MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-1128

Synonym-based Word Frequency Analysis to Support the Development and Presentation of a Public Health Quality Improvement Taxonomy in an Online Exchange

Jamie Pina[a], Kelley Chester[b], Diana Danoff[c], Mark Koyanagi[d]

[a,b,c,d] *RTI International, Center for Advancement in Health Information Technology*

Abstract/Objective

Word frequency analysis has not been fully explored as an input to public health taxonomy development. We used document analysis, expert review, and user-centered design to develop a taxonomy of public health quality improvement concepts for an online exchange of quality improvement work (www.phqix.org). Online entries were made searchable using a faceted search approach. To present the most relevant facets to users, we analyzed 334 published public health quality improvement documents using word frequency analysis to identify the most prevalent clusters of word meanings. We reviewed the highest-weighted concepts and identified their relationships to quality improvement details in our taxonomy. The meanings were mapped to items in our taxonomy, and presented in order of their weighted percentages in the data. Using this combination of methods, we developed and sorted concepts in the faceted search presentation so that relevant search criteria were accessible to users of the online exchange. Word frequency analysis may be a useful method to incorporate in other taxonomy development and presentation when relevant data is available.

Keywords:

Taxonomy, word frequency analysis, public health, public health informatics, information exchange

Methods

Taxonomy is a form of classification that creates a normalized or hierarchical organization of concepts or terms [1]. We applied document analysis, expert review, and user-centered design methods to identify appropriate elements of an original taxonomy for public health quality improvement activities. Word frequency analyses can be used as an input in developing taxonomies [2]. We applied word frequency analysis to 334 (*N*=334) public health quality improvement documents including reports and summaries from public health agencies throughout the United States. Using NVivo Version 9 by QSR, we analyzed the documents to identify the most frequently recurring word-meaning clusters. We reviewed the documents using synonym identification, which finds highly recurring words and their synonyms (words with a very close meaning) throughout the texts and ranks them based on the recurrence of word meanings across the entire body of data. We compared this list of ranked synonym clusters to the elements in our taxonomy and mapped our taxonomy elements to the clusters. We reviewed 50% (*n*=100) of the highest-weight clusters. We then created high-level categories for the display of elements in our taxonomy based on the ranked synonym cluster. These categories were later used to sort and organize the presentation of data elements in our taxonomy.

Results

We analyzed 50 of the top synonym clusters and identified 12 main categories for our taxonomy data. They are presented on the website's search results view in order according to the weighted percentages identified through the word frequency analysis.

Conclusion

When a large body of searchable text is available and time or resource constraints suggest that traditional qualitative or thematic analysis not possible, applying word frequency analysis to a body of text may provide an alternative form of analysis in taxonomy development. Knowledge of the presence and recurrence of word meanings in a body of related text facilitates the generation of categories that provide structure and meaning to the taxonomy. In this application of word frequency analysis, categories of taxonomy data elements presented to users according to the weighted percentages found in word frequency analyses appear to align with user expectations. By comparing the results of word frequency analysis to a taxonomy developed using other methods, analysts may validate their choices for data elements within the taxonomy. In the field of public health in the United States, where successful information retrieval leads to improved public health performance, the development of taxonomies supplemented by word frequency analysis may support broader public health goals.

References

[1] Gilchrist A. Thesauri, taxonomies and ontologies - an etymological note. Journal of Documentation 2003: 59 (1): 7-18.

[2] Kreis C and Gorman, P. Word frequency analysis of dictated clinical data: a user-centered approach to the design of a structured data entry interface. Proc AMIA Annu Fall Symp, 1997: p. 724-8.

Address for correspondence

Jamie Pina, PhD, MSPH
Research Scientist
Center for the Advancement of Health IT
RTI International
1440 Main Street, Suite 310
Waltham, MA 02451-1623
(781) 434-1778
jpina@rti.org
Skype: jamie.pina.rti
Twitter: @jamiepina