Semi-Automated Identification of Biomedical Literature: A Proof of Concept Study

Gaelen P. Adam (gaelen_adam@brown.edu)
Brown University School of Public Health https://orcid.org/0000-0002-1103-9205
Dimitris Pappas
Institute for Language and Speech Processing
Haris Papageorgiou
Institute for Language and Speech Processing
Evangelos Evangelou
University of Ionnina Medical School
Thomas A. Trikalinos
Brown University School of Public Health

Methodology

Keywords: evidence synthesis, systematic review methods, literature identification, abstract screening, text mining, machine learning

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-560637/v1

Abstract

Background: The typical approach to literature identification involves two discrete and successive steps: (i) formulating a search strategy (i.e., a set of Boolean queries) and (ii) manually identifying the relevant citations in the corpus returned by the query. We have developed a literature identification system (Pythia) that combines the query formulation and citation screening steps and uses modern approaches for text encoding (dense text embeddings) to represent the text of the citations in a form that can be used by information retrieval and machine learning algorithms.

Methods: Pythia incorporates a set of natural-language questions with machine-learning algorithms to rank all PubMed citations based on relevance. Pythia returns the 100 top-ranked citations for all questions combined. These 100 articles are exported, and a human screener adjudicates the relevance of each abstract and tags words that indicate relevance. The “curated” articles are then exploited by Pythia to refine the search and re-rank the abstracts, and a new set of 100 abstracts is exported and screened/tagged, until convergence (i.e., no other relevant abstracts are retrieved) or for a set number of iterations (batches). Pythia performance was assessed using seven systematic reviews (three prospectively and four retrospectively). Sensitivity, precision, and the number needed to read were calculated for each review.

Results: The ability of Pythia to identify the relevant articles (sensitivity) varied across reviews from a low of 0.09 for a sleep apnea review to a high of 0.58 for a diverticulitis review. The number of abstracts that a reviewer had to read to find one relevant abstract (NNR) was lower than in the manually screened project in four reviews, higher in two, and had mixed results in one. The reviews that had greater overall sensitivity retrieved more relevant citations in early batches, but neither study design, study size, nor specific key question significantly affected retrieval across all reviews.

Conclusions: Future research should explore ways to encode domain knowledge in query formulation, possibly by incorporating a "reasoning" aspect to Pythia to elicit more contextual information and leveraging ontologies and knowledge bases to better enrich the questions used in the search.

Background

Evidence derived from systematic reviews and meta-analyses continues to accumulate exponentially. This is due to the continued increase in the rate at which new information is generated (1–3). For example, Williamson and colleagues reported that in 2000, 490,000 new records were added to PubMed, and that number has increased nearly every year since, so that, by 2017, 1,110,000 new records were being added annually (4). Thus, there is imperative to develop novel and efficient methods and processes of literature identification for evidence synthesis purposes.

The typical approach to literature identification involves two discrete and successive steps: (i) formulating a search strategy (normally a combination of a set of Boolean queries) and (ii) human/manual screening of the relevant citations in the corpus returned by the search, preferably by at least two independent reviewers (5). Even though this helps the literature identification process be transparent and replicable, it does not ensure that it is comprehensive. A 2017 study of 58 systematic review searches in multiple databases estimated that less than a quarter of reviews (23%) identify all relevant articles, and only 40% of reviews identify 95% of the relevant citations (6). Additionally, this process leads to a substantial time and cost burden for the review team, which must exclude many irrelevant studies to identify relevant ones.

Increasingly, systematic reviewers are looking to technology to improve this two-step process. A series of tools have been developed that leverage text mining (i.e., measuring the frequency of term/phrase occurrences in a corpus of texts), and machine learning to identify relevant search terms, to structure search strategies, or to classify retrieved abstracts as relevant or not (7–9). In this paper, we describe the development and evaluation of a system (Pythia) that (i) unifies the search query formulation and citation screening steps and (ii) uses modern deep learning and information retrieval methods to increase the efficiency and effectiveness of literature identification for systematic reviews.

Methods

Pythia takes a set of natural-language questions—derived by splitting the review’s key questions (i.e., research questions) into their component phrases—and uses modern approaches for semantic text encoding to represent the text of the citations in a form that can be used by deep neural networks.

Pythia returns the top-ranked citations for each question amounting to a total of 100 citations (e.g., if there are two questions, it selects the top 50 for each question; if there are 10 questions, it selects only the top 10 for each question). A citation can appear in the results for more

than one question. A human then screens the 100 citations, annotating specific terms that indicate relevance in the relevant abstracts by the specific aspect of the eligibility criteria (see Fig. 1 for an example of an annotated abstract). Based on these screening decisions and the annotations of relevant terms, Pythia refines its search and returns the next top 100 unscreened citations, to be screened/tagged, until convergence is achieved (i.e., no other relevant abstracts are retrieved) or for a set number of iterations (batches). For this project, we limited each project to 10 batches.

Creating the Dataset

The literature collection searched by Pythia was constructed using the metadata of all PubMed records (ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline/). For each article, we concatenated the title and abstract of each citation and indexed the resulting concatenated text, along with the publication year of the paper. We discarded any citation without an abstract, leaving a subset of approximately 21 million articles (of the original 31 million articles in PubMed).

Selecting the Citations for Screening

Pythia uses the natural-language questions provided by the review team to produce a list of 100 potentially relevant citations. This set is screened by a human, who annotates each citation as relevant or irrelevant and tags words or phrases as indicative of relevance based on the eligibility criteria, using the Population, Intervention, Comparator, Outcome (PICO) framework. Pythia then extracts a set of positive key phrases from the articles annotated as included and a set of negative key phrases from the articles annotated as rejected. Each key phrase is executed as a search, and Pythia retrieves 200 articles, using the BM25 retrieval algorithm (10). Any abstracts that had been previously retrieved are excluded, and the retrieved articles are ranked using a deep-learning model designed to retrieve both citations and snippets (11). Pythia is designed to penalize the score of each article containing a negative key-phrase and increase the score of any article containing a positive key-phrase, thus taking into account the feedback provided by users. Finally, Pythia returns the top 100 citations for human screener evaluation.

Experimental Setup

The first 100 articles are exported and screened manually, with the human screener adjudicating the relevance of each abstract and tagging keywords that indicate relevance by the specific aspect of the eligibility criteria in those abstracts that are deemed potentially relevant. These “curated” articles are used by Pythia to refine the search and re-rank the full corpus of abstracts. Pythia then exports a new set of 100 top-ranked abstracts to be screened and tagged. For this project, we chose to limit each review to 10 cycles of article export and manual screening, representing 1000 citations screened per review. Because the same citation might be identified (and screened) for more than one key question, the total number of unique citations ranged from 800 to 1,000 across projects.

Evaluation

To evaluate the Pythia, we selected a convenience sample of seven systematic reviews on a variety of clinical topics undertaken by the Brown Evidence-based Practice Center in the last five years (12–18). At the time the project began, three reviews were completed and four were ongoing. By the time of final analysis, all reviews were completed, though one was still undergoing peer review (15). All reviews followed the methods set out in the EPC Methods Guide (19). The search strategies for all seven reviews were conducted by a trained medical librarian and peer reviewed, using the 2015 PRESS assessment form (20). After a series of pilot round to ensure consistency, the review team double-screened all retrieved citations for relevance.

In three of these reviews, Pythia was used prospectively, with a human annotator screening 10 batches of 100 citations each. Pythia reranked all articles in the database after each batch was screened and provided another set of 100 for screening.

In the other four reviews, Pythia was evaluated retrospectively, with automatic annotation based on labels given to the abstracts screened in the original review's manual screening. This created a list of known positive and negative articles. To estimate the importance of the human-generated PICO tags in the prospective annotation, the machine automatically extracted "key phrases" from relevant and irrelevant articles to use in the analysis. We examined three settings, namely “no key phrases”, “positive key phrases only”, and “positive and negative key phrases”.

- In the “no key phrases” setting, Pythia did not use the key phrases of the citations in ranking articles.
- In the “positive key phrases only” setting, Pythia used only the positive key phrases to increase the score of all articles that share a key phrase with a known positive article (key phrases extracted from known relevant citations).
- In the “positive and negative key phrases” setting, Pythia penalized any citation that shared a key phrase with a known negative article (key phrases extracted from known irrelevant citations) and increased the score of any citation that shares a key phrase with a known

positive article.

With this design, we may have missed relevant articles that were not included in the provided list (i.e., reviewers of the original review never saw that article). Therefore, we expect that the retrospective scores would improve with human inspection. The studies screened as relevant were compared to the studies with PubMed identifiers (PMIDs) included in the final report. Any studies screened in through the prospective process, but not included in the original report searches, were assessed for eligibility by the original report's primary investigator. None was found to be eligible.

Performance Measures

For each systematic review, the final included citations that had a PubMed Identifier (PMID), indicating that they could be found in PubMed, were considered to be the reference standard (T). This set was divided by whether they were identified by Pythia (TP) or not (FN). The citations identified by Pythia (P) were divided into subgroups by whether they were included in the final report (TP) or not (FP). We omitted the number of citations correctly rejected (TN) because this number is extremely large (the source set included approximately 21 million citations) and is not of particular interest.

We were interested in two dimensions of classification performance: workload and sensitivity (i.e., recall). Sensitivity was defined as the ability to identify the truly relevant citations using Pythia (TP/T). To measure the workload, we defined precision as the proportion of citations screened that were relevant (TP/P). To make this number more intuitive, we use NNR, defined as the number of irrelevant citations that the reviewer had to screen for each relevant citation found (1/Precision). Our aim was to maximize sensitivity while minimizing workload.

For comparison, we report precision and NNR for the manual screening process as defined by the number of relevant articles included in the final report as a proportion of the total number of articles retrieved in the PubMed searches for each review. This does not include double-screening, so each abstract was only counted once. The sensitivity for the manual screening process by definition is 100%. Where P values are reported, they were calculated using the Fisher exact test.

Results

Details of the seven recently completed systematic reviews are presented in Table 1. All reviews were produced for the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Program and followed the AHRQ Methods Guide (19); all were published between the years 2017 and 2021. The review topics included a range of healthcare domains. Initial search sizes ranged from 1,738 citations (642 in PubMed) to 15,813 citations (9,741 in PubMed).

Table 1
Included Datasets

Prospective/ Retrospective	Systematic Review Topic and Title	Brief Description	Domain	Citations screened in full review, N (N from PubMed)	Screened in at title/ abstract/ keyword level (% of total screened)	Screened in based on full text (% of total screened); N (%) from PubMed	No. KQs	No. NLQs
Prospective	**Diverticulitis**: Management of Colonic Diverticulitis (13)	Benefits and harms of nonsurgical interventions for acute diverticulitis, imaging tests to diagnose acute diverticulitis, colorectal cancer screening, surgical and nonsurgical interventions to prevent recurrence	Gastroenterology	15,199 (7,981)	722 (4.75)	88 (0.6); 86 (1.1)	4	14
Prospective	**Sleep Apnea** : Continuous Positive Airway Pressure Treatment for Obstructive Sleep Apnea (15)	Long-term health outcomes with CPAP treatment and assesses the validity of surrogate measures (e.g., AHI) for clinical outcomes.	Sleep Medicine	15,333 (10,891)	1,593 (10.4)	71 (0.5); 70 (0.6)	2	7
Prospective	**Headaches in Pregnancy** : Management of Primary Headaches in Pregnancy (17)	Pharmacologic and nonpharmacologic interventions to prevent or treat attacks of primary headaches in women who are pregnant, attempting to become pregnant, postpartum, or breastfeeding.	Obstetrics	8,154 (6,587)	400 (4.9)	72 (0.8); 64 (9.7)	2	6
Retrospective	**Nonmelanoma Skin Cancer** : Treatments for Basal Cell and Squamous Cell Carcinoma of the Skin (16)	Comparative effectiveness and safety of all currently used therapeutic strategies for Basal and Squamous Cell Carcinoma	Oncology	15,813 (9,741)	534 (3.4)	125 (0.8); 78 (0.8)	2	2
Retrospective	**Tympanostomy Tubes**: Tympanostomy Tubes in Children With Otitis Media (18)	Effectiveness and harms of tympanostomy tubes in children with chronic otitis media with effusion and recurrent acute otitis media; necessity for water precautions in children with tympanostomy tubes; and treatments for otorrhea	Pediatrics	8,498*	509 (6.0)*	176 (2.0)*	5	8

KQ = key question; NLQ = natural language question. Numbers may vary slightly from those given in the final report due to the stage of the report at the time of this analysis.

* The PubMed search was done separately for this review, so only PubMed numbers are given.

Prospective/ Retrospective	Systematic Review Topic and Title	Brief Description	Domain	Citations screened in full review, N (N from PubMed)	Screened in at title/ abstract/ keyword level (% of total screened)	Screened in based on full text (% of total screened); N (%) from PubMed	No. KQs	No. NLQs
Retrospective	**Urinary Incontinence** : Nonsurgical Treatments for Urinary Incontinence in Women: A Systematic Review Update (12)	Comparative effectiveness and harms of nonpharmacological and pharmacological interventions for women with all forms of urinary incontinence	Urogynecology	7,840 (3,706)	723 (9.2)	244 (3.1); 96 (2.6)	4	24
Retrospective	**Venous Thromboembolism**: Venous Thromboembolism Prophylaxis in Major Orthopedic Surgery: Systematic Review Update (14)	Comparative effectiveness of different thromboprophylaxis interventions for patients undergoing major orthopedic surgery	Orthopedic surgery	1,738 (981)	455 (26.2)	56 (3.2); 53 (8.4)	6	8
KQ = key question; NLQ = natural language question. Numbers may vary slightly from those given in the final report due to the stage of the report at the time of this analysis.								
* The PubMed search was done separately for this review, so only PubMed numbers are given.								

Sensitivity and Precision/NNR

Table 2 provides the overall sensitivity, precision, and NNR for each review. The sensitivity of Pythia varied across reviews from 0.09 for the headaches in pregnancy review to 0.58 for the review on diverticulitis. For three reviews, the precision/NNR was better than the standard procedure of separate search and manual screening (diverticulitis, nonmelanoma skin cancer, and tympanostomy), in one review (sleep apnea) Pythia performed better than manual screening in terms of precision but the confidence intervals overlap. In one review (venous thromboembolism) manual screening had better precision/lower NNR than Pythia, and in another (headaches in pregnancy) the precision was better but the confidence intervals overlap. In one review (urinary incontinence), the sensitivity and precision/NNR were affected by whether or not Pythia was given key phrases. The search that used positive key phrases only favored Pythia, but the ones that used positive and negative key phrases favored manual screening. Sensitivity ranged from 0.28 for Positive key phrases only to 0.11 for no key phrases. We saw similar patterns across all four retrospective reviews, with the positive key phrases only searches having the best performance. This suggests that the tool performed better when it pulled key phrases from relevant citations and used those to increase the weight of citations that had those phrases.

Table 2
Sensitivity and precision/NNR.

Review	Iteration	Sensitivity (95% CI)	Precision (95% CI)	NNR (95% CI)	Precision for entire search* (95% CI)	NNR for entire search* (95% CI)
Prospective						
Diverticulitis	Overall	0.57 (0.46, 0.68)	0.05 (0.04, 0.07)	19 (14, 25)	0.01 (0.01, 0.01)	93 (75, 116)
HIP	Overall	0.09 (0.04, 0.19)	0.01 (0.00, 0.01)	155 (71, 421)	0.01 (0.01, 0.01)	104 (81, 135)
Sleep Apnea	Overall	0.16 (0.08, 0.26)	0.01 (0.01, 0.02)	80 (45, 160)	0.01 (0.01, 0.01)	156 (123, 199)
	Part 1	0.07 (0.02, 0.16)	0.01 (0.00, 0.03)	80 (34, 244)		
	Part 2	0.10 (0.04, 0.20)	0.01 (0.01, 0.03)	72 (35, 177)		
Retrospective						
NMSC	Positive and negative key phrases**	0.31 (0.21, 0.42)	0.03 (0.02, 0.05)	30 (20, 47)	0.01 (0.01, 0.01)	125 (100, 158)
	Positive key phrases only***	0.40 (0.29, 0.51)	0.04 (0.03, 0.06)	23 (17, 34)		
	No key phrases	0.24 (0.15, 0.35)	0.03 (0.02, 0.04)	38 (24, 63)		
Tymp.	Positive and negative key phrases**	0.48 (0.41, 0.56)	0.09 (0.07, 0.11)	11 (9, 14)	0.02 (0.02, 0.02)	48 (42, 56)
	Positive key phrases only***	0.48 (0.40, 0.55)	0.09 (0.07, 0.11)	11 (9, 14)		
	No key phrases	0.48 (0.41, 0.56)	0.09 (0.07, 0.11)	11 (9, 14)		
UI	Positive and negative key phrases**	0.22 (0.14, 0.31)	0.02 (0.01, 0.03)	46 (30, 74)	0.03 (0.02, 0.03)	39 (32, 48)
	Positive key phrases only***	0.28 (0.19, 0.38)	0.03 (0.02, 0.04)	36 (25, 54)		
	No key phrases	0.11 (0.06, 0.20)	0.01 (0.01, 0.02)	87 (49, 174)		
VTE	Positive and negative key phrases**	0.40 (0.26, 0.54)	0.02 (0.01, 0.03)	46 (30, 74)	0.05 (0.04, 0.07)	19 (14, 25)
	Positive key phrases only***	0.47 (0.33, 0.61)	0.03 (0.02, 0.04)	38 (26, 59)		
	No key phrases	0.38 (0.25, 0.52)	0.02 (0.01, 0.03)	48 (31, 78)		

NNR = number needed to read; HIP = headaches in pregnancy; NMSC = nonmelanoma skin cancer; Tymp. = tympanostomy tubes; UI = urinary incontinence; VTE = venous thromboembolism.

*includes citations from PubMed only; some of the citations with PMIDs may have been found through other databases, but most were probably identified in the PubMed search.

**key phrases extracted from the titles and abstracts of relevant and irrelevant citations and incorporated into the search.

*** key phrases extracted from the titles and abstracts of relevant citations only and incorporated into the search.

Sensitivity and Precision by Iteration (Batch)

Figure 2 shows the percent of relevant articles found across iterations. The pattern across the three settings of the retrospective reviews ("positive and negative key phrases", "no key phrases", and "positive key phrases only") is similar, so only the results for "positive key phrases only" are shown, as it had the best overall performance. The figure shows that across batches the cumulative relevant identified citations increased as a proportion of the total relevant studies in the final report. This suggests that the better performing reviews found more relevant studies in early batches. Some reviews (such as diverticulitis and tympanostomy) showed a pattern of leveling out after the first several batches, but others showed a steadier rise or one that started later (such as the venous thromboembolism and nonmelanoma skin cancer reviews, respectively). For these reviews, more batches may have yielded a higher sensitivity.

Other factors affecting sensitivity

We found that neither study design, study size, nor key question (specific topic) statistically significantly affected sensitivity in all reviews (Appendix: Tables 1 through 5), but in specific reviews there were some statistically significant predictors of sensitivity. There is an indication that Pythia is more likely to find randomized controlled trials (RCTs) than other study designs, as the percentage of RCTs found is higher across most reviews, but for only one review (tympanostomy) was this difference statistically significant. There was also an indication that the type of question may affect retrieval. For three reviews, we found a statistically significant difference across the report's key questions. In the diverticulitis review, this was driven by a large percentage of the antibiotic treatment studies (key question 2); Pythia identified 19 of the 25 studies included for that key question (76%) and the colonoscopy studies (key question 3; 16 of 19 total studies; 84%) compared to those for other questions. In the tympanostomy tubes review, the difference was driven by the retrieval of a very high percentage of the treatment of otorrhea studies (13 studies identified of 14 total; 93%). Finally, in the VTE review, Pythia identified a large percentage of the efficacy (key question 5) studies (9 studies identified of 11 total; 82%) compared to those for other questions.

Concept Maps and Natural Language Queries

Because of the generally low sensitivity of the Pythia, we created concept maps to represent the implicit domain knowledge that a reviewer brings to the search and screening process (Appendix: Figs. 1 and 2). These maps make it clear that the natural language questions we relied on in this project are unlikely to be sufficiently descriptive to ensure adequate sensitivity. This is because the computer cannot replicate the process by which a human translates the key questions into a series of Boolean queries, which involves an understanding of the domain and conceptualization of topics. Without this baseline understanding and contextual knowledge, Pythia could not produce enough relevant citations to allow the machine learning algorithms to identify all relevant articles in the corpus.

Discussion

In the traditional method, a human (usually a trained medical librarian) creates a search strategy based on the systematic review's population and intervention (or exposure), often with other concepts that may include outcomes, study designs, language, location, and other key insights. The librarian identifies a comprehensive set of synonyms and controlled vocabulary terms for each concept of interest and then combines them into one or more queries using Boolean logic. These queries are then manually executed in each database, the results are deduplicated, and members of the review team double-screen each unique citation.

In this project, we sought to combine searching and screening into a single process, thereby saving time and increasing the likelihood that all citations would be identified. After testing Pythia prospectively and retrospectively across several completed systematic review projects, we found that while the burden was decreased in most reviews, the sensitivity of Pythia to retrieve the relevant abstracts was unstable and generally low, ranging from 8 to 58 percent. We believe that this result is, at least in part, due to the translation and formulation of the key questions as natural language questions, where large amounts of domain knowledge are implicit. An expert can easily infer this knowledge, but machines cannot adequately encode and parse it. Future research should explore ways to encode this domain knowledge in query formulation, possibly by incorporating a "reasoning" aspect to Pythia to elicit more contextual information and leveraging ontologies, such as the Unified Medical Language System (21), and knowledge bases, electronic resources that contain curated information from data repositories (22), to better enrich the questions used in the search. Automating and optimizing the process of translating the key research questions to natural language questions by exploiting deep seq2seq models is another option to pursue.

The evaluation of Pythia was limited by the small number of prospective reviews (3) and the fact that the prospective analyses included labels by only a single screener, which may have reduced sensitivity. In all three reviews, two to five relevant citations identified by the tool were screened out by the single human screener. These omissions appear to be at random (there is no pattern as to the types of studies mislabeled) but may, nevertheless, have affected the machine learning algorithm, likely reducing sensitivity. This is particularly true in the reviews with only a small number of relevant studies, such as headaches in pregnancy. The retrospective analyses (n = 4) were also limited

by the fact that we could provide labels for only the abstracts that had also been screened by the original review team. Future evaluations should include more prospective reviews, as well as double screening and consensus adjudication of conflicts to better reflect real-world practice.

This work builds on ongoing research in the automation of specific steps in evidence synthesis, specifically the steps involved in identifying relevant literature. In their 2015 systematic review of text mining for systematic review literature identification, O'Mara-Eves et al. evaluated 44 studies that reported on text mining tools for reducing screening workload. They concluded that the field is rapidly evolving and that text mining and machine learning approaches have a role to play in the reduction of screening burden without a large sacrifice in recall (8). However, to our knowledge, this is one of the first projects to leverage text mining and machine learning to look at combining the searching and screening steps. Due to its overall low sensitivity, mixed improvements in precision, and overall variability across projects, Pythia is not ready for widespread use for any comprehensive synthesis. However, these results suggest that there is potential for a system like this to aid in systematic review search development, aiding in the efficient generation of a set of relevant citations that can be used in search strategy design. Further refinement and revisions of Pythia may yield a system with better or more consistent performance across topics and questions or a way to establish when the system is performing well and therefore may be useful.

Conclusions

Two widely used guides for systematic review methodology, the Cochrane Handbook (23) and the AHRQ Methods Guide (19), recommend that searches be as comprehensive as possible within time and budget constraints. As the body of literature to be screened increases, tools that leverage machine learning have become increasingly useful in prioritizing screening based on the likelihood of relevance. The system described in this report (Pythia) aims to use active machine learning technology to combine the search and screening steps of systematic review, thereby improving comprehensiveness and reducing screening burden. Based on the findings of our evaluation, this technology has promise. Future work should focus on improving recall, specifically in terms of improving the ability of the system to parse and contextualize natural language queries.

Abbreviations

AHRQ Agency for Healthcare Research and Quality

EBM Evidence-based Medicine.

FN false negative

HIP Headaches in Pregnancy

T Total included

KT key terms

NMSC Nonmelanoma Skin Cancer

NNR number needed to read

P positive

PMID PubMed identifier

RCT randomized controlled trial

TP true positive

UI Urinary Incontinence

VTE Venous Thromboembolism

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The annotation tool is available at: https://inception-project.github.io/. We deployed it in ATHENA RC Clarin: https://www.clarin.gr/en

Competing interests: The authors declare that they have no competing interests

Funding: This work was funded under contract R03 HS027247-01, Agency for Healthcare Research and Quality, US Department of Health and Human Services. The authors of this article are responsible for its content. Statements in the article do not necessarily represent the official views of or imply endorsement by Agency for Healthcare Research and Quality or the Department of Health and Human Services. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funder.

Authors' Contributions: TT, HP, VE conceptualized and designed the system. DP implemented it. GA and VE tested it. GA computed analyses and wrote the manuscript. All authors read and commented on the manuscript.

Acknowledgements: Not applicable

References

1. Elliott JH, Mavergames C, Becker L, Meerpohl J, Thomas J, Gruen R, et al. The efficient production of high quality evidence reviews is important for the public good. Bmj. 2013;346:f846.
2. Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. Bmj. 2013;346:f139.
3. Wallace BC, Dahabreh IJ, Schmid CH, Lau J, Trikalinos TA. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. J Comp Eff Res. 2013;2(3):273–82.
4. Williamson PO, Minter CIJ. Exploring PubMed as a reliable resource for scholarly communications services. J Med Libr Assoc. 2019;107(1):16–29.
5. Institute of Medicine Committee on Standards for Systematic Reviews of Comparative Effectiveness R. In. Eden J, Levit L, Berg A, Morton S, editors Finding What Works in Health Care: Standards for Systematic Reviews. Washington (DC): National Academies Press (US). Copyright. 2011 by the National Academy of Sciences. All rights reserved.; 2011.
6. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017;6(1):245.
7. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA, editors. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; 2012: ACM.
8. 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.
9. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform. 2014;51:242–53.
10. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. Nist Special Publication Sp. 1995;109:109.
11. Pappas D, McDonald R, Brokos G-I, Androutsopoulos I, editors. AUEB at BioASQ 7: document and snippet retrieval. Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2019: Springer.
12. Balk E, Adam GP, Kimmel H, Rofeberg V, Saeed I, Jeppson P, et al. AHRQ Comparative Effectiveness Reviews. Nonsurgical Treatments for Urinary Incontinence in Women: A Systematic Review Update. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018.
13. Balk EM, Adam GP, Cao W, Danko K, Bhuma MR, Mehta S, et al. AHRQ Comparative Effectiveness Reviews. Management of Colonic Diverticulitis. Rockville (MD): Agency for Healthcare Research and Quality (US); 2020.
14. Balk EM, Ellis AG, Di M, Adam GP, Trikalinos TA. AHRQ Comparative Effectiveness Reviews. Venous Thromboembolism Prophylaxis in Major Orthopedic Surgery: Systematic Review Update. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.
15. Balk ET, Protocol T: Continuous Positive Airway Pressure Treatment for Obstructive Sleep Apnea in Medicare Eligible Patients. 2020.
16. Drucker A, Adam GP, Langberg V, Gazula A, Smith B, Moustafa F, et al. AHRQ Comparative Effectiveness Reviews. Treatments for Basal Cell and Squamous Cell Carcinoma of the Skin. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.
17. Saldanha IJ, Roth JL, Chen KK, Zullo AR, Adam GP, Konnyu KJ, et al. AHRQ Comparative Effectiveness Reviews. Management of Primary Headaches in Pregnancy. Rockville (MD): Agency for Healthcare Research and Quality (US); 2020.
18. Steele D, Adam GP, Di M, Halladay C, Pan I, Coppersmith N, et al. AHRQ Comparative Effectiveness Reviews. Tympanostomy Tubes in Children With Otitis Media. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.

19. Relevo R, Balshem H. Finding evidence for comparing medical interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews; 2011.
20. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. J Clin Epidemiol. 2016;75:40–6.
21. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med. 1993;32(4):281.
22. Gerdesköld C, Toth-Pal E, Wårdh I, Nilsson GH, Nager A. Use of online knowledge base in primary health care and correlation to health care quality: an observational study. BMC Med Inform Decis Mak. 2020;20(1):294.
23. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf M-I, et al. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6. 0: Cochrane; 2019.

Figures

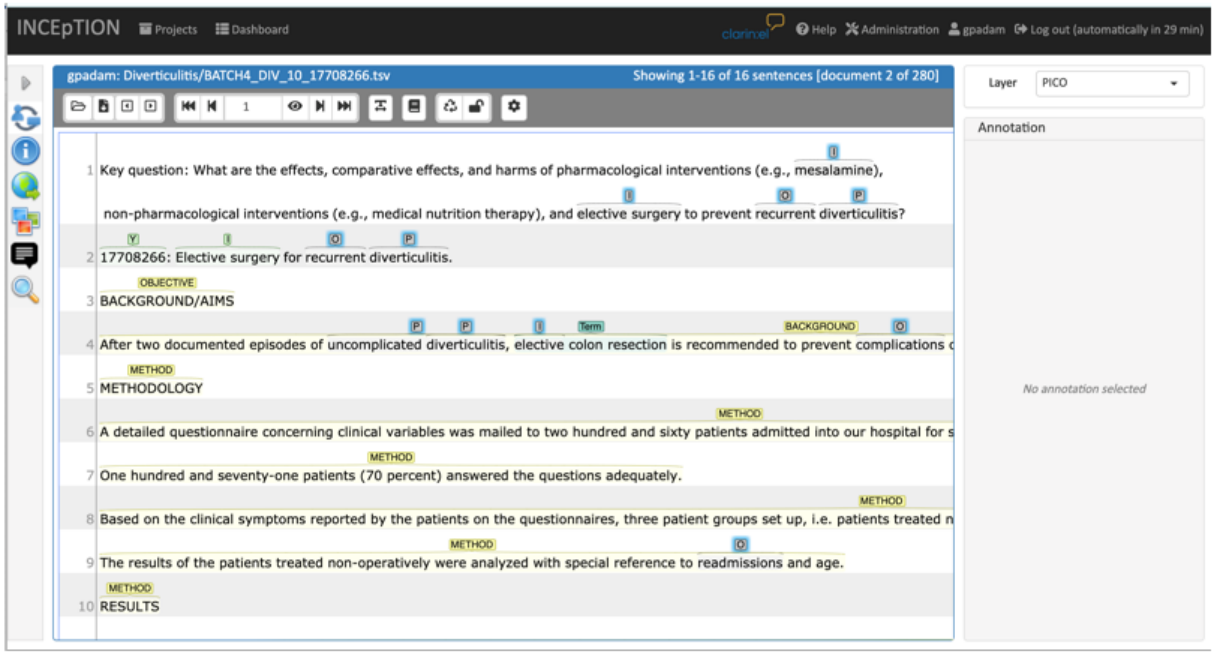

Figure 1

Example of an annotated abstract

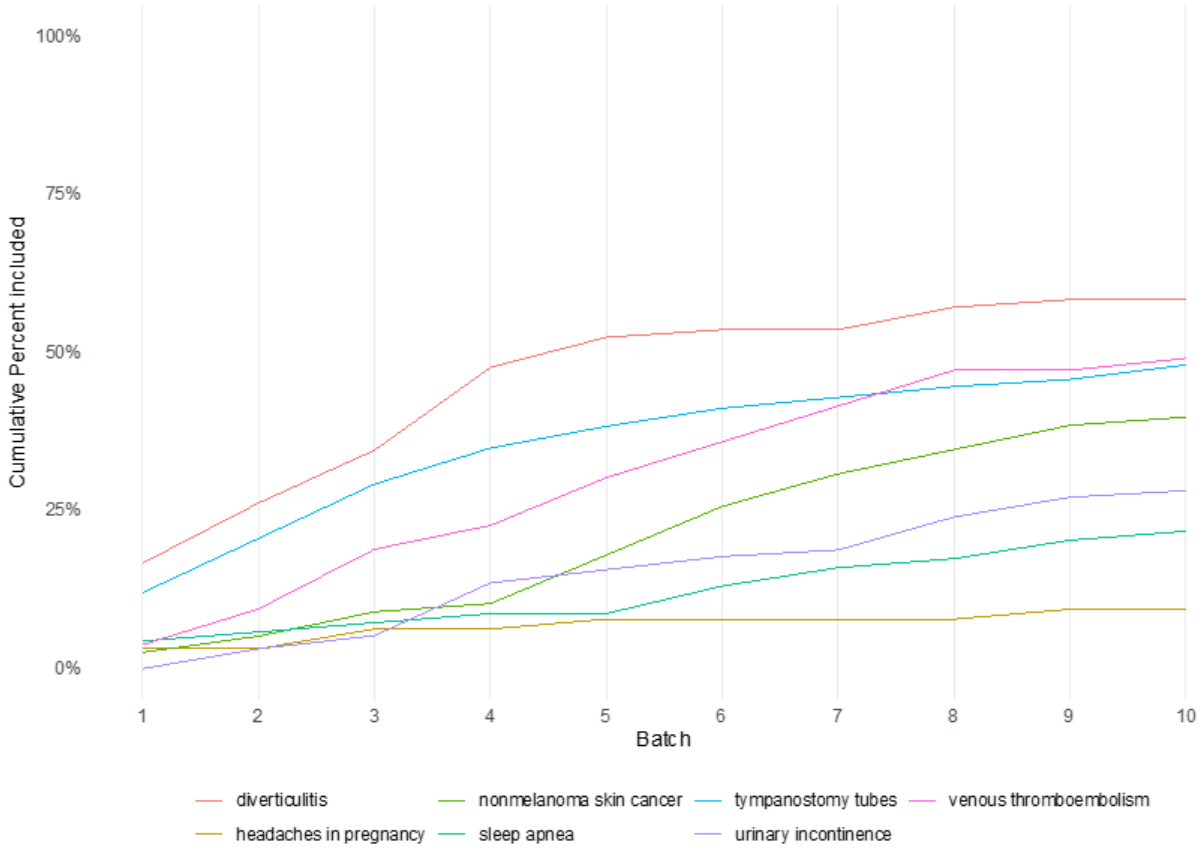

Figure 2

Cumulative sensitivity for each project by iteration (batch).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Semiautomation.additional.files.docx