] Anna Noel-Storr,[2] Iain J Marshall,[3] Aaron M Cohen,[4] Neil R Smalheiser,[5] and James Thomas[6]

[1]College of Computer and Information Science, Northeastern University, Boston MA, USA, [2]Radcliffe Department of Medicine, University of Oxford, Oxford, UK, [3]Department of Primary Care and Public Health Sciences, King's College London, London, UK, [4]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, USA, [5]Department of Psychiatry and Psychiatric Institute, University of Illinois College of Medicine, Chicago, IL, USA and [6]EPPI-Centre, Department of Social Science, University College London, London, UK

Corresponding Author: Byron C Wallace, 202 WVH, 360 Huntington Avenue, Boston, MA 02115. E-mail: byron@ccs.neu.edu. Phone: 413-512-0352. Fax: 617-373-5121

Received 10 January 2017; Revised 10 April 2017; Accepted 18 May 2017

ABSTRACT

Objectives: Identifying all published reports of randomized controlled trials (RCTs) is an important aim, but it requires extensive manual effort to separate RCTs from non-RCTs, even using current machine learning (ML) approaches. We aimed to make this process more efficient via a hybrid approach using both crowdsourcing and ML.
Methods: We trained a classifier to discriminate between citations that describe RCTs and those that do not. We then adopted a simple strategy of automatically excluding citations deemed very unlikely to be RCTs by the classifier and deferring to crowdworkers otherwise.
Results: Combining ML and crowdsourcing provides a highly sensitive RCT identification strategy (our estimates suggest 95%–99% recall) with substantially less effort (we observed a reduction of around 60%–80%) than relying on manual screening alone.
Conclusions: Hybrid crowd-ML strategies warrant further exploration for biomedical curation/annotation tasks.

Key words: machine learning, evidence-based medicine, crowdsourcing, human computation, natural language processing

BACKGROUND AND SIGNIFICANCE

Randomized controlled trials (RCTs) remain the best tool for assessing the efficacy of treatments. Systematic reviews are rigorous syntheses of all RCTs addressing a particular clinical question, and are considered the most reliable form of evidence.[1] However, finding RCTs remains surprisingly difficult. Controlled vocabularies of major research databases often lack the concept "RCT" or apply it inconsistently.[2–4] Consequently, those who undertake systematic reviews often manually read ("screen") thousands of irrelevant records to find a small set of relevant RCTs. This is partly why systematic reviews are time-consuming and expensive to conduct, a problem exacerbated by the rapid growth of the published evidence base.[5]

To increase the efficiency of reviewing, Cochrane (a global network of health science researchers) has invested effort into systematically identify RCTs and is compiling a comprehensive database of controlled trials. Part of this effort is the *Embase screening project*, now part of "Cochrane Crowd" (http://crowd.cochrane.org/). This project enlists "citizen scientists" to screen records identified from a high recall search of Embase for reports of randomized or quasi-randomized trials. To date, this volunteer crowd has identified >30 000 trial reports; these have in turn been indexed in Cochrane's database.

Cochrane crowd volunteers undergo mandatory training, which involves screening 20 practice records (feedback is provided). In practice, an agreement protocol ensures that all records are assessed multiple times. These safeguards increase the already substantial manual effort required of crowd workers. The prospect of automating the process of RCT screening via machine learning (ML) is therefore appealing. ML has been used extensively to reduce the workload in screening for individual systematic reviews,[6,7] but hybrid crowd-ML screening systems have not been explored in depth.

It is currently unclear whether ML systems can achieve sufficient recall to be used as a sole means of RCT identification. Cohen et al.[2] report high areas under the curve for this task, but given the low prevalence of RCTs, any particular threshold used for classification could still yield a substantial number of missed trials or produce an abundance of false positives. An alternative strategy is to use ML to *reduce* workload by screening out "obvious" non-RCTs and deferring to humans for the rest. In this brief communication, we describe a simple hybrid ML-crowdsourcing approach to facilitate exhaustive RCT identification. We report results from both retrospective (simulated) and prospective experiments: this system is currently in place. Our results show that considerable workload reduction can be achieved via semi-automation without substantially sacrificing recall.

EXPERIMENTAL SETUP

Retrospective simulation

Data

Three categories of workers have contributed to the Embase screening project: *novices, experts* (individuals who screened >1000 abstracts and realized at least 95% recall and specificity with respect to a reference standard), and *resolvers*. Individuals in these categories are associated with different costs and screening accuracies (the latter 2 types being more experienced and thus pricier, but more accurate). We define 1 unit of cost as a single screening decision provided by a novice screener (we assume experts and resolvers are twice and 4 times as expensive, respectively).

We used a dataset comprising 61 365 citations. Each of these came attached with multiple screening decisions collected as part of the Embase screening project in 2014–2015. In total, we had 131 070 individual assessments (~2.1 decisions per citation, on average). Most of these (97 512) were from expert screeners; 29 383 labels were from novices, and the remainder (4175) were from resolvers. We note that these data were collected toward the beginning of the project and hence contained a higher proportion of "expert" screeners than is currently the case.

To induce a "ground truth" label for each citation in the dataset, we followed the conservative strategy validated by Cochrane, which proceeds as follows. If only novices are involved, 3 consecutive consistent judgments are required. If at least 1 screener is an expert, then only 2 consecutive agreements are required. Any designation as *unsure* sends an abstract to a resolver, as do any disagreements. In these cases, the resolver's label is taken as ground truth. After aggregation, we had 2531 RCTs (positive instances) and 58 834 non-RCTs (negative instances).

RCT classifiers

For our ML model (which classifies abstracts as reports of RCTs or not), we used a linear kernel support vector machine induced over n-gram representations of abstracts.[8] We trained this model on a dataset external to Embase; the model and dataset are described at length elsewhere[2] (we used the model variant that depends on citation information only, ie, no Medical Subject Heading terms). Note that this model was trained on records published between 1987 and 2012, whereas the Embase dataset comprises records published between 2015 and 2016. Thus there can be no overlap between training and testing data here. This simplifies analysis, avoiding evaluation of the model using the same data on which it was trained. However, it may degrade classifier performance if the population of studies in Embase differs from that in the original training set.

Simulation strategy

We aimed to evaluate the effects of a hybrid crowd-ML approach to abstract screening on accuracy and annotator effort. We adopted a simple strategy concerning when to rely on the model and when to defer to human expertise. We first predicted the probability of all abstracts being RCTs using the pretrained RCT classifier described above. Where the model predicted with sufficiently high confidence that an abstract was *not* an RCT (defined here as a predicted probability ≤ 0.1), we excluded it without manual review. The remaining articles (probability >0.1) received a definitive classification via crowd manual review.

We then tallied the simulated number of crowd labels required (a proxy for effort) using the usual process compared to the hybrid crowd-ML approach. We calculated the effective labor reduction and the resultant precision/recall using this hybrid approach, as compared to the "ground truth" assessments currently in place. We also explored varying the confidence threshold parameter.

We used number of labeling decisions as a proxy for cost, and we treated a single novice label as a "unit" of cost. We assumed that expert labels are twice as expensive (thus 1 expert label incurs 2 cost units) and that resolver labels are 4 times as expensive.

Prospective evaluation

The strategy reported above is now being used prospectively for RCT identification. We therefore provide a preliminary evaluation of its performance *in practice*, by quantifying the labor savings and estimating recall.

RESULTS

We first report the empirical performance of a simple and conservative heuristic strategy in which we simply trusted classifier exclusion decisions for records that received a probability ≤ 0.1 of being an RCT, and sending the remainder to the crowd for standard assessment. To evaluate this, we used a set of approximately 158 000 Embase records screened by Cochrane Crowd. We held back 30% of records and trained the classifier on the remaining approximately 111 000 records, subsequently scoring the other approximately 47 000 records with the induced classifier. The receiver operating characteristic curve achieved on the held-out examples is shown in the left-hand subplot in Figure 1. Note that this model is (1) trained directly on the population of documents being screened by Cochrane crowdworkers and (2) independent from the pretrained RCT classifier used in our simulation experiments below.

Table 1 shows that 77.4% (36 709/47 445) of the records screened by the crowd had a predicted probability of being an RCT of <0.1. If we accept the classifier assessment in such cases (taking no further action with these records), we would lose a small number of RCTs while drastically reducing workload.

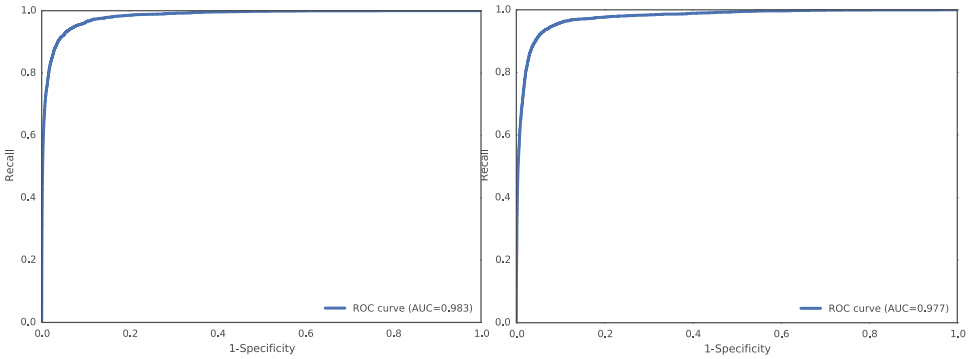

Figure 1. Left: Receiver operating characteristic curve showing the performance of our RCT classifier, trained on a subset of the Embase dataset. Right: Receiver operating characteristic curve showing the performance of our pretrained RCT classifier[2] on the entire Embase dataset.

Table 1. Distribution of "ground truth" RCTs and non-RCTs within ranges of classifier confidence (N gives the number of abstracts that fall into each range)

Probability	RCT	Non-RCT	*N*	Cumulative recall	Percent "screened"
0.9 to 1.0	1511	210	1721	0.633	3.6
0.8 to < 0.9	269	242	511	0.746	4.7
0.7 to < 0.8	150	270	420	0.809	5.6
0.6 to < 0.7	110	323	433	0.855	6.5
0.5 to < 0.6	92	396	488	0.893	7.5
0.4 to < 0.5	71	573	644	0.923	8.9
0.3 to < 0.4	63	912	975	0.950	10.9
0.2 to < 0.3	55	1635	1690	0.972	14.5
0.1 to < 0.2	47	3807	3854	0.992	22.6
<0.1	19	36 690	36 709	1	100

Table 2. First 3 columns: number of labels acquired from each type of labeler using the manual, hybrid, and totally automated approaches (with 2 different thresholds shown). Second 2 columns: precision and recall with respect to identifying RCTs. The manual measures have asterisks; we assume these are "ground truth" by construction (see text for discussion)

	Novice	Expert	Resolver	Precision	Recall
Manual	29,376	97,512	1,895	1.0*	1.0*
Hybrid	3,884	12,218	4,175	0.99	0.96
Classifier-only (threshold = 0.5)	0	0	0	0.99	0.71
Classifier only (threshold = 0.1)	0	0	0	0.27	0.96

Retrospective simulation results

The right-hand subplot in Figure 1 shows the receiver operating characteristic curve of the pretrained classifier[2] on the entire Embase set.

The cost incurred in practice using the manual approach was 241 100 units; under simulation, using the hybrid approach, it was 35 860 (a reduction of nearly 7-fold). We provide the total label count breakdowns for the respective strategies in Table 2, together with precision and recall.

The approach above used a single fixed confidence threshold, $t = 0.1$, to decide when to trust the model to automatically exclude abstracts and when to defer to humans. Adjusting t will trade recall against labor. We simulated adjusting t, and report the results in Figure 2. This shows the recall achieved vs the simulated total effort expended using values of t ranging from 0.0 (completely manual, since this means we never trust the model) to 0.5.

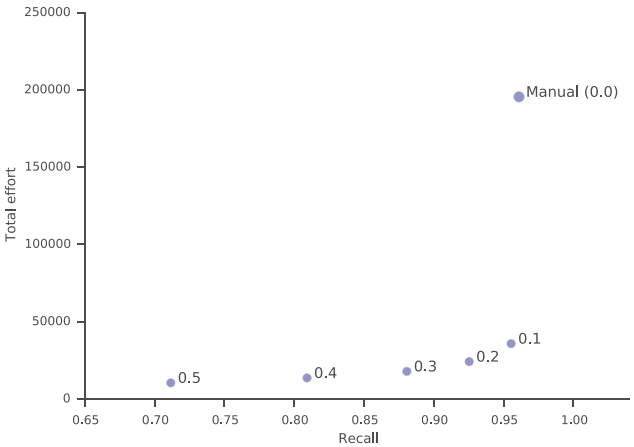

Figure 2. A scatterplot of recall vs (simulated) total expended effort for varying values of the confidence threshold *t*. As noted in the text, effort is modeled as unit costs, where 1 novice screening decision = 1 unit, 1 expert decision = 2 units, and 1 resolver decision = 4 units.

Prospective results

We conclude by providing a *prospective* assessment of the simple hybrid strategy explored retrospectively above (ie, deferring to the classifier when sufficiently confident). Starting from April 2016, we began to withhold from screening citations that received a predicted probability of being an RCT of ≤0.1. In total, 17 495 such citations were identified and held aside (ie, not assigned to crowd workers). Out of the remaining 4920 citations that received a predicted probability of >0.1, the crowd-labeling process described above identified 1563 as being reports of RCTs. Holding aside low-probability citations reduced the crowd workload by 78%.

To assess the performance of this hybrid strategy in practice, domain experts at Cochrane manually assessed the 17 495 citations not routed to the crowd to identify any false negatives. In total, 32 of the 17 495 citations were deemed RCTs (false negatives), which gives a specificity of 99.8% and an overall recall of 98%. Many of these do not explicitly state that the trial was randomized, so this is a conservative estimate.

In sum, when deployed prospectively, integrating ML into the crowd workflow was found to prospectively reduce workload by 78%, while still identifying 98% of all RCTs.

DISCUSSION AND LIMITATIONS

Our results support the use of a hybrid ML-crowd approach for the task of identifying reports of RCTs, and perhaps for supporting biomedical annotation tasks more generally. The simple hybrid strategy we propose is currently in use by Cochrane and has resulted in a more efficient workflow. Our prospective results suggest that this efficiency comes at a modest cost in recall.

Setting a confidence threshold t that codifies when to trust the machine and when to defer to human expertise is an important practical consideration for our approach. From Figure 2, one can observe that thresholds greater than 0 substantially reduce labor, but aggressive thresholding severely degrades recall.

The reported results are pessimistic with respect to the performance of the hybrid strategy, because it is possible that the few "false negatives" the hybrid strategy produced were in fact not false negatives (and conversely, apparent "false positives" may have been true positives). Even with this imperfect evaluation, we observe that the hybrid approach achieved nearly perfect precision and above 95% recall. This is promising, considering the accompanying reduction in effort/cost. Relying on the classifier alone with threshold = 0.1 resulted in a high recall (0.96) but low precision (0.27), while setting the threshold to 0.5 increased precision but dramatically reduced recall.

We note several limitations. First, considered against the likely number of RCTs remaining to be located, even a 1%–2% drop in recall might still represent an unacceptably large loss, eg, assuming we need to find another 500 000 RCTs, this level of recall might still miss 500–1000 trials. However, it seems unlikely that the manual process is infallible. Furthermore, database searching is not the only way that reviewers identify studies for inclusion, and the time saved in not needing to search major databases could be invested in other tasks, eg, checking bibliographies and trial registries.

An additional limitation is that the hybrid strategy relies on significantly more "resolvers" than does the current approach. However, our results suggest that this is worthwhile even when factoring in the higher cost of resolver time. Recruiting additional resolvers would be feasible, given the size of the Cochrane network. Finally, the hybrid strategy we used here is naïve; it entails simply trusting the classifier when it is very confident that a given abstract is not an RCT. More sophisticated strategies for assigning abstracts to workers[9,10] may bring further benefits, but may complicate deployment.

CONCLUSIONS

We have presented a simple, effective hybrid machine learning and crowdsourcing approach to identifying reports of RCTs, in which we defer to crowdworkers when the classifier is not sufficiently certain. We deployed this approach in practice and found that using our hybrid approach reduced the number of manual screening decisions needed by 78% while maintaining an estimated recall of 98%. This work thus demonstrates the potential of hybrid crowd–machine learning systems for scaling biomedical annotation tasks.

FUNDING

BCW's contribution to the work was supported by the Agency for Healthcare Research Quality, grant R03-HS025024, and from the National Institutes of Health/National Cancer Institute, grant UH2-CA203711. IJM acknowledges support from the UK Medical Research Council, through its Skills Development Fellowship program, grant MR/N015185/1. JT and ANS acknowledge support from Cochrane via the Transform project and its discretionary fund. AMC and NS were supported by National Institutes of Health grant R01-LM010817.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

BCW, ANS, JT, and IJM conceived of the project and approach. AMC and NS developed/trained and shared the initial machine learning classifier. BCW and IJM coded and executed all simulation experiments. JT and ANS ran and evaluated prospective experiments. JT ported BW's code/model for use in the prospective setup. BCW drafted the first version of manuscript; all authors provided edits and feedback. BCW drafted major revisions in response to *JAMIA* reviewer comments; all authors reviewed and approved these.

REFERENCES

1. Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann NY Acad Sci*. 1993;703(1):156–65.
2. Cohen AM, Smalheiser NR, McDonagh MS, *et al*. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *J Am Med Inform Assoc*. 2015;22(3):707–17.
3. McKibbon KA, Wilczynskil NL, Haynes RB. Retrieving randomized controlled trials from Medline: a comparison of 38 published search filters. *Health Info Libr J*. 2009;26(3):187–202.
4. Wieland LS, Robinson KA, Dickersin K. Understanding why evidence from randomised clinical trials may not be retrieved from Medline: comparison of indexed and non-indexed records. *Brit Med J*. 2012;344:2008–12.
5. Bastian H, Glaszio P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*. 2010;7(9):1–6.
6. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioninformatics*. 2010;11(1):55.
7. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc*. 2006;13(2):206–19.
8. Joachims T. Text categorization with support vector machines: learning with many relevant features. In *European Conference on Machine Learning (ECML)*. Berlin: Springer; 1998.
9. Wallace BC, Small K, Brodley CE, Trikalinos TA. Who should label what? Instance allocation in multiple expert active learning. In *SIAM International Conference on Data Mining*. Phoenix, AZ: Society for Industrial and Applied Mathematics; 2011.
10. Nguyen AT, Wallace BC, Lease M. Combining crowd and expert labels using decision theoretic active learning. In *Human Computation and Crowdsourcing (HCOMP)*. San Diego, CA: Association for the Advancement of Artificial Intelligence (AAAI); 2015.