Text mining is a potentially valuable technique for analyzing large unstructured datasets to identify meaningful patterns.

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).
- Sensitivity of the filter was determined to be 98.3%, while specificity was 82.7%.
- 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

- 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%.
- Positive predictive value (PPV) was 76.1%, while negative predictive value (NPV) was 98.8% (Table 2).

References