APPLICATION OF TEXT MINING TO THE DEVELOPMENT OF A GEOGRAPHIC SEARCH FILTER TO FACILITATE EVIDENCE RETRIEVAL IN OVID MEDLINE

Popoff E,[1] Cheung A,[1] Szabo SM[1] [1] Broadstreet HEOR, Vancouver, Canada

BROADSTREET

Background

- Text mining is a potentially valuable technique for analyzing large unstructured datasets to identify meaningful patterns.
- A recent application of text mining was in machine learning algorithms developed to classify abstracts in order to automate systematic literature reviews.[1]
- Given the increasing volume of published research in bibliographic databases like MEDLINE, efficient retrieval of relevant evidence is crucial and represents an opportunity to integrate text mining tools.

Objective

This study aimed to develop and validate a geographic search filter for accurately identifying research from the United States (U.S.) in Ovid MEDLINE.

Methods

- U.S. and non-U.S. citations with a valid PUBMED ID were collected from bibliographies of reviews by the U.S. Preventive Services Task Force, which publishes evidence-based recommendations in various disease areas.
- U.S. citations were defined as having:
  - U.S.-based author affiliations, and
  - U.S.-based publishing location and/or grant funding.
- Citations were partitioned by U.S./non-U.S. status and randomly divided 3:1 to a training set to identify search terms for the filter, and testing set for its validation.
- Punctuation and commonly occurring words such as conjunctions were removed.
- Using text mining, common one- and two-word terms in title and abstract fields were identified, and frequencies compared between U.S. and non-U.S. citations.
- A preliminary search filter was developed by combining terms related to U.S. citations in title and abstract fields.
- For validation, the filter was run on Ovid MEDLINE. Citations picked up by the filter were matched with the citations in the testing set to calculate its sensitivity and specificity.
- Analyses used the tidytext package in R.

References

1. Popoff, Evan & Jansen, Jeroen & Besada, M & Cope, Shannon & Kanters, Steve. (2018). PRM94 - Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews. Value in Health. 21. S371. 10.1016/j.jval.2018.09.2215.

Disclosures

The authors of this study did not receiving any funding for this work.
Contact: epopoff@broadstreetheor.com

Results

- 21,915 citations were collected; 16,436 were assigned to the training set (n=5,902 U.S.; n=10,534 non-U.S.).
- Within the training set, the range of publication years, number of disease areas covered, and number of journals covered among U.S. and non-U.S. citations were larger in the non-U.S. group, corresponding to its larger number of citations (Table 1).
- Among U.S. citations, common U.S.-related terms included (expressed as ratio of frequency in U.S. to non-U.S. citations):

  U.S. populations
  - "African American" (18.0), "Americans" (15.5), "Medicare beneficiaries" (12.0), and "Veterans" (4.6)

  U.S. geographic terms
  - "Baltimore" (20.1) and "United States" (6.1)

Table 1. Characteristics of U.S. and non-U.S. citations identified in the training set

Description	U.S. citations	Non-U.S. citations
Number of citations	5,902	10,534
Publication years	1985-2019	1964-2020
Number of diseases and conditions covered	60	61
Most common diseases/conditions	Cardiovascular disease (9.6%), Obesity (6.7%)	Cardiovascular disease (9.1%), BRCA1/2 cancer (5.1%)
Number of journals	1,012	1,970

Table 2. Validation testing for the search filter

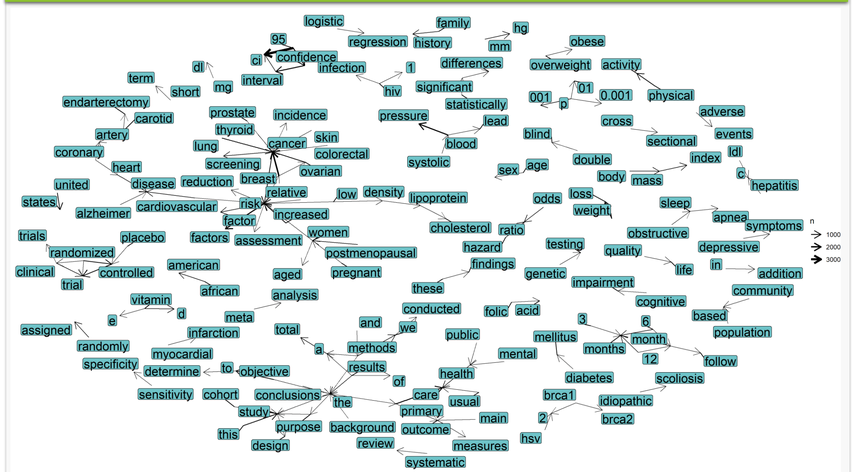
	U.S. citations	Non-U.S. citations	
Picked up by search	True positive (TP) **1,934**	False positive (FP) **609**	PPV = TP/(TP + FP) = **76.1%**
Not picked up by search	False negative (FN) **34**	True negative (TN) **2,902**	NPV = TN/(FN + TN) = **98.8%**
	Sensitivity = TP/(TP + FN) = **98.3%**	Specificity = TN/(FP + TN) = **82.7%**	

Results (cont'd)

- Among non-U.S. citations, common terms were:

  Non-U.S. geographic terms
  - "Japan" (0.04), "French" (0.05), "Edinburgh" (0.06),"Swedish" (0.06).
- Figure 1 displays a directed word graph depicting connecting words appearing 200 or more times amongst U.S. citations included in the training set. Amongst the most common word connections were: "95 confidence interval", "risk factor", and "breast cancer."
- The testing set consisted of 5,479 citations for use in validating the filter (n=1,968 U.S.; n=3,511 non-U.S.).
- Sensitivity of the filter was determined to be 98.3%, while specificity was 82.7% (Table 2).
- Positive predictive value (PPV) was 76.1%, while negative predictive value (NPV) was 98.8% (Table 2).

Fig 1. Directed word graph of U.S. citations in the training set



Conclusions

- In this study, a MEDLINE-based search filter was developed and validated to streamline the systematic identification of evidence from U.S. studies.
- The filter demonstrated excellent sensitivity and negative predictive value, while also having satisfactory specificity and positive predictive value.
- Periodic updates will be necessary to reflect changes in MEDLINE's controlled vocabulary.
- Future work could include refinement to improve sensitivity and specificity, and application of these methods to other jurisdictions.