JAMIA Open, 3(3), 2020, 338–341
doi: 10.1093/jamiaopen/ooaa042
Advance Access Publication Date: 28 October 2020
Application Notes

Application Notes

New improved Aggregator: predicting which clinical trial articles derive from the same registered clinical trial

Neil R. Smalheiser and Arthur W. Holt

Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois, USA

Corresponding Author: Neil R. Smalheiser, MD, PhD, Department of Psychiatry, University of Illinois at Chicago, 1601 W. Taylor St. MC912, Chicago, IL 60612, USA; neils@uic.edu

Received 5 May 2020; Revised 15 June 2020; Editorial Decision 3 August 2020; Accepted 2 September 2020

ABSTRACT

Objectives: To identify separate publications that report outcomes from the same underlying clinical trial, in order to avoid over-counting these as independent pieces of evidence.

Materials and Methods: We updated our previous model by creating larger, more recent, and more diverse positive and negative training sets consisting of article pairs that were (or not) linked to the same ClinicalTrials.gov trial registry number. Features were extracted from PubMed metadata; pairwise similarity scores were modeled using logistic regression and used to form clusters of articles that are likely to arise from the same registered clinical trial.

Results: Articles from the same trial were identified with high accuracy (F1 = 0.859), nominally better than the previous model (F1 = 0.843). Predicted clusters showed a low error rate of splitting of 8–11% (ie, when 2 articles belonged to the same trial but were assigned to different clusters). Performance was similar whether only randomized controlled trial articles or a more diverse set of clinical trial articles were processed.

Discussion: Metadata are surprisingly accurate in predicting when 2 articles derive from the same underlying clinical trial.

Conclusion: We have continued confidence in the Aggregator tool which can be accessed publicly at http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/RCT_Tagger.cgi.

Key words: evidence-based medicine, clinical trials, systematic reviews, information retrieval, informatics

LAY SUMMARY

An important type of bias in evidence-based medicine occurs when 1 clinical trial gives rise to multiple publications that are erroneously counted as arising from independent sources of evidence. We have previously modeled and implemented a free web-based tool, Aggregator, that clusters clinical trial articles together if they are likely to report findings from the same underlying registered clinical trial. We have now extended the modeling using larger and more diverse training examples. Our evaluation shows that the new model maintains the accuracy of the existing tool for clinical trial articles broadly, and this gives new confidence that the tool may be useful to users who are conducting evidence syntheses.

BACKGROUND AND SIGNIFICANCE

The road from clinical trial to clinical practice is long and slippery. New treatments are tested in clinical trials (of varying size and design), which may be registered in formal clinical trial registries (or not), and which may be published in the peer-reviewed literature (or not).[1] Systematic reviews and meta-analyses are written to assess the published evidence in a standardized and comprehensive manner, and to reach conclusions concerning efficacy and safety that are as free from bias as possible.[2] An important type of bias occurs when 1 clinical trial gives rise to multiple publications that are erroneously counted as arising from independent sources of evidence, which can lead to inaccurate estimates of efficacy.[3,4] This is a challenging problem requiring close reading of the full text (plus other information such as writing to authors), made worse by the fact that multiple publications often do not cite each other, may have completely non-intersecting sets of authors, and often do not mention clinical trial registry numbers.[5,6]

OBJECTIVES

We have previously modeled the problem of deciding whether 2 randomized controlled trial (RCT) articles indexed in PubMed belong to the same clinical trial, and created a public tool called Aggregator that clusters together RCT articles that are likely to arise from the same registered trial.[7] However, we decided to update and extend the existing model and code and evaluations, for 4 reasons: first, the original training data were collected in 2013, whereas the number of trial-linked publications has grown substantially since then. Instead of ~450 training examples, now we have ~4500 examples. Second, the original negative training pairs came from different registered trials but were matched on both condition and intervention narrowly, which is possibly too restrictive and not optimally robust. Third, the original clustering algorithms were coded in MATLAB, which is proprietary software and had compatibility issues as operating systems and Python were updated. Fourth, the original evaluations were carried out only for RCT articles; we wished to check whether the tool is as accurate for clinical trial articles more broadly.

MATERIALS AND METHODS

The detailed "Materials and Methods" section are described in Supplementary File S1. References cited therein are placed in the reference list of the main article.

RESULTS

Aggregator has been implemented as 1 piece of an overall suite of informatics tools that can accelerate the process of writing systematic reviews. To access Aggregator, a user first enters a PubMed query into the RCT Tagger tool (http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/RCT_Tagger.cgi). RCT Tagger assigns a value $0 \le P \le 1$ to each article which estimates the probability that it is a randomized controlled trial (RCT).[8] The user will then see a new Aggregator tab and have the option to process all retrieved articles that have RCT confidence scores $\ge$0.9. In this manner, articles which lack NLM indexing can still be identified and considered (eg, recently published articles that have not yet been indexed for publication type). The articles are processed pairwise and clustered in real time, and displayed to show clusters containing 2 or more articles that are likely to arise from the same registered trial. We have used simple hierarchical agglomerative clustering, which begins by identifying the most similar article pairs and placing them into the same cluster. Other articles are examined in random order and placed into an existing cluster if their average similarity to the articles in that cluster is high enough; then the threshold similarity is progressively lowered and the remaining articles are examined again until a threshold similarity value of 0.5 is reached, meaning (roughly) that an article added to a cluster at that point has a probability of 0.5 of belonging to that cluster. This method worked well in the past[7] and was retained here.

Pairwise model

Performance of the pairwise model is shown in Figure 1 as a function of the confidence score. Trained and optimized on 75% of the training data and tested on the remaining 25%, we obtained precision = 0.89, recall = 0.82, accuracy = 0.86, and F1 = 0.86. This performance is nominally slightly better than our earlier model.[7]

Error analysis

False-positive predictions comprised 37% of all errors. Most false-positives involved pairs of articles sharing 2–6 authors, researching the same or similar conditions at the same location. For example, PMID's 20701787 and 22895351 receive a predicted probability of 69.3% from the model. The publications share 3 authors, share affiliation country, share 3 grant numbers, and indeed are derived from very similar trials, NCT00673309 and NCT00675714. We believe that the false-positive predictions are still informative, as they bring together trials that may be directly related (eg, a phase I and phase 2 trial on the same therapy) and generally involve the same research team.

False-negative predictions comprised 63% of all errors. Positive article pairs could receive predicted probability estimates below 50% when the most highly weighted features were not a match. For example, article pair PMID 29145839 and 26188189 derive from the same trial but only received a probability score of 40.7% because they lack any shared names in the author list. In another example, PMID 29726951 and 26581681 share 2 authors, a grant number, an affiliation, and a substance, yet only have a model-predicted probability of 46.2% because they lack several highly weighted features (eg, a high PubMed-related article ranking or all-capitalized words).

Evaluation of clustering

To evaluate the quality of the clustering solution, we repeated the series of 20 PubMed queries carried out on various conditions and 22 queries carried out on similar conditions plus specific interventions that were used to evaluate the earlier version of Aggregator (ref.[7]; also see "Materials and Methods" section). The most important error is "splitting," that is, the proportion of articles that belong to the same trial but are predicted to reside in distinct clusters. This needs to be minimized so that users will not falsely regard different studies as an independent. The purity parameter measures "lumping" of articles that arise from distinct studies, but are clustered together; this is less important since it should be easier for systematic reviewers to manually separate articles that are placed together in 1 cluster than to identify and link related articles that are placed in separate clusters. As shown in Table 1, performance overall is excellent, though better average performance was seen for the condition + intervention queries than observed for the queries based on condition alone. This probably reflects the fact that the condition + intervention articles are a more topically homogeneous set, and thus resembled more closely the positive versus training sets that were used in the model.

We note that the original Aggregator model was evaluated and implemented only on RCT articles,[7] even though the original (and new) models were trained on a broader set of clinical trial articles. We repeated the evaluation by running the same queries but instead of limiting the articles to randomized controlled trials (publication type [PT]), they were limited to the broader category of clinical trial (PT). This increased the number of retrieved articles by ~50% (conditions queries: average number of retrieved articles limited to RCT = 1167 vs 2308 limited to Clinical Trial; conditions + intervention queries:

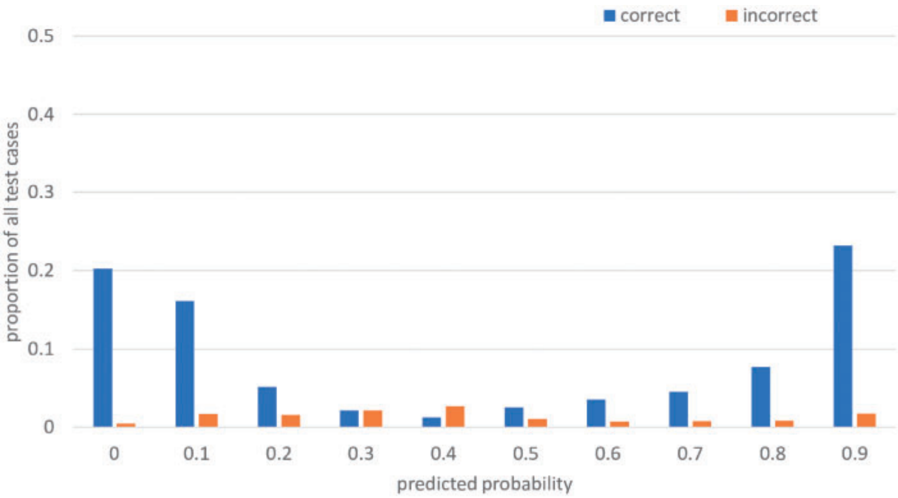

Figure 1. Performance of the pairwise similarity model. Shown are the proportions of correct and incorrect predictions, as a function of the confidence score for each pair of articles for the hold-out test samples that were not used for model fitting. Most correct predicted probability estimates were very definitive (ie, ≤0.1 or >0.8). In contrast, the incorrect estimates were scattered between 0 and 1, but particularly below 0.5. This suggests that the biggest limitation to performance is due to features missing from articles, causing some positive pairs to receive low predicted probability estimates.

Table 1. Clustering performance of Aggregator

	Split	Purity	F1
RCT articles retrieved and clustered			
Conditions queries, all retrieved articles	0.11	0.91	0.90
Conditions queries, only NCT-containing articles	0.083	0.90	0.91
Condition + intervention query, all retrieved articles	0.086	0.94	0.93
Condition + intervention query, only NCT-containing articles	0.079	0.93	0.93
Clinical trial articles retrieved and clustered			
Conditions queries, all retrieved articles	0.11	0.90	0.89
Conditions queries, only NCT-containing articles	0.10	0.89	0.89
Condition + intervention query, all retrieved articles	0.085	0.93	0.92
Condition + intervention query, only NCT-containing articles	0.077	0.92	0.92

Note: We used Aggregator either to cluster all articles retrieved by these searches, or only clustered the subset of articles that contained NCT numbers. The clustering algorithm generally performed better when both condition and intervention were queried.

Abbreviations: RCT: randomized controlled trial.

average number of retrieved articles limited to RCT = 314 vs 566 limited to Clinical Trial). As shown in Table 1, the clustering performance was similar whether only RCT articles or the broader set of clinical trial types were evaluated.

DISCUSSION

Multiple publications that report clinical outcomes from the same clinical trial can bias the apparent effect size calculated in systematic reviews and meta-analyses.[3,4] This is a well-recognized problem that is generally handled by time-consuming manual effort.[5,6] In our analysis, we defined a trial as one given a unique registry number in ClinicalTrials.gov or other trial registries. The updated model containing more extensive training data had nominally better performance than the earlier version,[7] and should be easier to maintain into the future since it does not rely on proprietary software.

The Aggregator tool is free and open for public use without the need for registration or password. However, it has several limitations: the model is only designed to evaluate clinical trial articles that are topically similar and PubMed indexed. It was not specifically designed to detect cases of plagiarism (by different authors), or situations where trial organizers have deliberately attempted to obscure relationships among their publications, for example, by using ghostwriters.

Perhaps most importantly, the current web tool implementation only allows users to process articles that are predicted likely to be RCTs, even though our model is trained on a broader set of clinical trial articles (ie, any MEDLINE-indexed PT that includes the word "trial" as well as Multicenter Study). As shown in this report, the clustering performance of Aggregator is similar whether only RCTs or all clinical trial articles are considered. Our team is currently developing automated probabilistic publication type taggers to identify not only RCTs, but a variety of other publication types including the broader category of clinical trials, whether or not they have received formal MEDLINE indexing. When this tagger is available, we will use it to extend the types of trial articles that can serve as input to Aggregator.

Apart from forming clusters, the updated pairwise similarity model could potentially be used to create a tool in which, for any given clinical trial article, a ranked list of the most similar articles are displayed. Since features such as shared authors and all-capitalized trial acronyms are important in the pairwise model, such a ranking would be quite different from text-based rankings. We are also planning to employ the pairwise model as part of a larger effort to match registered clinical trials to their most similar articles.[9–11]

FUNDING

This work was supported by NIH Grants R01LM010817 and P01AG039347. The study sponsor had no role in study design; in the collection, analysis, and interpretation data; in the writing of the report; or in the decision to submit the paper for publication.

AUTHOR CONTRIBUTIONS

NRS: Conceived of the study, designed the tool, and wrote the initial draft of the paper. AWH: Carried out, optimized, and evaluated the model, and assisted in writing the paper.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 2012; 344: d7292.
2. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; 126 (5): 376–80.
3. Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997; 315 (7109): 635–40.
4. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol* 2000; 53 (2): 207–16.
5. von Elm E, Poglia G, Walder B, Tramèr MR. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA* 2004; 291 (8): 974–80.
6. Wilhelmus KR. Redundant publication of clinical trials on herpetic keratitis. *Am J Ophthalmol* 2007; 144 (2): 222–6.
7. Shao W, Adams C, Cohen A, *et al*. Aggregator: a machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. *Methods* 2015; 74: 65–70.
8. 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.
9. King G, Langche Z. Logistic regression in rare events data. *Polit Anal* 2001; 9 (2): 137–63.
10. Bashir R, Bourgeois FT, Dunn AG. A systematic review of the processes used to link clinical trial registrations to their published results. *Syst Rev* 2017; 6 (1): 123.
11. Dunn AG, Coiera E, Bourgeois FT. Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of ClinicalTrials.gov. *J Clin Epidemiol* 2018; 95: 94–101.