Journal of Biomedical Informatics 79 (2018) 32–40

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A shared latent space matrix factorisation method for recommending new trial evidence for systematic review updates

Didi Surian[a,*], Adam G. Dunn[a], Liat Orenstein[b], Rabia Bashir[a], Enrico Coiera[a], Florence T. Bourgeois[b,c]

[a] Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[b] Computational Health Informatics Program, Boston Children's Hospital, Boston, United States
[c] Department of Pediatrics, Harvard Medical School, Boston, United States

ARTICLE INFO

Keywords:
Systematic reviews
Clinical trials
Information retrieval
Matrix factorisation

ABSTRACT

Background: Clinical trial registries can be used to monitor the production of trial evidence and signal when systematic reviews become out of date. However, this use has been limited to date due to the extensive manual review required to search for and screen relevant trial registrations. Our aim was to evaluate a new method that could partially automate the identification of trial registrations that may be relevant for systematic review updates.

Materials and methods: We identified 179 systematic reviews of drug interventions for type 2 diabetes, which included 537 clinical trials that had registrations in ClinicalTrials.gov. Text from the trial registrations were used as features directly, or transformed using Latent Dirichlet Allocation (LDA) or Principal Component Analysis (PCA). We tested a novel matrix factorisation approach that uses a shared latent space to learn how to rank relevant trial registrations for each systematic review, comparing the performance to document similarity to rank relevant trial registrations. The two approaches were tested on a holdout set of the newest trials from the set of type 2 diabetes systematic reviews and an unseen set of 141 clinical trial registrations from 17 updated systematic reviews published in the Cochrane Database of Systematic Reviews. The performance was measured by the number of relevant registrations found after examining 100 candidates (recall@100) and the median rank of relevant registrations in the ranked candidate lists.

Results: The matrix factorisation approach outperformed the document similarity approach with a median rank of 59 (of 128,392 candidate registrations in ClinicalTrials.gov) and recall@100 of 60.9% using LDA feature representation, compared to a median rank of 138 and recall@100 of 42.8% in the document similarity baseline. In the second set of systematic reviews and their updates, the highest performing approach used document similarity and gave a median rank of 67 (recall@100 of 62.9%).

Conclusions: A shared latent space matrix factorisation method was useful for ranking trial registrations to reduce the manual workload associated with finding relevant trials for systematic review updates. The results suggest that the approach could be used as part of a semi-automated pipeline for monitoring potentially new evidence for inclusion in a review update.

1. Background

Systematic reviews of clinical trials are at the foundation of evidence-based medicine and should represent comprehensive, high quality, and up to date syntheses of trial evidence. With the rapid growth of the scientific literature, identifying relevant evidence and keeping systematic reviews up to date is increasingly difficult. Studies examining the timing of systematic reviews suggest that reviews are updated on average every 5.5 years, though a substantial proportion should be updated within 2 years [1–4]. Performing systematic reviews is time and resource intensive and even determining when a systematic review needs updating often requires completion of the searching and screening steps of the systematic review process. To facilitate this assessment, a number of tools and guidelines have been developed that aim to identify when new relevant research becomes available, or estimate the risk that the results of the systematic review may have substantially changed due to new evidence [3–10].

These approaches rely on bibliographic databases, which are limited

* Corresponding author.
E-mail address: didi.surian@mq.edu.au (D. Surian).

https://doi.org/10.1016/j.jbi.2018.01.008
Received 20 September 2017; Received in revised form 9 December 2017; Accepted 22 January 2018
Available online 02 February 2018
1532-0464/

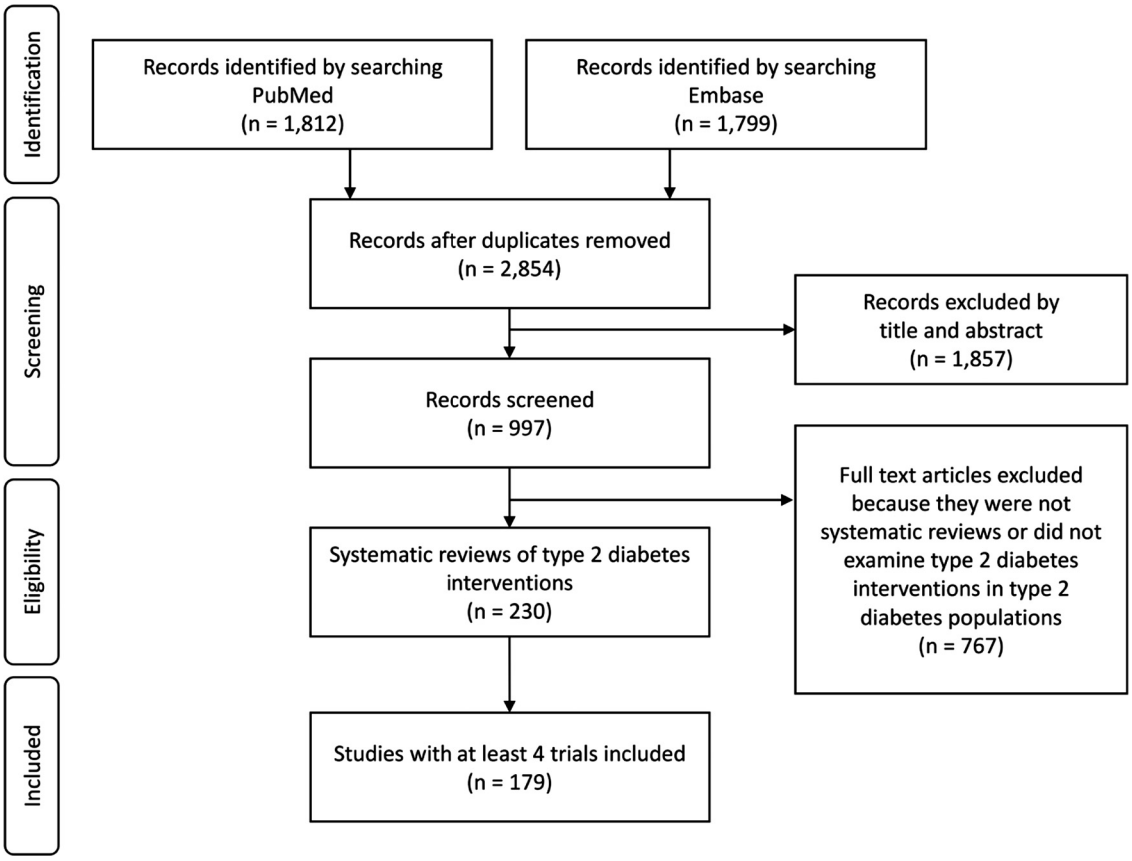

Fig. 1. From 2854 unique articles identified in the search, 230 systematic reviews of type 2 diabetes were identified and 179 were included in the experiments.

due to publication and reporting biases that affect the timing and completeness of the results [11]. About half of all trials remain unpublished two years after trial completion, and of those that are published, around half have missing or changed outcomes [12,13]. As a consequence, bibliographic databases may not provide a complete and timely source of relevant trial evidence for systematic reviews. As various policies and mandates are making prospective trial registration standard practice, clinical trial registries are an increasingly comprehensive and timely source of new research evidence, and in many cases may provide a more complete and less biased record than bibliographic databases [14]. However, the vast majority of methods aiming to support the identification of relevant studies for a systematic review operate over bibliographic databases rather than trial registries [15] and systematic reviews often fail to incorporate any clinical trial registries to identify relevant trials [16]. New methods for identifying relevant trials in clinical trial registries could help determine when systematic reviews need to be updated and support living systematic reviews and automated systematic review updates [17–19].

Our aim was to evaluate a new method to partially automate the identification of trials that may be relevant for systematic review updates given the existing trials in a systematic review. This process could serve to signal when a systematic review becomes out of date, based on the amount and type of new evidence that is detected.

2. Related work

A number of semi-automated methods have been proposed to identify relevant trials for inclusion in systematic reviews and improve the efficiency of the searching and screening processes [15,20,21]. The methods typically use the words or concepts included in the text of published articles to find similarities that are then used to distinguish relevant from irrelevant articles. Some work has also been done to directly extract information on populations, interventions, comparators, and outcomes [22,23], which can then be used to match search queries. Several approaches have included the use of active learning [24,25], while others have examined representations that use neural network based vector space models [26]. Far less work has been performed on identifying trials from the information stored in clinical trial registries or on linking clinical trial registries to bibliographic databases [14,27–29]. However, some methods have shown that it is possible to identify meaningful clusters of similar trials within registries [30–32], especially in relation to populations [33], and ClinicalTrials.gov data has been used in predicting black box warnings [34].

Matrix factorisation has the potential to support the identification of relevant trials for inclusion in systematic reviews. The approach has a long history of use in addressing problems in link prediction [35–37], for example in building systems that recommend books, music, or new social connections to users. This process, commonly referred to as “the item prediction problem”, aims to predict the presence or absence of links between the nodes of a graph where the vertices represent users and items, and edges that connect vertices are weighted according to preference scores. Matrix factorisation produces a mapping between users and items into a low-dimensional representation (latent factors) to model the user-item affinity in vector space.

Past work on the use of matrix factorisation for collaborative filtering focused on increasing prediction accuracy by including neighbourhood information [38]. Later, Koren et al. [39] proposed SVD++, a matrix factorisation approach that unified neighbourhood and latent factors. Guo et al. [40] proposed TrustSVD, an extension of SVD++ that incorporates social trust information to help mitigate data sparsity and the cold start problem. TrustSVD includes factorisation of two matrices that share a same latent space—meaning a matrix of user-item preference scores and another matrix that defines trust information among users.