Journal of Clinical Epidemiology 103 (2018) 22−30

ORIGINAL ARTICLE

Extending PubMed searches to ClinicalTrials.gov through a machine learning approach for systematic reviews

Corrado Lanera[a], Clara Minto[a], Abhinav Sharma[b,1], Dario Gregori[a], Paola Berchialla[c], Ileana Baldi[a,*]

[a]*Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic and Vascular Sciences, University of Padova, Via Loredan 18, Padova 35131, Italy*

[b]*Department of Biological Sciences and Bioengineering (BSBE), IIT, Kanpur, India*

[c]*Department of Clinical and Biological Sciences, University of Torino, Via Santena 5bis, Torino 10126, Italy*

Accepted 29 June 2018; Published online 5 July 2018

Abstract

Objectives: Despite their essential role in collecting and organizing published medical literature, indexed search engines are unable to cover all relevant knowledge. Hence, current literature recommends the inclusion of clinical trial registries in systematic reviews (SRs). This study aims to provide an automated approach to extend a search on PubMed to the ClinicalTrials.gov database, relying on text mining and machine learning techniques.

Study Design and Setting: The procedure starts from a literature search on PubMed. Next, it considers the training of a classifier that can identify documents with a comparable word characterization in the ClinicalTrials.gov clinical trial repository. Fourteen SRs, covering a broad range of health conditions, are used as case studies for external validation. A cross-validated support-vector machine (SVM) model was used as the classifier.

Results: The sensitivity was 100% in all SRs except one (87.5%), and the specificity ranged from 97.2% to 99.9%. The ability of the instrument to distinguish on-topic from off-topic articles ranged from an area under the receiver operator characteristic curve of 93.4% to 99.9%.

Conclusion: The proposed machine learning instrument has the potential to help researchers identify relevant studies in the SR process by reducing workload, without losing sensitivity and at a small price in terms of specificity.

Keywords: Systematic review; Meta-analysis; Clinical trial registry; Indexed search engine; Machine learning; Text mining

1. Introduction

In medical practice and research, the highest level of evidence is represented by systematic reviews (SRs) [1]. An SR is the synthesis and evaluation of all relevant literature on a specific topic, aimed to make the available knowledge more accessible to physicians, care providers, and decision makers [2]. Conducting an SR is not an easy task because it must follow specific guidelines and protocols to ensure reproducibility of the methods. After the definition of review questions, researchers should accurately identify evidence from articles, studies, and any other relevant documentation. This selection process consists of an active search through online and offline literature repositories and the identification of evidence from a large amount of irrelevant information [3]. In the search phase, researchers use keyword combinations to create queries that are able to filter documentations in large medical databases. This operational step is prone to potential bias related to the source of information, specificity, and completeness of search strings. After application of queries, researchers manually complete the study selection process by a screening of titles, abstracts, and full texts and assess the papers' eligibility. Finally, they describe the process using a PRISMA flow diagram [4].

The increasing number of Web repositories and the development of new scientific topics make the SR process even more complex [5]. Researchers can retrieve information using search engines, such as PubMed or Embase, that

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest: None.

[1] The work was performed during an internship at the Unit of Biostatistics, Epidemiology, and Public Health, Department of Cardiac, Thoracic, and Vascular Sciences, University of Padova, Via Loredan 18, 35131 Padova, Italy.

* Corresponding author. Tel.: +39 049 8275403; fax: +39 02 700445089.

E-mail address: ileana.baldi@unipd.it (I. Baldi).

https://doi.org/10.1016/j.jclinepi.2018.06.015

What is new?

Key findings

- Our study proposes a new classifier that extends PubMed searches to clinical trial registries with high discrimination ability and sensitivity.

What this adds to what was known?

- Text mining and machine learning techniques can be used to support comprehensive systematic reviews.

What is the implication and what should change now?

- The proposed machine learning instrument can help researchers identify relevant studies in the systematic review process by reducing workload without losing sensitivity and at a small price in terms of specificity.

are organized in hierarchical branching structures (MeSH and Emtree, respectively), facilitating paper categorization and specific searches. This logical and hierarchical structure has important implications in the literature search process. First, it facilitates article retrieval by reducing or eliminating potential bias related to the differences in wording, language, and brand names. Second, even if not exhaustive, MeSH or Emtree structures are useful for limiting the number of records to the relevant ones, especially when the study topic is broad.

Despite their essential role in collecting and organizing published medical literature, indexed search engines are often unable to cover all relevant knowledge. A meta-analysis based only on their sources may provide biased estimates due to the exclusion of relevant unpublished information [6]. Furthermore, trial findings can influence the probability of publication and the presence of selective reporting outcomes [7]. The World Health Organization stated that unreported studies could lead to a misleading picture of the risks and benefits of a treatment, leading to the use and consumption of ineffective or harmful products [8]. For this reason, SRs should be based on a wide literature data set, which is essential for clinicians and patients to have a reliable and complete picture of their condition and make informed decisions. Among alternative informative sources, current literature recommends the inclusion of clinical trial registries such as ClinicalTrials.gov [9,10]. ClinicalTrials.gov is an international Web-based platform organized by the US National Library of Medicine providing access to more than 263,373 clinical trials from 202 countries. Studies are registered and regularly updated by the principal investigator, and records are never removed from the site. On ClinicalTrials.gov, clinicians and patients can retrieve complete information about the disease, intervention, study design and phase, location, and contacts, as well as the links to published papers. Some records also include the results of the study, such as the main characteristics of the population, incidence of adverse events, and collected outcomes. Clinical trial registries are important literary sources contributing to an updated evidence-based medical practice and may contain data that cannot be found in published papers [11]. It has been estimated that 50% of results reported in ClinicalTrials.gov were not initially available elsewhere, while some other information on serious adverse events was not always reported in the corresponding publication [12,13]. In a recent study, Baudard et al. found that adding clinical trial registries to the search base of SRs that did not originally search such registries identified an additional 122 trials for 41 SRs, which affected the strength of evidence of the SRs [10]. Despite their relevant role, clinical trial registries are seldom used as sources of studies for SRs, probably due to difficulties in record management. The main limitations are related to the absence of hierarchical order, poor interfaces, a limited number of synonyms, and the impossible combination of different queries. In ClinicalTrials.gov, the search strategy is based only on retrieval of one or more text words in the fields Condition/Disease, Title, Brief Description, Interventions, Locations, and Country. Text word variations include a limited number of synonyms, but no hierarchical order or subcategories are used. Recently, the Clinical Trials Transformation Initiative proposed a solution to improve the usability of data included in ClinicalTrials.gov by creating a database for aggregate analysis (AACT) and categorization of clinical trials based on clinical specialty. However, this classification is limited to the definition of Disease/Condition and is not consistent with the original MeSH classification, which does not allow for differentiation between clinical specialties.

This study aims to (1) provide an instrument based on text mining and machine learning (ML) techniques that can perform an automated literature search on clinical trial registries and (2) evaluate the usability and effectiveness of the proposed instrument. To reach our objectives, we present a case study based on results reported by Baudard et al. [10].

2. Materials and methods

2.1. Data sources

To create and test the instrument for automated literature search, we used two different data sources. First, we used information reported in the article "Impact of searching clinical trials registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses" [10]. That study identified additional trials not included in original SRs, through a manual search of the International Clinical Trials Registry Platform (ICTRP).

Specifically, the authors adapted and applied to the ICTRP the search strings of 14 SRs on the effectiveness of pharmacological treatments for several health conditions (i.e., atrial fibrillation, psoriasis, colorectal cancer, gastric cancer, Alzheimer's disease, Parkinson's disease, diabetes, rheumatoid arthritis, and hypertension). Then, they verified the consistency of the retrieved records with the inclusion criteria listed in the original papers and included relevant trials in a final estimation of treatment effectiveness. For our purpose, we used the same 14 SRs listed in the study by Baudard et al. [10]. This information allowed us to recreate search strings for PubMed and compare the results of automated searches with those reported by the authors of each SR. Second, we used the full database of ClinicalTrials.gov downloaded from the website of the Clinical Trials Transformation Initiative. The database was organized in pipe-delimited files with data on each single study, such as identifier (NCT number), location, start date, and sample size. Data could be reported as a number, string (i.e., text), date, or Boolean (i.e., true and false).

2.2. Training data sets

We created a training data set for each of the 14 SRs described previously. Each training data set included positive and negative records. Positive records were papers included in the original SRs, whereas negative records were a sample of papers off topic. Positive records were identified by running the original query in PubMed. When the search strategy did not allow us to retrieve all relevant papers, missing citations were manually included in the training set. On the other hand, negative records were retrieved by adding the Boolean operator "NOT" to the original query. In other words, we identified off-topic papers by subtracting records of the original search strategy from the complete PubMed database. Negative records were filtered by "Text availability: abstract," "Article types: Clinical trial," "Species: Humans," and "Languages: English". Because PubMed allows users to download up to 200 citations at a time, "Sort by: Best Match" option was selected to avoid any potential bias in the selection of papers based on Entrez Date. Then, negative records were downloaded in groups of 200 every time to achieve a ratio of at least 20 negative records to each positive one. The description of search strings and retrieved records is reported briefly in Table 1 (a more detailed description is reported as Supplementary Material Table S1). Finally, the first author, year, title, and abstract from each positive and negative paper were collected and included in the training set.

2.3. Testing data sets

A snapshot of the whole ClinicalTrials.gov was taken on January 5, 2017. This was composed of a set of pipe-delimited files from which we extracted the following information:

unique identifier (NCT number);
brief title;
official title;
brief summary;
detailed description;
study type (nature of investigation, such as interventional or observational);
overall recruitment status;
month and year of study start (enrollment of first participant);
month and year of primary completion (examination of final participant);
allocation;
number of arms;
study phase;

Table 1. Results of PubMed search strategies for the 14 systematic reviews included in Baudard et al.'s study [10]

Systematic review	Health condition	Positive records	Negative records
Yang et al. 2014 [14]	Atrial fibrillation	18	400
Meng et al. 2014 [15]	Psoriasis	9	200
Segelov et al. 2014 [16]	Colorectal cancer	13	400
Li et al. 2014 [17]	Gastric cancer	6	200
Lv et al. 2014 [18]	Colorectal cancer	12	400
Wang et al. 2015 [19]	Alzheimer's disease	32	800
Zhou et al. 2014 [20]	Parkinson's disease	9	200
Liu et al. 2014 [21]	Type 2 diabetes mellitus	23	600
Douxfils et al. 2014 [22]	Venous thromboembolic events	13	400
Kourbeti et al. 2014 [23]	Rheumatoid arthritis	75	1,600
Li et al. 2014 [24]	Primary hypertension	9	200
Cavender et al. 2014 [25]	Venous thromboembolic events	14	400
Chatterjee et al. 2014 [26]	Venous thromboembolic events	18	400
Funakoshi et al. 2014 [27]	Solid cancers	43	1,000

Final training data sets included the sum of positive and negative citations. Details of search strings are available in Table S1 (Supplementary Web Appendix).

minimum age for participant eligibility;

interventional study model (otherwise, the strategy for assigning interventions to participants);

inclusion of drug product subject to the US FDA (Federal Food, Drug, and Cosmetic Act).

We used the brief title, official title, and detailed description as textual information to perform our testing search. The other information was used to identify trials (NCT numbers) and to include filters similar to those applied in the study by Baudard et al. [10]. Specifically, Baudard et al. limited ICTRP results to clinical trials whose overall status was either completed or terminated. Moreover, we applied additional filters using fields consistently with the inclusion and exclusion criteria described in the 14 SRs, replicating the selection filters used in the study by Baudard et al. [10] and in the original SRs (see 2.5 Procedure Workflow for further details). Overall, 233,609 trials were finally included in the testing data set.

2.4. Text mining

The text mining strategy consisted of (1) text preprocessing, (2) training of the ML classifier, and (3) estimation of the performance of the classifier on the testing data set. We also considered an option to handle the unbalanced data in the training set. Text preprocessing steps converted the textual data into numbers. A support-vector machine (SVM) model, which is one of the most widely used classifiers for text mining [28], was chosen as the classifier and was trained using fivefold cross-validation. In each of the training data sets, the ratio of positive to negative samples was at least 1:20 by construction. Data of this type are known as unbalanced data. Hence, on the side of the straight application of the defined procedure, we also used the data handling strategy random undersampling (RUS), which randomly removes cases from the majority samples (in our case, the negative samples) to make the classes more balanced [29]. We applied the RUS strategy to obtain a final positive:negative ratio of the class samples of 35:65 according to Khoshgoftaar et al.'s study [30]. In this way, we had 28 data sets overall, two for each SR, that is, the original one and the one after the application of the RUS.

2.5. Procedure workflow

For each of the 14 SRs, the title and the abstract of the retrieved records were merged, and text preprocessing steps were applied in the following order: conversion to lowercase, removing nonwords, stemming words, stripping white space, and building the sequences of every two adjacent words from the original text (bigrams). Furthermore, a document-term matrix (DTM) was created with this collection of tokens (i.e., a unit of textual information), and the matrix was filled with a term frequency (TF) weighting scheme. The sparsity of all 14 DTMs was very high, ranging from 99% to 100%. The top 4% of the features were selected according to term frequency-inverse document frequency (TF-IDF) rank as a tribute to (a double application of) Pareto's rule, that is, that 80% of the effects come from 20% of the causes. These selected features were retained. The SVM was fivefold cross-validated, and within the cross-validation step, the balancing strategy of RUS and the 35:65 positive:negative ratio were applied. Next, reweighting with TF-IDF was applied.

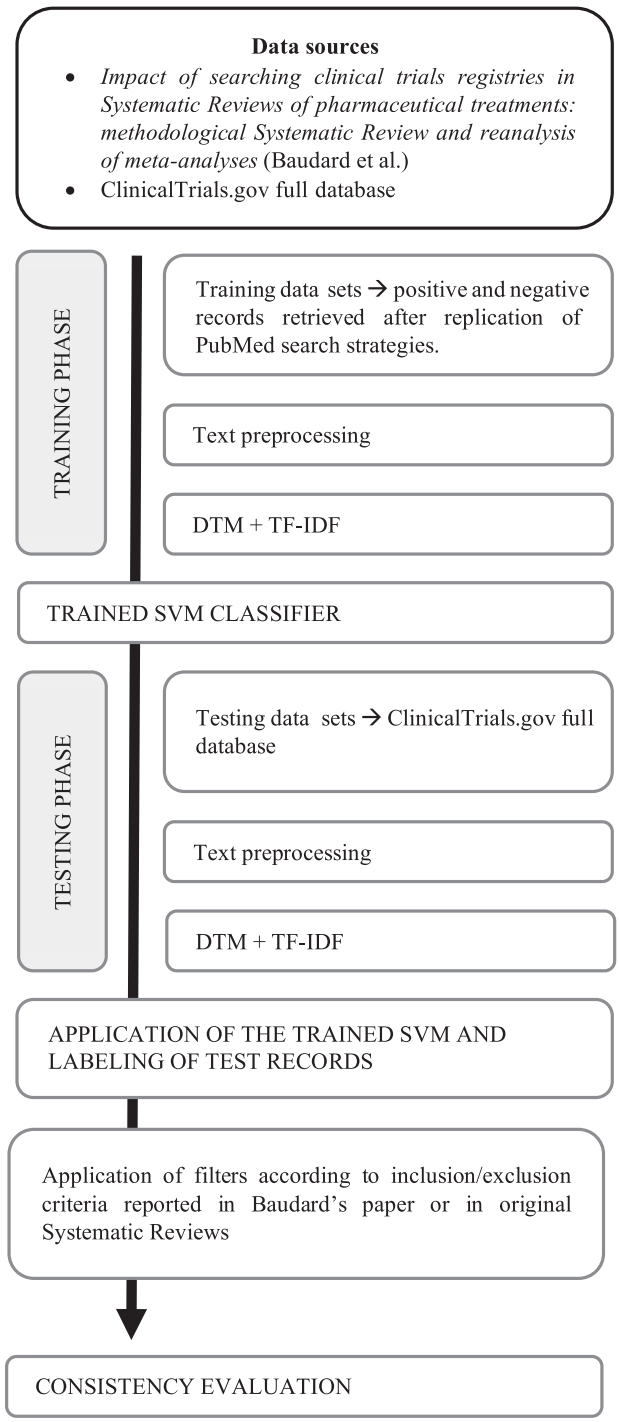

Fig. 1. General procedure workflow.

The testing ClinicalTrials.gov data set went through the same text preprocessing strategy in the same order, and then DTM was created with the TF weighting scheme initially. Furthermore, it was adapted with the same features retained from the training data set and was reweighted with the TF-IDF weighing scheme with the same retained IDF weights of the corresponding training DTM, which were retained when applied on the whole training data set.

Each cross-validated SVM model was applied to the corresponding testing data set for each SR. The procedure workflow is briefly described in Fig. 1. Analyses were carried out in R, version 3.4.2, [31] with the packages caret, tm, stringr, and unbalanced [32–35].

To compare the consistency between the manual search in the study by Baudard et al. [10] and this automated search, we replicated the selection filters used in the study by Baudard et al. [10] and in the original SRs. Thus, positive citations identified by automated search were limited by adding all the following filters: (1) recruitment status, defined as completed or terminated; (2) interventional design; (3) start date before the search on the ICTRP; (4) primary completion date before the search on the ICTRP as reported in the study by Baudard et al. [10]; and (5) specific filters based on inclusion criteria reported in the original SRs. The goodness and robustness of our results were evaluated by verifying the inclusion of the additional clinical trials previously identified by Baudard et al.

3. Results

The performance results of the most suitable filter are reported in Table 2. The sensitivity was 100% in all SRs except one (87.5%), and the specificity ranged from 97.2% to 99.9%. The area under the receiver operator characteristic curve, which measures the ability of the instrument to distinguish relevant articles from off-topic articles, ranged from 93.4% to 99.9%. The performance of the procedures in which an RUS strategy was implemented was similar (data are not shown). Table 3 reports the numbers of predicted positive citations before and after the application of a selection of filters. It also compares our results with the results of Baudard et al.'s manual search results on the ICTRP. As shown in the table, filters progressively reduced the number of citations (predicted positives), without excluding additional clinical trials identified in the study by Baudard et al. [10] (true positives).

The only false negative (1 of 8 positives) pertained to an SR on the role of biological therapy in metastatic colorectal cancer [16] and referred to the study with ClinicalTrials.gov identifier NCT00079066.

Notably, the total number of records from our automated search (predicted positives) was lower than the number of records from the manual search in half the cases, with a mean of 472 and a maximum of 2119 records compared with 572 and 2680, respectively, retrieved in the study by Baudard et al. [10].

4. Discussion

The time requirement and the need for the involvement of different professionals make an SR a very labor-intensive process [36]. The quality of the results depends on the extent to which the identified literature is accurate and comprehensive of all available knowledge on a specific

Table 2. Number of training (PubMed) and testing (ClinicalTrials.gov) positive and negative records as well as the number of predicted positives and the relevant statistics for each systematic review considered

Systematic review	Training positives	Training negatives	Testing positives	Testing negatives	Predicted positives	AUC	PPV	SENS	SPEC	LR+	LR−
Yang et al. 2014 [14]	18	400	5	233,604	1,718	0.9963	0.0029	1	0.9927	136.9863	0
Meng et al. 2014 [15]	9	200	4	233,605	462	0.9990	0.0087	1	0.9980	500.0000	0
Segelov et al. 2014 [16]	13	400	8	233,601	1,595	0.9341	0.0044	0.875	0.9932	128.6765	0.1259
Li et al. 2014 [17]	6	200	3	233,606	1,635	0.9965	0.0018	1	0.9930	142.8571	0
Lv et al. 2014 [18]	12	400	3	233,606	1,429	0.9969	0.0021	1	0.9939	163.9344	0
Wang et al. 2015 [19]	32	800	5	233,604	1,901	0.9959	0.0026	1	0.9919	123.4568	0
Zhou et al. 2014 [20]	9	200	3	233,606	1,011	0.9978	0.0030	1	0.9957	232.5581	0
Liu et al. 2014 [21]	23	600	30	233,579	2,178	0.9954	0.0138	1	0.9908	108.6957	0
Douxfils et al. 2014 [22]	13	400	10	233,599	378	0.9992	0.0265	1	0.9984	625.0000	0
Kourbeti et al. 2014 [23]	75	1,600	25	233,584	1,843	0.9961	0.0136	1	0.9922	128.2051	0
Li et al. 2014 [24]	9	200	2	233,607	6,558	0.9860	0.0003	1	0.9719	35.5872	0
Cavender et al. 2014 [25]	14	400	7	233,602	149	0.9997	0.0470	1	0.9994	1,666.6667	0
Chatterjee et al. 2014 [26]	18	400	17	233,592	771	0.9984	0.0220	1	0.9968	312.5000	0
Funakoshi et al. 2014 [27]	43	1,000	11	233,598	3,851	0.9918	0.0029	1	0.9836	60.9756	0

Abbreviations: AUC, area under the receiver operator characteristic curve; LR+, positive likelihood ratio; LR−, negative likelihood ratio; PREV, prevalence of positive in ClinicalTrials.gov; PPV, positive predictive value; SENS, sensitivity; SPEC, specificity.

topic. Also, the reliability of an SR is determined by the inclusion of up-to-date contents [37]. Our study proposes a classifier that can extend PubMed searches to clinical trials registries. This tool reduces the effort and time expenditure of an SR without losing accuracy and sensitivity.

Other researchers have highlighted how ML could make the standard SR process more efficient [38]. They focused on a ''living'' SR, considering as the starting point the existence of an initial SR provided by humans. Accordingly, we have provided an instrument that is also usable for the ''living'' step of updating an SR data set, but it is specially tailored to the more complex and tricky step of contributing to the base data set definition/extraction for new sources of data (work left to humans in the study by Thomas et al. [38]). Our procedure showed high performance in detecting true-positive citations of interest in completely different sources of data from the original one regarding the way meta-data are stored, the way information is accessed, and the structure of the information. It left out only 1 of 133 human-detected positive citations from 14 independent SRs. From this starting point, we have also highlighted how, with simple and quick filtering, the number of false positives can be easily and drastically reduced without affecting the sensitivity of the procedure. In this way, the work left to humans can be reduced and quite limited on the first run of the ''living'' update of the SR, that is, the part of data set definition that was completely based on human effort until now.

Other studies have shown how an ML approach for the classification of information based on clinical text could be very effective [39], including when tested on databases different (and not subsampled) from the original one [40]. On the other hand, to our knowledge, no other study was conducted on this wide range of differentiated data sets with hundreds of thousands of entries.

The strict procedure that we followed allowed all the test sets to be blinded both from the training ones at every stage and from all the training procedures, making us confident in the quality of the results themselves. In an SR, both very specific positive and very specific negative sets can be selected to create a high-quality training set. This characteristic together with the ability of the SVMs to distinguish the well-separated type of data and the high proportion of (few) positive records against a huge number of negative ones have led to the nearly perfect results in sensitivity, which is the main characteristic of interest in this endeavor.

Our study demonstrates the usefulness of ML when scientific literature is not reported in indexed search engines. This is the case of clinical trial registries such as ClinicalTrials.gov, whose interfaces are usually not sophisticated. Their limited functionality has an important impact on process workload and often requires the application of long search strings, multiple searches, and the screening of a high number of nonspecific records. Moreover, when a researcher wants to use the same query on different search engines and registries, he must adapt each singular term and string according to the specific requirement of each platform. In the case of registries, an adaptation from common search engines (PubMed or Embase) is even more complex because of the frequent absence of text functionalities such as truncation or brackets. The use of ML could allow a more accurate and easier translation of queries by reducing the number of irrelevant records.

The main strength of the study is the robustness of the training and testing procedure, which was designed to be stable and unbiased. Furthermore, an R package and a companion GUI are under development (preliminary version publicly available at https://github.com/UBESP-DCTV/costumer). They are intended to be a user-friendly tool for health care researchers, who will only have to provide (a) the set of citations finally retained, (b) a personalized set of negative citations or the search string used on PubMed (to automatically identify a suitable set of random negative citations), and (c) an optional set of false positives already known from a previous run or directly the set of filters to be applied on nontextual meta-data. The first part of feature (c) highlights the usability of the package for a very quick update of the SR, for example, after the first run (for which the false positives must be manually identified).

Our study has some limitations. First, we adopted a defined ML algorithm and used only one strategy for managing the unbalanced data. We acknowledge that other techniques, such as convolutional neural networks, are effective at achieving slightly better F-scores [41] over more traditional approaches to biomedical text classification, such as SVM, especially when there is significant label imbalance. Nevertheless, convolutional neural networks typically take at least an order of magnitude more time than traditional classifiers, especially compared with SVM [42]. Hence, we decided to start our investigation by considering SVM only. We are already working on testing both a wider range of ML techniques and more methods for unbalanced data sets. Nevertheless, the performance with the choice adopted in terms of the number of positives, number of true positives, and number of negatives, as well as in terms of computational speed, is already good, and we do not expect more improvement, although small relative increases in specificity can still have a big impact on absolute numbers of false positives. Moreover, filters were manually applied after automatic searches and were not yet included in this ML instrument. The reason for this choice was that inclusion/exclusion criteria are rarely reported in the title, abstract, or description. Thus, it was not possible to make a more accurate automatic selection of trials. That said, similar studies were able to reach a very high level of sensitivity at the cost of a discrete specificity [43]. Explanations for our results lie in the choice of the training set and of the task itself, that is, positives and negatives were highly

Table 3. The number of predicted positives and true positives in manual and automated searches after filter application

	Manual search		Automated search					
			Filter 0		Filter 1		Filter 2	
Systematic review	Predicted positives	True positives	None	Predicted positives	Study_type	Predicted positives	Overall_status	Predicted positives
Yang et al. 2014 [14]	12	1	-	1,718	interventional	1,341	completed OR terminated	759
Meng et al. 2014 [15]	26	1	-	462	interventional	399	completed OR terminated	282
Segelov et al. 2014 [16]	684	2	-	1,595	interventional	1,432	completed OR terminated	836
Li et al. 2014 [17]	201	1	-	1,635	interventional	1,545	completed OR terminated	376
Lv et al. 2014 [18]	665	1	-	1,429	interventional	1,294	completed OR terminated	727
Wang et al. 2015 [19]	227	1	-	1,901	interventional	1,690	completed OR terminated	1,191
Zhou et al. 2014 [20]	3	1	-	1,011	interventional	793	completed OR terminated	534
Liu et al. 2014 [21]	1,661	21	-	2,178	interventional	2,028	completed OR terminated	1,587
Douxfils et al. 2014 [22]	76	1	-	378	interventional	270	completed OR terminated	190
Kourbeti et al. 2014 [23]	581	4	-	1,843	interventional	1,449	completed OR terminated	1,023
Li et al. 2014 [24]	909	2	-	6,558	interventional	5,483	completed OR terminated	3,629
Cavender et al. 2014 [25]	71	1	-	149	interventional	130	completed OR terminated	87
Chatterjee et al. 2014 [26]	217	1	-	771	interventional	509	completed OR terminated	279
Funakoshi et al. 2014 [27]	2,680	2	-	3,851	interventional	3,762	completed OR terminated	2,147

Records of the manual search are those retrieved on the International Clinical Trials Registry Platform by Baudard et al. [10]. Records of the automated search are those retrieved on ClinicalTrials.gov using our machine learning instrument. Predicted positives are a pool of citations resulting from manual search strings or from automated searches. True positives are clinical trials added by Baudard et al. to each systematic review. The description of filters reports data element entries and the number of retrieved records. Filters were applied sequentially from filter 0 to filter 5.

separated by design: the set of positives is the (human-filtered) output of a PubMed query string and not of a ''rule-free'' human selection from the whole of PubMed, as was done, for example, for the classification task performed in the study by Marshall et al. [43]. The same applies to the negatives, which were sampled from the output of the negative search with the PubMed query string used for positives. As a result, positives shared a similar word characterization, which is easily identified by SVM and can lead to a near-perfect sensitivity and an excellent specificity.

Following the recommended paradigm for model validation [44,45], this predictive tool underwent internal validation through cross-validation and external validation on an independent data source. This aspect, in conjunction with the broad range of health conditions analyzed, strongly argues in favor of the credibility of the proposed instrument.

Table 3. Continued

Automated search						
Filter 3		Filter 4		Filter 5		All
Start_before	Predicted positives	Primary_completion before/within	Predicted positives	Specific filters	Predicted positives	True positives
April 2014	705	April 2014	588	allocation = randomized number_of_arms ≠ 1	457	1
August 2013	243	August 2013	202	allocation = randomized number_of_arms ≠ 1	144	1
May 2012	770	May 2012	588	allocation = randomized number_of_arms ≠ 1	274	2
February 2014	837	February 2014	695	allocation = randomized number_of_arms ≠ 1 phase = 2 OR 3	289	1
February 2014	716	February 2014	583	allocation = randomized number_of_arms ≠ 1 minimum_age ≥ 18 years	243	1
December 2013	1,118	December 2013	972	allocation = randomized number_of_arms ≠ 1 intervention_model = parallel OR crossover	729	1
February 2013	468	February 2013	372	allocation = randomized number_of_arms ≠ 1	263	1
January 2013	1,317	January 2013	1,112	allocation = randomized number_of_arms ≠ 1 minimum_age ≥ 18 years phase ≥ 3	622	21
December 2013	174	December 2013	150	allocation = randomized number_of_arms ≠ 1	116	1
June 2013	941	June 2013	756	allocation = randomized number_of_arms ≠ 1 is_fda_regulated = true	409	4
January 2014	3,412	January 2014	2,771	allocation = randomized number_of_arms ≠ 1	2,119	2
April 2014	84	April 2014	74	allocation = randomized number_of_arms ≠ 1	60	1
March 2014	263	January 2001–March 2014	207	allocation = randomized number_of_arms ≠ 1	169	1
February 2014	2,111	January 2004–February 2014	1,699	allocation = randomized number_of_arms ≠ 1 phase = 2 OR 3	711	2

Acknowledgments

Authors' contributions: I.B. and P.B. designed the study; C.M. managed the systematic reviews; C.L. and A.S. performed the analysis; C.M. and A.S. wrote the manuscript; and all authors contributed to result interpretation and approved the final manuscript.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.06.015.

References

[1] Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, et al. Explanation of the 2011 Oxford centre for evidence-based medicine (OCEBM) levels of evidence (background document) 2016. Available at: https://www.cebm.net/2016/05/ocebm-levels-of-evidence/. Accessed January 13, 2018.

[2] CRD's guidance for undertaking reviews in health care 2009. Available at: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf. Accessed July 23, 2018.

[3] Hirschman L, Burns GAPC, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database (Oxford) 2012;2012:bas020.

[4] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic

reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6:e1000100.
[5] Balan PF, Gerits A, Vanduffel W. A practical application of text mining to literature on cognitive rehabilitation and enhancement through neurostimulation. Front Syst Neurosci 2014;8:182.
[6] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence−publication bias. J Clin Epidemiol 2011;64:1277−82.
[7] Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009;MR000006.
[8] WHO. Major research funders and international NGOs to implement WHO standards on reporting clinical trial results 2017. Available at: http://www.who.int/mediacentre/news/releases/2017/clinical-trial-results/en/. Accessed January 13, 2018.
[9] Hughes S, Cohen D, Jaggi R. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. BMJ Open 2014;4:e005535.
[10] Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ 2017;356:j448.
[11] Halfpenny NJ, Thompson JC, Quigley JM, Scott DA. Clinical trials registries for systematic reviews − an alternative source for unpublished data. Value Health 2015;18:A12.
[12] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database — update and key issues. N Engl J Med 2011;364:852−60.
[13] Tang E, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles. BMC Med 2015;13:189.
[14] Yang Q, Qi X, Li Y. The preventive effect of atorvastatin on atrial fibrillation: a meta-analysis of randomized controlled trials. BMC Cardiovasc Disord 2014;14:99.
[15] Meng Y, Dongmei L, Yanbin P, Jinju F, Meile T, Binzhu L, et al. Systematic review and meta-analysis of ustekinumab for moderate to severe psoriasis. Clin Exp Dermatol 2014;39:696−707.
[16] Segelov E, Chan D, Shapiro J, Price TJ, Karapetis CS, Tebbutt NC, et al. The role of biological therapy in metastatic colorectal cancer after first-line treatment: a meta-analysis of randomised trials. Br J Cancer 2014;111:1122−31.
[17] Li D-H, Pan Z-K, Ye F, An H-X, Wu J-X. S-1-based versus 5-FU-based chemotherapy as first-line treatment in advanced gastric cancer: a meta-analysis of randomized controlled trials. Tumour Biol 2014;35:8201−8.
[18] Lv Z-C, Ning J-Y, Chen H-B. Efficacy and toxicity of adding cetuximab to chemotherapy in the treatment of metastatic colorectal cancer: a meta-analysis from 12 randomized controlled trials. Tumour Biol 2014;35:11741−50.
[19] Wang J, Yu J-T, Wang H-F, Meng X-F, Wang C, Tan C-C, et al. Pharmacological treatment of neuropsychiatric symptoms in Alzheimer's disease: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2015;86:101−9.
[20] Zhou C-Q, Zhang J-W, Wang M, Peng G-G. Meta-analysis of the efficacy and safety of long-acting non-ergot dopamine agonists in Parkinson's disease. J Clin Neurosci 2014;21:1094−101.
[21] Liu X, Xiao Q, Zhang L, Yang Q, Liu X, Xu L, et al. The long-term efficacy and safety of DPP-IV inhibitors monotherapy and in combination with metformin in 18,980 patients with type-2 diabetes mellitus−a meta-analysis. Pharmacoepidemiol Drug Saf 2014;23:687−98.
[22] Douxfils J, Buckinx F, Mullier F, Minet V, Rabenda V, Reginster J-Y, et al. Dabigatran etexilate and risk of myocardial infarction, other cardiovascular events, major bleeding, and all-cause mortality: a systematic review and meta-analysis of randomized controlled trials. J Am Heart Assoc 2014;3:e000515.
[23] Kourbeti IS, Ziakas PD, Mylonakis E. Biologic therapies in rheumatoid arthritis and the risk of opportunistic infections: a meta-analysis. Clin Infect Dis 2014;58:1649−57.
[24] Li ECK, Heran BS, Wright JM. Angiotensin converting enzyme (ACE) inhibitors versus angiotensin receptor blockers for primary hypertension. Cochrane Database Syst Rev 2014;CD009096.
[25] Cavender MA, Sabatine MS. Bivalirudin versus heparin in patients planned for percutaneous coronary intervention: a meta-analysis of randomised controlled trials. Lancet Lond Engl 2014;384:599−606.
[26] Chatterjee S, Sardar P, Giri JS, Ghosh J, Mukherjee D. Treatment discontinuations with new oral agents for long-term anticoagulation: insights from a meta-analysis of 18 randomized trials including 101,801 patients. Mayo Clin Proc 2014;89:896−907.
[27] Funakoshi T, Latif A, Galsky MD. Safety and efficacy of addition of VEGFR and EGFR-family oral small-molecule tyrosine kinase inhibitors to cytotoxic chemotherapy in solid cancers: a systematic review and meta-analysis of randomized controlled trials. Cancer Treat Rev 2014;40:636−47.
[28] Jindal R, Malhotra R, Jain A. Techniques for text classification: literature review and current trends. Webology 2015;12:1.
[29] Liu AC. The effect of oversampling and undersampling on classifying imbalanced text datasets. Austin: The University of Texas at Austin; 2004.
[30] Khoshgoftaar TM, Seiffert C, Van Hulse J, Napolitano A, Folleco A. Learning with limited minority class data. Mach. Learn. Appl. 2007 ICMLA 2007 Sixth Int. Conf. On, IEEE 2007;:348−53.
[31] R Development Core Team. R: a language and environment for statistical computing 2008: Vienna, Austria: R Foundation for Statistical Computing.
[32] Wing J, Kuhn M, Weston S, Williams A, Keefer C, Engelhardt A, et al. caret: Classification and regression training 2017. Available at: https://cran.r-project.org/web/packages/tm/tm.pdf. Accessed July 23, 2018.
[33] Feinerer I, Hornik K. tm: Text mining package 2017. Available at: https://cran.r-project.org/web/packages/tm/tm.pdf. Accessed July 23, 2018.
[34] Wickham H. stringr: Simple, consistent wrappers for common string operations 2017. Available at: https://cran.r-project.org/web/packages/stringr/stringr.pdf. Accessed July 23, 2018.
[35] Pozzolo AD, Caelen O, Bontempi G. unbalanced: Racing for unbalanced methods selection 2015. Available at: https://cran.r-project.org/web/packages/unbalanced/unbalanced.pdf. Accessed July 23, 2018.
[36] Khabsa M, Elmagarmid A, Ilyas I, Hammady H, Ouzzani M. Learning to identify relevant studies for systematic reviews using random forest and external information. Mach Learn 2016;102:465−82.
[37] CRD's guidance for undertaking reviews in health care. Centre for reviews and dissemination. Austin: The University of Texas at Austin; 2009.
[38] Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 2017;91:31−7.
[39] Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc 2015;22:155−65.
[40] Connolly B, Matykiewicz P, Bretonnel Cohen K, Standridge SM, Glauser TA, Dlugos DJ, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. J Am Med Inform Assoc 2014;21:866−70.
[41] Rios A, Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. ACM-BCB 2015;2015:258−67.
[42] Majumder S, Balaji N, Brey K, Fu W, Menzies T. 500+ Times faster than deep learning (a case study exploring faster methods for text mining stackoverflow). Conference proceeding arXiv preprint arXiv:1802.05319 2018. Available at: https://arxiv.org/pdf/1802.05319.pdf. Accessed July 23, 2018.
[43] Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Res Synth Methods 2018;. https://doi.org/10.1002/jrsm.1287.
[44] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245−7.
[45] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453−73.