Journal of Clinical Epidemiology 108 (2019) 86−94

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses

Ivan Lerner[a,b,c], Perrine Créquit[a,b,c,d], Philippe Ravaud[a,b,c,d,e], Ignacio Atal[a,b,c,*]

[a]Centre de Recherche Epidémiologie et Statistique Paris Sorbonne Cité, INSERM U1153, Paris, France
[b]Université Paris Descartes − Sorbonne Paris cité, Paris, France
[c]Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, Centre d'Epidémiologie Clinique, Paris, France
[d]Cochrane France, Paris, France
[e]Department of Epidemiology, Mailman School of Public Health, Columbia University New York, USA

Accepted 4 December 2018; Published online 7 December 2018

Abstract

Objectives: We aimed to develop and evaluate an algorithm for automatically screening citations when updating living network meta-analysis (NMA).

Study Design and Setting: Our algorithm learns from the initial screening of citations conducted when creating an NMA to automatically identify eligible citations (i.e., needing full-text consideration) when updating the NMA. We evaluated our algorithm on four NMAs from different medical domains. For each NMA we constructed sets of initially screened citations and citations to screen during an update that took place 2 years after the conduct of the NMA. We encoded free text of citations (title and abstract) using word embeddings. On top of this vectorized representation, we fitted a logistic regression model to the set of initially screened citations to predict the eligibility of citations screened during an update.

Results: Our algorithm achieved 100% sensitivity on two NMAs (100% [95% confidence interval 93−100] and 100% [40−100] sensitivity), and 94% (81−99) and 97% (86−100) on the remaining two others. For all NMAs, our algorithm would have spared to manually screen 1,345 of 2,530 citations, decreasing the workload by 53% (51−55), while missing 3 of 124 eligible citations (2% [1−7]), none of which were finally included in the NMAs after full-text consideration.

Conclusion: For updating an NMA after 2 years, our algorithm considerably diminished the workload required for screening, and the number of missed eligible citations remained low.

Keywords: Automatic screening; Network meta-analysis; Live cumulative network meta-analysis; Machine learning; Natural language processing; Word embeddings

1. Background

Systematic reviews (SRs) are the core of evidence synthesis in biomedical research. They are based on a comprehensive search strategy that aims to collect an exhaustive set of studies for a given medical question. Often, multiple competing treatments are available for a given medical condition; however, SRs only provide a fragmented panorama of the evidence for all treatments [1]. Network meta-analyses (NMAs) [2] provide part of the solution by allowing for simultaneous comparison of multiple treatments for a given condition.

In addition, the evidence synthesis needs to be updated regularly to maintain clinically relevant results. Indeed, half of SRs are published more than 14 months after the last search date [3]; therefore, 7% of reviews are out-of-date by the time they are published [4]. In addition, less than

Conflict of interest statement: The authors declare that they have no competing interests.

Funding: This work was partially funded by the grant N°2016-02/058/AB-KA from the Institut National du Cancer (INCa).

Authors' contributions: I.L. contributed to study design, data preprocessing and analysis, results interpretation, and writing. I.A. contributed to study design, results interpretation, and writing. P.C. contributed to study design, results interpretation, and writing. P.R. contributed to study design and results interpretation. All authors read and approved the final manuscript.

Availability of data and material: The datasets generated and analyzed during the present study are available in a git repository, https://gitlab.com/lerner.ivan/automatic_screening_NMA.

* Corresponding author. Centre d'Epidémiologie Clinique, Hôpital Hôtel-Dieu, 1, place du parvis Notre Dame, Paris, France 75004. Tel.: +33-1-42-34-87-65; fax: +331 42 34 87 90.

E-mail address: ignacio.atal-ext@aphp.fr (I. Atal).

https://doi.org/10.1016/j.jclinepi.2018.12.001

What is new?

Key findings

- Using data from four network meta-analyses we showed that automatic screening can successfully be applied for updating network meta-analysis (NMA), considerably diminishing the workload without missing any finally included citations.

What this adds to what was known?

- We showed that representing citations using word embeddings, a numerical representation of words based on the idea that words with similar meaning occur in similar contexts, improved significantly the prediction of eligible citations when updating NMAs.

What is the implication and what should change now?

- The integration of automatic screening for the systematic updating of living NMAs should be considered. It will contribute to the comparative effectiveness research objective and provide a complete and up-to-date overview of the evidence.

half of SRs are updated [5]. The Cochrane handbook for SRs suggests that SRs should be updated every 2 years [6]; however, updating SRs is challenging because of the increasing number of publications [7]. A recently developed type of NMA, live cumulative NMA [8], also called living NMA, aims at being a unique access point to an up-to-date overview of all existing evidence on all available treatments for a precise health condition. Living NMAs are based on large and exhaustive searches of a wide panel of databases and frequent updates.

An SR is based on a search for citations and two screening stages. First, citations (i.e., titles and abstracts) are retrieved from electronic databases such as MEDLINE using search equations. Second, these citations are manually screened to select eligible citations. Finally, full texts for all eligible citations are retrieved and manually screened to select included citations. The screening process is one of the most time-consuming tasks when conducting SRs [9] and thus an important barrier to updating the synthesis of evidence.

Efforts for automated screening based on machine learning have been developed in recent years [10–12]. The automation of screening may save a large amount of work but may lose accuracy compared with human updating. Machine-learning techniques applied to automatic screening were suggested to save 30–78% of the workload but miss 4–5% of relevant studies [12,13].

Automation of screening relies on automatic analysis of free text. In natural language processing, word embeddings [14] were designed to overcome the limitations of the basic representation of words. Classically, words are represented according to their position in the list of all words mentioned in the corpus, without notion of distance between words. Conversely, word embeddings were conceived to provide numerically close representations of words that are semantically and syntactically close based on the context in which they appear. For example, the words ''bronchoscopy'' and ''cystoscopy'' will be represented by close numerical vectors because they share similar contexts, such as ''the patient underwent bronchoscopy/cystoscopy before the operation.'' Word embeddings have been found useful in tasks such as topic modeling [15] and feature extraction for classification of text using machine learning [16,17].

2. Objective

We aimed to develop and evaluate an automated screening algorithm for updating NMAs of randomized controlled trials using vectorized representation of text based on word embeddings and machine learning.

3. Materials and methods

Our algorithm learns from the initial screening of citations conducted when creating an NMA to automatically identify eligible citations when updating the NMA. We replicated the initial and update screening phases for four NMAs from different medical fields. For each NMA we constructed sets of initially screened citations and sets of citations to screen for an update 2 years after the initial screening. We then built an automatic screening algorithm that learned to discriminate between eligible and ineligible citations based on the sets of initially screened citations, separately for each NMA. Finally, we evaluated the performance of the algorithm over each set of citations to screen for the update. Fig. 1 summarizes the different stages of the workflow and represents the inputs and outputs of the system.

3.1. Data on screening process

We used data from four NMAs [1,18–20] in the fields of pneumology, urology, oncology, and psychiatry, with more than 1,000 screened citations each. For each NMA, we disposed of the search equations, the titles of eligible citations after title and abstract screening, and the titles of finally included citations after full-text screening. We used the search equations to newly search electronic databases (MEDLINE, EMBASE, CENTRAL, and PsycINFO) to retrieve all screened citations. As the last date of search, we used December 31 of the year preceding the actual last date of search for the NMA to have all citations published within each year. We replicated updates of NMAs by artificially introducing a cut-off time separating citations by

publication year. For each NMA we constructed sets of initially screened citations and sets of citations to screen if an update was conducted 2 years after the initial screening. For example, *Khoo* et al. originally included citations until June 1, 2015: we considered citations published between January 1, 2013, and December 31, 2014, as the set of citations to screen if an update was conducted (test set), and all citations published before December 31, 2012, as the set of initially screened citations (training set).

3.2. Automatic screening

To automate the screening process, we trained a machine-learning algorithm to classify eligible and ineligible citations after title and abstract screening (Fig. 1). We represented free text of citations (title and abstract) using word embeddings. We compared the performances of classification with a baseline in which free text in citations was represented using a term frequency–inverse document frequency (tf–idf) matrix.

3.2.1. Citation representation based on word embeddings

For each citation, we represented the title and abstract using embedded word vectors [21], whereby each word was encoded into a 200-dimensional numerical vector. We used word vectors from a previous study [22] that

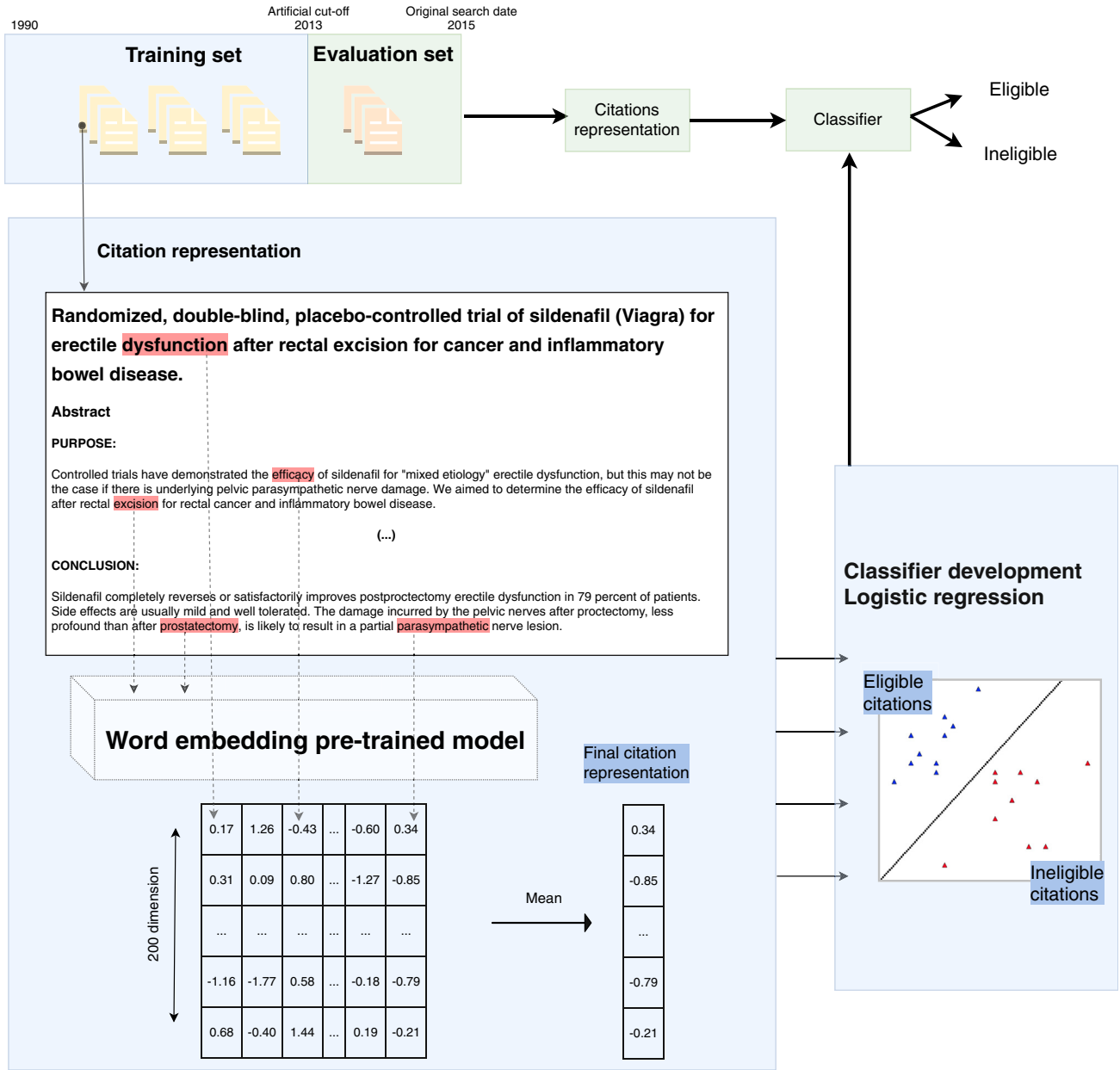

Fig. 1. Workflow of automatic screening using word embeddings. Summary of the different stages of the workflow and detailed representation of the inputs and outputs of the system.

trained a Skip-gram model [21] over all the available biomedical literature from PubMed and PubMed Central until 2013, enriched for common words with a Wikipedia corpus. For each NMA as a corpus, the 30 most frequent words and words appearing less than five times across all citations were not encoded nor were words not corresponding to pretrained word vectors, which included stop words. We then represented each citation using the average of its word vectors. For each NMA, we applied principal component analysis (PCA) to the vectorized representation of screened citations to visualize eligible and ineligible citations in a two-dimensional plot.

3.2.2. Citation representation based on tf–idf as a baseline

For each NMA as a corpus, we excluded the 30 most frequent words and words appearing less than five times across all citations, as well as common English stop words. We tokenized text and applied the Porter Stemmer Algorithm to reduce inflected or derived words to their stem. We then vectorized citations based on *tf–idf*.

3.2.3. Classifier

For each NMA we fitted a logistic regression model with L2 regularization to the set of initially screened citations to predict their eligibility after screening according to their vectorized representation. Each fitted model was then used to automatically identify eligible citations in the set of citations to screen during the update. Models were fitted using the stochastic gradient descent algorithm with exponential decay. We used a weighted loss function along with oversampling of eligible citations at a 1:1 ratio during training to cope with class imbalance. The weighted loss function penalized more classification error of eligible citations than those of noneligible citations. We searched for optimal hyperparameters on development sets that were built by sampling 20% of the set of initially screened citations. The hyperparameters optimized included the learning rate, the regularization term, and the positive weight. We selected the models with the best sensitivity, and if models had equal sensitivity, we selected those with the best specificity.

3.2.4. Evaluation

We assessed the performance of the algorithm to accurately classify eligible and ineligible citations in the sets of citations to screen during an update. Performance was measured in terms of sensitivity, specificity, missed studies, and workload saving, overall and for each NMA. Sensitivity corresponded to the ratio of the number of correctly labeled eligible citations to the total number of eligible citations. Specificity corresponded to the ratio of the number of correctly labeled ineligible citations to the total number of ineligible citations. Missed studies corresponded to the ratio of the number of inaccurately labeled eligible citations to the total number of eligible citations. Workload saving corresponded to the ratio of the number of correctly labeled ineligible citations to the total number of citations. We assessed whether eligible citations that were misclassified by the algorithm were finally included in the NMA.

3.2.5. Sensitivity analysis

We assessed the robustness of our results by repeating the analysis with earlier cut-off time—3 years and 4 years—for separating sets of initially screened citations and sets of citations to screen for an update. In this regime less eligible and noneligible citations were available for training the algorithm.

3.3. Implementation

Algorithms were implemented in Python using TensorFlow [23] and scikit-learn [24]. The code and dataset are available on open source at https://gitlab.com/lerner.ivan/automatic_screening_NMA. The code for analysis is available as one Jupyter Notebook in our GitHub repository (https://gitlab.com/lerner.ivan/automatic_screening_NMA/blob/master/sysReviewFromVectorized/scan_save_eval.ipynb).

3.4. Statistical analysis

Descriptive data are presented with number (percentage) and 95% confidence intervals calculated by the Clopper–Pearson method using the statsmodels [25] library in Python. We assessed the statistical significance of the difference in sensitivity and specificity between word embeddings and baseline (*tf–idf* representation) by calculating Fisher's exact test.

4. Results

4.1. Screening process

Our study included four NMAs in different fields of medicine (Table 1), which altogether totaled 14,853 screened citations. We present in Fig. 2 the evolution over time of the number of eligible and ineligible studies for each NMA. The NMAs presented diverse paces of publications or number of eligible citations published during the year. The Bateman et al., Chen et al., and Créquit et al. studies each showed a peak in pace of publication, with more than 10 eligible citations published each year during the peak. The time between this peak and the last date of search varied across NMAs. Conversely, the pace of publication of the Khoo et al. was more stable over time. For the Bateman et al. and Chen et al. studies, the artificial cut-off times introduced (in January 2011 and 2010, respectively) to create the sets of initially screened citations and citations to screen for a 2-year update took place at the end of an intense publication cycle. The cut-off introduced in 2013 for Créquit et al. took place in the middle of an intense publication cycle.

Table 1. NMA characteristics after replicating the search equations

First author	Field	Databases	Total screened	Eligible	Proportion of eligible citations (%)	Last date of search
Bateman et al., 2015	Pneumology	MEDLINE, EMBASE	4,219	400	9	12/31/2012
Chen et al., 2015	Urology	MEDLINE, EMBASE	1,662	256	15	12/31/2011
Créquit et al., 2016	Oncology	MEDLINE, EMBASE, CENTRAL	3,373	113	3	12/31/2014
Khoo et al., 2015	Psychiatry	MEDLINE, EMBASE, PsycINFO	5,599	75	1	31/12/2014

Abbreviation: NAM, network meta-analysis.

For each NMA, we retrieved citations from electronic databases with the original search equations, and we identified eligible citations using data we disposed from the original screening process. We present for each NMAs the electronic databases, the total number of citations, the number of eligible citations, the ratio of the number eligible to total citations, and the last date of search.

4.2. Automatic screening

4.2.1. Citation representation

The median length of citations (i.e., titles and abstracts) was 294 words, which for all citations totaled 2,341,517 words. The size of the vocabulary was 18,669 (Bateman et al.), 15,935 (Créquit et al.), 11,535 (Chen et al.), and 18,821 words (Khoo et al.). The proportion of vectorized words with word embeddings was 84% (Bateman et al.), 82% (Créquit et al.), 92% (Chen et al.), and 82% (Khoo et al.). Eligible citations after vectorized representation using word embeddings and dimensionality reduction with PCA seemed to be spatially close (Fig. 3). Although PCA in two dimensions explained only 32–38% of the variability, citations were partially separated between eligible and ineligible by the encoding scheme only.

4.2.2. Classifier evaluation

For two of the four NMAs, logistic regression on top of a word embeddings representation achieved 100% sensitivity. For Créquit et al. and Khoo et al., it achieved 100% (95% confidence interval 93–100) and 100% (40–100) sensitivity, and 58% (54–62) and 78% (75–81) specificity. For Chen et al., it achieved 94% (81–99) sensitivity and 59% (52–66) specificity, missing two eligible citations, none of which were finally included in the NMA after full-text consideration. For Bateman et al., it achieved 97% (86–100) sensitivity and 33% (30–36) specificity, missing one eligible citation, which was not finally included in the NMA after full-text consideration. For three of four NMAs, using word embeddings representation was significantly superior to tf–idf in terms of specificity ($P < 0.05$), and for all NMAs, using word embeddings seemed to be superior to tf–idf in terms of sensitivity although the differences were not statistically significant (Table 2). We expected the sensitivity to be systematically high because all models were developed to have high sensitivity regardless of the text representation. For all NMAs, our algorithm would have spared screening manually 1,345 of 2,530 citations, decreasing the workload by 53% (51–55), while missing 3 of 124 eligible citations (2% [1–7]).

4.2.3. Sensitivity analysis

The algorithm had similar performances when trained to predict the eligibility of citations during an update happening 4 years after the initial screening. Indeed, it decreased the workload by 56% (55–58) while missing 7 of 269 (3% [1–5]) eligible citations (Table S1 and S2).

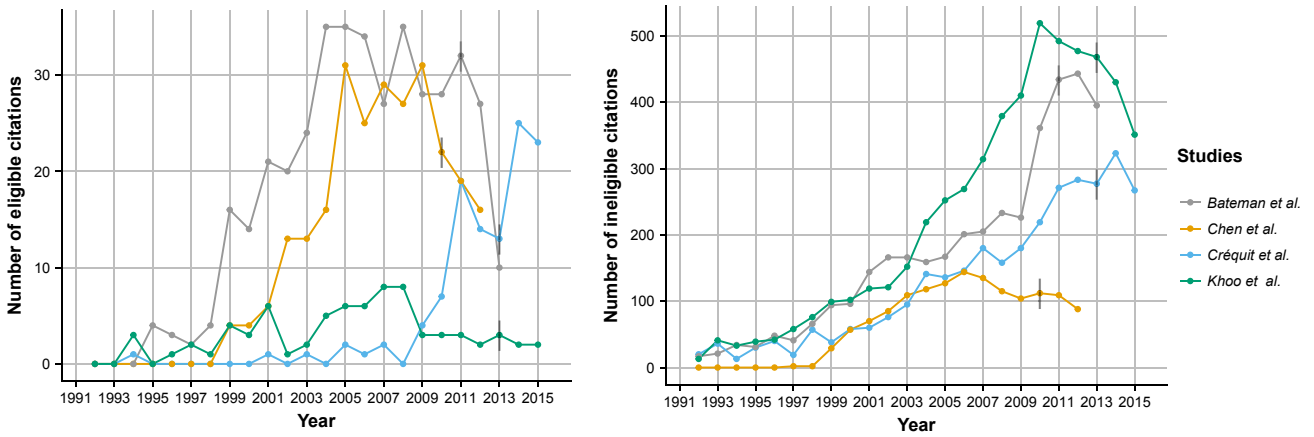

Fig. 2. Pace of publication of eligible and ineligible citations. Number of eligible (left) and ineligible (right) citations published each year between 1990 and 2015. Gray horizontal lines represent the cut-offs introduced in time to separate sets of initially screened citations and sets of citations to screen if an update was conducted after this cut-off for each network meta-analysis.

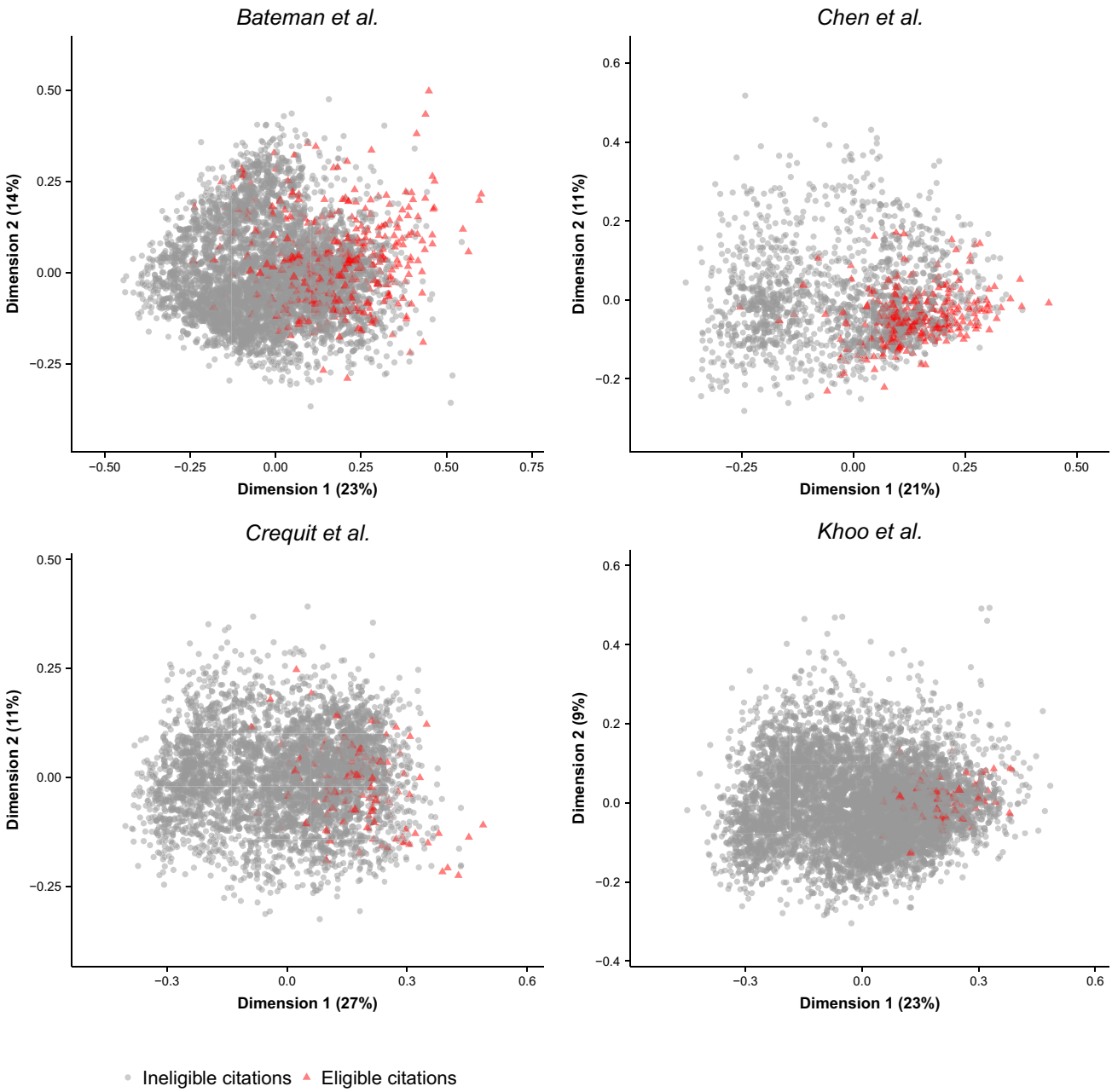

Fig. 3. Visualizing citations using principal component analysis. Citations are represented by the average of their word vectors, then reduced to two dimensions by principal component analysis. Red triangles represent eligible citations, and gray circles represent ineligible citations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

5. Discussion

In this study we evaluated algorithms for automatically screening citations when updating NMAs 2 years after the conduct of the initial NMA. Our results showed that a model of logistic regression on top of a word embedding representation of the title and abstract achieved good discriminative properties for this task. Our model achieved high sensitivity; it missed 3 of 124 eligible citations (2% [1–7]) and still was able to maintain substantial specificity, decreasing the workload by 53% (51–55). These performances may have been mostly due to the embedded representation of citations.

Our automatic classification method missed three eligible citations across all NMAs, but none of them was finally included in the NMA after full-text consideration. Indeed, in our study, we labeled citations to train our classifier according to their eligibility after the screening of title and abstracts only and not after final inclusion after full-text consideration. Using eligible citations as labels for training the algorithm allowed us to have a "safety net" regarding missed citations. The results of the analysis for these

Table 2. Automatic screening when updating 2 y after the initial conduct of the NMA

	Number of citations to screen after 2 y of update							
		Eligible						
First author	Total	Manual	Correctly predicted	Missed	Total predicted positive	Ineligible spared to screen	Sensitivity (95% CI)	Specificity (95% CI)
Word embeddings representation								
Bateman et al.	875	37	36	1	596	278	0.97 (0.86–1.00)	0.33 (0.30–0.36)
Chen et al.	232	35	33	2	113	117	0.94 (0.81–0.99)	0.59 (0.52–0.66)
Créquit et al.	638	48	48	0	297	341	1.00 (0.93–1.00)	0.58 (0.54–0.62)
Khoo et al.	785	4	4	0	176	609	1.00 (0.40–1.00)	0.78 (0.75–0.81)
tf–idf representation								
Bateman et al.	875	37	35	2	651	222	0.95 (0.82–0.99)	0.26 (0.24–0.30)
Chen et al.	232	35	32	3	157	72	0.91 (0.77–0.98)	0.37 (0.30–0.40)
Créquit et al.	638	48	44	4	317	317	0.92 (0.80–0.98)	0.54 (0.50–0.58)
Khoo et al.	785	4	4	0	533	252	1.00 (0.40–1.00)	0.32 (0.29–0.36)

Abbreviations: CI, confidence interval; NAM, network meta-analysis; tf–idf, term frequency–inverse document frequency.

For each NMA, we constructed sets of initially screened citations and of citations to screen during an update by introducing an artificial cut-off for time based on publication year of the screened citations. We evaluated the performance of logistic regression on top of both tf–idf and word embeddings representation when the update took 2 y after the conduct of the initial NMA. Sensitivity corresponded to the ratio between the number of correctly labeled eligible citations and the total number of eligible citations. Specificity corresponded to the ratio between the number of correctly labeled ineligible citations and the total number of ineligible citations. Loss of studies corresponded to the ratio between the number of inaccurately labeled eligible citations and the total number of eligible citations. Total predicted positive are all citations classified as eligible by the algorithm. Ineligible citations spared from screening are ineligible citations correctly predicted. We calculated 95% CIs with the Clopper–Pearson method.

NMA would not have been affected by the loss of these citations. Conversely, logistic regression on top of a tf–idf representation missed nine eligible citations, of which one was finally included in the NMA after full-text consideration [26]. The proportion of citations considered as eligible citations after title and abstract screening varies considerably from one reviewer to another, and for NMAs such as Chen et al. having a high ratio of eligible studies (15%), these labels could be too noisy and lower specificity of the algorithm.

A recent study investigated machine-learning algorithm to update three SRs in the field of rheumatology, using a support vector machine (SVM) with a term-frequency bag-of-word representation of citations [13]. They reported a mean sensitivity of 96% while reducing the number of citations to be screened by a mean of 78%. Our results confirmed the possibility to achieve high sensitivity for automatic screening not only when updating conventional MAs but also NMAs, for which it would be more difficult for a text mining framework to automatically identify the names of the interventions considered in the NMA as a feature for classification. In addition, our algorithm shows similar performance when applied to different fields of medicine but also when applied to NMAs where initial screening conditions differ, such as the percentage of eligible citations. We showed in our study that word embeddings could be a better method for representing citations to feed machine learning algorithms compared with tf–idf. Techniques based on other features than free text were proposed to alleviate the burden of screening, such as ranking based on co-citation metrics [27]; however, their performance decreased when citations included a large diversity of authors (50% of workload saving with 21% loss of studies). Semisupervised approaches [11] or active learning [15] are known to be more competitive with fewer screened citations available, for instance, when conducting the initial screening of an SR. However, when updating SRs, more training data are available, and classical supervised approaches are therefore possible.

A strength of our study is that we evaluated our algorithm by replicating the context of the update of an NMA and did not train and tested our classifier to discriminate citations regardless of the date of publication (e.g., using cross-validation). In addition, our algorithm achieved good performances in different fields of medicine. These performances were established with NMAs and not simply SRs, which are based on complex search equations because several interventions need to be considered. Finally, we used pretrained word embeddings to take advantage of knowledge of the free-text structure previously extracted from a very large dataset from the biomedical literature. Word embeddings provided a simple and computationally efficient representation of citations; they also proved useful for distinguishing eligible and ineligible citations (Fig. 3). We also showed that they provide better features than tf–idf for automatic screening using logistic regression.

Our study shows several limitations. First, in the context of an NMA aiming at comparing all available treatments for a particular condition (such as a live cumulative NMA), new treatments may become available with time,

which requires updating search equations. Our algorithm was evaluated only when search equations are not modified over time. However, an updated search equation would include additional terms (e.g., corresponding to the new treatments to include in the NMA), thereby implying a larger amount of citations to screen. Our algorithm can still be applied to the subset of these citations retrieved by the initial search equation and retrained afterward with the new search equation. This study also lacked comparisons with other classification algorithms or uses of more sophisticated text representation. There may be room for improvements in citation representation; for example, a previous study [17] showed that combining a tf−idf representation of unigrams with word vectors may increase classification accuracy. One could investigate representations that account for word order such as paragraph embeddings [28]. Features based on co-citations metrics could be incorporated into the model to account for other sources of information than free text. Our logistic regression model did not allow for building nonlinear hypotheses to discriminate citations, and using more complex models such as SVMs or gradient boosting machines [29] may increase discrimination performance. However, the use of word embeddings with a simple linear model may provide performance comparable to the best-performing existing algorithms in many text classification tasks [16].

NMAs are a useful framework to address the comprehensive and up-to-date synthesis of biomedical evidence globally. Indeed, NMAs by their construction already enable comparison of all available treatments. Comparing all available treatments while staying up-to-date would fulfill the conditions for directly operable synthesis of evidence in everyday clinical practice. These objectives were recently introduced by living NMAs [8]. Sharing a similar vision as Thomas et al. [30], efforts will be made to directly connect machine-learning algorithms with electronic databases via their application programming interface for a pipeline of search equations followed by automatic screening before manual screening.

6. Conclusion

When updating an NMA after 2 years, our screening algorithm based on word embeddings considerably diminished the workload of screening, and missed eligible citations remained low. Machine-learning algorithms may greatly reduce the time needed to update NMAs. Reviewers may use these methods to update NMAs more regularly, thereby reinforcing their validity and clinical relevance.

Acknowledgments

The authors thank Tania Martin for providing the data on screening process. They also thank Laura Smales for language revision of the article. This work was partially funded by the grant N°2016-02/058/AB-KA from the Institut National du Cancer (INCa).

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.12.001.

References

[1] Créquit P, Trinquart L, Yavchitz A, Ravaud P. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. BMC Med 2016;14:8.

[2] Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009;181:488−93.

[3] Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008;61:531−6.

[4] Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007;147:224−33.

[5] Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA 1998;280:278−80.

[6] Higgins J, Green S, Scholten R. Cochrane Handbook for Systematic Reviews of Interventions VersionIn: Chapter 3: maintaining reviews: updates, amendments and feedback, 5 2008.

[7] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7:e1000326.

[8] Créquit P, Trinquart L, Ravaud P. Live cumulative network meta-analysis: protocol for second-line treatments in advanced non-small-cell lung cancer with wild-type or unknown status for epidermal growth factor receptor. BMJ Open 2016;6:e011841.

[9] Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA 1999;282:634−5.

[10] Paynter R. EPC methods: an exploration of the use of text-mining software in systematic reviews. Rockville, MD: Agency for Healthcare Research and Quality (US); 2016. Available at https://www.ncbi.nlm.nih.gov/books/NBK362044/.

[11] Kontonatsios G, Brockmeier AJ, Przybyła P, McNaught J, Mu T, Goulermas JY, et al. A semi-supervised approach using label propagation to support citation screening. J Biomed Inform 2017;72:67−76.

[12] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4:5.

[13] Shekelle PG, Shetty K, Newberry S, Maglione M, Motala A. Machine learning versus standard techniques for updating searches for systematic reviews: a diagnostic accuracy study. Ann Intern Med 2017;167:213−5.

[14] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [Internet]. arXiv [cs.CL] 2013. Available at http://arxiv.org/abs/1301.3781. Accessed July 1, 2017.

[15] Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. J Biomed Inform 2016;62:59−65.

[16] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:

Volume 2, Short Papers. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017:427−31.

[17] Balikas G, Amini MR. An empirical study on large scale text classification with skip-gram embeddings. Archiv 2016. Available at https://arxiv.org/abs/1606.06623. Accessed July 1, 2017.

[18] Bateman ED, Esser D, Chirila C, Fernandez M, Fowler A, Moroni-Zentgraf P, et al. Magnitude of effect of asthma treatments on asthma quality of life questionnaire and asthma control questionnaire scores: systematic review and network meta-analysis. J Allergy Clin Immunol 2015;136:914−22.

[19] Chen L, Staubli SEL, Schneider MP, Kessels AG, Ivic S, Bachmann LM, et al. Phosphodiesterase 5 inhibitors for the treatment of erectile dysfunction: a trade-off network meta-analysis. Eur Urol 2015;68:674−80.

[20] Khoo AL, Zhou HJ, Teng M, Lin L, Zhao YJ, Soh LB, et al. Network meta-analysis and cost-effectiveness analysis of new generation antidepressants. CNS Drugs 2015;29:695−712.

[21] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Archiv 2013. Available at https://arxiv.org/abs/1310.4546. Accessed July 1, 2017.

[22] Sampo Pyysalo Filip Ginter Hans Moen Tapio Salakoski Sophia Ananiadou. Distributional semantics resources for biomedical text processing 2013. Available at http://bio.nlplab.org/pdf/pyysalo13literature.pdf. Accessed July 1, 2017.

[23] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al, Xiaoqiang Zheng Google Research*. TensorFlow: large-scale machine learning on heterogeneous distributed systems 2015. Available at http://download.tensorflow.org/paper/whitepaper2015.pdf. Accessed July 1, 2017.

[24] Fabian P, Gaël V, Alexandre G, Vincent M, Bertrand T, Olivier G, et al. scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825−30.

[25] Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. Proc. of the 9th Python in Science Conf. (SCIPY 2010). Available at https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf. Accessed July 1, 2017.

[26] Levy B, Spira A, Becker D, Evans T, Schnadig I, Ross Camidge D, et al. A randomized, phase 2 trial of docetaxel with or without PX-866, an irreversible oral phosphatidylinositol 3-kinase inhibitor, in patients with relapsed or metastatic non−small-cell lung cancer. J Thorac Oncol 2014;9:1031−5.

[27] Janssens AC, Gwinn M. Novel citation-based search method for scientific literature: application to meta-analyses. BMC Med Res Methodol 2015;15:84.

[28] Le Q, Mikolov T. Distributed representations of sentences and documents. International Conference on Machine Learning 2014;:1188−96.

[29] Dalal SR, Shekelle PG, Hempel S, Newberry SJ, Motala A, Shetty KD. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. Med Decis Making 2013;33:343−55.

[30] 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.