

Citations alone were enough to predict favorable conclusions in reviews of neuraminidase inhibitors

Xujuan Zhou^a, Ying Wang^a, Guy Tsafnat^a, Enrico Coiera^a, Florence T. Bourgeois^{b,c,d}, Adam G. Dunn^{a,*}

^aCentre for Health Informatics, Australian Institute of Health Innovation, The University of New South Wales, Sydney, NSW 2052, Australia

^bDivision of Emergency Medicine, Boston Children's Hospital, Boston, MA, USA

^cDepartment of Pediatrics, Harvard Medical School, Boston, MA, USA

^dChildren's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA

Accepted 12 September 2014; Published online 22 October 2014

Abstract

Objectives: To examine the use of supervised machine learning to identify biases in evidence selection and determine if citation information can predict favorable conclusions in reviews about neuraminidase inhibitors.

Study Design and Setting: Reviews of neuraminidase inhibitors published during January 2005 to May 2013 were identified by searching PubMed. In a blinded evaluation, the reviews were classified as favorable if investigators agreed that they supported the use of neuraminidase inhibitors for prophylaxis or treatment of influenza. Reference lists were used to identify all unique citations to primary articles. Three classification methods were tested for their ability to predict favorable conclusions using only citation information.

Results: Citations to 4,574 articles were identified in 152 reviews of neuraminidase inhibitors, and 93 (61%) of these reviews were graded as favorable. Primary articles describing drug resistance were among the citations that were underrepresented in favorable reviews. The most accurate classifier predicted favorable conclusions with 96.2% accuracy, using citations to only 24 of 4,574 articles.

Conclusion: Favorable conclusions in reviews about neuraminidase inhibitors can be predicted using only information about the articles they cite. The approach highlights how evidence exclusion shapes conclusions in reviews and provides a method to evaluate citation practices in a corpus of reviews. © 2015 Elsevier Inc. All rights reserved.

Keywords: Neuraminidase inhibitors; Bibliometrics; Evidence synthesis; Reviews as a topic; Citation analysis; Supervised machine learning

1. Introduction

Variation in the inclusion of evidence can lead to flawed or unreliable conclusions in reviews and other peer-reviewed articles [1,2]. The resulting disagreement across reviews or guidelines may erode trust in evidence-based medicine and reduce the quality of clinical decision making. There has been considerable disagreement, for example, across reviews about the clinical use of neuraminidase inhibitors for the prophylaxis and treatment of influenza: some reviewers strongly recommend the use of these drugs [3–5], whereas others conclude that they provide only modest benefit and question the ability to draw

any meaningful conclusions from the limited available evidence [6,7].

For oseltamivir, the most commonly prescribed neuraminidase inhibitor, the evidence supporting its use has been mired in controversy [8]. Certain data used to support claims made by the company producing the drug were not released to the public [8], and concerns have been raised about the conflicts of interest held by members of the World Health Organization advisory panel that recommended stockpiling the drug in case of a pandemic [9].

Differences in the way evidence is selected for inclusion in literature reviews that could affect the conclusions are described as reference or inclusion bias [2,10]. These biases come in many forms, including the preferential inclusion of studies with positive outcomes and statistically significant results [11–13], from high-impact journals or authors with financial conflicts of interest [14–17], or disproportionate levels of self-citation [18,19]. Citation network analyses have been used to examine the incidence and potential

Funding: This study was supported by National Health and Medical Research Council Program Grant 568612 and Project Grant 1045065. The sponsors had no role in the study.

Conflict of interest: None.

* Corresponding author. Tel.: +61 2 9385 8699; fax: +61 2 9385 8692.

E-mail address: a.dunn@unsw.edu.au (A.G. Dunn).

What is new?

- It is possible to accurately predict favorable conclusions in reviews about neuraminidase inhibitors using only information about what is included in their reference lists.
- Citation network analyses have been used to identify biases in reviews by comparing the selection of evidence across a corpus of reviews, but little is known about the effects of citation biases on review conclusions and recommendations.
- Machine learning classification methods applied across a corpus of reviews may be used to identify primary studies that are overrepresented or underrepresented in reviews with favorable conclusions.
- Extensions to the approach presented here may provide new methods for automatically evaluating the entire evidence base of interventions for which systemic citation bias is suspected.

implications of differences in evidence selection [1,20–24]. However, little is known about how the biases in evidence selection may affect the conclusions of reviews.

Supervised machine learning has been used to examine analogous problems. A Bayesian classifier was found to be capable of predicting conclusions that individual decision makers reached, based solely on the articles to which they were exposed [25]. Another example using four types of classifiers showed that machine learning could predict which articles should be screened for inclusion in systematic reviews [26].

We sought to measure the association between the primary articles cited in reviews of neuraminidase inhibitors and the likelihood of a conclusion favoring the use of the drugs for influenza, evaluating classifiers trained to predict conclusions based only on the reference lists of the reviews. The classifiers were used to identify the citations that best distinguish favorable reviews from all others, revealing how the inclusion of specific primary evidence may have influenced conclusions.

2. Methods

2.1. Study data

Reviews were identified in PubMed by searching all English-language articles published since January 2005 for keywords “influenza” and at least one of “neuraminidase inhibitors,” “oseltamivir,” “zanamivir,” “peramivir,” or “laninamivir,” or their synonyms, in the title, abstract, or keywords and then restricting the set to include only articles for which the publication type was a review. The final

search was performed in May 2013 and identified 211 articles. Because there were no further restrictions on the form of the reviews, the set included narrative reviews that did not include explicit search criteria or the reporting of reasons for excluding some published articles. Of the 211 that were identified, 59 were excluded by consensus (Diana Arachi, Joel Hudgins, and F.T.B.) because they did not review the clinical use of neuraminidase inhibitors (these included reviews of drug development, manufacture, or drugs from other classes), producing a set of 152 reviews about the clinical use of neuraminidase inhibitors for influenza.

Two reviewers independently examined the full text of each review (blinded to the authors and affiliations, citations, journal and formatting, acknowledgments, and conflicts of interest) and rated each review as favorable, unfavorable, or neutral to the use of neuraminidase inhibitors for the prophylaxis or treatment of influenza. The two reviewers were guided by answering questions about the presentation of evidence in relation to efficacy, safety, and resistance and the presence of recommendations for clinical use. When the two evaluators independently agreed that a review was favorable, the review was classified as favorable—all other reviews were assigned to the alternative group. Among the 152 reviews, 93 (61%) were deemed to be favorable and 59 (39%) were assigned to the alternate group.

All primary articles cited by the reviews were retrieved and verified manually. Publication dates were also recorded for each article. There were 4,574 unique articles cited in the reviews. The total number of citations from the set of reviews to these articles was 10,086; 3,112 were cited once, 582 were cited twice, and 880 were cited three times or more. The most commonly cited article was cited 46 times—in 30% of the reviews (Fig. 1). Before applying machine learning to train and test classifiers, we examined the distribution of citations to identify which primary articles were overrepresented or underrepresented among favorable reviews. To do this, while accounting for publication dates,

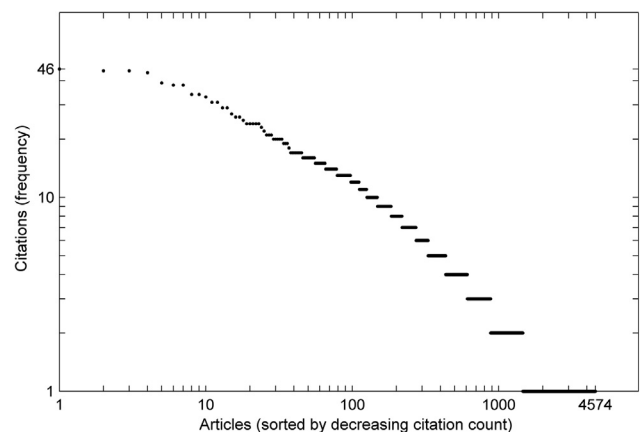


Fig. 1. The distribution of citations to each of 4,574 unique articles from the 152 reviews on the clinical use of neuraminidase inhibitors.

we calculated the proportional occurrence of articles across the two classes. The number of reviews that cited each article was calculated as a proportion of the number of reviews published after the publication date of the article. These proportions were calculated for favorable and other reviews separately and then compared in Fisher's exact tests (without applying a correction for multiple testing to determine significance).

2.2. Classifier training and feature selection

Supervised machine learning was used to train classifiers that could distinguish between two classes: reviews with favorable conclusions and all other reviews. The data set used to train the classification algorithms was limited to the presence or absence of a citation to each of the 4,574 unique articles cited by at least one of the reviews.

Even when addressing the same clinical question, reviews tend to cite a high proportion of articles not cited by any other review, resulting in a very sparse feature space where many features are uninformative, and unlikely to help distinguish between the two classes. As an initial step in removing features likely to be uninformative, we restricted the feature space on which the classifiers were trained. This produced three feature spaces: (1) the set of all 4,574 cited articles as features; (2) the 1,462 articles that were cited at least twice; and (3) the 880 articles that were cited at least three times.

Three types of classifier construction methods were tested, and their performances compared to find the most accurate model for the task. The methods that were chosen were (1) nonlinear support vector machines (SVMs) with radial basis function (RBF) kernels [27,28], (2) naïve Bayes classifiers [29–31], and (3) the k nearest neighbor (KNN) algorithm [32]. For each of the three classifier construction methods, the optimize selection strategy (a combination of forward selection and backward elimination) [33,34] was used in training to identify the sets of features that together best distinguished between favorable reviews and all others (Fig. 2).

2.3. Validation and performance evaluation

The comparative performances of the classifiers were evaluated using the typical metrics for data classification: [1] accuracy: the percentage of reviews classified correctly into a given category in relation to the total number of reviews tested; [2] precision: the percentage of true positives detected in relation to total number of reviews classified for a category; [3] recall: the percentage of true positives detected in relation to the actual number of reviews in the category; and [4] F_1 measure: the harmonic mean of precision and recall.

A stratified 10-fold cross-validation was applied to evaluate how well the classifiers could predict the conclusions of unseen reviews. The performances of the classifiers are

then given as the average value of the metrics on the testing subset across each of the 10 runs. RapidMiner was used to train and validate the classifiers [35].

3. Results

3.1. Citation distributions

Twelve articles were found to be relatively overrepresented in the citations of favorable reviews, and 19 articles were found to be relatively underrepresented in the citations of favorable reviews (Fig. 3). These articles were cited between 4 and 29 times across both classes and had publication dates between 1983 and 2011. Further details of these articles and their proportional citations in the set of reviews are provided in the Appendix at www.jclinepi.com.

3.2. Classifier construction and testing

The SVM method produced the most accurate classifier, achieving 96.2% accuracy (0.96 precision, 0.98 recall, and 0.97 F_1) starting from feature space 2, the set of articles with at least two citations (Table 1). The feature selection method eliminated the vast majority of the 1,462 features, identifying a set of 24 articles, and increased the accuracy by 25.7% from 70.5% (Table 2). The best performance was achieved with kernel parameters $\gamma = 0.1$, 0.01, and 1.0 for input feature spaces 1, 2, and 3, respectively (Table 1).

The best performing naïve Bayes classifier achieved 95.5% accuracy ($F_1 = 0.97$), also starting from feature space 2 (Table 1). The feature selection method improved the accuracy by 31.8% from 63.7%, producing a set of 21 features (Table 2).

For the KNN classifier, the performance was primarily determined by the choice of k and the distance metric applied [36]. The most accurate KNN classifiers were achieved for feature space 1 using $k = 1$, for feature space 2 using $k = 3$, and for feature space 3 using parameter $k = 5$ (Table 1). Feature selection reduced the features to 21, 20, and 17 and increased the accuracy from 60.5% to 92.1%, 63.3% to 89.6%, and 63.2% to 87.6% (Table 2).

3.3. Selected features

The two best performing classifiers selected 24 and 21 features, and the two had 10 articles in common. A closer examination of the articles that were part of the overlap between the two best performing classifiers suggests why these articles were particular to unfavorable and neutral reviews. For example, among the overlapping group, Nguyen et al. [37] reported a case of a child who died after testing positive for an H1N1 strain resistant to treatment by oseltamivir and zanamivir. (The article was cited in 32% of the unfavorable and neutral reviews that could have cited it and 0% of favorable reviews.) Another study by Burch et al. [38] concluded that there were no clinically

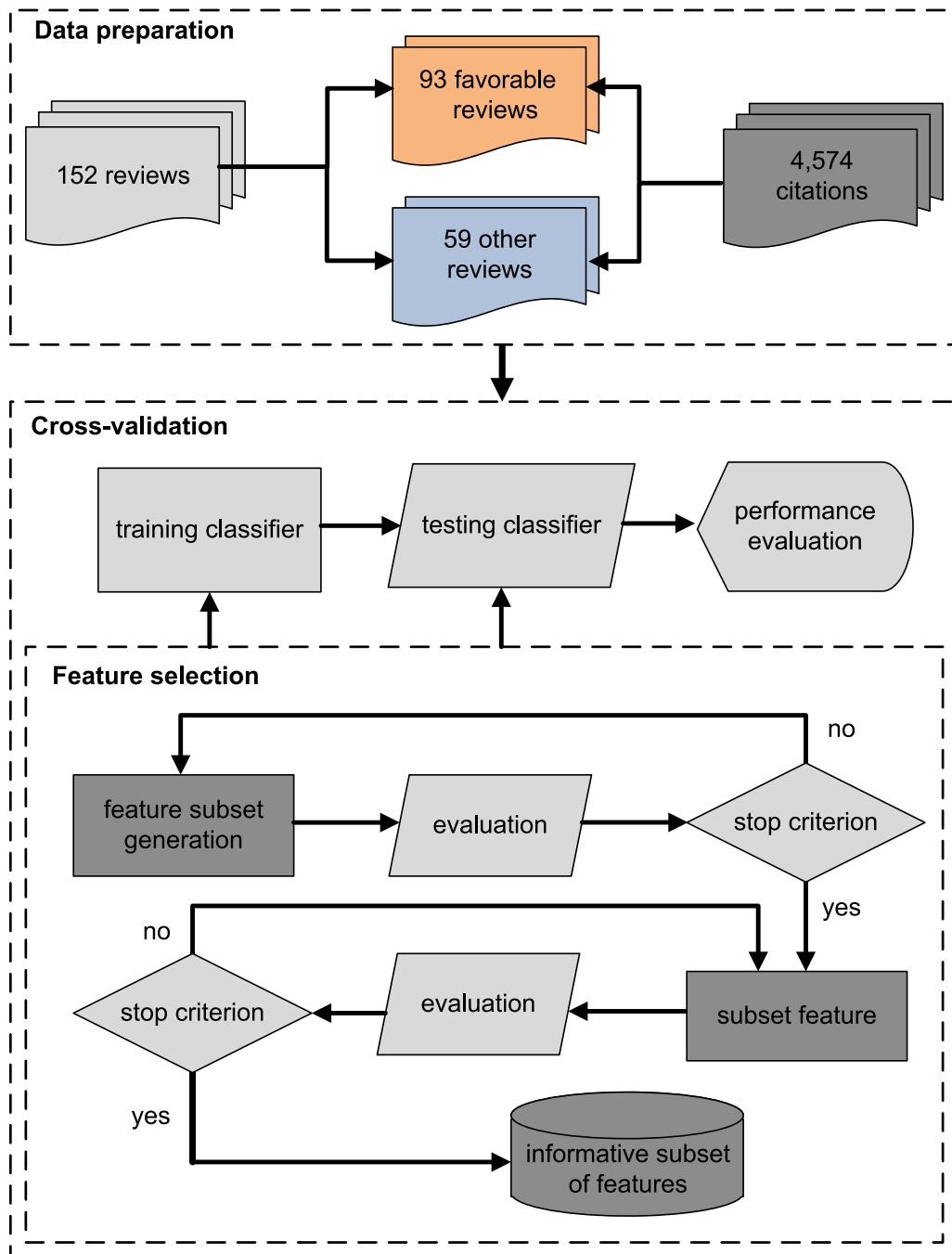


Fig. 2. The training and validation processes represented as a flow diagram that describes the process for producing a classifier that predicts favorable conclusions using cited articles as features.

significant economic benefits of oseltamivir and zanamivir use in healthy adults and uncertainty in the evidence for at-risk groups. (The article was cited in 6% of unfavorable and neutral reviews that could have cited it and 0% of favorable reviews.)

Alternatively, Adisasmito et al. [39], Kiso et al. [40], or Tumpey et al. [41] signaled a favorable conclusion in the reviews that cited them. In 2010, Adisasmito et al. [39] reported that oseltamivir significantly reduces mortality in patients with H5N1. Also in 2010, Kiso et al. [40] reported

that laninamivir would be “highly effective for the treatment and prophylaxis of infection with H5N1 influenza viruses, including oseltamivir-resistant mutants.” In 2002, Tumpey et al. [41] suggested that oseltamivir and zanamivir would be effective against reemergent strains of the virus that caused a pandemic in 1918. None of these three studies reported the results of a randomized controlled trial, and all studies included disclosures of funding from the pharmaceutical companies manufacturing neuraminidase inhibitors.

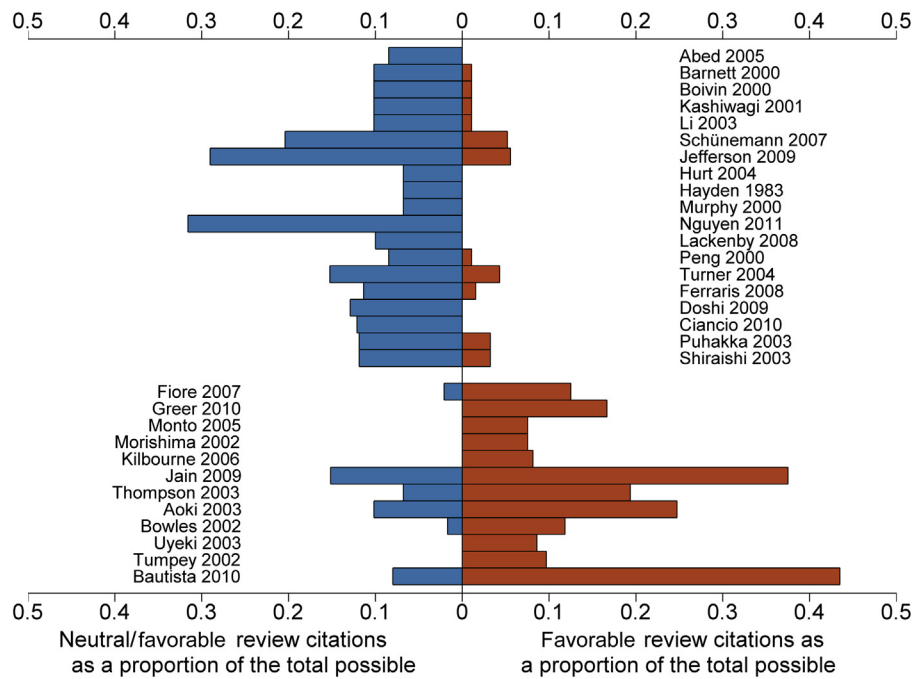


Fig. 3. Articles that were underrepresented (top) or overrepresented (bottom) in the citations of favorable reviews ($P < 0.05$ in a Fisher's exact test) are illustrated as the number of citations as a proportion of reviews published after the publication date of the article. The articles are ranked by P -value from those most underrepresented at the top to those most overrepresented at the bottom.

4. Discussion

We developed classifiers that were able to predict favorable conclusions among reviews of neuraminidase inhibitors. The most accurate classifier was able to correctly predict 96.2% of review conclusions, using only information about the articles that were cited. The results of these experiments suggest that the differences in the reference lists of reviews about neuraminidase inhibitors are

consistent enough to accurately distinguish between favorable reviews and all others.

Previous work examining the consequences of discrepancies in evidence selection has focused on specific cases where research consensus appeared to have been affected by the amplification or avoidance of specific evidence [1] or otherwise examined the structure of citation networks to consider the flow, or lack of flow, in the translation

Table 1. The classification results for different classifiers predicting the conclusions of reviews

Classifier	Precision	Recall	F ₁	Accuracy (%) (95% CI)
Naïve Bayes				
Including all citations	0.93	0.99	0.96	94.2 (90.5, 97.9)
Articles cited more than once	0.95	0.99	0.97	95.5 (92.2, 98.8)
Articles cited more than twice	0.90	0.99	0.94	91.6 (87.2, 96.0)
KNN ($k = 1$)				
Including all citations	0.95	0.97	0.96	92.1 (87.8, 96.4)
Articles cited more than once	0.81	0.89	0.85	78.9 (72.4, 85.4)
Articles cited more than twice	0.81	0.95	0.87	82.4 (76.3, 88.5)
KNN ($k = 3$)				
Including all citations	0.92	0.96	0.94	91.7 (87.3, 96.1)
Articles cited more than once	0.87	0.98	0.92	89.6 (84.7, 94.5)
Articles cited more than twice	0.89	0.91	0.90	87.5 (82.2, 92.8)
KNN ($k = 5$)				
Including all citations	0.89	0.95	0.92	89.0 (84.0, 94.0)
Articles cited more than once	0.82	0.98	0.89	84.4 (78.6, 90.2)
Articles cited more than twice	0.87	0.95	0.91	87.6 (82.4, 92.8)
SVM RBF				
Including all citations ($\gamma = 0.1$)	0.97	0.96	0.96	95.5 (92.2, 98.8)
Articles cited more than once ($\gamma = 0.01$)	0.96	0.98	0.97	96.2 (93.2, 99.2)
Articles cited more than twice ($\gamma = 1.0$)	0.94	0.95	0.75	93.0 (88.9, 97.1)

Abbreviations: CI, confidence interval; KNN, k nearest neighbor; SVM, support vector machine; RBF, radial basis function.

Table 2. Prediction accuracy on three feature spaces with and without feature selection

Classifier	Accuracy without FS (%) (95% CI)	Accuracy with FS (%) (95% CI)	Difference (%)	No. of selected features
Naïve Bayes				
Including all citations	65.0 (57.4, 72.6)	94.2 (90.5, 97.9)	29.2	21
Articles cited more than once	63.7 (56.1, 71.3)	95.5 (92.2, 98.8)	31.8	21
Articles cited more than twice	63.1 (55.4, 70.8)	91.6 (87.2, 96.0)	28.5	24
KNN				
Including all citations ($k = 1$)	60.5 (52.7, 68.3)	92.1 (87.8, 96.4)	31.6	21
Articles cited more than once ($k = 3$)	63.3 (55.6, 71.0)	89.6 (84.7, 94.5)	26.3	20
Articles cited more than twice ($k = 5$)	63.2 (55.0, 70.9)	87.6 (82.4, 92.8)	24.4	17
SVM RBF				
Including all citations ($\gamma = 0.1$)	69.1 (61.7, 76.5)	95.5 (92.2, 98.8)	26.4	22
Articles cited more than once ($\gamma = 0.01$)	70.5 (63.2, 77.8)	96.2 (93.2, 99.2)	25.7	24
Articles cited more than twice ($\gamma = 1.0$)	67.3 (59.8, 74.8)	93.0 (88.9, 97.1)	25.7	21

Abbreviations: FS, feature selection; CI, confidence interval; KNN, k nearest neighbor; SVM, support vector machine; RBF, radial basis function.

and synthesis of clinical evidence [21,22]. Our work extends these findings by examining the relationship between evidence selection and conclusions in a large corpus, including 10,086 citations to 4,574 unique articles in 152 reviews.

There have been a number of concerns around the available scientific evidence supporting the clinical use of neuraminidase inhibitors. In particular, there appear to have been delays in the release of clinical trial results around the efficacy and safety of oseltamivir [42], difficulties in accessing unpublished clinical trial data [43,44], and biases in the synthesis of evidence in systematic reviews on neuraminidase inhibitors [45]. Most reviews we identified were favorable to the use of neuraminidase inhibitors in clinical practice. Among the reviews graded as favorable, citations to primary studies describing the emergence of drug resistance were systematically underrepresented relative to reviews that were graded as neutral or unfavorable.

The value of this approach is not only in the ability to predict the conclusions of an unseen review based only on what is in the reference list; it suggests a new way to examine how differences in the selection of evidence across a body of literature might contribute to the risk of a flawed research consensus. Once classifiers are trained using the appropriate corpus, it may be possible to apply this technique to evaluate unseen reviews individually, either at the point of submission or in postpublication review. As a tool for evidence surveillance, this method may be extended to automatically identify interventions for which the systematic underrepresentation of groups of primary studies indicates a risk in the quality of evidence synthesis.

Current limitations in the method include the number of manual steps in the process, as both the verification of citations and the evaluation of conclusions required substantial human input. However, there are attempts to improve automated and open access to citation information online [46]. An extension of this work in unsupervised machine learning (such as clustering) may also reduce the need for human input in the grading of reviews and further improve

automation for the purpose of evidence surveillance. A second limitation is that we tested the predictive abilities of the classifiers using testing sets comprising reviews from within the same period in which the classifiers were trained. Substantial changes in the primary literature may degrade the performance of the classifiers on new reviews. A third limitation of the work is that we did not formally analyze the content of the selected features relative to other features that might have been chosen. Therefore, although they were useful for predicting favorable conclusions in unseen reviews, any interpretation of the content could be considered post hoc speculation. Finally, the results do not reveal a deliberate bias in the selection of evidence, just an association between what was cited in the reviews and the direction of the conclusions. Reviewers' predispositions for or against a favorable conclusion may have affected their selection of evidence, but an alternative explanation might be that the heterogeneity of the clinical questions and review designs may have affected the conclusions drawn. Because many of the narrative reviews did not report inclusion and exclusion criteria, we were precluded from examining how heterogeneity in the populations, interventions, and outcomes may have explained the differences in evidence selection.

5. Conclusion

The results of these experiments demonstrate that for reviews about the clinical use of neuraminidase inhibitors, the reference lists of the reviews are strong predictors of the conclusions. For reviews about neuraminidase inhibitors, there is cause for concern about the high proportion of favorable reviews and the associated exclusion of primary evidence that does not support the use of the drugs in clinical practice. The approach described here may be of value to authors and editors seeking to identify combinations of primary articles that are routinely excluded or amplified among groups of reviews.

Supplementary data

Supplementary data related to this chapter can be found at <http://dx.doi.org/10.1016/j.jclinepi.2014.09.014>.

References

- [1] Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 2009;339:b2680.
- [2] Gøtzsche PC. Reference bias in reports of drug trials. *BMJ* 1987;295:654–6.
- [3] Moscona A. Neuraminidase inhibitors for influenza. *N Engl J Med* 2005;353:1363–73.
- [4] Smith JR, Rayner CR, Donner B, Wollenhaupt M, Klumpp K, Dutkowski R. Oseltamivir in seasonal, pandemic, and avian influenza: a comprehensive review of 10-years clinical experience. *Adv Ther* 2011;28(11):927–59.
- [5] Monto AS. The risk of seasonal and pandemic influenza: prospects for control. *Clin Infect Dis* 2009;48(Suppl 1):S20–5.
- [6] Jones M, Del Mar C. Safety of neuraminidase inhibitors for influenza. *Expert Opin Drug Saf* 2006;5(5):603–8.
- [7] Jefferson T, Jones M, Doshi P, Mar CD. Neuraminidase inhibitors for preventing and treating influenza in healthy adults: systematic review and meta-analysis. *BMJ* 2009;339:b5106.
- [8] Doshi P. Neuraminidase inhibitors—the story behind the Cochrane review. *BMJ* 2009;339:b5164.
- [9] Cohen D, Carter P. WHO and the pandemic flu “conspiracies”. *BMJ* 2010;340:c2912.
- [10] Bax L, Moons KG. Beyond publication bias. *J Clin Epidemiol* 2011;64:459–62.
- [11] Moscati R, Jehle D, Ellis D, Fiorello A, Landi M. Positive-outcome bias: comparison of emergency medicine and general medicine literatures. *Acad Emerg Med* 1994;1(3):267–71.
- [12] Jannot A-S, Agoritsas T, Gayet-Ageron A, Perneger TV. Citation bias favoring statistically significant studies was present in medical research. *J Clin Epidemiol* 2013;66:296–301.
- [13] Nieminen P, Rucker G, Miettunen J, Carpenter J, Schumacher M. Statistically significant papers in psychiatry were cited more often than others. *J Clin Epidemiol* 2007;60:939–46.
- [14] Etter J-F, Stapleton J. Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *J Clin Epidemiol* 2009;62:831–7.
- [15] Egger M, Smith GD. Meta-analysis bias in location and selection of studies. *BMJ* 1998;316:61–6.
- [16] Okike K, Kocher MS, Torpey JL, Nwachukwu BU, Mehlman CT, Bhandari M. Level of evidence and conflict of interest disclosure associated with higher citation rates in orthopedics. *J Clin Epidemiol* 2011;64:331–8.
- [17] Perneger TV. Citation analysis of identical consensus statements revealed journal-related bias. *J Clin Epidemiol* 2010;63:660–4.
- [18] Kulkarni AV, Aziz B, Shams I, Busse JW. Author self-citation in the general medicine literature. *PLoS One* 2011;6(6):e20885.
- [19] Falagas ME, Kavvadia P. “Eigenlob”: self-citation in biomedical journals. *FASEB J* 2006;20(8):1039–42.
- [20] Greenberg SA. Understanding belief using citation networks. *J Eval Clin Pract* 2011;17:389–93.
- [21] Fiorentino F, Vasilakis C, Treasure T. Clinical reports of pulmonary metastasectomy for colorectal cancer: a citation network analysis. *Br J Cancer* 2011;104:1085–97.
- [22] Kho ME, Brouwers MC. The systematic review and bibliometric network analysis (SeBriNA) is a new method to contextualize evidence. Part 1: description. *J Clin Epidemiol* 2012;65:1010–5.
- [23] Kho ME, Brouwers MC. Application of the systematic review and bibliometric network analysis (SeBriNA) methodology contextualizes evidence: part 2: rituximab for non-Hodgkin’s lymphoma. *J Clin Epidemiol* 2012;65:996–1009.
- [24] Trikalinos TA, Moorthy D, Chung M, Yu WW, Lee J, Lichtenstein AH, et al. Concordance of randomized and nonrandomized studies was unrelated to translational patterns of two nutrient-disease associations. *J Clin Epidemiol* 2012;65:16–29.
- [25] Lau AYS, Coiera EW. A Bayesian model that predicts the impact of Web searching on decision making. *J Am Soc Inf Sci Tech* 2006;57(7):873–80.
- [26] Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif Intell Med* 2012;55:197–207.
- [27] Cortes C, Vapnik V. Support-vector networks. *Machine Learn* 1995;20(3):273–97.
- [28] Scholkopf B, Smola A. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA, USA: MIT Press; 2001.
- [29] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; 2006.
- [30] Hand DJ, Yu KM. Idiot’s Bayes—not so stupid after all? *Int Stat Rev* 2001;69(3):385–98.
- [31] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learn* 1997;29(2–3):103–30.
- [32] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theor* 1967;13(1):21–7.
- [33] Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell* 1997;97(1):273–324.
- [34] Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Machine Learn Res* 2003;3:1157–82.
- [35] Mierswa I, Wurst M, Klinkenberg R, Scholz M and Euler T. YALE: rapid prototyping for complex data mining tasks. In: *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Philadelphia, PA, USA: ACM; 2006: 935–40.
- [36] Latourrette M. Toward an explanatory similarity measure for nearest-neighbor classification. Springer Berlin Heidelberg: Machine learning ECML 2000; 2000:238–45.
- [37] Nguyen HT, Fry AM, Loveless PA, Klimov AI, Gubareva LV. Recovery of a multidrug-resistant strain of pandemic influenza A 2009 (H1N1) virus carrying a dual H275Y/I223R mutation from a child after prolonged treatment with oseltamivir. *Clin Infect Dis* 2010;51:983–4.
- [38] Burch J, Paulden M, Conti S, Stock C, Corbett M, Welton N, et al. Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation. *Health Technol Assess* 2009;13:1–265.
- [39] Adisasmito W, Chan PK, Lee N, Oner AF, Gasimov V, Aghayev F, et al. Effectiveness of antiviral treatment in human influenza A (H5N1) infections: analysis of a Global Patient Registry. *J Infect Dis* 2010;202:1154–60.
- [40] Kiso M, Kubo S, Ozawa M, Le QM, Nidom CA, Yamashita M, et al. Efficacy of the new neuraminidase inhibitor CS-8958 against H5N1 influenza viruses. *PLoS Pathog* 2010;6(2):e1000786.
- [41] Tumpey TM, Garcia-Sastre A, Mikulasova A, Taubenberger JK, Swayne DE, Palese P, et al. Existing antivirals are effective against influenza viruses with genes from the 1918 pandemic virus. *Proc Natl Acad Sci U S A* 2002;99:13849–54.
- [42] Krumholz HM, Jackevicius CA, Ross JS. Tamiflu: 14 flu seasons and still questions. *BMJ* 2013;346.
- [43] Jefferson T, Doshi P. Multisystem failure: the story of anti-influenza drugs. *BMJ* 2014;348:g2263.
- [44] Jefferson T, Jones M, Doshi P, Spencer EA, Onakpoya I, Heneghan CJ. Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. *BMJ* 2014;348:g2545.
- [45] Dunn AG, Arachi D, Hudgins J, Tsafnat G, Coiera E, Bourgeois FT. Financial conflicts of interest and conclusions about neuraminidase inhibitors for influenza: an analysis of systematic reviews. *Ann Intern Med* 2014;161:513–8.
- [46] Shotton D. Publishing: open citations. *Nature* 2013;502:295–7.