

Education for Information: Interdisciplinary Journal of Information Studies, 2020, 36(1):101-105. DOI: 10.3233/EFI-190347

Identifying empirical studies for mixed studies reviews: The mixed filter and the automated text classifier

Reem El Sherif^a, Alexis Langlois^b, Xiao Pandu^b, Jian-Yun Nie^b, James Thomas^c, Quan Nha Hong^c and Pierre Pluye^a

- a. Department of Family Medicine, McGill University, Montréal, QC, Canada
- b. Recherche appliquée en linguistique informatique, Université de Montréal, Montréal, QC, Canada
- c. EPPI-Centre, Department of Social Science, UCL Institute of Education, University College London, England, UK

ABSTRACT

Mixed studies reviews include empirical studies with diverse designs (qualitative, quantitative and mixed methods). To make the process of identifying relevant empirical studies for such reviews more efficient, we developed a mixed filter that included different keywords and subject headings for quantitative (e.g., cohort study), qualitative (e.g., focus group), and mixed methods studies. It was tested for six journals from three disciplines. We measured precision (proportion of retrieved documents being relevant), sensitivity (proportion of relevant documents retrieved), and specificity (proportion of non-relevant documents not retrieved). Records were coded before applying the filter and compared with retrieved records, and descriptive statistics were performed, suggesting the mixed filter has high sensitivity, but lower precision and specificity (close to 50%). Next, based on the success of the filter, we developed an automated text classification system that can automatically select empirical studies in order to facilitate systematic mixed studies reviews. Several algorithms were trained and validated with 8,050 database records that were previously manually categorized. Decision trees had the best results and surpassed the accuracy of the filter by 30% when using full-text documents. This algorithm was then adapted into an online format that can be used by researchers to analyze their bibliography and categorize records into “empirical” and “nonempirical”.

Keywords: Bibliographic database, systematic review, information retrieval, empirical research, search filter, automated classifier

1. MIXED STUDIES REVIEW

In mixed studies reviews, diverse empirical research (qualitative, quantitative, and mixed methods) is reviewed concurrently, to develop a breadth and depth of understanding and corroboration of scientific knowledge (Pluye & Hong, 2014; Pluye et al., 2016). Mixed studies reviews can address complex research questions, and are therefore becoming increasingly popular in all health disciplines (Shaw et al., 2014). There has been significant methodological advancement of mixed studies reviews in the last decade, and a toolkit for researchers designing, conducting and reporting systematic mixed studies reviews has been developed and is accessible in an open-access format (Pluye et al., 2018).

As in other reviews, the first key step of a mixed studies review is the identification of potentially relevant studies in bibliographic databases. However, due to the high number of potentially irrelevant scientific publications that may be retrieved, this may be a very time consuming and labour intensive step for reviewers (Bjork et al., 2009; Gough et al., 2012; Jinha, 2010). While there are search strategies for retrieving some specific designs such as randomized controlled trials, there is no search filter to retrieve common types of empirical papers (i.e., papers reporting qualitative, quantitative and mixed methods studies) for mixed studies reviews. Our goal, therefore, was to develop a filter and, subsequently, an online tool that facilitates the identification of empirical records for information professionals, researchers, students, and educators that are conducting mixed studies reviews.

2. BIBLIOGRAPHIC DATABASE FILTER: THE MIXED FILTER

Our search of the literature yielded no search filter to retrieve studies with diverse qualitative, quantitative and mixed methods research designs. Thus, three health librarians reviewed the literature on bibliographic database filters and proposed a mixed filter consisting of a collection of search terms to identify common empirical studies in bibliographic databases. This mixed filter included a combination of keywords and subject headings for quantitative (e.g., cohort study), qualitative (e.g., focus group), and mixed methods. It was developed for Ovid MEDLINE, adapted for different bibliographic databases (Embase, PsycINFO, and CINAHL), and pilot tested in a systematic mixed studies review (Pluye et al., 2019). It is available for free via the 'Identify potential relevant studies' page of the mixed studies reviews wiki (<http://toolkit4mixedstudiesreviews.pbworks.com>) (Pluye et al., 2018).

We then evaluated the performance of this filter in terms of sensitivity, specificity and precision (El Sherif et al., 2016). The mixed filter was tested in six journals from three disciplines that include studies with complex research questions and diverse research designs: Primary Care, Medical Informatics, and Public Health and Epidemiology. We selected two journals from each discipline, one

with a high impact factor and thus a higher proportion of more frequently cited empirical research, and one with a lower impact factor. We focused on articles published between 2008 and 2013, to ensure we obtained a manageable sample of database records with at least 250 records per journal. For each journal, the primary author coded database records as empirical (relevant) when they described a research question or objective, data collection, analysis, and results. The mixed filter was then applied to each journal and the author identified how many of the empirical records were retrieved or missed. Descriptive statistics were performed, and we measured precision (proportion of retrieved relevant documents), sensitivity (proportion of relevant documents retrieved), and specificity (proportion of non-relevant documents not retrieved).

The overall performance across all six journals suggested a high sensitivity of 89.5% but lower precision and specificity, respectively of 60.4% and 54.5%. This is very promising: sensitivity is key for systematic mixed studies reviews where the goal is to achieve a comprehensive and exhaustive retrieval of database records (retrieve almost all relevant records). The results of this project indicate that the mixed filter is useful for conducting a mixed studies review.

3. DEVELOPMENT AND PERFORMANCE OF AN AUTOMATED TEXT CLASSIFICATION

Automated text classification is a method that automatically classifies texts into pre-defined categories (Sebastiani, 2002). It has been explored in systematic reviews and shown to reduce the time needed to screen records by more than 50% without any loss of relevant studies (Thomas, 2013). This can allow the identification of potential relevant studies using algorithms, and screening potential relevant studies (Thomas et al., 2011). A recent systematic review examining the use of text mining in the screening of records for systematic reviews concluded that it can reduce the workload by between 30% and 70% (O'Mara-Eves et al., 2015).

We used the mixed filter as a baseline to develop an automated text classification system that can automatically classify empirical studies in order to facilitate mixed studies reviews. To test the performance of this system several algorithms were trained and validated with 8,050 database records that were previously manually categorized (Langlois et al., 2018). The efficiency of each of the algorithms was measured using sensitivity, precision, specificity and accuracy. Decision trees had the best results and surpassed the accuracy of the filter by 30% when using full-text documents. Results also showed that selection of relevant features can be improved by mixing observable terms with concepts from a meta-thesaurus.

4. ATCER: THE AUTOMATED TEXT CLASSIFIER OF EMPIRICAL RESEARCH

Based on the above-mentioned results, the most performant algorithm (Decision trees) was used to implement an online classifier that can be used by students and researchers to categorize records from their bibliography into “empirical” and “nonempirical”. The online tool, the Automated Text Classifier of Empirical Research (ATCER), analyzes the titles and abstracts of the bibliography and provides a “probability” percentage for the record being empirical or not empirical (Fig. 1). It is freely available online: <https://babel.iro.umontreal.ca>. By default, records are deemed ‘empirical’ when the result is 50% and above, and ‘non-empirical’ when the result is below 50%, which saves half of the selection-related time/resource. However, the user can decide to modify the cut-off threshold depending on the number of retrieved records, available resources and timeline of their mixed studies review.

The usability of ATCER was pilot tested and reviewed with the help of six researchers with experience conducting systematic mixed studies reviews using an existing test bibliography or their own bibliography. All uploaded bibliographies are stored in a secured server and will be used for continuous improvement of the algorithm. A future study will explore the performance of ATCER using a larger sample of records to suggest cut-off thresholds for the probability of a study being empirical.

5. CONCLUSION

To our knowledge, no other tool exists to reduce the workload and increase efficiency of the study selection step of mixed studies reviews. As information professionals are encouraged and expected to participate in these reviews, we believe the mixed filter and ATCER may provide concrete help during the process. We encourage researchers and students to use these tools while conducting their reviews and provide constructive feedback on how they can be improved.

ACKNOWLEDGEMENTS

We would like to acknowledge health librarians who developed the mixed filter: Genevieve Gore, Francesca Frati and Vera Granikov. Reem El Sherif holds a Doctoral Research Award from the Canadian Institute of Health Research (CIHR). Quan Nha Hong holds a Postdoctoral Research Bursary from the ‘Fonds de recherche du Québec – Santé’ (FRQS). Pierre Pluye holds a Senior Research Scholarship Award from the FRQS. The development and implementation of ATCER were sponsored by the Method Development platform of the Quebec SPOR SUPPORT Unit.


REFERENCES

Bjork, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: yearly volume and open access availability. *Information Research: An International Electronic Journal*, 14(1).

- El Sherif, R., Pluye, P., Gore, G., Granikov, V., & Hong, Q. N. (2016). Performance of a mixed filter to identify relevant studies for mixed studies reviews. *Journal of the Medical Library Association: JMLA*, 104(1), 47.
- Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London. Sage.
- Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3), 258-263.
- Langlois, A., Nie, J. Y., Thomas, J., Hong, Q. N., & Pluye, P. (2018). Discriminating between empirical studies and nonempirical works using automated text classification. *Research Synthesis Methods*, 9(4), 587-601.
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1), 5.
- Pluye, P., El Sherif, R., Granikov, V., Hong, Q. N., Vedel, I., Galvao, M. C. B. et al. (2019). Health outcomes of online consumer health information: A systematic mixed studies review with framework synthesis. *Journal of the Association for Information Science and Technology*.
- Pluye, P., & Hong, Q. N. (2014). Combining the Power of Stories and the Power of Numbers: Mixed Methods Research and Mixed Studies Reviews. *Annual Reviews of Public Health*, 35, 29-45.
- Pluye, P., Hong, Q. N., Bush, P., & Vedel, I. (2016). Opening-up the definition of systematic literature review: the plurality of worldviews, methodologies and methods for reviews and syntheses. *Journal of Clinical Epidemiology*, 73, 2-5.
- Pluye, P., Hong, Q. N., Granikov, V., & Vedel, I. (2018). The wiki toolkit for planning, conducting and reporting mixed studies reviews. *Education for Information*, 34(4), 277-283.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.
- Shaw, R. L., Larkin, M., & Flowers, P. (2014). Expanding the evidence within evidence-based healthcare: thinking about the context, acceptability and feasibility of interventions. *Evidence Based Medicine*, ebmed-2014-101791.
- Thomas, J. (2013). Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation? *OA Evidence-Based Medicine*, 1(2), 12.
- Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1), 1-14.

TABLES AND FIGURES

Fig. 1. Screenshot from the results in the ATCER website.

 **UNITÉ DE SOUTIEN**
SRAP QUÉBEC

PUBLICATION
CLASSIFIER

Home Analyze About | Sign Out [REEM] FR

Show 10 entries

Search:

Title ↑↓	Category ↑↓	EMP-Prob ↑↓	NON-EMP ↑↓	Time ↑↓
Erratum: Correlations of Family Medicine Clerkship Evaluations and Objective Structured Clinical Examination Scores and Residency Director's Ratings (Family Medicine (1999) 31:2 (90-94))	Nonempirical	4.82	95.18	2018-09-12 @ 12:35
The prevalence and management of diabetes in Thai adults: The International Collaborative Study of Cardiovascular Disease in Asia	Nonempirical	46.56	53.44	2018-09-12 @ 12:35
Development of underwater and hyperbaric medicine as a medical specialty in Turkey	Nonempirical	6.06	93.94	2018-09-12 @ 12:35
Structured clinical teaching strategy	Empirical	92.85	7.15	2018-09-12 @ 12:35
The inadequacy of musculoskeletal knowledge after foundation training in the United Kingdom	Empirical	82.85	17.15	2018-09-12 @ 12:35
The watched structure clinical examination (WASCE) as a tool of assessment	Empirical	88.11	11.89	2018-09-12 @ 12:35