HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2016 February 04.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2015 November ; 2015: 940–947. doi:10.1109/

BIBM.2015.7359810

A Prototype for a Hybrid System to Support Systematic Review Teams: A Case Study of Organ Transplantation

Tanja Bekhuis^{*,†}, Eugene Tseytlin^{*}, and Kevin J. Mitchell^{*}

Tanja Bekhuis: tcb24@pitt.edu; Eugene Tseytlin: tseytlin@pitt.edu; Kevin J. Mitchell: kjm84@pitt.edu

^{*}Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, USA

[†]Department of Dental Public Health, School of Dental Medicine, University of Pittsburgh, USA

Abstract

We describe a prototype for a hybrid system designed to reduce the number of citations needed to re-screen (NNRS) by systematic reviewers, where citations include titles, abstracts, and metadata. The system obviates the need for screening the entire set of citations a second time, which is typically done to control human error. The reference set is based on a complex review about organ transplantation (N=10,796 citations). Data were split into 50% training and test sets, randomly stratified for percentage eligible citations. The system consists of a rule-based module and a machine-learning (ML) module. The former substantially reduces the number of negative citations passed to the ML module and improves imbalance. Relative to the baseline, the system reduces classification error (5.6% vs 2.9%) thereby reducing NNRS by 47.3% (300 vs 158). We discuss the implications of de-emphasizing sensitivity (recall) in favor of specificity and negative predictive value to reduce screening burden.

1. Introduction

Rapid growth in both the cost of health care and scientific information means that any effort to find out what constitutes best health care is urgent and difficult. Rigorous methods have emerged to find and weigh the evidence in research reports. These methods are the basis for *systematic reviews*. A rigorous systematic review can answer comparative effectiveness questions regarding best therapeutic regimens, drugs, devices, or policies. In principle, the value added of a systematic review (SR) as opposed to a traditional one is that the entire process is transparent, reproducible, and free of bias.

Unfortunately, production is slow and may even fail. For example, Tricco et al report that 19% of the protocols published in the respected Cochrane Library fail to reach fruition as full reviews [1]. Of those that are published as reviews, the average time to completion is 2.4 years. Worse, this estimate ignores time spent exploring the literature to assess significance of possible review questions, and then time spent developing a protocol [2].

The situation is exacerbated by international and national standards meant to ensure production of quality reviews, if followed [3-6]. This is because following methodological standards usually entails carrying out very laborious and often tedious procedures. For example, during the screening phase, it is common for teams to screen thousands of citations (with titles, abstracts, and metadata)—twice—to control human error. To grasp the labor

involved, consider a recent review about information technology for management of medications [7]. It involved screening 40,582 citations. In general, the rate of SR production is inadequate because the research literature is growing exponentially and global demand for reviews is high (Figures 1 and 2) [8].

In this paper, we describe two modules in the Evidence in Documents, Discovery, and Analysis (EDDA) System currently under development. The modules are configured to reduce the number of citations needed to re-screen (NNRS). We combine a rule-based module to filter out negative citations for reports judged as ineligible, followed by a machine-learning module that automates classification of the positive and negative citations that remain. The rationale for this approach is quite different from our earlier work [2, 9, 10] and the work of others in this domain (see the comprehensive review by O'Mara-Eves et al [11]). Instead of focusing on optimal sensitivity (recall), we now focus on identifying the preponderantly large number of negative citations for reports marked as ineligible for full-text review. The negative citations usually swamp the positive ones because the original search strategies are designed to be very sensitive rather than specific [12]. However, once the retrieval set exists, the real task is to reduce the number of citations for a second round of screening, as well as flag inconsistent reviewer judgments for quality control. Note that the labor associated with the first round of screening can be divided among several teammates. In a worst-case scenario, a solo reviewer can screen the citations. In either case, we assume that the input for the EDDA system is a set of citations screened once not twice by humans.

2. Methods

In a comparative study of classifiers, Bekhuis and Demner-Fushman wrote that ignoring class imbalance is a potential problem [2]. If not redressed, imbalance can impair performance of a machine-learning classifier, especially when the majority class (in this case, the negative or ineligible citations) is much larger than the minority class (the positive or eligible citations). Additionally, imbalance in the presence of noise may be particularly problematic [13].

Imbalance and noise are typical of datasets derived from systematic reviewer judgments. Consider, search precision for systematic reviews is about 3% [14]; thus, the negative class represents about 97%. Furthermore, noise from labeling errors is introduced by overly inclusive reviewer judgments. For example, if a citation is 'empty,' i.e., has no abstract, reviewers are likely to label it as positive and then read the full text in the next stage. In this way, they save a study from being discarded too soon.

For this study, we considered randomly undersampling the majority class to improve imbalance, but then confirmation or disconfirmation of negative judgments would be impossible for citations not in the subsample. This is a risky strategy as eligible studies falsely labeled as negative by humans, but correctly labeled by a computer, would be lost. We therefore chose to filter out citations with negative labels based on a suite of logical rules in a preprocessing step, saving the correctly classified negative citations and passing the remaining citations (misclassified citations and all the true positives) to a Bayesian classifier. Thus, the input dataset for machine learning was smaller and better balanced.

We developed a rule-based module for preprocessing and used the same experimental setup for machine learning described in our previous study [9].

2.1. Reference set and features

We used one of the reference datasets that we built (see Bekhuis et al [9]). It was derived from a published SR on organ transplantation and mycophenolic acid [15]. The SR is particularly complex as it entailed 5 key questions. The first addressed whether monitoring mycophenolic acid in serum or plasma in patients who received a solid organ transplant reduced transplant rejections and adverse events. Interestingly, reviewers did not sort the reports by key question until after they screened citations and read full-text reports.

To build the reference set with labels, we re-ran MEDLINE [16] and Embase [17] searches that appeared in the review, limiting to records added no later than the reported search date. This limit precluded retrieving citations for studies that could have been eligible, but were not seen by the reviewers. MEDLINE and Embase searches return citations rather than full texts, which mirrors the experience of the review team. We recovered 93% (10796 / 11642) of the citations screened by the review team. We labeled citations as include (positive) or exclude (negative) based on a flowchart, tables, and reference list appearing in the review. Thus, the labels reflect the consensus judgments of reviewers, each of whom has domain expertise.

We split the data into randomly stratified halves, reserving 50% for training and 50% for testing. Training and test sets each had N=5398 citations; n = 244 or 243 (4.5%) positive citations, respectively. However, for machine-learning *after* rules were applied, the training set had N=1075 citations, n=244 (22.7%) positive citations; the test set had N= 1119 citations, n = 243 (21.7%) positive citations.

2.2. Baseline

We used the test results for the organ transplantation SR reported in [9] as a baseline. However, in that study we averaged over two tests (A|B and B|A), where A and B refer to randomly stratified halves of the data. To be analogous to our test of the suite of rules (see below), we used results from the B|A test, as this represents a test on half the data. Based on the confusion matrix from the B|A test, we computed additional performance metrics, namely, specificity and negative predictive value.

2.3. Rule-based module

We developed logical rules to exclude the negative citations in 10% of the training set (500 negative and 25 positive citations). Subsequent to analyzing errors, we either revised rules or added new ones. We then used the entire training set to check performance for each rule, as well as incrementally to evaluate its added value. If error diagnostics after training suggested further revisions, we used the 10% subset again. This iterative cycling is typical of rapid development and is described by Pustejovsky and Stubbs [18], although they refer to their splits as dev-train, dev-test, and final test. To assess validity of the entire suite of rules, we ran an independent test just once on the held-out test set.

The first author, who is an experienced evaluator of SRs and a methodologist, developed cascading exclusionary rules by analyzing the objectives in the organ transplant SR, as well as excerpts pertaining to eligibility criteria that would have appeared in the protocol, i.e., the information that reviewers would have known when they screened citations. Then she classified the information according to the PICO⁺ model (see below). Domain specific rules covered organ transplantation, blood or serum, mycophenolic acid, physiologic monitoring, and various outcomes. Rules to exclude assumed two forms: (1) if exclusionary evidence exists, then exclude; and (2) if key inclusionary evidence is missing, then exclude. The rules are displayed below Table 1.

To implement rules, we used the Jess Rule Engine (Jess v.71p2). Jess is a scripting environment written in Java by Friedman-Hill at Sandia National Laboratories [19]; it is freely available for academic research. Jess integrates the Java programming environment with a forward-chaining production system.

Rule engines manage both code and data as malleable entities. Data called *Facts* populate working memory and are available for matching. Rules may be added, disabled, or removed dynamically; they assume the form of if-then statements. If the left-hand side of a rule is matched by a subset of data in working memory, the rule fires to transform the data or alter the reasoning path.

For the organ transplantation review, facts in working memory were derived from information stored in a set of categories. Categories correspond to the well-known model for clinical research questions, namely Population (or Patients or Participants), Intervention, Comparator, Outcome, Setting (or site), and Time (PICOST⁺). The plus sign indicates that we enriched the model with categories for study design, publication type, and demographics, information important to most review teams. For the organ transplant SR, categories for setting (S) and time (T) were not relevant. We therefore focused on PICO⁺ categories to guide annotation of citations and information extraction (IE).

The rule system depends on a template and terminology customized for the review topic, which we developed in Protégé v. 4.3.0. The template (IE model) has slots for PICO⁺ categories. Some of the slots are generic and probably useful for most reviews, e.g., age, ethnicity, and study design, whereas others are specific to the review type. For example, intervention and outcome are useful for clinical research questions. We constructed the terminology by pulling relevant branches from publicly available resources, such as MeSH [20], NCI Thesaurus [21], and the EDDA Study Designs and Publications terminology [22]. Together the template and terminology are represented as a Web Ontology Language (OWL) file. The Protégé OWL plugin enables visualization during construction of the customized terminology and debugging of the rules system.

After development of the PICO⁺ terminology for the organ transplantation review, we processed citations using the template to extract terminology matches for each citation. The processor has a simple Java API (PICOExtractor) callable by the Jess Reasoner. The PICOExtractor resolves synonymy and returns preferred terms for each PICO⁺ category.

Additionally, hierarchical information for each term enables resolution of hyponymous relationships.

After application of the suite of rules to exclude negative citations, false positives occur because the final rule is to include by default. However, the set of exclusionary rules significantly reduces the number of negative citations to pass to the machine-learning step and reduces imbalance—the goal of this module.

2.4. Machine-learning module

The input dataset for machine learning after rule-based preprocessing consisted of misclassified citations: false negatives (FNs) and false positives (FPs) plus true positives (TPs). We extracted alphanumeric⁺ features described in our previous research [9] and used the same setup for grid optimization of the parameters for the complement naïve Bayes (cNB) classifier followed by training and testing. The cNB classifier is suitable for imbalanced data [23].

3. Results

3.1 Comparative performance of rules

Because we are concerned with identifying negative citations, the performance metrics of interest are specificity (SP) and negative predictive value (NPV). SP is the true negative rate, i.e., TNs divided by (TNs + FPs). Negative predictive value (NPV) is the ratio of true negatives (TNs) divided by (TNs + FNs). Training performance for rules individually and incrementally, as well as for the test of the entire suite is displayed in Table 1.

Per rule (training)—Median SP = .19 (range: .00 to .54). The number of TNs ranged from 0 for the species rule to 2782 for the physiological monitoring rule. Median NPV = .97 (range: .00 to 1.000). The least accurate rule was for species and the most accurate rule for organ transplantation.

Incremental (training)—The value added by a rule is the difference in SP percentage points between two steps. Value added ranged from 0 % points for the species rule to 22 % points for the blood rule. NPV is high throughout, although it drifts down from .996 in the first step to .964 in the final step. Final SP = .84 is a 3.7-fold improvement over the first step where SP = .23. NPV deteriorated by 3.2 % points between the first and last step. Results for the final step (after all 9 rules had fired) are almost equivalent to the independent test results, which suggest the suite of exclusionary rules is valid for this SR.

Independent test—SP = .83 and 4279 negative citations were correctly classified. The suite of 9 rules was very accurate (NPV = .966).

Resultant dataset for ML module—The input dataset for the ML module test has 5398 - 4279 = 1119 citations, which is a 21 % subset of the initial test set. Because all citations except for the TNs are passed to the ML module, the number of citations labeled as positive remains the same, while imbalance is reduced: 4.5% (243 / 5398) positive citations at baseline vs 21.7% (243 / 1119) for the ML module.

3.2 Test results for baseline, machine-learning module, and hybrid system

Test results are displayed in Table 2.

Relative to the baseline, SP is better for the hybrid system by 3.3 % points (99.07% vs 95.77%; $z=10.56$, $P < .0002$, two-tail). The difference for NPV is not statistically significant (97.89% vs 98.37%; $z = -1.77$, NS). Classification error (CE) is better for the hybrid system (2.93% vs 5.56%); the difference is statistically significant ($z = -6.78$, $P < .0002$, two-tail). The hybrid system reduces the screening burden (NNRS) by 47.3% (158 vs 300 citations).

As an intermediate step, performance for the cNB classifier on the reduced dataset is worse than all baseline measures except for NNRS. Presumably, this is because the citations passed to the ML module are those that rules fail to classify correctly and are harder to classify. However, once the relatively large set of TNs correctly classified by the rule-based module is taken into account, one can see that overall performance is better for the hybrid system relative to the baseline.

Table 3 displays confusion matrices for the baseline (matrix 3a), the ML module (matrix 3b), and the hybrid-system test results (matrix 3c).

4. Discussion

We have demonstrated that a hybrid system can potentially reduce the number of citations needed to re-screen. It is therefore possible to reduce the labor associated with the second round of screening citations.

Moreover, we focused on SP and NPV rather than sensitivity (recall). We reasoned that sensitive search strategies for SRs developed by information professionals already include the studies of interest. Thus, performance metrics suitable for an information retrieval task, such as recall, precision, and the harmonic mean (F), are not appropriate for evaluating the hybrid system we described here.

Given a set of citations with labels as input, we identified the majority of negative citations using a rule-based module to reduce imbalance and improve performance in the machine-learning module. If the hybrid system were in production, the review team would be presented with just 3% of the citations to re-screen. If we assume the test results are our best estimate for a production system, this would mean the reviewers would re-screen $158 \times 2 = 316$ citations identified by the hybrid system vs $300 \times 2 = 600$ citations from the baseline setup with no rule-based preprocessing. (The multiplier accounts for test results based on 50% of the data.)

The question naturally arises as to whether the reduction in the classification error relative to the baseline is worth the extra effort of building a custom terminology and a suite of rules. As biomedical evidence continues to grow exponentially, very large sets of retrieved citations are more likely. However, performance metrics expressed as percentages hide the impact of improved performance; a few percentage points can be meaningful if humans perceive screening the reduced number of citations as less of a burden.

In a production system, assuming the hybrid system generalizes to other types of reviews and is scalable, we could offer both quality assurance and control. By confirming judgments regarding the relatively large set of ineligible citations (TNs), as well as the typically smaller set of eligible citations (TPs), quality assurance is possible. By naming misclassified studies (FNs and FPs), i.e., where the computer disagrees with the review team or solo reviewer, quality control is possible. In other words, the misclassified citations would need to be reconsidered. In the end, the reduction of screening burden during the second pass would be substantial—for this case study, the reduction is about 97%—and the reviewers would receive informative feedback.

Regarding performance of the suite of exclusionary rules, it was somewhat surprising that 9 simple rules could identify 83% of the negative citations. Additionally, the rules appear to be quite valid, as the results on the training and test set were almost equivalent. For this case study, the generic rules for study designs and publication types together found 38% of the negative citations and were very accurate. Note that we retained the generic rule for species even though it had no value in the training set because we were unsure of how it would perform in the test set. To evaluate this rule in future research, we will analyze how many citations are missing information on species. With the exception of the outcome rule, exclusionary rules do not fire unless a PICO⁺ category has information. It is possible that very accurate generic rules, i.e., those that are useful across a broad range of SR types, could be used as a standalone preprocessor for citations *without* labels. However, the set of generic rules would change depending on the eligibility criteria of a given SR. This is an open research question that we intend to address in future research. We also expect to recode the rules in Java. Although Jess is free to academic researchers, its integration could be problematic in a production system.

In the future, we will assess generalizability of the hybrid system to SRs in the EDDA database that vary in type (described in [9]). Additionally, for the hybrid system to scale, we need to further automate customization of the terminology. Finally, the PICOST⁺ model will not be appropriate for some review topics, such as diagnostic test accuracy. Thus, we will implement other models to guide customization of the template.

5. Conclusion

Based on a SR about organ transplantation and mycophenolic acid, the hybrid system described in this case study seems promising. If we can demonstrate generalizability and scalability of the hybrid system in future research, a production system could significantly reduce the labor associated with re-screening citations. It could also offer both quality control and assurance. By flagging misclassified citations, quality control is possible. By confirming a large number of judgments to exclude citations, as well as a relatively small number of judgments to include citations, quality assurance is possible.

Acknowledgments

We are very grateful to Dr. Dina Demner-Fushman, Lister Hill National Center for Biomedical Communications, US National Library of Medicine, for her advice. This research was supported by the US National Library of Medicine of the National Institutes of Health, award number 5R00LM010943. The content is solely the

responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Tricco AC, Brehaut J, Chen MH, Moher D. Following 411 Cochrane protocols to completion: a retrospective cohort study. *PLoS One*. 2008; 3:e3684. [PubMed: 18997866]
2. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif intell Med*. 2012; 55:197–207. [PubMed: 22677493]
3. Institute of Medicine of the National Academies. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: 2011.
4. Chandler J, Churchill R, Higgins J, Lasserson T, Tovey D. Methodological standards for the conduct of new Cochrane Intervention Reviews, v.2.3. Methodological Expectations of Cochrane intervention Reviews (MECIR). 2013
5. Patient-Centered Outcomes Research Institute. PCORI Methodology Standards. Washington, DC: Dec. 2012 p. 1-16.
6. Agency for Healthcare Research and Quality (AHRQ). Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality; 2011.
7. McKibbin KA, Lokker C, Handler SM, Dolovich LR, Holbrook AM, O'Reilly D, Tamblyn R, Hemens BJ, Basu R, Troyan S, Roshanov PS. The effectiveness of integrated health information technologies across the phases of medication management: a systematic review of randomized controlled trials. *J Am Med Inform Assoc*. Jan-Feb;2012 19:22–30. [PubMed: 21852412]
8. 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(9):e1000326. [PubMed: 20877712]
9. Bekhuis T, Tseytlin E, Mitchell KJ, Demner-Fushman D. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS One*. 2014; 9:e86277. [PubMed: 24475099]
10. Bekhuis T, Demner-Fushman D. Towards automating the initial screening phase of a systematic review. *Stud Health Technol Inform*. 2010; 160:146–50. [PubMed: 20841667]
11. 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. [PubMed: 25588314]
12. Lefebvre, C.; Manheimer, E.; Glanville, J. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: Wiley; 2008. Chapter 6: Searching for Studies.
13. Yan Hulse J, Khoshgoftaar T. Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*. 2009; 68:1513–1542.
14. Sampson M, Tetzlaff J, Urquhart C. Precision of health care systematic review searches in a cross-sectional sample. *Research Synthesis Methods*. 2011; 2:119–125. [PubMed: 26061680]
15. Oremus, M.; Zeidler, J.; Ensom, MHH.; Matsuda-Abedini, M.; Balion, C.; Booker, L.; Archer, C.; Raina, P. Evidence Reports/Technology Assessments, No 164 AHRQ Publication No 08-E006. Vol. 2012. Rockville, MD: US Agency for Healthcare Research and Quality; 2008. Utility of monitoring mycophenolic acid in solid organ transplant patients. Available at <http://www.ncbi.nlm.nih.gov/books/INBK38475> [Accessed August 17, 2015]
16. National Center for Biotechnology Information. [Accessed August 6, 2015] PubMed.gov: US National Library of Medicine, National Institutes of Health. <http://www.ncbi.nlm.nih.gov/pubmed/>
17. Elsevier, BV. [Accessed August 6, 2015] Embase. <http://www.embase.com/>
18. Pustejovsky, J.; Stubbs, A. Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications. Sebastopol, CA: O'Reilly Media; 2013.
19. Friedman-Hill, E. JESS in Action. Greenwich, CT: Manning; 2003.
20. US National Library of Medicine at NIH. Bethesda, MD: MeSH: Medical Subject Headings. Available at <http://www.nlm.nih.gov/mesh/> [Accessed August 7, 2015]
21. US National Cancer Institute at NIH. Bethesda, MD: NCI Thesaurus. Available at <http://ncit.nci.nih.gov/> [Accessed August 7, 2015]

22. Bekhuis, T.; Tseytlin, E.; Faith, A. [Accessed August 7, 2015] EDDA Study Designs and Publications, v. 1.3. National Center for Biomedical Ontology (NCBO) BioPortal. Available at <http://bioportal.bioontology.org/ontologies/EDDA>
23. Rennie, J.; Shih, L.; Teevan, J.; Karger, D. Proceedings of the Twentieth International Conference on Machine Learning (ICML). Vol. 20. Washington, DC: 2003. Tackling the poor assumptions of naive Bayes text classifiers.

Figure 1.

Growth in systematic reviews over 25 years (N=233,785 MEDLINE citations). This is an underestimate because reviews not in MEDLINE appear in other repositories, such as Embase. Source: MEDLINE via Pubmed.gov, retrieved May 9, 2015.

Figure 2.
Global production of systematic reviews of biomedical research. Source: gopubmed.com,
retrieved May 9,2015.

Table 1
Performance of rules to exclude negative citations

Train	Per rule		Incremental	
Rule ^a	SP ^b (TN ^c)	NPV ^d	SP (TN)	NPV
1. PT (Publication type)	.23 (1188)	.996	.23 (1188)	.996
2. SD (Study design)	.34 (1760)	.998	.38 (1953)	.995
3. SC (Species)	.00 (0)	.000	.38 (1953)	.995
4. OT (Organ transplantation)	.02 (104)	1.000	.40 (2037) ^e	.996
5. CT (Cell transplantation)	.01 (54) ^e	.931	.40 (2060) ^e	.994
6. MA (Mycophenolic acid)	.01 (43) ^e	.896	.41 (2099)	.992
7. BL (Blood)	.40 (2083)	.970	.63 (3270)	.979
8. PM (Physiological monitoring)	.54 (2782)	.976	.80 (4135)	.975
9. OU (Outcome) ^f	.19 (998)	.916	.84 (4323)	.964
Test			Entire suite of rules	
			.83 (4279)	.966

^a Some study designs are also indexed as publication types and therefore appear in both PT and SD rules;

^b SP = specificity;

^c TN = true negative;

^d NPV = negative predictive value;

^e SP varies when expressed with greater precision;

^f outcomes include patient survival, transplant rejection, adverse events, etc., specified in the published review.

The following exclusionary rules fire in the order presented:

1. If PT NOT null, exclude if PT EQ ((literature review or any of its subclasses) OR (meta-analysis or its subclass) OR abstract OR comment OR editorial OR (letter or its subclass) OR opinion paper)
2. If SD NOT null, exclude if SD EQ ((literature review or any of its subclasses) OR (meta-analysis or its subclass) OR (case study or its subclasses))
3. If SC NOT null, exclude if NOT *Homo sapiens*
4. If population type NOT null, exclude if NOT OT
5. If population type NOT null, exclude if CT
6. If intervention / comparator NOT null, exclude if NOT MA
7. If intervention / comparator NOT null, exclude if NOT BL
8. If intervention / comparator NOT null, exclude if NOT PM
9. If OU NULL, exclude

Table 2
Comparative performance

	SP ^a (%)	NPV ^b (%)	CE ^c (%)	NNRS ^d (FP + FN)
Baseline ML ^e ; cNB ^f classifier; grid optimization of cNB parameters; alphanumeric ⁺ features; based on previously reported test results	95.77	98.37	5.56	300
ML after rules applied Input dataset size reduced; TNs ^g identified by rule-based module are held out from this test	94.52	88.27	14.12	158
Hybrid system Overall performance (ML results for reduced input dataset adjusted by adding in TNs from rule-based module)	99.07	97.89	2.93	158

^a SP=specificity;

^b NPV=negative predictive value;

^c CE=c lassification error;

^d NNRS=number citations needed to re-screen (sum of false positives and false negatives);

^e ML=machine learning;

^f cNB=complement naïve Bayes (Weka classifier);

^g TN=true negative

Table 3**Confusion matrices**

3a. Baseline ML test (based on previously published results)		
TN=4937	FN =82	5019
FP = 218	TP=161	379
5155	243	5398

3b. ML module test (input dataset size reduced by omitting TNs identified by the rule-based module)		
TN=828	FN=110	938
FP = 48	TP=133	181
876	243	1119
3c. Hybrid system (adjusted by adding in TNs to matrix 3b identified by the rule-based module)		
TN=4279+828 = 5107	FN=110	5217
FP = 48	TP=133	181
5155	243	5398