A Research Protocol for a Systematic Review of Automatic Literature Screening in Medical Evidence Synthesis

Yuelun Zhang
Peking Union Medical College Hospital https://orcid.org/0000-0001-7990-9003
Siyu Liang
Chinese Academy of Medical Sciences and Peking Union Medical College
Yunying Feng
Chinese Academy of Medical Sciences and Peking Union Medical College
Qing Wang
Tsinghua University
Feng Sun
Peking University Health Science Centre
Shi Chen
Peking Union Medical College Hospital
Yiying Yang
Chinese Academy of Medical Sciences and Peking Union Medical College
Xin He
Chinese Academy of Medical Sciences and Peking Union Medical College
Huijuan Zhu
Peking Union Medical College Hospital
Hui Pan (panhui20111111@163.com)
Department of Endocrinology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 1 Shuaifuyuan, Dongcheng District, Beijing, China

---

Protocol

Keywords: Evidence-based practice, artificial intelligence, natural language process, protocol, systematic review, diagnostic test accuracy

Posted Date: August 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-62316/v1

Abstract

Background: Systematic review is an indispensable tool for optimal evidence collection and evaluation in evidence-based medicine. However, the explosive increase of the original literatures makes it difficult to accomplish critical appraisal and regular update. Artificial intelligence (AI) algorithms have been applied to automate the literature screening procedure in medical systematic reviews. In these studies, different algorithms were used and results with great variance were reported. It is therefore imperative to systematically review and analyse the developed automatic methods for literature screening and their effectiveness reported in current studies.

Methods: An electronic search will be conducted using PubMed, Embase and IEEE Xplore Digital Library databases, as well as literatures found through supplementary search in Google scholar, on automatic methods for literature screening in systematic reviews. Two reviewers will independently conduct the primary screening of the articles and data extraction, in which nonconformities will be solved by discussion with a methodologist. Data will be extracted from eligible studies, including the basic characteristics of study, the information of training set and validation set, the function and performance of AI algorithms, and summarised in a table. The risk of bias and applicability of the eligible studies will be assessed by the two reviewers independently based on Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2). Quantitative analyses, if appropriate, will also be performed.

Discussion: Automating systematic review process is of great help in reducing workload in evidence-based practice. Results from this systematic review will provide essential summary of the current development of AI algorithms for automatic literature screening in medical evidence synthesis, and help to inspire further studies in this field.

Registration: PROSPERO registration number CRD42020170815 (28 April 2020).

Background

Systematic review synthesizes the results of multiple original publications to provide clinicians with comprehensive knowledge and current optimal evidence in answering certain research questions. Major steps of a systematic review are: defining a structured review question, developing inclusion criteria, searching in the databases, screening for relevant studies, collecting data from relevant studies, assessing the risk of bias critically, undertaking meta-analyses where appropriate, and assessing reporting biases.[1–3] Systematic review aims to provide a complete, exhaustive summary of current literature relevant to a research question with an objective and transparent approach. In the light of these characteristics, systematic reviews, in particular those combining high quality evidence, which used to be at the very top of the medical evidence pyramid[4] and now become regarded as an indispensable tool for evidence viewing,[5] are widely used by reviewers in the practice of evidence-based medicine.

However, conducting systematic reviews for clinical decision making is time-consuming and labour-intensive, as the reviewers are supposed to perform a thorough search to identify any literatures that may

be relevant, read through all abstracts of searched literatures, and identify the potential candidates for further full-text screening.[6] For original researches, the median time from the publication to their first inclusion in a systematic review ranged from 2.5 to 6.5 years.[7] It usually takes over a year to publish a systematic review from the time of literature search.[8] However, advances in clinical research are likely to make these evidences be out of date within several years. With the explosive increase of original research articles, reviewers have found difficulty identifying most relevant evidence in time, let alone updating systematic reviews periodically.[9] Therefore, researchers are exploring automatic methods to improve the efficacy of evidence synthesis while reducing the workload on systematic reviews.

Recent progresses in computer science show a promising future that more intelligent works can be accomplished with the aid of automatic technologies, such as pattern recognition and machine learning. Being seen as a subset of artificial intelligence (AI), machine learning utilizes algorithms to build mathematical models based on training data in order to make predictions or decisions without being explicitly programmed.[10] Various machine learning studies have been introduced in the medical field, such as diagnosis, prognosis, genetic analysis, and drug screening, to support clinical decision making.[11–14] When it comes to automatic methods for systematic reviews, models for automatic literature screening have been explored to reduce repetitive work and save time for reviewers.[15,16]

To date, limited researches have been focused on automatic methods used for biomedical literature screening in systematic review process. Automated literature classification systems[17] or hybrid relevance rating models[18] were tested in specific datasets, yet further extension of review datasets and performance improvement are required. To address this gap in knowledge, this article describes the protocol for a systematic review aiming at summarizing existing automatic methods to screen relevant biomedical literature in the systematic review process.

Methods

Objectives

The primary objective of this review is to assess the diagnostic accuracy of AI algorithms (index test) compared with gold-standard human investigators (reference standard) for screening relevant literatures from original literatures identified by electronic search in systematic review. The secondary objective of this review is to describe the time and work saved by AI algorithms in literature screening. Additionally, we plan to conduct subgroup analyses to explore the potential factors that associate with the accuracy of AI algorithms.

Study registration

We prepared this protocol following the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P).[19] This systematic review has been registered on PROSPERO (Registration number: CRD42020170815, 28 April 2020).

Review question

Our review question was refined using PRISMA-DTA framework, as detailed in Table 1. In this systematic review, “literatures” refer to the subjects of the diagnostic test (the “participants” in Table 1), and “studies” refer to the studies included in our review.

Table 1
Review question.

Item	Description
“Participants”*	Original publications and literatures identified by electronic literature search
Index test	Automatic literature screening models using artificial intelligence algorithms
Reference standard	Traditional literature screening by human investigators
Outcome	Primary outcome: diagnostic accuracy, measured by sensitivity, specificity, precision, NPV, PPV, NLR, PLR, DOR, F-measure, accuracy, and AUC of automatic literature screening models Secondary outcomes: labour and time saving, mainly evaluated by the percentage of searched literatures that the reviewers do not have to read (because they have been screened out by the automatic literature screening models)
*The “participants” in our review refer to the original publications and literatures identified in a systematic literature search, rather than human participants or patients in traditional systematic reviews.	
Abbreviations: *AUC, area under curve; DOR, diagnostic odds ratio; NLR, negative likelihood ratio; NPV, negative predictive value; PLR, positive likelihood ratio; PPV, positive predictive value*.	

Inclusion and exclusion criteria

We will include studies in medical research that reported a structured study question, described the source of the training or validation sets, developed or employed AI models for automatic literature screening, and used the screening results from human investigators as the reference standard.

We will exclude traditional clinical studies in human participants, editorials, commentaries or other non-original reports. Pure methodological studies in AI algorithms without application in evidence synthesis will be excluded as well.

Information source and search strategy

An experienced methodologist will conduct searches in three major public electronic medical and computer science databases, including PubMed, Embase and IEEE Xplore Digital Library, for publications ranged from January 2000 to present. We set this time range because to the best of our knowledge, AI algorithms prior to 2000 are unlikely to be applicable in evidence synthesis.[20] In addition to the literature search, we will also find more relevant studies through checking the reference lists of studies identified by

electronic search. Related abstracts and preprints will be searched in Google scholar. There are no language restrictions in searches. We will use both free text words and MeSH/EMTREE terms to develop strategies related to three major concepts: systematic review, literature screening, and AI. Multiple synonyms for each concept will be incorporated into the search. Details of the search strategies are shown in Table 2.

Table 2
Search strategy.

Concept	Search terms
Systematic review	#1 ("medical evidence" OR PICO OR PECODR OR "intervention arms" OR "experimental methods" OR "study design parameters" OR "Patient oriented Evidence" OR "eligibility criteria" OR "evidence based medicine" OR "clinically important elements" OR "evidence based practice" OR "results from clinical trials" OR "research results" OR "clinical evidence" OR "Meta Analysis" OR "Clinical Research" OR "medical abstracts" OR "clinical trial literature" OR "clinical trial characteristics" OR "clinical trial protocols" OR "clinical practice guidelines" OR "systematic review")
Literature screening	#2 (extract* OR classif* OR identif* OR retriev* OR detect* OR judg* OR determin* OR decid* OR sort* OR infer* OR interpret* OR includ* OR exclud* OR filter OR filtering OR select*)
Artificial intelligence	#3 ("Artificial Intelligence" OR "natural language" OR "language processing" OR "Knowledge Acquisition" OR "Knowledge Representation" OR "Support Vector Machine" OR svm OR Gaussian OR Bayes OR Bayesian OR "Cluster" OR Clustering OR "Hidden Markov" OR "conditional random field" OR "Random Forest" OR (Graphical AND model) OR Regression OR "feature engineering" OR "zero-shot learning" OR "few-shot learning" OR "reinforcement learning" OR "transfer learning" OR (unsupervised OR supervised OR semi-supervised OR distant-supervised OR self-supervised) OR "neural network" OR "neural networks" OR (neural AND algorithm*) OR (neural AND machine) OR (network AND algorithm*) OR (network AND machine) OR (automatic AND network) OR (automatic AND networks) OR (automatic AND algorithm*) OR (automatic AND model) OR (automatic AND models) OR (automatic AND machine) OR (automatic AND learning) OR (automatic AND method) OR (learning AND network) OR (learning AND networks) OR (learning AND algorithm*) OR (learning AND machine) OR (learning AND method) OR (deep AND network) OR (deep AND networks) OR (deep AND algorithm*) OR (deep AND model) OR (deep AND models) OR (deep AND machine) OR (deep AND learning))
Combined concepts	#1 AND #2 AND #3

Abbreviations: *SVM, support vector machine.*

Study selection

Literatures with titles and abstracts from online electronic databases will be downloaded and imported into EndNote X9.3.2 software (Thomson Reuters, Toronto, Ontario, Canada) for further process after removing duplications.

All studies will be screened independently by 2 authors based on the titles and abstracts. Those which do not meet the inclusion criteria will be excluded with specific reasons. Disagreements will be solved by discussion with a methodologist if necessary. After the initial screening, the full texts of the potentially

relevant studies will be independently reviewed by the two authors to make decisions on final inclusions. Conflicts will be resolved in the same way as they were initially screened. Excluded studies will be listed and noted according to PRISMA-DTA flowchart.

Data collection

A data collection form will be used for information extraction. Data from the eligible studies will be independently extracted and verified by two investigators. Disagreements will be resolved through discussion and consultation with the original publication. We will also try to contact the authors to collect the missing data. If one study did not report detailed accuracy data or did not provide enough data that are essential to calculate the accuracy data, this study will be omitted from the quantitative data synthesis.

The following data will be extracted from the original studies: characteristics of study, information of training set and validation set, the function and performance of AI algorithms. The definitions of variables in data extraction are shown in Table 3.

Table 3
Definitions of variables in data extraction.

Variable	Definitions
Study characteristics	
Year	Year of publication
Authors	Last name of authors
Study type	Article, abstract, or systematic review
Journal, conference	Name of journal or conference
Training set information	
Training set	Name of dataset used for training
Area	General medicine, detailed disease, or specific intervention
Source	Name of electronic databases searched for building training set
Time range	Time range of training set
Type of publication	Abstract, or full-text
Number of all literatures	Number of all literatures in training set
Number of included literatures	Number of included literatures identified by the step of screening in training set
Training method	Supervised, semi-supervised, or unsupervised
Validation set information	
Validation set	Name of dataset used for validation
Area	General, disease, or intervention
Source	Name of electronic database searched for building validation set
Time range	Time range of validation set
Type of publication	Abstract, or full-text
Number of all literatures	Number of all literatures in validation set
Number of included literatures	Number of included literatures identified by the step of screening in validation set
Golden standard	Process of screening by human investigators

Abbreviations: *AUC, area under curve; DOR, diagnostic odds ratio; NLR, negative likelihood ratio; NPV, negative predictive value; PLR, positive likelihood ratio; PPV, positive predictive value.*

Variable	Definitions
AI algorithm information	
Model name	Name of model
Model type	Classification, regression, ranking or others
Model performance	Including but not limited to sensitivity, specificity, precision, NPV, PPV, NLR, PLR, DOR, F-measure, accuracy, and AUC
Cost saving	Decreased number of screened literatures by human investigators

Abbreviations: *AUC, area under curve; DOR, diagnostic odds ratio; NLR, negative likelihood ratio; NPV, negative predictive value; PLR, positive likelihood ratio; PPV, positive predictive value.*

Risk of bias assessment, applicability, and levels of evidence

Two authors will independently assess risk of bias and applicability with a checklist based on Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).[21] The QUADAS-2 contains 4 domains, respectively regarding patient selection, index test, reference standard, and flow and timing risk of bias. The risk of bias is classified as “low”, “high”, or “unclear”. Studies with high risk of bias will be excluded in the sensitivity analysis.

In this systematic review, the “participants” are literatures rather than human subjects. The index test is AI model used for automatic literature screening. Therefore, we will slightly revise the QUADAS-2 to fit our research context (Table 4). We deleted one signal question in the QUADAS-2 “was there an appropriate interval between index test and reference standard”. The purpose of this signal question in the original version of the QUADAS-2 is to judge the bias caused by the change of disease status between the index test and the reference test. The “disease status”, or the final inclusion status of one literature in our research context, will not change, thus there is no such concerns.

Table 4
The revised QUADAS-2 tool for risk of bias assessment.

Domains	Signal questions	Answers
"Patient" (literature) Selection	**Risk of Bias**	
	Was a consecutive or random sample of literatures enrolled	Yes/No/Unclear
	Was a case-control design avoided	Yes/No/Unclear
	Did the study avoid inappropriate exclusions	Yes/No/Unclear
	Could the selection of literatures have introduced bias	Low/High/Unclear Risk
	Concerns regarding applicability	
	Is there concern that the included literatures do not match the review question	Low/High/Unclear Risk
Index Test (AI algorithms in literature screening)	**Risk of Bias**	
	Were the index test results interpreted without knowledge of the results of the reference standard	Yes/No/Unclear
	If a threshold was used, was it pre-specified	Yes/No/Unclear
	Could the conduct or interpretation of the index test have introduced bias	Low/High/Unclear Risk
	Concerns regarding applicability	
	Is there concern that the index test, its conduct, or interpretation differ from the review question	Low/High/Unclear Risk
Reference Standard (results of screening by human investigators)	**Risk of Bias**	
	Is the reference standard likely to correctly classify the target condition	Yes/No/Unclear
	Were the reference standard results interpreted without knowledge of the results of the index test	Yes/No/Unclear
	Could the reference standard, its conduct, or its interpretation have introduced bias	Low/High/Unclear Risk
	Concerns regarding applicability	
	Is there concern that the target condition as defined by the reference standard does not match the review question	Low/High/Unclear Risk
Flow and Timing	**Risk of Bias**	

Domains	Signal questions	Answers
	Did all literatures receive a reference standard	Yes/No/Unclear
	Did literatures receive the same reference standard	Yes/No/Unclear
	Were all literatures included in the analysis	Yes/No/Unclear
	Could the literature flow have introduced bias	Low/High/Unclear Risk

The levels of the evidence body will be evaluated by the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) framework.[22]

Diagnostic accuracy measures

We will extract the data of per study in a two-by-two contingency table from the formal publication text, appendices, or by contacting the main authors to collect sensitivity, specificity, precision, negative predictive value (NPV), positive predictive value (PPV), negative likelihood ratio (NLR), positive likelihood ratio (PLR), diagnostic odds ratios (DOR), F-measure, and accuracy with 95% CI. If the outcomes cannot be formulated in a two-by-two contingency table, we will extract the reported performance data. If possible, we will also assess the area under the curve (AUC), as the two-by-two contingency table may not be available in some scenarios.

Qualitative and quantitative synthesis of results

We will qualitatively describe the application of AI in literature screening. If there were adequate details and homogeneous data for the quantitative meta-analysis, we will combine the accuracy of AI algorithms in literature screening using the random-effects Rutter-Gatsonis hierarchical summarised receiver operating characteristic curve (HSROC) model which was recommended by the Cochrane Collaboration for combining the evidence for diagnostic accuracy.[23] The effect of threshold will be incorporated in the model in which heterogeneous thresholds among different studies will be allowed. The combined point estimates of accuracy will be retrieved from the summarised receiver operating characteristic curve (ROC).

Subgroup analyses and meta-regression will be used to explore the between-study heterogeneity. We will explore the following predefined sources of heterogeneity: (1) AI algorithm type, (2) study area of validation set (targeted specific diseases, interventions, or a general area); (3) searched electronic databases (PubMed, EMBASE, or others), (4) proportion of eligible to original studies (the number of eligible literature identified in the screening step divided by the number of original literature identified during the electronic search). Furthermore, we will analyse the possible sources of heterogeneity from both dataset and methodological perspectives in HSROC as covariates following the recommendations from the Cochrane Handbook for Diagnostic Tests Review.[23] We regarded the factor as a source of heterogeneity if the coefficient of the covariate in the HSROC model was statistically significant. We will

not evaluate the reporting bias (e.g. publication bias) since the hypothesis underlying the commonly used methods, such as funnel plot or Egger's test, may not be satisfied in our research context. Data were analysed using R software, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) with two-tailed probability of type I error of 0.05 ($\alpha = 0.05$).

Discussion

Systematic review has developed rapidly within the last decades and plays a key role in enabling the spread of evidence-based practice. Systematic review, though costing less than primary research in money expenditure, is still time-consuming and labour-intensive. Conducting systematic review begins with electronic database searching for a specific research question, then at least two reviewers read each abstract of searched records to identify potential candidate literatures for full-text screening. Only 2.9% searched records are relevant and included in the final synthesis on average,[24] typically, reviewers have to find the proverbial needle in the haystack of irrelevant titles and abstracts. Computational scientists have developed various algorithms for automatic literature screening. Developing an automatic literature screening instrument will be source-saving and improve the quality of systematic review by liberating reviewers from repetitive work. In this systematic review, we aim to describe the development process and algorithms used in various AI literature screening systems, in order to build a pipeline for the update of existing tools and creation of new models.

The accuracy of automatic literature screening instruments varied widely in different algorithms and review topics.[17] The automatic literature screening systems can reach a sensitivity as high as 95%, despite at the expense of specificity, since reviewers try to include every publication relative to the topic of review. As the automatic systems may have a low specificity, it is also important to evaluate how much reviewing work the reviewers can saved in the step of screening. We will not only assess the diagnostic accuracy of AI screening algorithms compared with human investigators, but also collect the information of work saved by AI algorithms in literature screening. Additionally, we plan to conduct subgroup analyses to identify potential factors that associate with the accuracy and efficacy of AI algorithms.

As far as we know, this will be the first systematic review to evaluate AI algorithms for automatic literature screening in evidence synthesis. Few systematic reviews have focused on the application of AI algorithms in medical practice. The literature search strategies in previous published systematic reviews rarely use specific algorithms as search terms. Most of them generally use words such as "artificial intelligence", "machine learning" in strategies, which may lose the studies that only reported one specific algorithms. In order to include AI-related studies as much as possible, our search strategy contained all of the AI algorithms commonly used in the past 50 years, and it was reviewed by an expert in machine learning. The process of literature screening can be assessed under the framework of the diagnostic test. Findings from this proposed systematic review will provide comprehensive and essential summary of the application of AI algorithms for automatic literature screening in evidence synthesis. The proposed systematic review may also help to improve and promote the automatic methods in evidence synthesis in the future by locating and identifying the potential weakness in the current AI models and methods.

List Of Abbreviations

Abbreviations	Full term
AI	Artificial intelligence
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
PRISMA-P	Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols
SVM	Support vector machine
GRADE	Grading of Recommendations, Assessment, Development and Evaluations
NPV	Negative predictive value
PPV	Positive predictive value
NLR	Negative likelihood ratio
PLR	Positive likelihood ratio
DOR	Diagnostic odds ratio
AUC	Area under the curve
HSROC	Hierarchical summarised receiver operating characteristic curve
ROC	Receiver operating characteristic curve

Declarations

Ethics approval and consent to participate

This research is exempt from ethics approval because the work is carried out on published documents.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This study will be supported by the Undergraduate Innovation and Entrepreneurship Training Program (Number 202010023001). The sponsors have no role in study design, data collection, data analysis, interpretations of findings, and decisions for dissemination.

Authors' contributions

H Pan conceived this research. This protocol was designed by YL Zhang, SY Liang, and YY Feng. YY Yang, X He, Q Wang, F Sun, S Chen, and HJ Zhu provided critical suggestions and comments on the manuscript. YL Zhang, SY Liang, and YY Feng wrote the manuscript. All authors read and approved the final manuscript. H Pan is the guarantor for this manuscript.

Acknowledgements

We thank Professor Siyan Zhan (Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, siyan-zhan@bjmu.edu.cn) for her critical comments in designing this study.

References

1. Higgins J, Thomas J, Chandler J, et al. Cochrane Handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. *Reference Source.* 2020.
2. Mulrow CD, Cook D. *Systematic reviews: synthesis of best evidence for health care decisions.* ACP Press; 1998.
3. Armstrong R, Hall BJ, Doyle J, Waters E. 'Scoping the scope' of a cochrane review. *Journal of Public Health.* 2011;33(1):147-150.
4. Paul M, Leibovici L. Systematic review or meta-analysis? Their place in the evidence hierarchy. *Clin Microbiol Infect 2014 Feb;20(2):97-100 doi: 101111/1469-069112489.* 2014(1469-0691 (Electronic)):97-100.
5. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evidence Based Medicine.* 2016;21(4):125.
6. Bigby M. Evidence-Based Medicine in a Nutshell: A Guide to Finding and Using the Best Evidence in Caring for Patients. *Archives of Dermatology.* 1998;134(12):1609-1618.
7. Bragge P, Clavisi O, Turner T, Tavender E, Collie A, Gruen RL. The global evidence mapping initiative: scoping research in broad topic areas. *BMC medical research methodology.* 2011;11(1):92.
8. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. *Journal of clinical epidemiology.* 2008;61(6):531-536.
9. Shojania K, Sampson M, Ansari M, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med 2007 Aug 21;147(4):224-33 doi:.* 2007(1539-3704 (Electronic)):224-233.
10. Bishop CM. *Pattern recognition and machine learning.* springer; 2006.

11. Wang L-y, Chakraborty A, Comaniciu D. Molecular diagnosis and biomarker identification on SELDI proteomics data by ADTBoost method. Paper presented at: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference2006.
12. Cetin MS, Houck JM, Vergara VM, Miller RL, Calhoun V. Multimodal based classification of schizophrenia patients. Paper presented at: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)2015.
13. Sun Y, Loparo K. Information Extraction from Free Text in Clinical Trials with Knowledge-Based Distant Supervision. Paper presented at: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)2019.
14. Li M, Lu Y, Niu Z, Wu F-X. United complex centrality for identification of essential proteins from PPI networks. *IEEE/ACM transactions on computational biology and bioinformatics.* 2015;14(2):370-380.
15. Whittington C, Feinman T, Lewis SZ, Lieberman G, Del Aguila M. Clinical practice guidelines: Machine learning and natural language processing for automating the rapid identification and annotation of new evidence. *Journal of Clinical Oncology.* 2019;37.
16. Turner MD, Chakrabarti C, Jones TB, et al. Automated annotation of functional imaging experiments via multi-label classification. *Frontiers in Neuroscience.* 2013(7 DEC).
17. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association.* 2006;13(2):206-219.
18. Rúbio TR, Gulo CA. Enhancing academic literature review through relevance recommendation: Using bibliometric and text-based features for classification. Paper presented at: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI)2016.
19. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ.* 2015;350:g7647.
20. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev.* 2015;4:78.
21. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529-536.
22. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008;336(7650):924-926.
23. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 09 0 London: The Cochrane Collaboration.* 2010.
24. Sampson M, Tetzlaff J, Urquhart C. Precision of healthcare systematic review searches in a cross-sectional sample. *Research Synthesis Methods.* 2011;2(2):119-125.

Supplementary Files

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

- PRISMAPchecklist.pdf