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

Ivan Lerner, Perrine Créquit, Philippe Ravaud, Ignacio Atal

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

Ivan Lerner, MSc[1-3]; Perrine Créquit, MD, PhD[1-4]; Philippe RAVAUD, MD, PhD[1-5]; Ignacio ATAL, PhD[1-3]

[1] Centre de Recherche Epidémiologie et Statistique Paris Sorbonne Cité, INSERM U1153, Paris, France

[2] Université Paris Descartes – Sorbonne Paris cité, Paris, France

[3] Centre d'Epidémiologie Clinique, Assistance Publique-Hôpitaux de Paris, Hôpital Hôtel-Dieu, Paris, France

[4] Cochrane France, Paris, France
[5] Department of Epidemiology, Mailman School of Public Health, Columbia University New York, US

Word count:

Abstract: 223; Manuscript: 2971

Corresponding author: Ignacio Atal, PhD

Centre d'Epidémiologie Clinique,
Hôpital Hôtel-Dieu,
1, place du parvis Notre Dame, 75004 Paris, France
Tel: (+33) 1 42 34 87 65
E-mail: ignacio.atal-ext@aphp.fr

ABSTRACT

Objective

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% [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 1345 of 2530 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

What is new ?

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

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], so 7% of reviews are out-of-date by the time they are published[4]. In addition, less than 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 by 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 as compared with human updating. Machine-learning techniques applied to automatic screening were suggested to save 30% to 78% of the workload but miss 4% to 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 modelling[15] and feature extraction for classification of text using machine learning[16,17].

OBJECTIVE

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

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. Figure 1 summarizes the different stages of the workflow and represents the inputs and outputs of the system.

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 1000 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 PsychINFO) 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 6/1/2015: we considered citations published between 1/1/2013 and 12/31/2014 as the set of citations to screen if an update was conducted (test set), and all citations published before 12/31/2012 as the set of initially screened citations (training set).

Automatic screening

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

Citation representation based on word embeddings

For each citation, we represented the title and abstract by 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 trained a Skip-gram model[21] over all the available biomedical literature from PubMed and PMC 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 5 times across all citations were not encoded, nor were words not corresponding to pre-trained word vectors which included stop words. We then represented each citation by 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 2-D plot.

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 5 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.*

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 by 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 non-eligible 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.

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.

Sensitivity analysis

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

Implementation

Algorithms were implemented in python by 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*).

Statistical analysis

Descriptive data are presented with number (%) and 95% confidence intervals (CIs) 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.

RESULTS

Screening process

Our study included four NMAs in different fields of medicine (Table 1), which altogether totalled 14,853 screened citations. We present in Figure 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.

Automatic screening

Citation representation

The median length of citations (i.e., titles and abstracts) was 294 words, which for all citations totalled 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 (Figure 3). Although PCA in two dimensions explained only 32% to 38% of the variability, citations were partially separated between eligible and ineligible by the encoding scheme

only.

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% [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 out 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 since all models were developed to have high sensitivity regardless the text representation. For all NMAs, our algorithm would have spared screening manually 1345 of 2530 citations, decreasing the workload by 53% [51-55], while missing 3 of 124 eligible citations (2% [1-7]).

Sensitivity analysis

The algorithm had similar performances when trained to predict the eligibility of citations during an update happening four 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).

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 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 as compared to 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). Semi-supervised 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 a SR. However when updating SRs, more training data is 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 trained and tested our classifier to discriminate citations regardless of the date of publication (eg., by 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 pre-trained 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 (Figure 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 afterwards 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 to the model in order to account for other source of information than free text. Our logistic regression model did not allow for building non-linear 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.

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.

FIGURES

Figure 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.

Figure 2. Pace of publication of eligible and ineligible citations

Number of eligible (left) and ineligible (right) citations published each year between 1990 and 2015. Grey 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.

Figure 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 grey circles ineligible citations.

TABLES

Table 1. Network meta-analysis characteristics after replicating the search equations

For each network meta-analysis (NMAs) 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.

Table 2. Automatic screening when updating two years after the initial conduct of the network meta-analysis

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 years 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% confidence intervals with the Clopper-Pearson method.

LIST OF ABBREVIATIONS

NMA: network meta-analysis.

SR: systematic review.

PCA: principal component analysis.

SVM: support vector machines.

tf-idf: term frequency - inverse document frequency

DECLARATIONS

Availability of data and material:

The datasets generated and analysed during the current study are available in a git repository, https://gitlab.com/lerner.ivan/automatic_screening_NMA.

Competing interests: 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: IL contributed to study design, data pre-processing and analysis, results interpretation and writing. IA contributed to study design, results interpretation and writing. PC contributed to study design, results interpretation and writing. Philippe Ravaud contributed to study design and results interpretation. All authors read and approved the final manuscript.

Acknowledgements:

We thank Tania Martin for providing the data on screening process. We thank Laura Smales for language revision of the manuscript.

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. doi:10.1186/s12916-016-0555-0

2. Ioannidis JPA. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ.

2009;181: 488–493.

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: 531–536.

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: 224–233.

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: 278–280.

6. Higgins J, Green S, Scholten R. Chapter 3: Maintaining reviews: updates, amendments and feedback. Cochrane handbook for systematic reviews of interventions version. 2008;5.

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: 634–635.

10. Paynter R. EPC Methods: An exploration of the use of text-mining software in systematic reviews. 2016.

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; doi:10.1016/j.jbi.2017.06.018

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; doi:10.7326/L17-0124

14. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [Internet]. arXiv [cs.CL]. 2013. Available: http://arxiv.org/abs/1301.3781

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. pp. 427–431.

17. Georgios Balikas M-RA. An empirical study on large scale text classification with skip-gram embeddings. Archiv. 2016; Available: https://arxiv.org/abs/1606.06623

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: 914–922.

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: 674–680.

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. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. Archiv. 2013; Available: https://arxiv.org/abs/1310.4546

22. Sampo Pyysalo Filip Ginter Hans Moen Tapio Salakoski Sophia Ananiadou. Distributional Semantics Resources for Biomedical Text Processing. 2013; Available: http://bio.nlplab.org/pdf/pyysalo13literature.pdf

23. Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, ´ Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, ´ Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng Google Research□. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015; Available: http://download.tensorflow.org/paper/whitepaper2015.pdf

24. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

25. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. of the 9th Python in Science Conference. researchgate.net; 2010; Available: https://www.researchgate.net/profile/Josef_Perktold/publication/264891066_Statsmodels_Econometric_and_Statistical_Modeling_with_Python/links/5667ca9308ae34c89a0261a8/Statsmodels-Econometric-and-Statistical-Modeling-with-Python.pdf

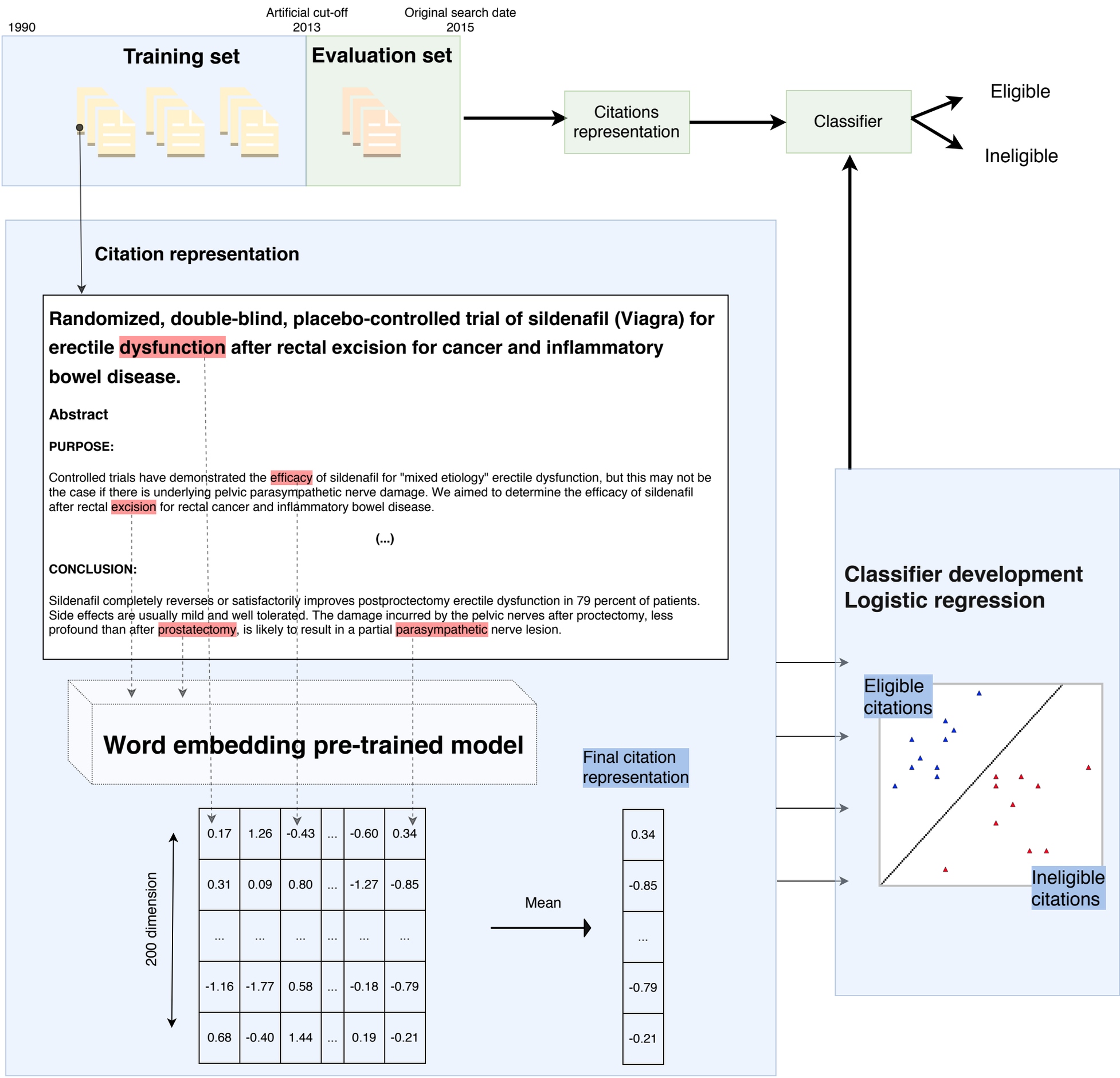
26. Levy B, Spira A, Becker D, Evans T, Schnadig I, Ross Camidge D, et al. A Randomized, Phase 2 Trial of Docetaxel with or without PX-866, an Irreversible Oral Phosphatidylinositol 3-Kinase Inhibitor, in Patients with Relapsed or Metastatic Non–Small-Cell Lung Cancer. J Thorac Oncol. 2014;9: 1031–1035.

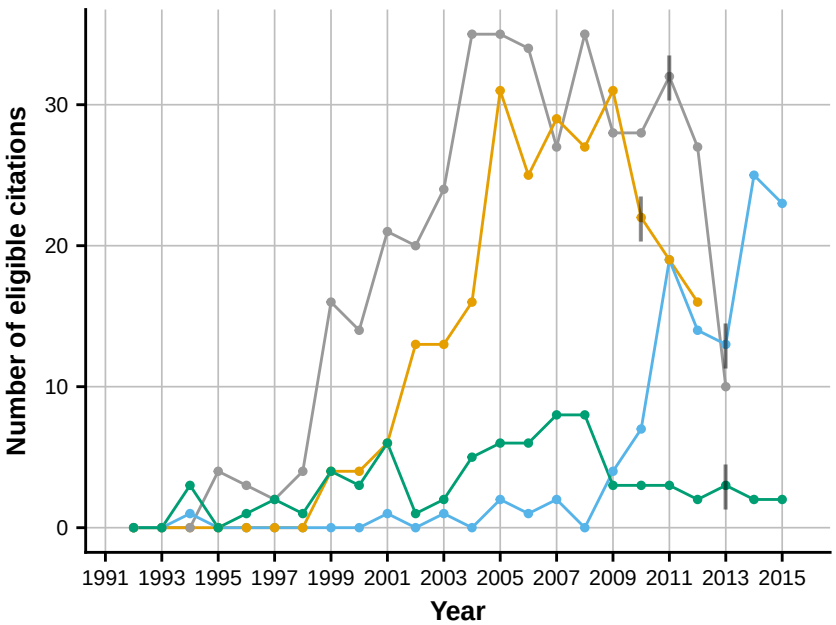
27. Janssens ACJW, Gwinn M. Novel citation-based search method for scientific literature: application to meta-analyses. BMC Med Res Methodol. 2015;15: 84.

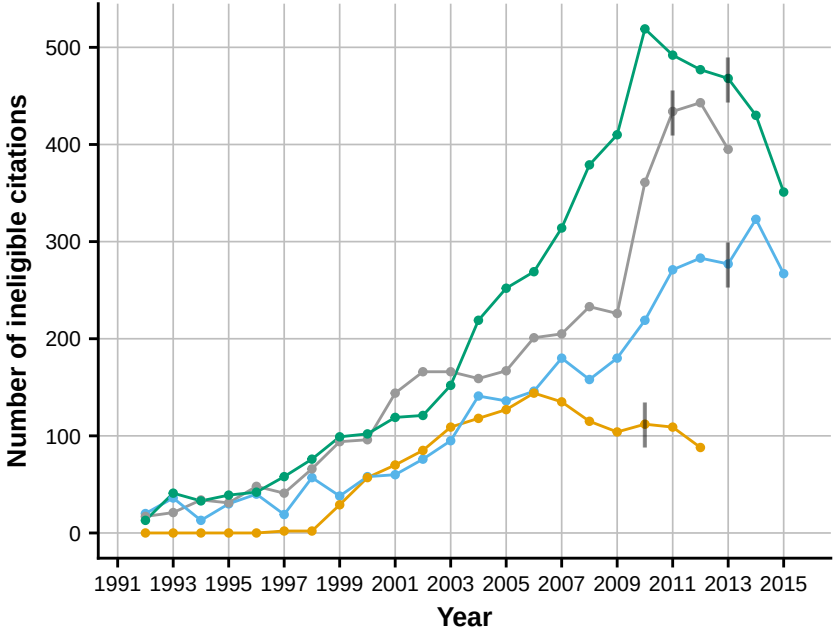
28. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. International Conference on Machine Learning. 2014. pp. 1188–1196.

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: 343–355.

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; doi:10.1016/j.jclinepi.2017.08.011


1990
Artificial cut-off
2013
Original search date
2015
Training set
Evaluation set
Citations representation
Classifier
Eligible
Ineligible
Citation representation
Randomized, double-blind, placebo-controlled trial of sildenafil (Viagra) for erectile dysfunction after rectal excision for cancer and inflammatory bowel disease.
Abstract
PURPOSE:
Controlled trials have demonstrated the efficacy of sildenafil for "mixed etiology" erectile dysfunction, but this may not be the case if there is underlying pelvic parasympathetic nerve damage. We aimed to determine the efficacy of sildenafil after rectal excision for rectal cancer and inflammatory bowel disease.
(...)
CONCLUSION:
Sildenafil completely reverses or satisfactorily improves postproctectomy erectile dysfunction in 79 percent of patients. Side effects are usually mild and well tolerated. The damage incurred by the pelvic nerves after proctectomy, less profound than after prostatectomy, is likely to result in a partial parasympathetic nerve lesion.
Word embedding pre-trained model
Final citation representation
200 dimension
0.17 1.26 -0.43 ... -0.60 0.34
0.31 0.09 0.80 ... -1.27 -0.85
... ...
-1.16 -1.77 0.58 ... -0.18 -0.79
0.68 -0.40 1.44 ... 0.19 -0.21
Mean
0.34
-0.85
...
-0.79
-0.21
Classifier development
Logistic regression
Eligible citations
Ineligible citations


Number of eligible citations
0
10
20
30
1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015
Year


Number of ineligible citations
0
100
200
300
400
500
1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015
Year


Studies
Bateman et al.
Chen et al.
Créquit et al.
Khoo et al.

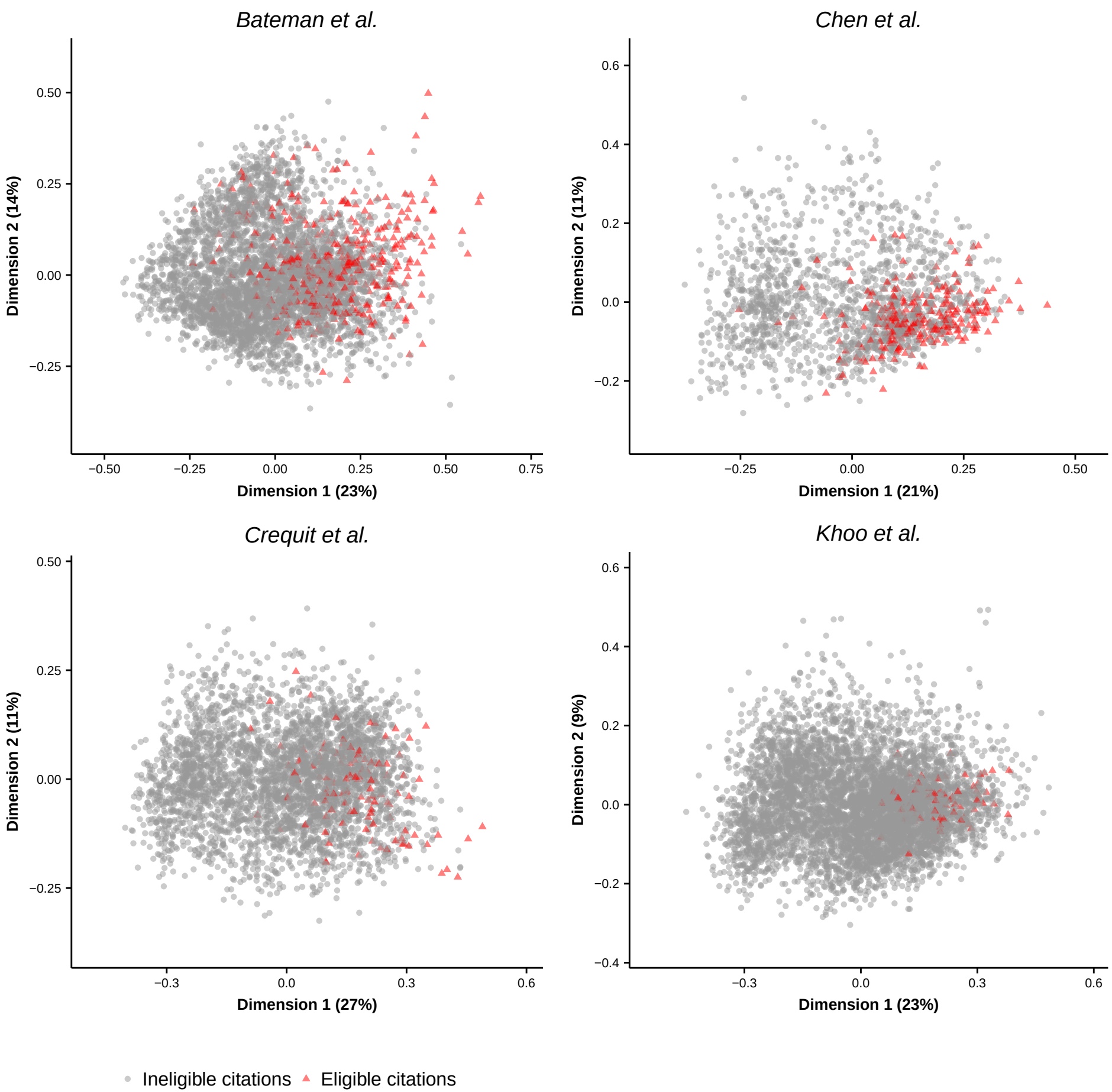

Bateman et al.
Chen et al.
Crequit et al.
Khoo et al.
Dimension 1 (23%)
Dimension 2 (14%)
Dimension 1 (21%)
Dimension 2 (11%)
Dimension 1 (27%)
Dimension 2 (11%)
Dimension 1 (23%)
Dimension 2 (9%)
Ineligible citations
Eligible citations

Table 1.

First author	Field	Databases	Total screened	Eligible	Proportion of eligible citations	Last date of search
Bateman et al. 2015	Pneumology	MEDLINE, EMBASE	4219	400	9 %	12/31/2012
Chen et al. 2015	Urology	MEDLINE, EMBASE	1662	256	15 %	12/31/2011
Créquit et al. 2016	Oncology	MEDLINE, EMBASE, CENTRAL	3373	113	3 %	12/31/2014
Khoo et al. 2015	Psychiatry	MEDLINE, EMBASE, PsycINFO	5599	75	1 %	31/12/2014

Table 2.

First author	Number of citations to screen after two years of update						Sensitivity (95% CI)	Specificity (95% CI)
	Total	Eligible				Ineligible spared to screen		
		Manual	Correctly predicted	Missed	Total predicted positive			
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)