doi:10.3233/978-1-61499-753-5-211

Automated Identification of National Health Survey Research Topics in the Academic Literature

Dean William YERGENS[a,b,1], Daniel James DUTTON[c] and Kirsten Marie FIEST[a,d,e]
[a] *Department of Critical Care Medicine, Cumming School of Medicine, University of Calgary, Canada*
[b] *Synthesis Research Inc., Calgary, Canada*
[c] *School of Public Policy, University of Calgary, Canada*
[d] *O'Brien Institute for Public Health, University of Calgary, Canada*
[e] *Hotchkiss Brain Institute, University of Calgary, Canada*

Abstract. National health surveys are routinely conducted to provide value data about a country's health status and the health services being consumed by the population. This information is used for surveillance, research, and the planning of healthcare services at local and national levels. Although these national health surveys are viewed as important resources for public and population health, there is limited information as to the type of research being conducted with these surveys. This study investigates, through the use of automated text data mining, an approach to identify and collate the type of academic literature being published using national health surveys.

Keywords. Text data mining, algorithm, epidemiology, literature review, national health survey

1. Introduction

National health surveys are routinely conducted to provide value data about a country's health status and the health services being consumed by the population. This information is used for surveillance, research, and the planning of healthcare services at local and national levels. Because of the availability and comprehensiveness of these surveys, this data is routinely analyzed by various academic and government institutions with the results being disseminated through the academic literature. Many countries engage in conducting national health surveys, as illustrated in Table 1, which not only allows for national comparisons, but for international comparisons as well.

The type of epidemiology and health services research being conducted with these national health survey datasets has not been described in the literature to our knowledge. Previous studies have examined the use of statistical methods in the biomedical literature[1], as well as individual journals[2]. Text data mining has also been

[1] Corresponding Author: Dean William Yergens, Department of Critical Care Medicine, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada; E-mail: dyergens@ucalgary.ca.

applied to examine the statistical methods utilized in the published medical literature[3] and within a Canadian national health survey[4].

The objective of this study was to develop an automated approach for determining the type of epidemiology and health services research being conducted using national health surveys and published in the academic literature, based strictly upon identifying specific keywords and phrases found in the paper's title.

Table 1. Examples of National Health Surveys

Survey	Country	Description
CCHS	Canada	Canadian Community Health Survey (CCHS) is a cross sectional survey, conducted since 2001, that collects health status, health care services utilization and health determinants data of the Canadian population. [5]
BRFSS	United States	Behavioral Risk Factor Surveillance System (BRFSS) is a health survey, conducted since 1984, collects health related risk behaviors, chronic health conditions, and use of healthcare services for the United States (US) population.[6]
NHANES	United States	National Health and Nutrition Examination Survey (NHANES), conducted since the early 1960s, assesses the health and nutritional status of adults and children in the US.[7]
KNHANES	South Korea	Korean National Health and Nutrition Examination Survey (KNHANES) is conducted to evaluate the health and nutritional status of the South Korean population. [8]

2. Methods

2.1. Literature Search Strategy

Literature searches were conducted for each of four national health surveys appearing in Table 1 using the PubMed bibliographical database. For the CCHS, our search strategy consisted of the phrase “Canadian Community Health Survey”; BRFSS consisted of “Behavioral Risk Factor Surveillance System”; NHANES consisted of the term “NHANES” or the phrase "National Health and Nutrition Examination Survey" with both of them excluding the terms “Korea*”; and the KNHANES search consisted of term “KNHANES” or the phrase "Korean National Health and Nutrition Examination Survey". All of the searches were limited to title and abstract only and were conducted on October 24, 2016. We did not assess if the national health survey was the major topic of the paper or was only referenced.

2.2. Data Management and Custom Software

All of the references were imported into a custom written Java-based literature reference management program (Synthesis). This software was created by DWY and is described in more detail elsewhere[4]. Synthesis is built upon the open-source Apache Lucene database and has the ability to manage textual documents for collating, managing, and performing Boolean queries based upon the imported references in the Lucene database. The Synthesis software is capable of taking a text definition file based upon keywords or phrases, Boolean operators, wildcards, and proximity searching and tag every reference that meets the user-defined criteria.

2.3. Topic Algorithm Development

To determine the research topics, DWY and KMF used the CCHS references to identify the main categories. Main categories consisted of keywords, phrases, and basic algorithms. This was an iterative process, which involved looking for commonly used words in the titles of the CCHS references and determining whether it was a suitable candidate to include within the main categories. This is described below.

To aid in the identification of commonly occurring words, a dynamic Word Cloud included in Synthesis was utilized. Once a topic/concept of interest was identified, a combination of Boolean logic (AND, OR, NOT) and wildcards was used to construct a statement that could automatically identify and tag the concept within Synthesis. An example of a statement defining the concept of Determinants would be: ‘title:determinants AND NOT title:"social determinants"’.

Once a concept was identified and determined suitable for inclusion, it was grouped into a higher-level main category. It should be noted that many frequently appearing words in the title were not suitable to be included as they were either too general and often did not reflect a topic categorization. Titles where the algorithm produced no categories were then marked as ‘Unclassified’. We did encounter several references in the Unclassified category that could benefit from more in depth analysis and rules. An example of this were references that simply had the Outcome and the Exposure as the title (e.g. Smoking and oral health status) which would indicate belonging to the Association category. However, to categorize these kinds of references, a list of potential domain area variables (e.g. smoking, oral health, etc.) would need to be constructed which we determined was outside of the scope of this paper.

Table 2. Category Definitions

Main Category	Category Concepts
Characteristics	Characteristics, Epidemiology, Determinants, Factor, Comorbidity, Consumption, Behavio(u)r, Burden, Unmet, Inequality, Inequity, Profile, Classification, Descriptive, Among
Association	Association, Between, Relationship, Differences, Comparison, Variation, Correlation, Disparities, Link
Estimates	Estimates, Prevalence, Incidence, Occurrence, Adjusted
Surveillance	Surveillance, Trends, Increase, Decrease, Change, Pattern, Update, Incremental, Screening, Rate
Risk	Risk
Utilization	Utilization, Usage, Access, Services, Treatment
Prediction	Prediction, Forecast, Impact, Adherence, Determinant
Evaluation	Evaluation, Validation, Accuracy, Reliability
Implementation	Implementation, Application, Planning, Management, Recommendation
Methodology	Methodology, Algorithm, Derive, Design, Develop
Spatial	Spatial, Geographical, Map
Unclassified	No categories identified

In total, eleven main categories were identified: Association, Characteristics, Estimates, Surveillance, Risk, Utilization, Implementation, Validation, Prediction, Methodology, and Spatial. Each of these main categories consisted of a variety of associated keywords/concepts as a main category could represent several associated concepts. For example, the Utilization main category consisted of derivatives of the following keywords: utilization, usage, access, services, and treatment. A list of the main categories and their associated sub-categories can be found in Table 2. It should be noted that a paper may be tagged with more than one category.

3. Results

Four separate literature searches were conducted. The search for the CCHS dataset resulted in 996 references, BRFSS 2289 references, NHANES 8286 references, and KNHANES 986 references. The 11 main concept algorithm definition file was applied within Synthesis to each of the four datasets (see results in Table 3).

The two most frequent main categories across all datasets were Characteristics and Association. Characteristics was identified in 25.0%, 34.1%, 19.0%, and 17.3% of the CCHS, BRFSS, NHANES, and KNHANES datasets, while Association was identified in 20.1%, 15.3%, 22.7%, and 33.4% of the datasets. The Estimates, Surveillance, and Risk main categories were the next most frequent groups, accounting for 7.5%, 7.5%, and 5.3% of the CCHS references, 10.2%, 11.7%, and 6.1% of the BRFSS references, 8.3%, 7.8%, and 8.9% of the NHANES references, and 8.9%, 9.0%, and 9.9% of the KNHANES references. The next natural grouping of main categories based upon percentages identified in the four datasets consisted of Utilization, Prediction, Evaluation, Implementation, Methodology, and Spatial. These main categories were represented in the low single digit percentage of all categories amongst the four datasets. References where a category could not be identified were labeled as Unclassified and percentages across the databases ranged from 9.7% (BRFSS), 12.9% (KNHANES), 18.5% (CCHS), to 22.6% (NHANES).

Table 3. Results from Topic Algorithm

Main Category	CCHS	BRFSS	NHANES	KNHANES
Characteristics	345 (25.0%)	1214 (34.1%)	2107 (19.0%)	247 (17.3%)
Association	278 (20.1%)	545 (15.3%)	2516 (22.7%)	478 (33.4%)
Estimates	104 (7.5%)	362 (10.2%)	914 (8.3%)	127 (8.9%)
Surveillance	103 (7.5%)	417 (11.7%)	867 (7.8%)	128 (9.0%)
Risk	73 (5.3%)	217 (6.1%)	989 (8.9%)	141 (9.9%)
Utilization	77 (5.6%)	166 (4.7%)	199 (1.8%)	23 (1.6%)
Prediction	54 (3.9%)	102 (2.9%)	419 (3.8%)	47 (3.3%)
Evaluation	27 (2.0%)	55 (1.5%)	204 (1.8%)	16 (1.1%)
Implementation	15 (1.1%)	57 (1.6%)	138 (1.2%)	17 (1.2%)
Methodology	27 (2.0%)	32 (0.9%)	200 (1.8%)	19 (1.3%)
Spatial	22 (1.6%)	43 (1.2%)	15 (0.1%)	2 (0.1%)
Unclassified	255 (18.5%)	346 (9.7%)	2499 (22.6%)	185 (12.9%)

4. Conclusion

This study provides an approach to using text data mining for categorizing research topics of national health surveys based upon the titles of academic publications. This study reports on four commonly used national health surveys from multiple countries and finds that title topic categorizations are relevantly consistent across all of datasets. This indicates that topic definitions could be applied to other health surveys outside of the CCHS, for which it was originally developed.

This study identifies three natural boundaries in the research being produced from national health surveys. The first grouping includes Characteristics and Associations, which account for roughly 40-50% of the research being published. The second group consists of Estimates, Surveillance, and Risk, accounting for 20-30% of the publications. The third group, includes the Utilization, Prediction, Evaluation, Implementation, Methodology, and Spatial main categories, which account for a small number of publications from these surveys.

The high percentage of references in the Characteristic and Association main category is expected. As national health surveys are readily available datasets, they provide a cost effective and timely solution for conducting much needed research in many different areas. The description of patient populations and associations between differing variables of interest is an important aspect of epidemiology and its application to population and public health. The finding that there were a low number of publications in the Implementation category is interesting, and we wonder if this is the result of actionable initiatives not being commonly reported in the academic literature. Knowing this information provides the opportunity to help guide future health policy into which areas should be strengthen and identifying gaps in the research.

There are several limitations in this study. First, the main categories identified are most likely only applicable to national health surveys or domains with a public or population health focus. We anticipate that other research areas have their own unique set of terminology and focus. An example of this could be clinical medicine, where survival analysis and outcome research may be more prevalent and require new categories. Second, we only analyzed titles to determine the research topic being investigated. Future research should also examine the abstract and full-text of each publication which could provide additional information to aid in categorizing the research topics.

References

[1] M. Scotch, M. Duggal, C. Brandt, Z. Lin Z, R. Shiffman, Use of statistical analysis in the biomedical informatics literature, *J Am Med Inform Assoc* Jan-Feb 17(1) (2010), 3-5.

[2] P.J. Becker, E. Viljoen, L. Wolmarans, C.B. IJsselmuiden, An assessment of the statistical procedures used in original papers published in the SAMJ during 1992, *S Afr Med J* Sep 85(9) (1995), 881-4.

[3] C. Meaney, R. Moineddin, T, Voruganti, M.A. O'Brien, P. Krueger, F. Sullivan, Text mining describes the use of statistical and epidemiological methods in published medical research, J Clin Epidemiol Jun 74 (2016), 124-32.

[4] D.W. Yergens, D.J. Dutton, S.B. Patten, An overview of the statistical methods reported by studies using the Canadian community health survey, *BMC Med Res Methodol* Jan 25 (2014), 14:15.

[5] Statistics Canada, Canadian Community Health Survey (CCHS), webpage: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=3359, Accessed Nov 6, 2016.

[6] Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System (BRFSS), webpage: http://www.cdc.gov/brfss/, Accessed Nov 6, 2016.

[7] Centers for Disease Control and Prevention, National Health and Nutrition Examination Survey, webpage: http://www.cdc.gov/nchs/nhanes/, Accessed Nov 6, 2016.

[8] Korean National Health & Nutrition Examination Survey, Survey Overview, Webpage: https://knhanes.cdc.go.kr/knhanes/eng/sub01/sub01_02.do, Accessed Nov 6, 2016.