Journal of Clinical Epidemiology 95 (2018) 94−101

Journal of Clinical Epidemiology

Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of ClinicalTrials.gov

Adam G. Dunn[a,*], Enrico Coiera[a], Florence T. Bourgeois[b,c]

[a]*Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney 2109, New South Wales, Australia*
[b]*Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA*
[c]*Department of Pediatrics, Harvard Medical School, Boston, MA, USA*

Accepted 14 December 2017; Published online 19 December 2017

Abstract

Objectives: Trial registries can be used to measure reporting biases and support systematic reviews, but 45% of registrations do not provide a link to the article reporting on the trial. We evaluated the use of document similarity methods to identify unreported links between ClinicalTrials.gov and PubMed.

Study Design and Setting: We extracted terms and concepts from a data set of 72,469 ClinicalTrials.gov registrations and 276,307 PubMed articles and tested methods for ranking articles across 16,005 reported links and 90 manually identified unreported links. Performance was measured by the median rank of matching articles and the proportion of unreported links that could be found by screening ranked candidate articles in order.

Results: The best-performing concept-based representation produced a median rank of 3 (interquartile range [IQR] 1−21) for reported links and 3 (IQR 1−19) for the manually identified unreported links, and term-based representations produced a median rank of 2 (1−20) for reported links and 2 (IQR 1−12) in unreported links. The matching article was ranked first for 40% of registrations, and screening 50 candidate articles per registration identified 86% of the unreported links.

Conclusion: Leveraging the growth in the corpus of reported links between ClinicalTrials.gov and PubMed, we found that document similarity methods can assist in the identification of unreported links between trial registrations and corresponding articles.

Keywords: Clinical trial registry; Bibliographic database; Publication bias; Reporting bias; Systematic review; Clinical trial reporting; Trial registration

1. Introduction

Clinical trial registries were established to track the conduct of clinical trials and make basic information about trials publicly available. A number of policies now mandate prospective registration for clinical studies of regulated interventions [1−3]. ClinicalTrials.gov is a US-based registry for clinical studies and is the largest single database of trial registrations. ClinicalTrials.gov also links registrations to published results by connecting to research articles indexed in bibliographic databases [4,5]. This linkage is achieved using a unique identifier (the National Clinical Trial [NCT] number) for each study. Publishers may include the NCT number in the abstract or full text of published articles and the metadata stored by PubMed, a bibliographic database that includes the details of more than 26 million biomedical articles.

Although the introduction of trial registries has been invaluable for monitoring trial reporting, a substantial proportion of articles remain disconnected from their registrations. In a 2012 study examining the quality of linking in ClinicalTrials.gov, 44% of registrations without linked publications were found to have corresponding published articles found by manual searches [6]. In a systematic review of studies that examine reported and unreported links between registrations and articles, the median proportion of registrations with reported links was 23% and the median proportion of unreported links (those that required manual searches) was 17%, with the remainder unpublished [7].

Conflicts of Interest: None.

Funding: This work was supported by the Agency for Healthcare Research and Quality (R03HS024798).

* Corresponding author. Tel.: +612-9850-2413.
E-mail address: adam.dunn@mq.edu.au (A.G. Dunn).

https://doi.org/10.1016/j.jclinepi.2017.12.007
0895-4356/

What is new?

Key findings

- Unreported links occur when articles reporting the results of a trial do not include the trial registration information in their bibliographic metadata.
- Document similarity measures can be used to replace the need to construct search strategies to identify unreported links from trial registrations to articles.

What this adds to what was known?

- The results confirm that approximately 45% of published trials have unreported links from ClinicalTrials.gov; 86% of unreported links could be found by screening the first 50 ranked candidate articles for each registration.

What is the implication and what should change now?

- Incomplete linking between ClinicalTrials.gov and PubMed hampers efforts to measure publication and outcome-reporting biases and to identify relevant trials for systematic reviews.
- Trial registrations can have their value enhanced if there is more effort devoted to linking them to published reports using automated methods because this will lead to more robust monitoring of trial reporting and improve the integrity of evidence synthesis.

The quality of this linkage between bibliographic databases and trial registries affects the time it takes to measure reporting biases. This includes determining which clinical studies remain unpublished [8–12] or comparing registered outcomes with what is reported in published articles [13–21]. Without comprehensive linking, research to evaluate reporting biases must instead rely on time-consuming manual searches to identify unreported links.

The presence of unreported links also limits the value of trial registries for systematic reviews. If links between registrations and articles were comprehensive, registrations could be more effectively used to automate the identification of trials for inclusion in systematic reviews [22,23], as well as provide early signals that a systematic review should be updated [24–33].

Our aim in this study was to evaluate whether we could use information contained in the recorded links between ClinicalTrials.gov registrations and PubMed articles to help identify unreported links. The long-term goal is to develop robust methods to identify all published research associated with a trial registration, whether or not links are provided in the registration record.

2. Methods

In the following experiments, we use similarities in the text from registrations in ClinicalTrials.gov and articles in PubMed. Using the set of reported links as a baseline, we test a series of different methods to represent the text as features and assign weights to each of the features. The resulting set of features are used to produce an automatic, weighted search query for use in PubMed, where the objective is to rank the matching article as high as possible in a list of candidate articles. The approach is expected to replace the need for an expert to construct search queries in PubMed for every registration without a link to a published article.

2.1. Study data

We included trial registrations in ClinicalTrials.gov for trials that were received by ClinicalTrials.gov on or after October 1, 2007, were marked as completed, and described as an interventional study. The date was selected to correspond to the passing of the Food and Drug Administration Amendments Act of 2007, which expanded the registration requirements for studies registered with ClinicalTrials.gov. A final search of ClinicalTrials.gov was conducted on April 14, 2017. Data extracted from registrations included titles, trial summaries, and conditions studied.

We next selected all articles indexed in PubMed that reported a clinical trial and were published on or after October 1, 2007. Articles were assumed to be reporting the results of trials if they included a ClinicalTrials.gov NCT number as a secondary source identifier or listed ‘‘clinical trial’’, ‘‘controlled clinical trial’’, or ‘‘randomized controlled trial’’ as a publication type and did not include ‘‘meta-analysis’’ or ‘‘review’’ as a publication type. A final search was conducted on April 14, 2017. Data extracted from each PubMed article entry included the title text, abstract text, and any NCT number stored as a secondary source identifier in the metadata. Where PubMed entries included NCT numbers as secondary source identifiers, we described these as reported links and created a data set comprising the set of registrations with known links from one or more articles.

We then created a second set of registrations for testing, comprising 90 registrations that had unreported links to trial articles identified by manually checking 200 registrations with trial completion dates between January 1, 2007, and December 31, 2015. The 200 registrations had no reported links to trial articles in PubMed at the time of the search. We manually searched PubMed and other bibliographic databases to identify articles that reported the results of the trials, following a search strategy previously described and common to studies examining