

2018

Characterizing the Information Needs of Rural Healthcare Practitioners with Language Agnostic Automated Text Analysis

Melissa Resnick

University of Texas Health Science Center at Houston, Melissa.P.Resnick@uth.tmc.edu

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthshis_dissertations

Part of the [Health Communication Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Resnick, Melissa, "Characterizing the Information Needs of Rural Healthcare Practitioners with Language Agnostic Automated Text Analysis" (2018). *UT SBMI Dissertations (Open Access)*. 39.
https://digitalcommons.library.tmc.edu/uthshis_dissertations/39

This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact nha.huynh@library.tmc.edu.

Footer Logo

Characterizing the Information Needs of Rural Healthcare Practitioners
with Language Agnostic Automated Text Analysis

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

By

Melissa P. Resnick, M.S., M.L.S.

University of Texas Health Science Center at Houston

2018

Dissertation Committee:

Trevor Cohen, MBChB, PhD¹, Advisor

Hua Xu, PhD¹

Magdala de Araújo Novaes, PhD^{2, 3}

Chiehwen Ed Hsu, PhD, MS, MPH, M. Law⁴

¹University of Texas, Health Science Center at Houston, School of Biomedical Informatics

²Telehealth Center, Clinics Hospital, Federal University of Pernambuco, Recife, Brazil

³Internal Medicine Department, Medical School, Federal University of Pernambuco, Recife, Brazil

⁴University of Maryland University College

Copyright by
Melissa P. Resnick
2018

Dedication

To the memory of my beloved husband, Joshua Harris Resnick, both a friend and lover, who offered the patience and support necessary for me to explore my curiosity; your belief in me has made this success possible, although not with us you left fingerprints of grace on my life.

To the memory of my beloved mother, Patricia Ann Horner, a strong and gentle soul who taught me to trust God, believe in hard work, and to always strive for the best; you are gone but your faith in me has made this journey possible.

Acknowledgements

I am deeply indebted to my mentors, collaborators, friends, and family who provided the encouragement critical to completing this research.

Special thanks to my primary mentor Trevor Cohen for his brilliant scientific knowledge and insight, patience, unending support, and for teaching me about the scientific process.

I owe extreme gratitude to Joel Resnick who provided me with so much motivation and nurture; and Sheila Resnick-Jones who taught me ingenuity.

I thank Magdala Novaes for welcoming to NUTES for three different visits, teaching me about the telehealth system in Brazil, and giving me access to data for my research. It is through her that I learned about my love for terminologies.

I'd like to thank Ed Hsu for his belief in my ability to embark on and finish my travel down the road to my Doctoral degree.

I'd like to thank Hua Xu for his insights into machine learning.

I would like to recognize Nilma Andrade, as she made my visits to NUTES go smoothly.

I especially thank Beatriz Alkmim for allowing me to access data set from the telehealth system in Minas Gerais.

I also thank Arthur Treuherz and Sueli Suga for the data sets from BIREME.

I was very lucky to have met and interacted with many other faculty, staff and colleagues at The University of Texas School of Biomedical Informatics, Núcleo de Telessaúde da Universidade Federal de Pernambuco, and the U.S. National Library of Medicine.

I am indebted to my friends, especially Anita and Emil Bonanno, Colleen and David Lawson, Dolores and Francis Shamenek, and Toni Whaley for their continuous encouragement and support during the rough times, and for celebrating during those joyous times.

Last, but not least, I am deeply grateful to my fiancé, Frank Shamenek, for his encouragement, confidence in me, love, and compassion.

Abstract

Objectives – Previous research has characterized urban healthcare providers' information needs, using various qualitative methods. However, little is known about the needs of rural primary care practitioners in Brazil. Communication exchanged during tele-consultations presents a unique data source for the study of these information needs. In this study, I characterize rural healthcare providers' information needs expressed electronically, using automated methods.

Methods – I applied automated methods to categorize messages obtained from the telehealth system from two regions in Brazil. A subset of these messages, annotated with top-level categories in the DeCS terminology (the regional equivalent of MeSH), was used to train text categorization models, which were then applied to a larger, unannotated data set. On account of their more granular nature, I focused on answers provided to the queries sent by rural healthcare providers. I studied these answers, as surrogates for the information needs they met. Message representations were generated using methods of distributional semantics, permitting the application of k-Nearest Neighbor classification for category assignment. The resulting category assignments were analyzed to determine differences across regions, and healthcare providers.

Results – Analysis of the assigned categories revealed differences in information needs across regions, corresponding to known differences in the distributions of diseases and tele-consultant expertise across these regions. Furthermore, information needs of rural nurses were observed to be different from those documented in qualitative studies of their

urban counterparts, and the distribution of expressed information needs categories differed across types of providers (e.g. nurses vs. physicians).

Discussion – The automated analysis of large amounts of digitally-captured tele-consultation data suggests that rural healthcare providers' information needs in Brazil are different than those of their urban counterparts in developed countries. The observed disparities in information needs correspond to known differences in the distribution of illness and expertise in these regions, supporting the applicability of my methods in this context. In addition, these methods have the potential to mediate near real-time monitoring of information needs, without imposing a direct burden upon healthcare providers. Potential applications include automated delivery of needed information at the point of care, needs-based deployment of tele-consultation resources and syndromic surveillance.

Conclusion – I used automated text categorization methods to assess the information needs expressed at the point of care in rural Brazil. My findings reveal differences in information needs across regions, and across practitioner types, demonstrating the utility of these methods and data as a means to characterize information needs.

Vita

1987.....Bachelor of Arts (cum laude), Biology, State
University of New York, University Center at Albany

1993.....Master of Science, Psychology, Rensselaer
Polytechnic Institute

2007.....Master of Library Science, Library Science,
City University of New York, Queens College

Publications

Field of Study

Biomedical Informatics

Table of Contents

Dedication	ii
Acknowledgements	iii
Abstract	iv
Vita	vi
Table of Contents	vii
List of Tables	viii
List of Figures	ix
Chapter 1: Introduction	1
Chapter 2: Literature Review	7
Chapter 3: Evaluation of Automated Text Categorization Methods for Categorizing Healthcare Providers' Information Needs	31
Chapter 4: Characterization of Healthcare Providers' Information Needs	45
Chapter 5: Analysis of Temporal Dynamics of Healthcare Providers' Information Needs	70
Chapter 6: Key Findings, Innovation, Contributions, Future Work and Conclusions.....	97
References	106
Appendix A: Map – States and Regions in Brazil	140
Appendix B: Map – State of Pernambuco	141
Appendix C: Map – State of Minas Gerais	142
Appendix D: Committee for the Protection of Human Subjects	143
Appendix E: Letter of Agreement – Statement of Authorization for Research.....	144
Appendix F: Alphabetical list of 109 unique categories from training data set	145
Appendix G: Outlier Detection – 90 categories x 4 tests.....	147

List of Tables

Table 1: Examples of Question-Answer Pairs	15
Table 2: Example of category assignment	36
Table 3: Rank of each of the five Models, by their top mean F1 measures.....	42
Table 4: Histogram of Category Assignment Frequency – MG	47
Table 5: Histogram of Category Assignment Frequency – PE.....	47
Table 6: Top 10 unique categories for each region, in order of frequency of assignment	49
Table 7: Unique Categories from Table 6 Top 10 Present in PE but not MG	50
Table 8: Unique Categories from Table 6 Top 10 Present in MG but not PE.....	50
Table 9: Select Top Level DeCS Categories with Select DeCS Terms.....	52
Table 10: Ranking of Requester Type	55
Table 11: PE Top 10 unique categories,	56
Table 12: MG Top 10 unique categories	60
Table 13: Categories that occurred in the Top 10 for certain provider types only	63
Table 14: Categories representing information needs of Urban versus Rural healthcare providers	65
Table 15: Thirteen categories selected for qualitative analysis	79
Table 16: Major / Minor Topics for Outlier Seasons of Categories	83
Table 17 Spearman’s r coefficient.....	93

List of Figures

Figure 1. Process of transforming raw data to data sets, culminating in category assignment.....	34
--	----

Chapter 1: Introduction

1.1 Overview

Primary care practitioners (PCPs) are an important component of the healthcare system in developing nations (Harzheim et al., 2006). They often provide the link between individuals of a rural community and the health care that they receive in secondary/tertiary centers (Atkinson, Fernandes, Caprara, & Gideon, 2005; Campos, Haddad, Wen, Alkmin, & Cury, 2009). As they lack specialist training, these healthcare providers at times have questions regarding a patient's treatment and care, leading to an information need. Studies of the information needs of healthcare providers in developed nations have shown that healthcare providers most frequently asked questions regarding treatment methods, diagnoses, and medications (Cheng, 2004; C. Cimino & Barnett, 1991; M. A. Clarke et al., 2013; Kourouthanassis, Mikalef, Ioannidou, & Pateli, 2015; Osheroff et al., 1991; Smith, 1996). However, little is known about the information needs of the healthcare providers in developing nations, such as Brazil. Thus, there is a need for further research in this area.

Furthermore, much of the research on information needs, thus far, has been done through questionnaires (Covell, Uman, & Manning, 1985; Huth, 1989; Kourouthanassis et al., 2015), observations (C. Cimino & Barnett, 1991; Covell et al., 1985), and interviews (J. W. Ely et al., 1999; Osheroff et al., 1991). These methods have limitations including, at times: biased reporting, lack of generalizability and small sample size (Kourouthanassis et al., 2015). In recent years, another source of data representing healthcare providers' information needs has presented itself. Some countries have set up a telehealth system connecting the healthcare providers to secondary/tertiary centers. It is through this system that the healthcare providers send their queries,

expressing their information needs. This digital communication between the healthcare providers and the secondary/tertiary centers provides a unique source of data that are amenable to automated analysis. Therefore, healthcare providers' queries now delivered electronically can be characterized so that real-time monitoring of information needs (rather than once-off studies) is viable in order to provide needed information at the point of care, without imposing a direct burden upon healthcare providers. In addition, the use of automated methods for the characterization of healthcare providers' information needs has the potential to provide information pertinent to needs-based deployment of tele-consultation resources and syndromic surveillance.

Syndromic surveillance involves data collection in order to monitor communities for epidemics and other public health problems (Thacker, Qualters, Lee, & Centers for Disease Control and Prevention, 2012). Several systems are currently utilized for this data collection. However, these traditional surveillance systems are plagued by a lag in time between an event and its notification (Milinovich, Williams, Clements, & Hu, 2014). Thus, some research has focussed on the creation of new surveillance systems or additional resources of data to supplement the existing surveillance systems.

Data are submitted to traditional surveillance systems by physicians, laboratories, and other healthcare providers (Choi, Cho, Shim, & Woo, 2016; Milinovich et al., 2014). These data often include symptoms, diagnoses, and chief complaints (Generous, Fairchild, Deshpande, Del Valle, & Priedhorsky, 2014; Kman & Bachmann, 2012). As they often describe features of clinical cases, the queries submitted electronically via the Brazilian tele-health system by healthcare providers may also be of value as a source of data to be used as a supplement to the Brazilian surveillance system. While the main focus of this project is on the utilization of tele-health communication as

a means to characterize information needs, I will also explore the possibility of using these data for the purpose of surveillance.

1.2 Hypothesis and Aims

My hypothesis is that automated analysis of healthcare providers' queries can assist in characterizing their information needs. The resulting characterized information needs can then be analyzed across provider types, as well as for seasonal and other variations. I evaluate this hypothesis with the following procedure.

Aim 1: Use semi-automated methods to categorize providers' queries.

Sub-Aim 1a: Develop evaluation/training set of manually indexed queries.

Sub-Aim 1b: Evaluate established methods of text categorization to determine the best approach for semi-automated indexing and/or categorization of these queries.

Aim 2: Characterize the information needs of the healthcare providers.

Sub-Aim 2a: Apply the methods from Aim 1 to determine these information needs.

Sub-Aim 2b: Evaluate the differences in information needs across clinician types.

Aim 3: Evaluate the data to characterize the temporal dynamics of these information needs.

Sub-Aim 3a: Adapt the methods developed in Aim 1 to determine changes in information needs across years and seasons.

Sub-Aim 3b: Evaluate changes in information needs as they pertain to concepts pertinent to public health in the region (e.g. “dengue”).

First, top-level categories in the DeCS terminology (the regional equivalent of MeSH) are assigned to a set of messages. Then, message representations are generated using distributional semantic models, such as Random Indexing (RI) (Kanerva, Kristoferson, & Holst, 2000), Reflective Random Indexing (RRI) (Cohen, Schvaneveldt, & Widdows, 2010) and Neural Word Embeddings (NWE) (Mikolov, Chen, Corrado, & Dean, 2013), permitting the application of k-nearest neighbor classification for text categorization. For each distributional model, k-NN is performed and evaluated with a set of parameters, to reveal the best-performing distributional semantic model for categorization with k-NN. Once identified, the best-performing distributional semantic model is applied to characterize the information needs of healthcare providers by categorizing two unannotated sets of their messages, one from a telehealth center in Pernambuco, Brazil, and a second one from a telehealth center in Minas Gerais, Brazil. This provides a general view of rural healthcare providers' information needs for Pernambuco, and Minas Gerais. The resulting categorization of information needs is then examined for those pertaining specifically to doctors, nurses, and community health workers (CHWs) for each of the two regions. As a concluding study, the categorized information needs are explored for changes across seasons and years.

1.3 Dissertation Structure

This dissertation is structured as follows. In Chapter 2, I review literature related to the field of and recent research in information needs, telehealth, as well as text categorization methods, and distributional semantics. In the following chapters I perform a series of experiments to evaluate my hypotheses. In Chapter 3, I describe the first experiment, in which I evaluate the performance of distributional semantic models (RI, RRI, and NWE) in text categorization with k-nearest neighbor (k-NN) classification. In chapter 4, I characterize healthcare providers' information needs by categorizing the two un-annotated data sets with k-NN classification using the best performing distributional semantic model discovered in the previous chapter. In addition, I explore the information needs of doctors, nurses and CHWs in both regions, Pernambuco and Minas Gerais. In Chapter 5, I employ statistical methods to analyze the healthcare providers' information needs for differences across years and seasons. The dissertation concludes with describing its significance and contributions to the field of information needs, and informatics.

1.4 Relevance

The purpose of this dissertation is to characterize information needs of rural healthcare providers' needs in two regions of Brazil with automated methods, and to determine if there are significant changes in these information needs across years and seasons. In previous research, qualitative methods have been employed. In this research, it has been demonstrated that automated methods can be used to characterize electronically-expressed information needs of rural healthcare providers, a previously untapped data source. These methods permit analysis of much larger data sets than prior work using qualitative methods. Little is known about the information needs of rural

healthcare providers. However, this research shows that information needs of healthcare providers in rural Brazil are different than those of their urban counterparts in developed countries. Furthermore, regional differences in rural healthcare provider information needs correspond to known differences in disease demographics in Brazil, providing support for the utility of these methods as applied to inter-provider communication as a means to assess information needs at scale. Combining methods of outlier detection with qualitative methods revealed that changes in automatically-assigned categories over time reflect changes in information needs and may be of utility for syndromic surveillance with further customization, and the availability of larger timestamped data sets.

Chapter 2: Literature Review

Many countries across the globe provide their citizens some type of access to health care. Often, it is the primary care practitioner (PCP) who delivers much-needed basic care at this point of access. Many times, this means that the PCPs work in rural areas, isolated from their fellow practitioners, and thus, from access to specialized information. This creates an information need for these PCPs. This isolation and lack of access to information is one reason that few healthcare providers work in the rural areas (Dee & Blazek, 1993). Thus, in order for these healthcare providers to provide quality care and remain in the rural areas, it is of paramount importance to meet their information needs.

In this current work, I have developed and evaluated automated methods to address PCP information needs, as expressed in electronic queries to more qualified colleagues. This research was conducted in the context of PCP information needs in rural Brazil. This setting is a paradigmatic example of resource-limited primary care practice with acute information needs on account of socioeconomic disparities.

2.1 Information Needs

There is no one definition for "information needs." For example, Smith (Smith, 1996) defines information needs as "the commodity used to help make patient care decisions." While Dorsch (Dorsch, 2000) provides a different definition: "knowledge needed for the health professional to carry out patient care and professional duties." Osheroff and colleagues (Osheroff et al., 1991) state that an "information need" is "the desire for information applicable to the decision-making process." The definition provided by Dorsch is the best fit for the current area of application and will be used for the remainder of this research.

2.1.1 Information Needs of Physicians

It has been demonstrated through questionnaires (Covell et al., 1985; Huth, 1989), interviews (C. Cimino & Barnett, 1991; Covell et al., 1985), and observations (J. W. Ely et al., 1999; Osheroff et al., 1991), that health professionals' information needs during patient care are often un-met (Coumou & Meijman, 2006; Smith, 1996; Woolf & Benson, 1989). These information needs are frequently expressed in the form of questions to other clinicians (Coumou & Meijman, 2006; J. W. Ely et al., 1999). In fact, numerous qualitative studies have shown that questions that arise during patient encounters concern patient care (J. W. Ely et al., 1999; John W. Ely et al., 2002; John W. Ely, Osheroff, Maviglia, & Rosenbaum, 2007). Further analysis of information needs revealed that most patient care questions refer to treatment methods, diagnoses, and medications (Cheng, 2004; C. Cimino & Barnett, 1991; Osheroff et al., 1991; Smith, 1996).

Once discovering that they have an information need, health professionals consult various sources in attempt to answer their questions. First, they consult other colleagues, followed by textbooks and print journals (Coumou & Meijman, 2006; J. W. Ely et al., 1999).

Health professionals also stated that they encountered barriers when satisfying their information needs. The most frequently cited barrier is lack of time (C. Cimino & Barnett, 1991; John W. Ely et al., 2007; Green & Ruff, 2005). Health professionals are very busy seeing patients and performing other professional duties (Dorsch, 2000). In addition, it is thought that health professionals have difficulty properly phrasing a question when searching the literature (John W. Ely et al., 2007; Green & Ruff, 2005). This, in turn, leads to other problems creating additional obstacles: excessive time required to find information; difficulty modifying the original question, which was often vague and open to interpretation; difficulty selecting an optimal search strategy;

failure of a seemingly appropriate resource to cover the topic; uncertainty about how to know when all the relevant evidence has been found so that the search can stop; and inadequate synthesis of multiple units of evidence into a clinically useful statement (John W. Ely et al., 2002). Although there is no easy fix for these difficulties, at least two solutions have been put forth. One solution is making use of the skills of a medical librarian. Dorsch (Dorsch, 2000) notes that a medical librarian can reduce the time needed for finding an answer to a question, as well as producing a comprehensive summary of the applicable information. A second solution is making use of electronic resources (Westbrook, Coiera, & Gosling, 2005). In fact, it has been shown that health professionals' ability to search databases improves with training (Garg & Turtle, 2003). Some research has also been done regarding the information needs of rural health professionals.

2.1.2 Information Needs of Rural Health Practitioners

Like their urban counterparts, rural health professionals have information needs (Dee & Blazek, 1993; Dorsch, 2000). Rural health professionals appear to have the same basic questions, those regarding patient care (Dee & Blazek, 1993; Dorsch, 2000). To answer their questions, they also consult their colleagues first, then their own personal libraries, consisting of textbooks and a limited number of journals (Dorsch, 2000). Rural health professionals also experience barriers to meeting their information needs. Although rural health professionals do experience lack of time, the similarity ends there. Dorsch (Dorsch, 2000) states that isolation, inadequate library access, lack of equipment, costs and inadequate Internet infrastructure all impede access to needed information. Clearly, much less is known about the information needs of rural health practitioners. In the proposed research I will develop methods to remedy this, in the context of information needs expressed by primary care practitioners in rural Brazil.

2.2 Healthcare in Brazil

As in many other countries outside the United States, the Brazilian constitution (Articles 196 – 200) "guarantees universal and equitable access to high-quality health care" (Alkmim et al., 2012; Assembléia Nacional Constituinte, 1988) to its citizens. Despite this guarantee, those in rural areas have less access to medical care than their urban counterparts (Brauer, 1992), and must travel hundreds of kilometers to access needed healthcare. Thus, healthcare is unevenly distributed throughout Brazil (Sachpazidis et al., 2005). The northeastern area of Brazil has more rural and poorer regions than the southeastern area of Brazil (Cufino Svitone, Garfield, Vasconcelos, & Araujo Craveiro, 2000). To provide these individuals access to healthcare, Brazil has put in place a primary care model of healthcare.

In the mid-1990's, Brazil began to replace a centralized healthcare system with a decentralized model (Atkinson et al., 2005; Paim, Travassos, Almeida, Bahia, & Macinko, 2011). This meant that the municipalities on the local level became responsible for developing their respective health activities. One important development from this model was the Family Health Program.

Under this program healthcare teams provide medical care to those in the rural areas far away from the clinical centers for that particular municipality (Atkinson et al., 2005; Paim et al., 2011). These healthcare teams are comprised of a physician, a nurse, and four to six community health workers (Atkinson et al., 2005; Harzheim et al., 2006; Paim et al., 2011). The community health worker or CHW plays an important role in the work of the healthcare team, and a critical role in the delivery of primary care.

The CHW is the link between the members of a community and the nurses and physicians of the primary care team (Bornstein & Stotz, 2008; Cardoso & Nascimento, 2010; dos Santos, Saliba,

Moimaz, Arcieri, & Carvalho, 2011; Galavote, do Prado, Maciel, & de Cássia Duarte Lima, 2011; Marzari, Junges, & Selli, 2011). Due to this important role, the Brazilian Ministry of Health mandates that CHWs have particular characteristics. Regardless of educational level, he/she is to: live in the community for a minimum of two years; be able to read and write; be at least 18 years of age; and be available to perform their duties (dos Santos et al., 2011; Kluthcovsky & Takayanagui, 2006). Once the CHW meets these criteria and is chosen, he/she is trained for up to three months (C. D. Johnson et al., 2013), and then becomes a member of the healthcare team. As a member of the healthcare team, they discuss various cases with the nurses and physicians to create a treatment plan (Peres, Caldas Júnior, da Silva, & Marin, 2011). As a member of the community, the CHW, then, implements these treatment plans. In addition, the CHW performs various other duties. In general, these involve health promotion and disease prevention (Avila, 2011; Bornstein & Stotz, 2008; Cardoso & Nascimento, 2010; de Holanda, Barbosa, & Brito, 2009; Kluthcovsky & Takayanagui, 2006). More specifically, these activities include, but are not limited to: primary care (chronic disease management, healthy pregnancy/child development), public health (screening, immunizations), and community care (health education groups, planning and performance) (C. D. Johnson et al., 2013). As part of the healthcare teams, the CHW assists in the treatment of various diseases.

These healthcare teams often treat such diseases as hypertension and diabetes alone (Harzheim et al., 2006). Sometimes, however, the healthcare teams need to send their patients to the clinical centers for a second opinion or to see a specialist, often requiring a great deal of traveling. But, the costs are prohibitive (C. Clarke, Bouland, Reed-Rowe, Friedman, & Meyer, 2011). To keep down these costs and to provide needed second opinions, Brazil has put in place a telehealth program.

2.3 Telehealth

Telehealth refers to the use of telecommunications technologies to make health and related services more accessible to healthcare providers in rural, remote, or otherwise underserved areas (Brauer, 1992; Kumar & Krupinski, 2008). Telehealth can also be used by primary care teams in rural areas to obtain second opinions from other healthcare providers, including specialists, in urban areas (Brixey & Brixey, 2017; Finn & Bria, 2009). Telehealth can be used to provide education at a distance, either through consultation with other healthcare providers (Joshi et al., 2011; Pan et al., 2008), or by obtaining continuing medical education (Pan et al., 2008; Ricci, Caputo, Callas, & Gagne, 2005; Sinha, 2000; Zollo, Kienzle, Henshaw, Crist, & Wakefield, 1999). As I will show below, some countries have set up telehealth programs to provide various services to healthcare providers and their patients.

In some countries, such as Australia, and Palau, telehealth is used to provide access to healthcare for people living in rural and remote areas. In Australia, for instance, those in rural areas only have access to healthcare provided by nurses with an occasional visit by a physician, but no access to specialists. To solve this dilemma, the Australian government paid for the infrastructure to support telecommunications technology. Since then, several telehealth projects have been started in order to improve access to healthcare specialists in urban areas (Ellis, 2004). These projects have improved access to specialists, such as dermatologists, psychiatrists, and ophthalmologists for these individuals (Ellis, 2004). Like those living in the rural areas of Australia, residents of Palau only have access to PCPs. In order to have access to a specialist, they must leave the island (C. Clarke et al., 2011). In 2010 a team from the University of California at San Diego visited Palau to assess the situation in order to determine how telehealth might be used to help provide specialty care to the residents of the island nation (C. Clarke et al., 2011). The team from San Diego has

been working with Palau since 2010 to meet the technological requirements needed for telehealth (C. Clarke et al., 2011).

Norway and the United States have used telehealth to provide other services. In Norway telehealth is used to provide advice from university hospitals to medical practitioners in regional hospitals. In 1989, Norway started a pilot project between the hospital at the University of Tromso and the regional hospital in northern Norway. Telehealth services were set up for remote diagnosis in pathology, radiology, dermatology, cardiology, and otorhinolaryngology (Rinde, Nordrum, & Nymo, 1993). By 1994, these services were provided on a permanent basis. In the late 1960s and early 1970s, the United States telehealth was used to provide psychiatric services, and medical advice to those working at first aid stations. By the early 21st century telecommunication technology had improved, making it easier to set up telehealth systems between countries (Rodrigues Netto et al., 2003). A system was set up between Johns Hopkins Hospital, Baltimore, USA and Sao Paulo, Brazil. The surgeons in the USA assisted the physicians in Brazil to perform two different surgeries with no ill effects (Rodrigues Netto et al., 2003). Around the same time, Brazil began setting up a telehealth system in order to provide various services to the healthcare providers within the country.

2.3.1 Telehealth in Brazil

The Federative Republic of Brazil is a union of 27 Federative Units: 26 states and one federal district, grouped into 5 regions: Northeast, North, West Central, Southeast, and South (see Appendix A). Various local telehealth projects began in the 1990's (Dias et al., 2015). By 2003, the Telehealth Center (NUTES), located within the Clinical Hospital of the Federal University Pernambuco (UFPE) in Recife, Pernambuco, Brazil (see Appendix B), was created (Diniz, Ribeiro

Sales, & de Araújo Novaes, 2016). By 2004 the telehealth service at NUTES provided medical second opinions to healthcare teams in the rural areas of Pernambuco (Barbosa, de A Novaes, & de Vasconcelos, 2003; de Araújo Novaes et al., 2005).

The Municipal Health Department also implemented telehealth services in Belo Horizonte, Minas Gerais, Brazil (see Appendix C) providing healthcare support and continuing education (de Melo et al., 2017; Rezende, Tavares, Alves, dos Santos, & de Melo, 2013; Ruas & Assunção, 2013). Tele-cardiology in particular was an early focus of services in Minas Gerais (Alkmim et al., 2012).

Due to the success of these projects, the nationwide Brazilian Telehealth Program was created (Dias et al., 2015; Maldonado, Marques, & Cruz, 2016), expanded and renamed Brazilian Telehealth Network Program (BTNP) (Dias et al., 2015; Maldonado et al., 2016). By 2016 telehealth centers had been integrated into universities in 23 of 27 states (Maldonado et al., 2016), providing both education and second opinion services to rural healthcare providers. The second opinion service is the focus of the current research, and functions as follows.

When the healthcare team needs a second opinion in order to provide care to a patient, they send their questions (in Portuguese) to the nurses and physicians at the telehealth centers. The appropriate health professional, a nurse or a physician, provides a second opinion or an answer (also in Portuguese) to the rural healthcare team. These questions and their corresponding answers (question-answer pairs) are collected for data sharing and reuse. Question-answer pairs from the telehealth services in Pernambuco and Minas Gerais were used for the current study (see Table 1 for examples). These queries represent a unique resource for the characterization of the day-to-day information needs of rural primary care practitioners in Brazil, of which very little is currently

known. However, in order to re-use these queries, they must first be meaningfully categorized.

This presents some unique challenges, which will subsequently be discussed.

MG Unannotated Data Set Question # 15312	Lista de problemas apresentados pelo paciente com a cronologia e evolução do quadro (incluindo tratamentos anteriores, caso realizados) PACIENTE DE 67 ANOS COM TONTURA FRAQUEZA E DESCONFORTO TÓRAX ANTERIOR Medicamentos em uso NÃO Resultados de exames complementares já solicitados ECG Descreva a dúvida que motivou sua solicitação de forma mais clara possível QUAL A CONDUTA FRENTE A ECG ALTERADO COM RITMO ATRIAL ECTÓPICO	List of problems presented by the patient with the chronology and evolution of the condition (including previous treatments, if performed) PATIENT OF 67 YEARS WITH DIZZINESS WEAKNESS AND DISCOMFORT PREVIOUS THORAX Medicines in use NO Results of complementary tests already requested ECG Describe the doubt that motivated your request as clearly as possible WHICH CONDUCTS AGAINST ECG ALTERED WITH ECOTOPIC ATRIAL RHYTHM
MG Unannotated Data Set Answer # 15312	boa noite! Apesar de eu não ter visto o ECG, provavelmente o ritmo atrial ectópico não é responsável pelas queixas de fraqueza e desconforto torácico apresentados pelo paciente. Assim, a busca pela etiologia de tais sintomas deve continuar, através da investigação clínica criteriosa.	good evening! Although I have not seen the ECG, ectopic atrial rhythm is probably not responsible for the complaints of weakness and chest discomfort presented by the patient. Thus, the search for the etiology of such symptoms should continue, through careful clinical investigation.
PE Unannotated Data Set Question # 675_ 2010_674	É possível uma criança menor de 2 anos que foi vacinada com a BCG desenvolver a forma de meningite tuberculosa?	Is it possible for a child under 2 years of age who has been vaccinated with BCG to develop tuberculous meningitis?
PE Unannotated Data Set Answer # 676_ 2010_674	A vacina BCG é recomendada para crianças menores de 02 anos, justamente por prevenir as duas formas mais graves, a Meningite e a TB miliar. Porém, pode ocorrer uma falha da vacina e a criança ser acometida da meningite tuberculosa.	The BCG vaccine is recommended for children younger than 2 years, precisely by preventing the two most serious forms, Meningitis and miliary TB. However, a vaccine failure may occur and the child may be affected by tuberculous meningitis.

Table 1: Examples of Question-Answer Pairs

2.4 Standardization and Information Reuse

Information sharing and reuse has become important in the field of informatics, due to the various systems available and in use (Musen, 1992). In fact, the Brazilian telehealth program faces the problem of representing biomedical knowledge from the primary care second opinion demands

generated by rural healthcare teams (Resnick et al., 2013). According to Musen (Musen, 1992), the application of standards, including controlled terminologies, is one step toward solving this problem. Thus, in the case of the Brazilian telehealth system, standardization of the question-answer pairs with a classification can assist with data sharing and reuse (Resnick et al., 2013).

In addition to information sharing and reuse, the use of standard terminologies assists in providing interoperability across: medical findings (diseases and lab results), medical tasks (surgery), and software programs (Musen, 1992). For example, a set of doctors at one hospital may represent "diabetes" as "diabetes mellitus type 2," while another set of doctors at a different hospital may represent this term as "adult-onset diabetes mellitus." A standard terminology can be used to link both of these representations for "diabetes" together.

In this case, standard terminologies can enable us to know that a query is expressing the same idea, but in different words. Using the example above, a physician's query may be, "What are the symptoms for type 2 diabetes mellitus?" He/she could also ask, "What symptoms would a patient with adult-onset diabetes mellitus experience?" Thus, using a standardized terminology would allow us to link these two queries about diabetes mellitus, although one query expressed it as "adult-onset diabetes mellitus," and the other as "type 2 diabetes mellitus." I will now review some widely-used terminologies and consider their applicability to the task at hand.

2.5 Terminology

A terminology, or vocabulary, is a set of terms representing a system of concepts for a particular subject field (Hammond, Jaffe, Cimino, & Huff, 2012). Today, many classifications exist. These include, but are not limited to, the International Classification of Diseases (ICD), International Classification of Primary Care (ICPC), and Medical Subject Headings (MeSH). As noted above,

each classification has been created to serve a particular purpose (J. J. Cimino, 1998). For instance, the ICD was originally used to classify diseases causing death (World Health Organization, 2018), but is now used for billing purposes (AAPC Client Services, 2013). ICPC, on the other hand, is used to classify reasons for encounter, diagnoses, and interventions in primary care (Ayankogbe, Oyediran, Oke, Arigbabu, & Osibogun, 2009). MeSH, introduced by the National Library of Medicine (NLM), is used to index and to search the biomedical literature in MEDLINE (Trieschnigg et al., 2009). As such, it covers a broad range of topics. Therefore, I have chosen DeCS, the Portuguese version of MeSH, to index the health professionals' queries from the Brazilian telehealth system.

2.5.1 MeSH

The National Library of Medicine's MEDLINE is the world's largest (Lowe and Barnett, 1994), and the most heavily used (Coletti & Bleich, 2001) bibliographic database. However, one must be able to access the information in this database or it would be rendered unusable. As briefly mentioned before, the medical subject headings (MeSH) vocabulary is used to index the citations in the MEDLINE database in order to make them accessible (J. J. Cimino, 1996; Coletti & Bleich, 2001; Lipscomb, 2000; Lowe & Barnett, 1994; Trieschnigg et al., 2009). As defined by Bachrach and Charen (Bachrach & Charen, 1978), indexing is "the process of assigning to articles analysed the headings from MeSH which best describe their content and substance." MeSH has attributes, which makes it a powerful tool for indexing and for searching the MEDLINE database.

MeSH is composed of a hierarchical structure, often referred to as the MeSH tree structure (Lowe & Barnett, 1994), in which each of the categories the subcategories are arranged from general to more specific (Bachrach & Charen, 1978). MeSH is polyhierarchical, meaning that a concept can

appear in more than one subcategory or branch of the tree (Coletti & Bleich, 2001). This structure allows the searcher to create the most specific search query, allowing for the best possible results. Another important attribute is currency. The concepts in MeSH reflect their usage in the literature (Lipscomb, 2000). The NLM adds new concepts and removes or modifies them as they change over time (Bachrach & Charen, 1978; Lowe & Barnett, 1994). MeSH is updated on a yearly basis (Coletti & Bleich, 2001; Lowe & Barnett, 1994). This improves the retrieval of the most current information in MEDLINE. MeSH has become available in other languages (Darmoni et al., 2013; Lipscomb, 2000), enabling non-English speakers to access the literature from MEDLINE. MeSH is also available in Brazilian Portuguese, which will be discussed in the next section.

2.5.2 DeCS

The Biblioteca Regional de Medicina (BIREME), under the administration of the Pan American Health Organization (PAHO), opened in 1969 (Neghme, 1975). The primary purpose of BIREME is to improve the access to health sciences information to health professionals in Latin American countries (Bonham, 1990). However, the literature generated in Latin America is missing from existing large data systems, such as Embase and MEDLARS (Bonham, 1990). To provide access to this information, BIREME introduced LILACS (“*Literatura Latinoamericana e Caribe em Ciências de la Saude*”, “Latin American and Caribbean Literature in Health Sciences”). LILACS holds the following document types: Theses, books or book chapters, conference papers, technical-science reports, and journal articles (BIREME, n.d.-b; Bonham, 1990). Like MEDLINE, the literature in LILACS is made more easily accessible through indexing.

The indexing terminology used for LILACS is called DeCS (“*Descritores em Ciências da Saúde*”, “Health Sciences Descriptors”) (BIREME, 2017; Bonham, 1990). DeCS was developed from and

modeled after MeSH (Biblioteca Virtual em Saúde, n.d.-a; Pereira & Montero, 2012). In addition to the MeSH terms, BIREME developed DeCS terminology for other areas: Public Health, Homeopathy, Science and Health, and Health Surveillance (BIREME, 2017; Bonham, 1990). These additional terms provide better access to the work being done by Latin American authors (Bonham, 1990).

Although MeSH and DeCS are primarily used for indexing biomedical literature, this study is using DeCS, to classify the Portuguese queries from the Brazilian telehealth service at NUTES. At the current time, the assignment of DeCS terms to the queries is a manual process. It involves reading the question, discerning the best representative concept, and searching DeCS for the best matching concept. Manual indexing of these queries does not provide a scalable solution to the problem of keeping track of provider needs. In addition, these queries are difficult to categorize because they are short without a great deal of semantic content. Categorization of these queries is challenging due to the fact that they are written in another language other than English. In the following section, I will describe some widely-used methods for automatic classification of medical text and consider their applicability to this task. These include both statistical approaches, and vocabulary-based approaches such as MetaMap, and MedLEE. Statistical approaches are language independent, and therefore readily applicable. In the case of MetaMap, the meta-thesaurus of the Unified Medical Language System (UMLS) has been translated into other languages (Mancini et al., 2011). Though developed for English language text, MedLEE has been used on machine-translated text and is therefore relevant to this discussion also (Castilla, Furuie, & Mendonça, 2007).

2.6 Approaches to Automated Indexing in the Biomedical Domain

Automated methods for indexing biomedical text can broadly be categorized as: vocabulary-based or statistical in nature, with statistical methods further subcategorized into geometric, or probabilistic approaches. Vocabulary-based methods, such as MetaMap Indexer (MMI) of the Medical Text Indexer (MTI) (Mork, Yepes, & Aronson, 2013), use surface similarities between the text to be indexed and concepts and their synonyms in the source indexing vocabulary. Statistical methods of automated indexing, often leverage an estimate of the similarity between documents. With geometrically motivated approaches to estimating inter-document similarity, terms and documents are represented by vectors in a high-dimensional space (Cohen & Widdows, 2009). New categories are then assigned based on the position of an unseen document within this high-dimensional space. An example of a geometrically motivated statistical method of text categorization is k-nearest neighbor classification (k-NN), in which indexing terms manually assigned to nearby documents are assigned to un-annotated documents in the space. Although these methods provide different mechanisms for automated indexing, one must remember that these categories are not mutually exclusive. The NLM's Medical Text Indexer is an example of an approach that uses statistical and vocabulary-based methods (Alan R. Aronson et al., 2007).

MetaMap is a program that was developed at NLM to map biomedical text to concepts in the metathesaurus of the Unified Medical Language System (UMLS) (A. R. Aronson, 2001; Alan R. Aronson & Lang, 2010; Tran, Luong, & Krauthammer, 2007). According to Aronson and Lang (Alan R. Aronson & Lang, 2010), a specific vocabulary, such as MeSH, can be specified when using MetaMap. Thus, MetaMap can be used for indexing text to MeSH. In addition, MetaMap is useful for decision support systems, management of patient records, information retrieval, and data mining (A. R. Aronson, 2001). Thus, this program could be useful for the Brazilian telehealth

service at NUTES. However, one of the limitations of MetaMap is that it works with English text only (Alan R. Aronson & Lang, 2010). According to personal correspondence with Aronson, this is due to the fact that: (i) the UMLS metathesaurus is mostly English, and (ii) the Specialist Lexicon is exclusively English. Therefore, MetaMap cannot be used for indexing the health practitioners' queries in Portuguese. A final approach to indexing discussed here is MedLEE.

MedLEE (medical language extraction and encoding system) is a natural language processing system that facilitates access to coded data by automatically extracting, structuring, and encoding information from medical text (Friedman, 1997; Jain & Friedman, 1997). This NLP system was originally used on radiological reports of the chest (Friedman, 1997). In another study, MedLEE was used, with relative success, to encode clinical narratives with SNOMED (Lussier, Shagina, & Friedman, 2001). By 2004, Friedman and colleagues modified MedLEE and used it to map an entire clinical document to concepts in the UMLS. In addition to these research endeavors, MedLEE also has other applications: Decision support, literature search, quality assurance, and outcomes analysis (Friedman, 1997).

Of relevance to the current project, Castilla and colleagues (Castilla et al., 2007) used MedLEE to analyze Portuguese chest x-ray reports for data extraction. Prior to MedLEE processing, these reports were machine translated from Portuguese to English using the software program SYSTRAN. The results from this study indicate that it is possible to extract data from Portuguese chest x-ray reports using machine translation with MedLEE (Castilla et al., 2007). Once again, MedLEE appears to be a good method for processing the questions posed by Brazilian health practitioners. Like MetaMap, though, MedLEE was developed to process English language medical text, rather than Portuguese medical text itself. There is one other difference between

medical text and the healthcare provider queries: Many of the queries use more vernacular language, as opposed to medical language, which may influence the results from MedLEE.

2.6.1 Supervised Machine Learning Approaches

Supervised machine learning is a process in which the machine infers a function from labeled data, which can be used for mapping new examples (Mohri, Rostamizadeh, & Talwalkar, 2012). There are several types of supervised machine learning methods. The Naïve Bayesian classifier, support vector machine (SVM) and k-NN are three widely used algorithms. SVM and Bayesian algorithms are briefly discussed here with additional detail given about k-NN, as this method is used for automated indexing in this research. SVMs derive a function that optimally separates positive and negative class instances based on the spatial location of their geometric (feature vector) representations. This function can then be used to categorize new examples. After pattern recognition, the new data can be classified into either one of two or binary categories, which is referred to as linear classification. It is called “linear classification” because the learned function is linear, i.e. a line (or hyperplane) in the space. SVMs can also perform a non-linear classification by mapping the input into high-dimensional spaces, which is done by using the kernel trick. In the Naive Bayesian method, a probabilistic function with weighted parameters derived from training data is used to classify new entities. A Naive Bayes classifier assumes that the presence or absence of a feature of a class is independent of the presence or absence of other features. A final method of supervised machine learning used in classification is k-NN, which is described below.

K-nearest neighbor is one of a number of supervised learning approaches to text classification. It is a statistical method of classifying an object, a query in this case, based on the closest training examples in the feature space. The query is assigned to the class most common among its k nearest

neighbors. In this instance, k is a positive integer and usually small. In addition, the neighbors are taken from a set of queries where the correct classification, or assignment of DeCS term, is known. This is called the training set. Once the system has been trained, it is tested on queries for which the classification is not known to determine how well the nearest neighbors can correctly classify the test queries.

Researchers have studied text categorization using k -NN combined with other methods. The NLM's Medical Text Indexer (MTI) is one example of an approach that uses statistical and vocabulary-based methods (Alan R. Aronson, Mork, Gay, Humphrey, & Rogers, 2004). In another approach, Huang and colleagues (Huang, Névéol, & Lu, 2011) use a combination of k -NN and a learning-to-rank algorithm to annotate biomedical articles with MeSH. "MeSH Now" incorporates k -NN and a learning-to-rank algorithm in the MeSH indexing process, as well as a multi-label classifier and results from MTI (Mao & Lu, 2017). Thus, k -NN remains a valuable technique for text categorization in this context, and more sophisticated methods for text categorization (such as learning-to-rank) tend to retain k -NN as a feature. For the current work an advantage of statistical approaches in general, is that they are language-agnostic.

Dense reduced-dimensional vector representations have been used as the basis for k -NN estimation. The dimensionality of such spaces tends to be much less than the number of terms in the vocabulary. There are many ways to generate such vectors. Random Indexing (RI) (Kanerva et al., 2000) is particularly convenient for this purpose, as it scales at a rate that is linear to the corpus size. The original form of RI begins by generating random vector representations of documents. Next, the document vectors are combined to form term vectors. Each term vector is generated by adding together the vectors for the documents that contain the term, and then normalizing the result. RI and its variants have many applications (see for example, Sahlgren

(Sahlgren, 2005)), one of which is indexing biomedical text with MeSH (Vasuki & Cohen, 2010; Wahle, Widdows, Herskovic, Bernstam, & Cohen, 2012).

Previous studies of healthcare providers' information needs were limited by sample size. For instance, studies using interviews and questionnaires used the following samples: (1) 303 surveys (Kourouthanassis et al., 2015), (2) 50 interviews and 814 surveys (Cheng, 2004), and (3) 47 interviews and surveys (Covell et al., 1985). In observational studies, 103 physicians participated in one study (J. W. Ely et al., 1999), while 24 physicians participated in another study (Osheroff et al., 1991). Automated methods, such as k-NN, provide the means to use large sample sizes to categorize information needs of healthcare providers.

2.7 Surveillance

Surveillance is defined as: the systematic, ongoing collection, management, analysis, and interpretation of data followed by the dissemination of these data to public health programs to stimulate public health action" (Thacker et al., 2012). This data collection provides a means by which to monitor communities for epidemics and other public health problems (Thacker et al., 2012). One method of data collection is through passive surveillance systems, in which data are submitted to the appropriate public health authority by physicians, laboratories, and other healthcare providers (Choi et al., 2016; Milinovich et al., 2014). However, these traditional surveillance systems are plagued by a lag in time between an event and its notification (Milinovich et al., 2014). Thus, some research has focussed on the creation of new surveillance systems or additional resources of data to supplement the existing surveillance systems. In the following section I will discuss some of the available surveillance systems.

2.7.1 Surveillance Systems

The event of the web marked the creation of many electronic web-based surveillance systems (Choi et al., 2016). Some of the first systems were produced at universities or institutions (BioCaster, HealthMap, GETWELL), non-governmental organizations (GOARN, MedISys, and ProMED-Mail) and governmental agencies (EpiSPIDER and GPHIN) (Choi et al., 2016). Since then, other electronic surveillance systems have emerged, such as the Electronic Surveillance System for the Early Notification of Community Based Epidemics (ESSENCE), which uses real-time data from the Department of Defense Military Health System (Bravata et al., 2004; Generous et al., 2014), and BioSense, which uses data from the Department of Veterans Affairs, the Department of Defense, retail pharmacies, and Laboratory Corporation of America (Borchardt, Ritger, & Dworkin, 2006; Generous et al., 2014). Laboratories also play an important role in surveillance. The Laboratory Response Network (LRN), comprised of over 120 laboratories, provides surveillance for a number of pathogenic and non-pathogenic disease agents, using clinical or environmental samples (Generous et al., 2014; Kman & Bachmann, 2012). The Centers for Disease Control (CDC) provides a surveillance system, Early Aberration Reporting System (EARS), which is used at the national, state, and local levels by public health departments for detecting aberrations in surveillance data (Hutwagner, Thompson, Seeman, & Treadwell, 2003).

2.7.2 Data Sources

Traditional surveillance systems depend on various types of data, such as symptoms, diagnoses, chief complaints, (Generous et al., 2014; Kman & Bachmann, 2012). Generous and colleagues (Generous et al., 2014) also note that non-traditional public health data such as school absenteeism rates, over-the-counter medication sales, 911 calls, veterinary data, and ambulance run data, have

been used for surveillance. Recently, other sources of data have been considered including: social media (e.g. Twitter, Facebook, health forums), and mobile applications (Bernardo et al., 2013; Generous et al., 2014; Hill, Merchant, & Ungar, 2013). Researchers have been studying the feasibility of using search log data as a complement to existing surveillance systems. A number of sources of search log data have been examined, including, but not limited to: (i) health websites (e.g. MedlinePlus) (Scott-Wright, Crowell, Zeng, Bates, & Greenes, 2006), (ii) health resources (e.g. UpToDate) (Callahan et al., 2015; Santillana, Nsoesie, Mekaru, Scales, & Brownstein, 2014), and (iii) search engines (Christaki, 2015; Hill et al., 2013; Hulth, Rydevik, & Linde, 2009; Polgreen, Chen, Pennock, Nelson, & Weinstein, 2008). Each of these sources of search log data are discussed below, beginning with health websites.

2.7.2.1 Health Websites

In 1998, NLM created MedlinePlus, in order to provide health information to consumers (Miller, Lacroix, & Backus, 2000; Scott-Wright et al., 2006). At the time of its inception, MedlinePlus supplied information about medications, homeopathic remedies, and diseases (Miller et al., 2000). In 2006, Scott-Wright and colleagues examined consumers' information needs via the search log data from MedlinePlus (Scott-Wright et al., 2006). The authors discovered that consumer searches were relatively similar from month to month, and that this similarity may assist in discovering media-induced or seasonal changes in consumer interest in health topics (Scott-Wright et al., 2006). Other health resources have also been studied.

2.7.2.2 Health Resources

One such health resource is UpToDate (Callahan et al., 2015; Santillana et al., 2014). UpToDate is a source of expert-authored health information provided by Wolters Kluwer, which includes

information about specific symptoms, disease management, drug usage recommendations, and treatments (Callahan et al., 2015). It is used on a subscription basis by institutions and individuals including physicians, researchers, and students (Callahan et al., 2015). Callahan and colleagues (Callahan et al., 2015) used the search log data from UpToDate to assess the feasibility of measuring a drug-safety alert response of the users of this health resource. The authors noted that the response to a Food and Drug Administration (FDA) alert for the drug Celexa appeared later in the search logs and persisted longer than the FDA drug alert (Callahan et al., 2015). In another study, the UpToDate search log data was analyzed for its ability to predict an outbreak of influenza like illness (ILI) (Santillana et al., 2014). Santillana and colleagues (Santillana et al., 2014) showed that the analysis of the search log data can be used to predict (ILI). Although these predictions are timely and less likely to be influenced by news reports, they lack specificity exhibited by traditional surveillance systems (Santillana et al., 2014). Thus, analysis of search log data of a health resource can provide insight into information needs of healthcare providers, as well as a possible means of disease surveillance. Data from search engines have also been considered for surveillance.

2.7.2.3 Search Engines

The Internet has become an important source of health information for consumers (Baker, Wagner, Singer, & Bundorf, 2003; Dickerson et al., 2004; Eysenbach, 2006; Seifter, Schwarzwald, Geis, & Aucott, 2010). Consumers frequently use search engines to look for their health information (Liang & Scammon, 2013; Seifter et al., 2010). The data from these searches can be analyzed over time to identify disease outbreaks, which, in turn, can supplement traditional surveillance systems (Eysenbach, 2006; Ginsberg et al., 2009; Seifter et al., 2010; Wilson & Brownstein, 2009). To that end, researchers have studied the practicality of using search log data for surveillance from search engines, such as Bing (White, Tatonetti, Shah, Altman, & Horvitz, 2013), Yahoo! (Cooper,

Mallon, Leadbetter, Pollack, & Peipins, 2005; Generous et al., 2014; Polgreen et al., 2008), and Google (Carneiro & Mylonakis, 2009; Chan, Sahai, Conrad, & Brownstein, 2011; Dugas et al., 2013; Generous et al., 2014; Ginsberg et al., 2009; Wilson & Brownstein, 2009).

Search engines, such as Yahoo!, are used for locating information on topics including, but not limited to, news reports, product information and health (Cooper et al., 2005). In their study, Cooper and colleagues (Cooper et al., 2005) used the search log data from Yahoo!, during the years 2001-2003, to test the correlation of Yahoo! cancer search activity with the volume of cancer news coverage and the periodicity and peaks in Yahoo! cancer search activity. The authors discovered that the Yahoo! search activity associated with specific cancers correlated with the increase in news reports (Cooper et al., 2005). In addition, they reported that the number of Yahoo! cancer searches tended to be higher during weekdays and cancer awareness months, but lower during the summer months (Cooper et al., 2005). The authors concluded that: (i) the news greatly influences online searches about cancer; and (ii) Internet search activity provides a tool for monitoring health information-seeking behavior (Cooper et al., 2005).

In another study Yahoo! search log data was used to investigate the relationship between searches for influenza and actual occurrences of influenza (Polgreen et al., 2008). Using the frequency of Yahoo! searches, the authors were able to predict an increase in positive influenza cultures one to three weeks in advance, and an increase in deaths from pneumonia and influenza up to five weeks in advance (Polgreen et al., 2008). Polgreen and colleagues (Polgreen et al., 2008) concluded that search-term surveillance from search log data may provide an additional tool for disease surveillance.

Search log data from other search engines has also been used for research in surveillance. In one study, searches from Yahoo!, Bing, and Google were collected to study adverse drug events (White et al., 2013). White and colleagues (White et al., 2013) examined these searches for the drug pairing of Paroxetine (an antidepressant) and Pravastatin (a cholesterol-lowering drug), whose interaction is reported to cause hyperglycemia. The findings indicate that there is potential value in utilizing the signals from search log data as a complement to signals from other sources for pharmacovigilance (White et al., 2013). Much of the current research has focussed on Google search log data, some of which will be discussed below.

In 2008, Google constructed a tool, Google Flu Trends (GFT), which uses Google search log data to predict national state and regional influenza activity in the United States (Olson, Konty, Paladini, Viboud, & Simonsen, 2013). The original GFT algorithm was based on linear regression models that used weekly counts of each of the 50 million most common search queries from the searches submitted in the United States between 2003 and 2007 (Ginsberg et al., 2009; Olson et al., 2013). However, a change was made to the GFT algorithm as the 2009 influenza pandemic was completely missed by the GFT system (Olson et al., 2013). The algorithm was updated to include surveillance data from the influenza pandemic and less restrictive criteria for the flu-related search terms (Cook, Conrad, Fowlkes, & Mohebbi, 2011). As the updated GFT algorithm shows high retrospective correlation with influenza surveillance data, it continues to be used in forecasting models for influenza incidence (Olson et al., 2013). In addition to influenza incidence, search log data has been used to study incidence of other diseases and conditions.

Both infectious and non-infectious conditions have been investigated. Infectious diseases have included: chicken pox (Pelat, Turbelin, Bar-Hen, Flahault, & Valleron, 2009), Lyme disease (Seifter et al., 2010), Malaria (Ocampo, Chunara, & Brownstein, 2013), tuberculosis (Zhou, Ye,

& Feng, 2011), and dengue (Althouse, Ng, & Cummings, 2011; Chan et al., 2011; Generous et al., 2014), to mention a few. Interest has widened to include non-infectious conditions, such as: urinary tract infections (Rossignol et al., 2013), kidney stones (Breyer et al., 2011), and suicide (Hagihara, Miyazaki, & Abe, 2012). Despite the positive results obtained with the use of search log data to predict outbreaks of infectious diseases and to describe incidence of non-infectious diseases, these studies do have limitations, as described below.

As research with search log data has progressed, limitations have become apparent. One of the major concerns has been privacy (Choi et al., 2016; Polgreen et al., 2008). Choi and colleagues (Choi et al., 2016) note that this concern relates to the precise health information that is connected to the individuals searching the Internet. In addition, Polgreen and colleagues (Polgreen et al., 2008) note that different searches across the same individuals in a small geographical area could also represent a privacy concern. A second limitation involves media coverage (A. K. Johnson, Mikati, & Mehta, 2016; Liang & Scammon, 2013; Polgreen et al., 2008). For instance, some searches may be due to news reports and not related to disease activity (Polgreen et al., 2008). In addition, publication of a journal article about a disease or infection may lead to searches having no relation to the occurrence of that disease or infection (Polgreen et al., 2008). In some research, geographic data is extracted from the IP address, which may not truly represent the actual geographic region (Polgreen et al., 2008).

As discussed above, consumer information needs have been studied as a resource for syndromic surveillance. However, little, if any, research has been performed using strictly healthcare providers' information needs as syndromic surveillance. In this research, I studied the use of healthcare providers' information needs express electronically in the Brazilian telehealth system as a source for syndromic surveillance.

Chapter 3: Evaluation of Automated Text Categorization Methods for Categorizing Healthcare Providers' Information Needs

Previous studies of healthcare providers' information needs are limited in sample size. For instance, prior studies using interviews and questionnaires had the following samples: (1) 303 surveys (Kourouthanassis et al., 2015), (2) 50 interviews and 814 surveys (Cheng, 2004), and (3) 47 interviews and surveys (Covell et al., 1985). In observational studies, 103 physicians participated in one study (J. W. Ely et al., 1999), while 24 physicians participated in another study (Osheroff et al., 1991). Automated methods, such as k-NN, provide the means to use large sample sizes to categorize information needs of healthcare providers.

As a robust machine learning method, k-NN performs well as a multi-label classifier even in the presence of noisy data (Gao, Yang, & Zhou, 2016) and class imbalance (Ganganwar, 2012; Liu, Cao, & Yu, 2014), making it a suitable method for our task. In cases of class imbalance, not only may some categories be more frequent than others, but a strong imbalance between positive and negative matches for each category may exist (Gibaja & Ventura, 2014). k-NN performs well with imbalanced data (Ganganwar, 2012) including cases with small numbers of positive examples (Yang & Liu, 1999), as this model views the local neighborhood, making the number of negative matches in the periphery of the semantic space irrelevant. On account of its long track record, its established utility in the context of the closely-related problem of MeSH term prediction and fit to the class distribution of the current problem, I applied the k-NN algorithm in the current research.

Dense reduced-dimensional vector representations have been used as the basis for k-NN

estimation. The dimensionality of such spaces tends to be much less than the number of terms in the vocabulary, enhancing the efficiency of nearest neighbor search. There are many ways to generate such vectors. Random Indexing (RI) (Kanerva et al., 2000) is particularly convenient for this purpose, as it scales at a rate that is linear to the corpus size. RI and its variants have many applications (see for example, (Sahlgren, 2005)), one of which is indexing biomedical text with MeSH (Vasuki & Cohen, 2010; Wahle et al., 2012). In this chapter, I describe the evaluation of k-NN with RI and its variants, along with neural word embeddings (Mikolov et al., 2013), for categorization of healthcare providers' information needs.

3.1 Materials

3.1.1 Training Data Set

The formative second opinion (FSO) data set consists of questions asked by healthcare teams during teleconsulting with health professionals in telehealth centers across Brazil (Haddad et al., 2015). According to Haddad and colleagues (2015), these most frequently asked questions were chosen based on pertinence and relevance to primary health care (PHC). For each question, a literature review was performed and used to write an answer based on the best scientific and clinical evidence (Haddad et al., 2015). The questions and answers were made freely available online at the telehealth program website designed as a Virtual Health Library (VHL) in Primary Health Care, run by Bireme (Biblioteca Virtual em Saúde, n.d.-b; Haddad et al., 2015). Before publication, each question-answer pair is given the following information: an ID number, bibliographic references that support the evidence in the answer, type of professional asking the question (doctor, nurse, community health worker, etc.), assigned DeCS terms, assigned ICPC-2

code, the name of the telehealth center responsible for the answer, and the publication date on the VHL PHC (Biblioteca Virtual em Saúde, n.d.-b).

In September 2015, a training set of 883 question-answer pairs from the FSO was obtained from BIREME in São Paulo, Brazil. All but 11 of these question-answer pairs had already been indexed or assigned DeCS terms. The 11 question-answer pairs lacking DeCS terms were eliminated, leaving a data set consisting of 872 question-answer pairs with between 1 and 7 preassigned DeCS terms. For each unique DeCS term, the English translation and corresponding DeCS tree numbers were collected from the DeCS database. Each tree number was assigned the corresponding MeSH top level subcategory along with the English subcategory name. For example, the DeCS term "Diabetes Mellitus" has the tree number C19.246 and thus, was assigned the top-level category C19 with category name "Endocrine System Diseases." Due to the polyhierarchical structure of DeCS, the terms can have more than one category. These top-level category names were used as categories for this study.

3.1.2 Unannotated Data Sets

The purpose of this study is to characterize the content of information needs expressed in two unannotated data sets. These are: (1) a data set containing 5,580 question-answer pairs and the corresponding months, years, and requesters, which was obtained from NUTES at the Federal University of Pernambuco (PE); and (2) a second data set containing 30,998 question-answer pairs and the corresponding requesters, which was obtained from NUTES at the Federal University of Minas Gerais (MG). Figure 1 provides an overview of the data sets employed in this study.

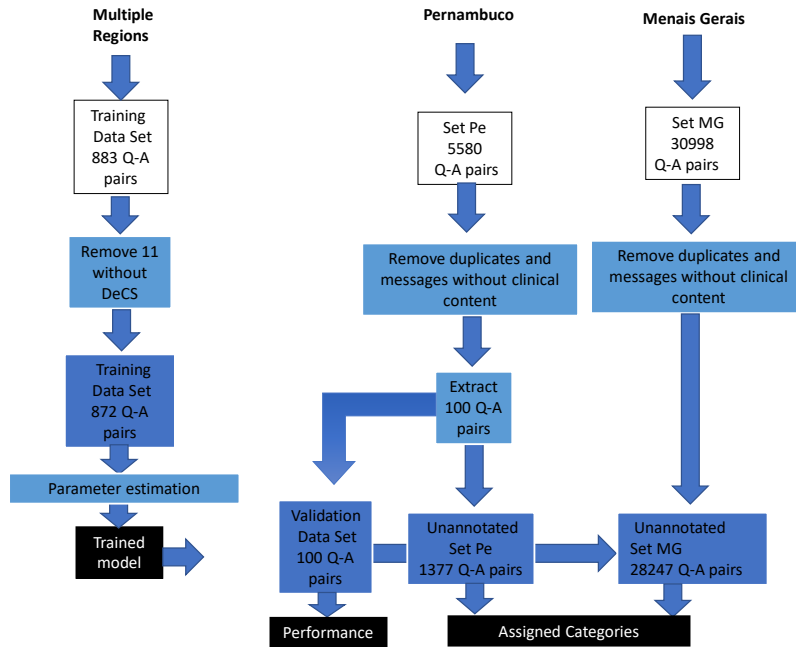


Figure 1: Process of transforming raw data to data sets, culminating in category assignment

These data sets were deidentified prior to this analysis, the study was approved by the Institutional Review Board (see Appendix D), and authorized by Universidade Federal de Pernambuco - Núcleo de Telessaúde – Hospital das Clínicas (see Appendix E). Due to the brevity of the questions, the answers were used for analysis. I focused my analysis on the answers provided in response to information needs, on account of the brevity of the question statements. Preliminary experiments showed that semantic information in the question was inadequate to provide a basis for classification, and a manual review of 100 question-answer pairs demonstrated that the topically important term from the question was present in most of the corresponding answers. Consequently, I focused my analysis on the answers to questions, as surrogates for the information needs they meet. For the first set (PE), of the 5,580 answers, 4,068 answers without significant clinical content (i.e., referring the requester to an unspecified webinar), and 35 duplicates were eliminated. 100 answers were randomly selected from the remaining 1,477 answers for validation purposes.

For the second set (MG), 2,751 duplicate and salutory answers were eliminated. Consequently, 972 answers were used for training and validation purposes, and the trained model was applied to the remaining 29,624 answers to characterize information needs.

3.1.3 External Corpus

Information from an external corpus is often used to generate pre-trained word representations in an effort to improve categorization performance. To this end, in late 2015, a corpus containing 126,176 abstracts from the LILACS database (BIREME, n.d.-a) was obtained from BIREME. These abstracts were in different languages including Portuguese, Spanish, and English. Most of these abstracts were tagged with their language type. The abstracts were separated by language, giving 92,453 tagged Portuguese language abstracts. These Portuguese language abstracts were used for this study.

3.2 Methods

3.2.1 Automated Text Categorization

Once the categories were assigned to the DeCS terms from the training set (see Table 2), automated methods for categorization were evaluated.

Answer ID #	Answer (Original Portuguese / Translated English)	DeCS Term(s) (Assigned by BIREME)	Category (Top-level DeCS categories)
sof_165	<p><i>Não foram encontradas bibliografias que tratem exclusivamente do uso de antibióticos como profilaxia para complicações bacterianas no caso de infecções respiratórias virais. No entanto uma revisão sistemática demonstrou que existe evidência insuficiente quanto ao benefício do uso de antibioticoterapia para infecções do trato respiratório superior do tipo resfriado comum em crianças ou adultos. Além disso antibióticos em adultos estão associados a efeitos colaterais significativos (A), assim como a um risco aumentado de resistência antimicrobiana.</i></p> <p>No bibliographies were found dealing exclusively with the use of antibiotics as prophylaxis for bacterial complications in the case of viral respiratory infections. However a systematic review has shown that there is insufficient evidence regarding the benefit of using antibiotic therapy for common cold-type upper respiratory tract infections in children or adults. In addition, antibiotics in adults are associated with significant side effects (A), as well as an increased risk of antimicrobial resistance.</p>	<p>Antibacterianos</p> <p>Antibioticoprofilaxia</p> <p>Infecções Bacterianas</p> <p>Infecções Respiratórias</p>	<p>Chemical Actions and Uses</p> <p>Therapeutics</p> <p>Bacterial Infections and Mycoses;</p> <p>Public Health</p> <p>Bacterial Infections and Mycoses;</p> <p>Respiratory Tract Diseases</p>

Table 2: Example of category assignment

3.2.2 Generating Vector Representations of Answers

There are many methods to derive reduced-dimensional vector representations of terms and documents (for reviews see (Turney & Pantel, 2010) and (Cohen & Widdows, 2009)). For this study, I utilized RI (Kanerva et al., 2000) on account of its scalability. Random Indexing (RI) is a method of distributional semantics that permits generation of a reduced-dimensional approximation of a document-by-term (or term-by-document) matrix without the need to generate

the full matrix explicitly. This is accomplished by superposing random vectors, which are high-dimensional (on the order of 1,000 dimensions) vectors, with a small number (on the order of 10) values set to 1 or -1 at random. These vectors have a high probability of being orthogonal, or close-to-orthogonal to one another, on account of the statistical properties of high-dimensional space (Kanerva, 1988; Sandin, Emruli, & Sahlgren, 2017; Widdows & Cohen, 2015). In this case, vector representations of answers were generated by superposing the random vectors for terms they contain (one random vector per term). This superposition operation was weighted in accordance with the type frequency (TF) and inverse-document frequency (IDF) of the terms concerned. In symbols the vector for an answer, $V(\text{answer})$, is generated as the sum of the t unique terms it contains, as follows:

$$V(\text{answer}) = \sum_i^t V_{\text{random}}(\text{term}^i) * TF(\text{term}^i) * IDF(\text{term}^i)$$

The result is a reduced-dimensional approximation of a TF-IDF weighted document-by-term space, constructed without the need to represent the unreduced space explicitly. For an accessible introduction to RI, I refer the interested reader to (Sahlgren, 2005).

These reduced-dimensional answer vectors can then be compared with one another using the cosine metric, resulting in a ranked list of neighbor answers for every cue answer in the set. These neighbor answers are used to assign categories to the cue answer using the k-NN algorithm, as discussed in the next section.

I also evaluated Reflective Random Indexing (RRI) (Cohen et al., 2010), an iterative variant of RI in which term and document vectors are superposed across multiple training cycles, as a way to incorporate background information from the LILACS corpus. I constructed document vectors from term vectors using (1) RI of the external corpus (RI-RI), and (2) RRI of the external corpus (RI-RRI, RRI). I evaluated the vector average of neural word embeddings (NWE) of the external corpus (av-NWE). NWE are generated using a neural-probabilistic approach to training of word vectors, in which the vectors (or embeddings) are the input weights of a neural network trained to predict terms in proximity to an observed term. Word embeddings were derived from LILACS using the Skipgram-with-Negative-Sampling algorithm of Mikolov and his colleagues (Mikolov et al., 2013). All representations were generated using the open source Semantic Vectors package for distributional semantics (Widdows & Cohen, 2010; Widdows & Ferraro, 2008). In building these vector spaces, I used a list of stop words in Portuguese from the Snowball project (Boulton, 2014), which consists of frequently occurring terms that carry little semantic content, and leveraged the Portuguese-specific tokenization and analysis routines provided by Apache Lucene (The Apache Software Foundation, n.d.). In all cases document vectors were generated as the normalized sum of word vectors. For RI, these were randomly generated word vectors. For the other models, these were word vectors that emerged from distributional modeling of the LILACS corpus. This vector sum was weighted using the log-entropy weighting metric (D. I. Martin & Berry, 2007) for RRI, and TFIDF weighting for the other models. For RI and its variants we used a dimensionality of 1000 as dimensionality on this order is typically used with this method and performance has been shown to deteriorate at lower dimensions (see for example (Sitbon, Bruza, & Prokopp, 2012)), while a dimensionality of 500 was used for NWE, as prior research on

biomedical term representations has not suggested moving to higher dimensionality than this is advantageous (Chiu, Crichton, Korhonen, & Pyysalo, 2016; Henry, Cuffey, & McInnes, 2018).

3.2.3 K-Nearest Neighbor Classification

After building the vector spaces, k-nearest neighbor classification (k-NN) was used to categorize the training data set, using leave-one-out cross-validation. First, k-NN was performed using each distributional model, where k equaled 1-5. In this case, each of the 872 answers, or cue answers, was classified by the other answers, or nearest neighboring answers. The semantic relatedness of the cue answers to its nearest neighbors was measured by the cosine of the angle between the cue answer and its nearest neighbors.

Using the DeCS terms, the corresponding categories were assigned to the answers. All duplicate categories were removed, so that each answer had only one copy of each of its assigned categories. In the next step, each of the nearest-neighbor categories was assigned a score of the sum of the cosine values of each of the nearest-neighboring answers in which it appears. Categories were then rank ordered using this score, and the top n categories, a model parameter explored in repeated cross-validation experiments, were used to predict those of the cue answer. These categories and corresponding cosines provided a means with which to evaluate performance.

The performance of the distributional models was measured utilizing a predetermined number of top categories from the nearest neighboring answers. At each level of k (1-5), for each distributional model, I evaluated the top number of categories from the nearest neighboring categories at 1-12, as measured by the highest cosine value. For example, when the number of top categories was set at three (3), the 3 categories with the highest cosine value were chosen for

comparison to and matching with the cue categories. In cases in which the cosine value of the last top category was equal to the cosine value of the next category or categories, I evaluated three different methods for sorting the categories before the final top category was selected: (1) alphabetical sorting, (2) reverse alphabetical sorting, and (3) random selection. In the case of alphabetical sorting, the categories were arranged alphabetically before the category was selected. In reverse alphabetical sorting, the categories were arranged in reverse alphabetical order before the category was chosen. In the case of random selection, the category was selected randomly from the categories with equal cosine values. Once the number of top categories was chosen, these categories were compared to the cue categories. The total number of cue categories (relevant), top number of nearest neighbor categories (retrieved), and matching categories (retrieved_and_relevant) were used to calculate precision and recall. Precision was calculated as follows: $\text{retrieved_and_relevant} / \text{retrieved}$. Recall was calculated as follows: $\text{retrieved_and_relevant} / \text{relevant}$. Once precision and recall were obtained, The F1 value was calculated as the harmonic mean of precision and recall ($\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$). In addition, I included the F2 value, which emphasizes recall over precision. The F2 metric is calculated as $\frac{5 \cdot \text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}$. The means of these metrics were calculated at both answer level (resulting in an average across answers) and category level (resulting in an average across categories). Precision, recall, F1 and F2 were selected, as these metrics have been widely used in prior research on categorization of biomedical text with MeSH (see for example, (Alan R. Aronson et al., 2004)). The best-performing distributional model was RI, and thus, it was applied to categorize the unannotated data.

3.3 Results and Discussion

3.3.1 Training Set Performance (Leave one out cross-validation)

3.3.1.1 Results

Table 3 shows the best results for each model (sorted by F1 score), with the parameters used to obtain them. RI, with the highest F1 value, performs better than RRI and its variants, whereas embeddings, with the lowest F1 value, has the worst performance. Each of these models performs the best at k-NN of 1, with each model utilizing 11 top categories, with the exception of embeddings, which utilizes 12 top categories.

Rank	Model	k-NN	Top Number of Categories	F1	F2	Precision	Recall	Accuracy
1	RI	1	11	0.4176	0.4270	0.4523	0.4499	0.3313
2	RI	5	3	0.4129	0.4334	0.4192	0.4650	0.3080
3	RI	2	4	0.4067	0.4434	0.3898	0.4928	0.3052
4	RI	4	3	0.4066	0.4262	0.4136	0.4568	0.3028
5	RI	3	3	0.4006	0.4189	0.4096	0.4479	0.2986
6	RI-RI	1	9	0.3953	0.4091	0.4211	0.4376	0.3122
7	RI-RRI	1	11	0.3820	0.3946	0.4073	0.4204	0.3021
8	RRI-RRI	1	10	0.3819	0.3945	0.4073	0.4203	0.3020
9	RI-RI	2	4	0.3797	0.4164	0.3588	0.4647	0.2831
10	RRI-RRI	4	3	0.3729	0.3923	0.3769	0.4216	0.2735
11	RI-RRI	4	3	0.3720	0.3909	0.3769	0.4198	0.2728
12	RI-NWE	1	10	0.3718	0.3813	0.3990	0.4028	0.2925
13	RI - RRI	5	3	0.3717	0.3914	0.3760	0.4212	0.2702
14	RRI-RRI	5	4	0.3701	0.4189	0.3335	0.4818	0.2651
15	RI - RI	3	3	0.3694	0.3856	0.3777	0.4112	0.2728
16	RI - RRI	2	4	0.3689	0.4044	0.3489	0.4510	0.2745
17	RI - RI	4	3	0.3688	0.3861	0.3761	0.4135	0.2709
18	RRI-RRI	2	4	0.3677	0.4036	0.3469	0.4503	0.2741
19	RRI-RRI	3	4	0.3663	0.4112	0.3330	0.4687	0.2665
20	RI - RI	5	4	0.3655	0.4131	0.3304	0.4747	0.2615
21	RI - RRI	3	3	0.3642	0.3834	0.3681	0.4125	0.2683
22	RI-NWE	4	3	0.3607	0.3785	0.3658	0.4057	0.2637
23	RI-NWE	5	3	0.3593	0.3766	0.3656	0.4035	0.2620
24	RI-NWE	2	9	0.3575	0.4226	0.3073	0.5133	0.2566
25	RI-NWE	3	4	0.3549	0.3979	0.3244	0.4529	0.2574

Table 3: Rank of each of the five Models, by their top mean F1 measures (across answers) and the corresponding F2, Precision, and Recall values. Best results are shown in boldface

Though k-NN of 1 often produces strong results, there is a danger of overfitting the data (Hastie, Friedman, & Tibshirani, 2001). Thus, I chose the next-best model, RI with k-NN of 5, top categories of 3, and an F1 value of 0.4129. As noted above, three sorting methods were tested in cases in which the cosine value of the last top category was equal to the cosine value of the next category or categories. It was discovered that there were negligible differences in performance for

these sorting methods. Thus, I chose random selection for this study.

3.3.1.2 Discussion

The performance of RI with these parameters is comparable to that documented with MTI (e.g. $F1=0.4192$ (National Library of Medicine, 2007), and $F1=0.409$ (Huang et al., 2011)), which provides support for using it. Tests for performance of text categorization systems use the performance of MTI in categorization of MEDLINE abstracts with MeSH as a benchmark. While these results are not directly comparable to the current results on account of the larger number of targets for prediction ($> 10,000$ MeSH terms) and much larger set of annotated abstracts available for training (> 10 million), this degree of accuracy is consistent with a system in practical use for a similar task which provides support for the decision to move forward with analysis of the larger data set.

In previous research, RRI has performed better than RI as a basis for categorizing biomedical text (Vasuki & Cohen, 2010). Therefore, it is surprising that RI performs better than the remaining models, including NWE, which have recently been applied effectively across a range of NLP tasks. This may be due to the fact that the medical concepts in the answers in the training data are expressed consistently, thus, eliminating the need to infer connections between related terms from an external corpus.

For error analysis, I reviewed 100 answers categorized by my system and found that for 40 of the 63 unique categories represented in these 100 answers, my system predicted the exact same category, producing a match. For 23 categories, the system did not predict the same category. In some instances, one of the predicted categories was a reasonable prediction, i.e. related to the topic represented by one of the unmatched cue categories. For example, "Therapeutics" was sometimes

predicted when the cue category "Health Services" was present and not matched. "Therapeutics" is related to "Health Services", as "Health Services" encompasses both "Diagnosis" and "Therapeutics". With 37% of unrelated assignments and performance comparable to MTI, I am confident in the category assignments.

3.4. Conclusion

In conclusion, RI with k-NN performed the best. Although k-NN of 1 and 11 top categories produced the best performance, I chose k-NN of 5 and 3 top categories to prevent overfitting of the data. RI with these parameters is chosen to categorize the unannotated data sets, as the performance is comparable to that of MTI (a system currently used for indexing MEDLINE abstracts with MeSH). In addition, qualitative analysis (described in detail in a subsequent chapter) demonstrates that categories assigned by the system to unannotated messages are reasonable in all cases. I am confident that RI with these parameters will perform well in categorizing the unannotated data sets, thus, determining healthcare providers' information needs, providing impetus to move on to this task.

Chapter 4: Characterization of Healthcare Providers' Information Needs

In the past, studies of healthcare providers' information needs have demonstrated that urban physicians have questions regarding patient care. Studies of health-related information needs in developed nations have shown that providers most frequently asked questions regarding treatment methods, diagnoses, and medications (Cheng, 2004; C. Cimino & Barnett, 1991; M. A. Clarke et al., 2013; Kourouthanassis et al., 2015; Osheroff et al., 1991; Smith, 1996). However, little is known about the information needs of rural healthcare providers in developing nations, such as Brazil.

These studies have been limited to qualitative studies using small sample sizes. For instance, in one study physicians filled out 303 surveys (Kourouthanassis et al., 2015), and in another study they participated in 47 interviews and filled out 47 surveys (Covell et al., 1985). In observational studies, 103 physicians participated in one study (J. W. Ely et al., 1999), while 24 physicians participated in another study (Osheroff et al., 1991). Automated methods, such as k-NN, provide the means to use large sample sizes to categorize information needs of healthcare providers. In this chapter, I describe the application of k-NN to the messages from the tele-health system to characterize rural healthcare providers' information needs. I also describe these information needs, both across regions and across provider types.

4.1 Methods

4.1.1 Categorization of the Unannotated Data Sets

The answers from the two unannotated sets were then categorized by applying the best-performing model (RI) to them. As with the training data set, the RI vector spaces of the unannotated data sets were built using: (1) the open source Semantic Vectors package for distributional semantics

(Widdows & Cohen, 2010; Widdows & Ferraro, 2008), (2) Portuguese stop word list (Boulton, 2014), and (3) TFIDF weighting. In this case, the word vectors emerged from distributional modeling of the training data set. These word vectors were summed to generate a document vector for each answer.

Once the vector spaces were built, k-NN 5 and 3 top categories were used to categorize the two unannotated data sets, using the categories assigned to the answers from the training data set. Next, each of the nearest-neighbor categories was assigned a score of the sum of the cosine values of each of the nearest-neighboring answers in which it appears. Categories were then rank ordered using this score, and the top 3 categories were chosen these categories were assigned to the unannotated answers, categorizing information needs. After these information needs were determined, information needs of the various healthcare providers were established.

The type of healthcare provider asking each question was determined from the metadata assigned to each question-answer pair. The categories assigned to each answer were assigned to the healthcare provider asking the corresponding question, thus, categorizing their information needs. The categories for answers to questions from each type of provider were then quantified.

4.2 Categorization of Information Needs

The distribution of category assignment is skewed toward a few highly prevalent categories (see Tables 4 and 5). As is apparent in the histograms in Tables 4 and 5, a small number of categories occur frequently, with many infrequently-assigned categories, which account for the remaining categorizations. The ten most frequently assigned categories comprise approximately 43% of all category assignments for PE, and approximately 38% of assignments made to MG. These are

presented in Table 6 for the purpose of comparative analysis of the prevailing concerns in each region.

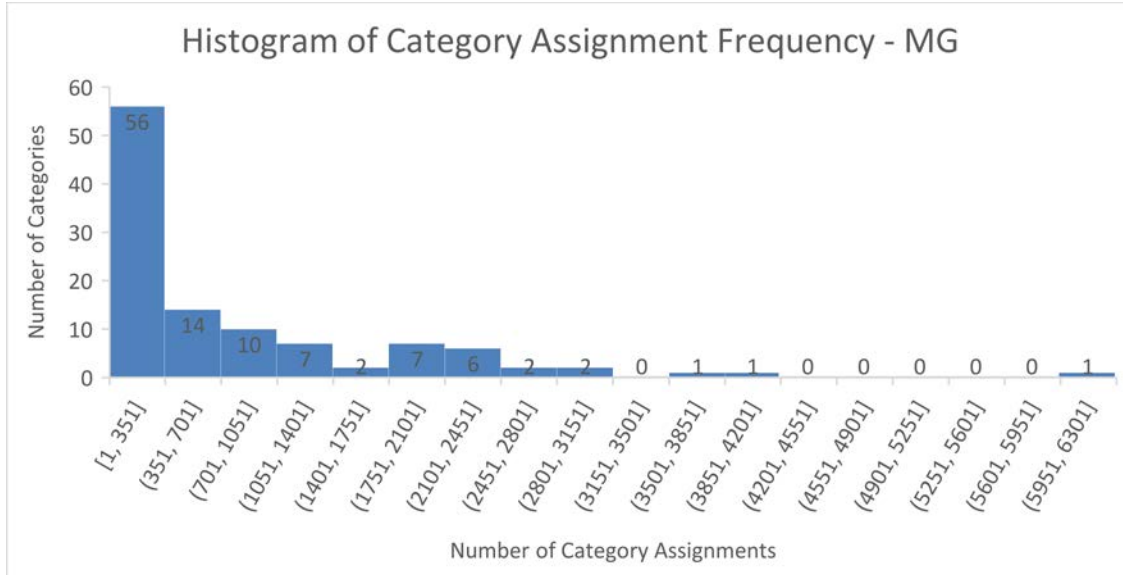


Table 4: Histogram of Category Assignment Frequency – MG

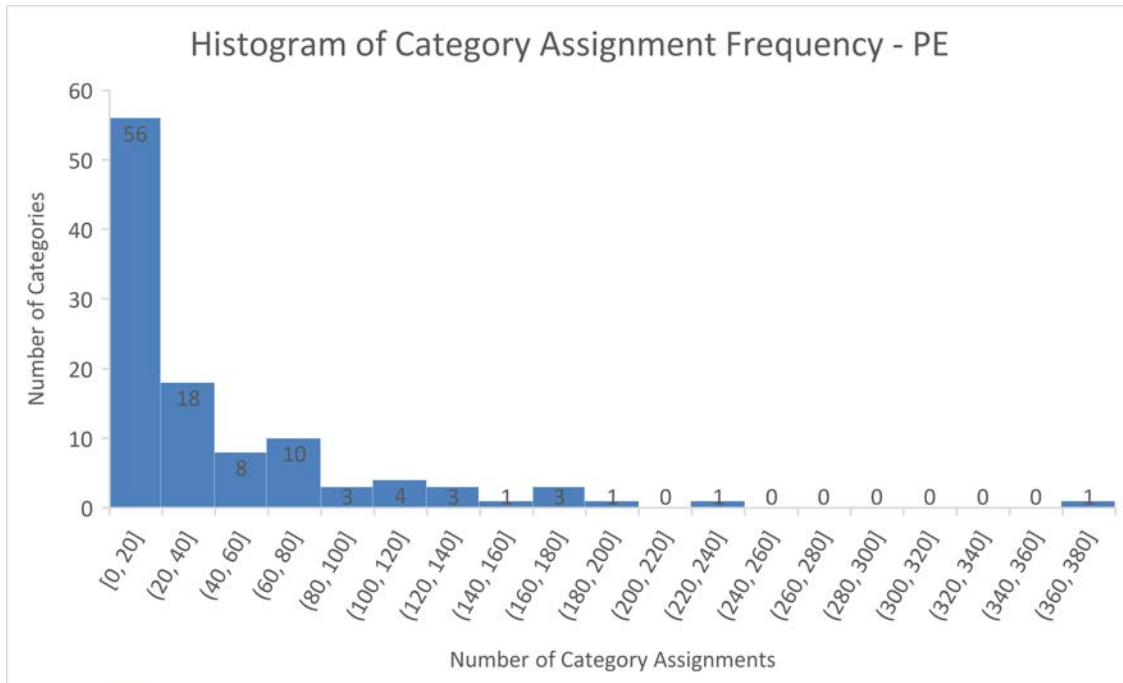


Table 5: Histogram of Category Assignment Frequency – PE

The top 10 PE categories include some categories that are ranked lower in the MG results. Specifically, "Therapeutics", "Female Urogenital Diseases and Pregnancy Complications" and "Reproductive Physiological Phenomena" are located within MG at Rank 11 (2.68%), Rank 12 (2.62%) and Rank 21 (1.93%), respectively. In contrast, the top 10 ranked MG categories include "Neoplasms", "Cardiovascular Diseases" and "Diagnosis" which are located within PE at Rank 13 (2.52%), Rank 18 (1.82%) and Rank 14 (2.45%), respectively (see Tables 7 and 8).

PE Category Name	PE # assignments	PE % assignments (n=4,129)	PE % answers (n=1,377)	RANK	MG% answers (n = 28,247)	MG % assignments (n = 84,687)	MG # assignments	MG Category Name
Persons	350	8.48%	25.42%	1	22.16%	7.39%	6260	Persons
Public Health	221	5.35%	16.05%	2	13.84%	4.62%	3909	Public Health
Bacterial Infections and Mycoses	186	4.50%	13.51%	3	13.04%	4.35%	3682	Skin and Connective Tissue Diseases
Health Services	164	3.97%	11.91%	4	10.91%	3.64%	3082	Bacterial Infections and Mycoses
Therapeutics	163	3.95%	11.84%	5	10.48%	3.50%	2960	Health Services
Female Urogenital Diseases and Pregnancy Complications	154	3.73%	11.18%	6	9.13%	3.04%	2578	Chemical Actions and Uses
Skin and Connective Tissue Diseases	134	3.25%	9.73%	7	8.83%	2.94%	2493	Health Personnel
Reproductive Physiological Phenomena	132	3.20%	9.59%	8	8.45%	2.82%	2387	Neoplasms
Health Personnel	128	3.10%	9.30%	9	8.23%	2.75%	2326	Cardiovascular Diseases
Chemical Actions and Uses	123	2.98%	8.93%	10	8.20%	2.73%	2316	Diagnosis

Table 6: “Top 10” unique categories for each region, in order of frequency of assignment

From left to right for PE (Pernambuco), and from right to left for MG (Minas Gerais), the columns proceed: “Category Name”; “# assignments”: the number of times a category was assigned to an answer; “% assignments”: the percentage of all assignments made up by this category; “% answers”: the percentage of all answers to which this category was assigned; and “Rank”.

PE Category Name	PE # assignments	PE % assignments (n=4,129)	PE % answers (n=1,377)	PE RANK	MG% answers (n = 28,247)	MG % assignments (n = 84,687)	MG # assignments	MG RANK
Therapeutics	163	3.95%	11.84%	5	8.04%	2.68%	2272	11
Female Urogenital Diseases and Pregnancy Complications	154	3.73%	11.18%	6	7.85%	2.62%	2217	12
Reproductive Physiological Phenomena	132	3.20%	9.59%	8	5.78%	1.93%	1633	21

Table 7: Unique Categories from Table 6 “Top 10” Present in PE but not MG

PE RANK	PE # assignments	PE % assignments (n=4,129)	PE % answers (n=1,377)	MG RANK	MG% answers (n = 28,247)	MG % assignments (n = 84,687)	MG # assignments	MG Category Name
13	104	2.52%	7.55%	8	8.45%	2.82%	2387	Neoplasms
18	75	1.82%	5.45%	9	8.23%	2.75%	2326	Cardiovascular Diseases
14	101	2.45%	7.33%	10	8.20%	2.73%	2316	Diagnosis

Table 8: Unique Categories from Table 6 “Top 10” Present in MG but not PE

Healthcare providers in both regions most often ask about "Persons" (e.g. adolescents) which isn't surprising, as this provides important contextual information. In both regions, information is needed about "Health Personnel". Providers' (nurses) questions about their responsibilities, and those of other healthcare professionals have been documented previously (Sanches, Alves, Lopes, & Novaes, 2012). "Public Health" relates to health promotion and disease prevention, important components of primary care in Brazil (Alkmim et al., 2012). "Health Services" include diagnosis and treatment of disease. The need for information about medication, is expressed by the category "Chemical Actions and Uses", and "Skin and Connective Tissue Diseases" needs concern dermatological conditions such as diabetic ulcers. Providers from both regions had concerns about "Bacterial Infections and Mycoses", encompassing such diseases as leprosy and tuberculosis, which impact the health of many Brazilians (Departamento de Informática do SUS, 2008). To assist with clarity of some of the categories contained in Table 9, top-level DeCS categories and their associated DeCS terms are presented in Appendix C. For viewing the similarities and differences between DeCS terms, I refer the interested reader to "DeCS Search" ("Search by Word", "Exact Descriptor") at (Biblioteca Virtual em Saúde, n.d.-a).

Top Level DeCS Category	DeCS Term (Portuguese)	DeCS Term (English)
Chemical Actions and Uses	Antibacterianos	Anti-Bacterial Agents
	Anticoagulantes	Anticoagulants
	Anticonvulsivantes	Anticonvulsants
	Antifúngicos	Antifungal Agents
Diagnosis	Acuidade Visual	Visual Acuity
	Autoexame de Mama	Breast Self-Examination
	Colposcopia	Colposcopy
	Monitorização Ambulatorial da Pressão	Arterial Blood Pressure Monitoring, Ambulatory
	Teste de Papanicolaou	Papanicolaou Test
	Triagem Neonatal	Neonatal Screening
Female Urogenital Diseases and Pregnancy Complications	Complicações na Gravidez	Pregnancy Complications
	Depressão Pós-Parto	Depression, Postpartum
	Pré-Eclâmpsia	Pre-Eclampsia
	Vaginite	Vaginitis
Health Services	Assistência Farmacêutica	Pharmaceutical Services
	Conduta do Tratamento Medicamentoso	Medication Therapy Management
	Cuidado do Lactente	Infant Care
	Serviços de Saúde Bucal	Dental Health Services
	Triagem	Triage
Public Health	Educação Alimentar e Nutricional	Food and Nutrition Education
	Estratégia Saúde da Família	Family Health Strategy
	Exposição a Praguicidas	Pesticide Exposure
	Nutrição da Criança	Child Nutrition
	Sistema Único de Saúde	Unified Health System
	Visita Domiciliar	Home Visit
Reproductive Physiological Phenomena	Gestação	Gestation
	Lactação	Lactation
	Menopausa	Menopause
Therapeutics	Hidratação	Fluid Therapy
	Higiene Bucal	Oral Hygiene
	Hospitalização	Hospitalization
	Vacinação	Vaccination

Table 9: Select Top Level DeCS Categories with Select DeCS Terms

However, there were some differences in the nature of these frequent categories across data sets. Only the PE “Top 10” includes “Therapeutics”, “Female Urogenital Diseases and Pregnancy Complications” and “Reproductive Physiological Phenomena”. It has been documented that healthcare providers, especially nurses, ask questions regarding obstetrics and gynecology (Sanches et al., 2012). This is evidenced by the requests for second opinions through the telehealth service about such topics as menopause, contraception, and sexually transmitted diseases. When dealing with complicated cases, PE healthcare providers may have additional questions related to treatment, as represented by "Therapeutics".

"Health Services" and "Therapeutics" are not entirely distinct from one another. As noted above, the category “Health Services” covers both diagnosis and treatment, whereas “Therapeutics” covers only treatment. Upon review of the answers coded with each of these categories, I noticed that "Health Services" includes neonatal screening (diagnosis) and vaccination (treatment), while "Therapeutics" includes contraception (treatment).

In contrast, only the MG “Top 10” includes “Neoplasms”, “Cardiovascular Diseases” and “Diagnosis”. The telehealth program in MG was originally set up for EKG referrals (Alkmim et al., 2012). Healthcare providers from MG also need information about cancer, particularly breast cancer, which has higher mortality rates in the south and southeast regions (MG) than in the north regions (PE) (Gonzaga et al., 2015).

Upon further analysis, PE answers assigned to the categories “Female Urogenital Diseases and Pregnancy Complications” and “Reproductive Physiological Phenomena” discuss such topics as pre-eclampsia and menopause respectively. In addition, MG answers assigned to "Neoplasms" and "Cardiovascular Diseases" discuss breast cancer and hypertension respectively. Hence, qualitative analysis provides evidence for the appropriateness of my system in assigning categories to the answers, as a means to characterize healthcare providers' information needs.

4.3 Categorization of Information Needs of Healthcare Providers Results and Discussion

4.3.1 PE: Categorization of Information Needs of Healthcare Providers Results

The PE data set is comprised of 12 unique types of requesters (see Table 10). Of the categories assigned by the model, nurses rank 1 of 12 with 64.45% (2,661 / 4,129) assignments, community health workers rank 2 of 12 with 10.68% (441 / 4,129) assignments, and doctors rank 4 of 12 with 7.992% (330 / 4,129) assignments. The cumulative total of the categories assigned to the three requesters, doctors, nurses, and community health workers, comprise 83.12% of the 4,129 assignments made by the model.

Pe Requester	Pe Raw number of times unique category assigned to unique Requester (n= 4,129)	Pe %	Rank	MG %	MG Raw number of times unique category assigned to unique Requester (n= 84,687)	MG Requester
Nurse	2661	64.447%	1	43.128%	36524	Nurse
Community Health Worker	441	10.681%	2	41.172%	34867	Doctor
Dental surgeon	360	8.719%	3	3.542%	3000	Dentist
Doctor	330	7.992%	4	3.015%	2553	Doctor (TeleConsultations)
Nursing Technician	147	3.560%	5	1.899%	1608	Nutritionist
Other Health Professionals	118	2.858%	6	1.892%	1602	Pharmacist
Other Professionals	21	0.509%	7	1.743%	1476	Physiotherapist
Primary Care Coordinator	21	0.509%	8	1.169%	990	Nurse (TeleConsultations)
Technology Professionals	12	0.291%	9	0.914%	774	Speech Therapist
Oral Health Attendant	9	0.218%	10	0.808%	684	Psychologist
Physiotherapist	6	0.145%	11	0.315%	267	Technical Nursing
No information	3	0.073%	12	0.103%	87	Physical educator
			13	0.099%	84	Biochemist
			14	0.064%	54	Social Worker
			15	0.050%	42	Carga
			16	0.032%	27	Biomedic
			17	0.014%	12	Occupational Therapist
			18	0.011%	9	Public Health Manager
			19	0.011%	9	Biologist
			20	0.011%	9	Psychiatrist
			21	0.007%	6	Community Health Worker
			22	0.004%	3	Regulator

Table 10: Ranking of Requester, based upon raw the raw number of times, in decreasing order, a unique category was assigned by the model to an answer of a question asked by Requester type

Doctor	Raw #	% (n=4129)	Rank	% (n=4129)	Raw #	Nurse	Raw #	% (n=4129)	Rank	% (n=4129)	Raw #	CHW
Persons	23	0.557%	1	5.231%	216	Persons	216	5.231%	1	1.090%	45	Persons
Female Urogenital Diseases and Pregnancy Complications	19	0.460%	2	3.439%	142	Public Health	142	3.439%	2	0.678%	28	Health Personnel
Skin and Connective Tissue Diseases	17	0.412%	3	3.391%	140	Bacterial Infections and Mycoses	140	3.391%	3	0.654%	27	Public Health
Public Health	16	0.388%	4	2.882%	119	Therapeutics	119	2.882%	4	0.484%	20	Health Services
Chemical Actions and Uses	13	0.315%	5	2.737%	113	Female Urogenital Diseases and Pregnancy Complications	113	2.737%	5	0.412%	17	Bacterial Infections and Mycoses
Neoplasms	12	0.291%	6	2.446%	101	Health Services	101	2.446%	6	0.412%	17	Therapeutics
Health Services	11	0.266%	7	2.155%	89	Reproductive Physiological Phenomena	89	2.155%	7	0.291%	12	Nervous System Diseases
Reproductive Physiological Phenomena	11	0.266%	8	2.131%	88	Skin and Connective Tissue Diseases	88	2.131%	8	0.291%	12	Reproductive Physiological Phenomena
Bacterial Infections and Mycoses	10	0.242%	9	1.986%	82	Investigative Techniques	82	1.986%	9	0.291%	12	Skin and Connective Tissue Diseases
Pathologic Processes	10	0.242%	10	1.962%	81	Diagnosis	81	1.962%	10	0.266%	11	Neoplasms

Table 11: PE Top 10 unique categories, ranked based upon the raw number of times, in decreasing order, the unique category was assigned

The two "Top 10" categories assigned only to doctors include "Chemical Actions and Uses" and "Pathologic Processes" (see Table 11). The two categories assigned only to nurses are "Investigative Techniques" and "Diagnosis". The two "Top 10" categories assigned only to community health workers include "Health Personnel", "Nervous System Diseases".

4.3.2 PE: Categorization of Information Needs of Healthcare Providers Discussion

As the data shows, the three healthcare providers have many information needs in common. For nurses and doctors, this is likely due to the fact that they both provide direct patient care. On the other hand, the Community Health Worker (CHW) is the link between the population and healthcare teams. Through home visits, they mentor, monitor, and educate the population, as well as promote access to the healthcare teams (Galavote et al., 2011). In addition, CHWs discuss cases with the healthcare team (Peres et al., 2011), possibly explaining the high percentage of similar information needs to the doctors and nurses. Despite the similarities in information needs, there are differences as well.

The two "Top 10" categories assigned only to doctors include "Chemical Actions and Uses" and "Pathologic Processes". "Chemical Actions and uses" covers medications, while "Pathologic Processes" cover dysmenorrhea and oligomenorrhea. As evidenced by the subjects encompassed by "Pathologic Processes", doctors also have information needs regarding female issues. When dealing with complicated cases, PE doctors may also have additional questions related to medications for treatment of these various conditions, as

represented by "Chemical Actions and Uses". Like doctors, nurses have their own information needs.

The two "Top 10" categories assigned only to nurses include "Investigative Techniques" and "Diagnosis". "Investigative Techniques" covers laboratory tests, while "Diagnosis" covers mammography. It has been noted that breast cancer has been increasing in the Northeast region of Brazil (Freitas-Junior, Gonzaga, Freitas, Martins, & de Cássia de Maio Dardes, 2012; Gonzaga et al., 2015). This possibly explains the need for more information regarding lab tests, such as biopsies, and mammography. CHWs also have specific information needs.

During home visits, CHWs encounter many disease conditions, creating information needs in such areas as "Nervous System Diseases". Stroke (a nervous system disease) is the third cause of death in both men and women (Leite et al., 2015). As the CHW provides referrals to and discusses cases with the healthcare teams, they often need information regarding the appropriate "Health Personnel" to treat these cases.

4.3.3 MG: Categorization of Information Needs of Healthcare Providers Results

The MG data set is comprised of 22 unique types of requesters (see Table 10). Of the categories assigned by the model, nurses rank 1 of 22 with (36524 / 84687) 43.128% assignments, doctors rank 2 of 22 with (34867 / 84687) 41.172% assignments, and community health workers rank 21 of 22 with (6 / 84687) 0.007% assignments. The cumulative total of the categories assigned to the three requesters, doctors, nurses, and

community health workers, comprise 84.307% of the 84687 assignments made by the model.

Note that for CHWs the model assigned only 6 (not 10) categories and comprises approximately 0.007% of the categories assigned.

Information needs expressed only by doctors involve "Nervous System Diseases", "Female Urogenital Diseases and Pregnancy Complications", "Cardiovascular Diseases" and "Neoplasms"(see Table 11). The four categories assigned only to nurses are "Therapeutics", "Health Personnel", "Investigative Techniques", and "Diagnosis". The three "Top 6" categories assigned only to community health workers include "Physiological Phenomena", "Health Services Management" and "Nutritional and Metabolic Diseases".

Doctor	Raw #	% (n=84687)	Rank	% (n=84687)	Raw #	Nurse	Raw #	% (n=84687)	Rank	% (n=84687)	Raw #	CHW
Persons	2342	2.765%	1	3.224%	2730	Persons	2730	3.224%	1	0.0012%	1	Persons
Skin and Connective Tissue Diseases	1443	1.704%	2	2.161%	1830	Skin and Connective Tissue Diseases	1830	2.161%	2	0.0012%	1	Endocrine System Diseases
Public Health	1437	1.697%	3	2.135%	1808	Public Health	1808	2.135%	3	0.0012%	1	Physiological Phenomena
Bacterial Infections and Mycoses	1355	1.600%	4	1.731%	1466	Health Services	1466	1.731%	4	0.0012%	1	Health Services
Chemical Actions and Uses	1162	1.372%	5	1.573%	1332	Bacterial Infections and Mycoses	1332	1.573%	5	0.0012%	1	Health Services Management
Cardiovascular Diseases	1144	1.351%	6	1.385%	1173	Therapeutics	1173	1.385%	6	0.0012%	1	Nutritional and Metabolic Diseases
Neoplasms	1135	1.340%	7	1.325%	1122	Health Personnel	1122	1.325%	7			
Endocrine System Diseases	1070	1.263%	8	1.237%	1048	Investigative Techniques	1048	1.237%	8			
Nervous System Diseases	987	1.165%	9	1.228%	1040	Chemical Actions and Uses	1040	1.228%	9			
Female Urogenital Diseases and Pregnancy Complications	984	1.162%	10	1.222%	1035	Diagnosis	1035	1.222%	10			

Table 12: MG Top 10 unique categories, ranked based upon the raw number of times, in decreasing order, the unique category was assigned

4.3.4 MG: Categorization of Information Needs of Healthcare Providers Discussion

The MG data shows that doctors and nurses have the most information needs in common, likely for the same reasons as they do in PE. The CHWs asked very few questions (2 out of 28247), and thus it is not possible to draw any inferences about their information needs. In addition, the small number of questions asked by MG CHWs is explained by policy differences across states. Like the healthcare providers from PE, MG healthcare providers also have different information needs.

The four "Top 10" categories assigned only to doctors include "Nervous System Diseases", "Female Urogenital Diseases and Pregnancy Complications", "Cardiovascular Diseases" and "Neoplasms". Doctors are concerned about "Cardiovascular Diseases", because they are the leading cause of death in Minas Gerais (Alkmim et al., 2012). In addition, those with cardiovascular diseases are at high risk for having a stroke (Vahedi & Amarenco, 2000), presenting another information need, "Nervous System Diseases". Doctors also need information about "Neoplasms". Brazil has experienced an increase in mortality from prostate, breast, lung, and colorectal cancers (Schmidt et al., 2011). Doctors have concerns about "female urogenital and pregnancy complications". According to Nazareth and colleagues (Nazareth et al., 2017), doctors in Minas Gerais see many high-risk pregnancies due to diabetes and high blood pressure.

Nurses also have their own specific information needs.

The four "Top 10" categories assigned only to nurses include "Therapeutics" (treatment), "Health Personnel" (healthcare providers), "Investigative Techniques" (lab tests), and

"Diagnosis" (electrocardiography). These information needs can represent such tasks as caregiving (treatment), monitoring and assessing the patient (lab tests, and electrocardiography), and understanding responsibilities of various healthcare providers ("Health Personnel"). This knowledge is important in order for nurses to perform their duties (Sanches et al., 2012).

4.4 Information Needs of Healthcare Providers Across Regions

Table 13 illustrates distinguishing features of information needs across providers and regions. Of note, nurses in both regions are distinguished by information needs concerning "Investigative Techniques" which usually correspond to screening tests, demonstrating their role in health promotion. These needs are different than those identified for urban nurses in previous research (Spath & Buttlar, 1996). In addition, doctors' information needs differ across regions, as well as from their urban counterparts. Community Health Workers (CHWs) are not represented in this table, as the MG data set only included a single CHW question-answer pair. It is unclear why CHWs were not represented in this data set, although anecdotally I have been told that the policies governing access to telehealth facilities vary across regions.

PE - Doctors				MG - Doctors		
%	Count of assignments made by model (n=4,129)	Category	Rank	Category	Count of assignments made by model (n=84,687)	%
0.557%	23	Persons	1	Persons	2342	2.765%
0.460%	19	Female Urogenital Diseases and Pregnancy Complications	2	Skin and Connective Tissue Diseases	1443	1.704%
0.412%	17	Skin and Connective Tissue Diseases	3	Public Health	1437	1.697%
0.388%	16	Public Health	4	Bacterial Infections and Mycoses	1355	1.600%
0.315%	13	Chemical Actions and Uses	5	Chemical Actions and Uses	1162	1.372%
0.291%	12	Neoplasms	6	<i>Cardiovascular Diseases</i>	1144	1.351%
0.266%	11	Health Services	7	<i>Neoplasms</i>	1135	1.340%
0.266%	11	Reproductive Physiological Phenomena	8	Endocrine System Diseases	1070	1.263%
0.242%	10	Bacterial Infections and Mycoses	9	<i>Nervous System Diseases</i>	987	1.165%
0.242%	10	Pathologic Processes	10	<i>Female Urogenital Diseases and Pregnancy Complications</i>	984	1.162%
PE – Nurses				MG - Nurses		
%	Count of assignments made by model (n=4,129)	Category	Rank	Category	Count of assignments made by model (n=84,687)	%
5.231%	216	Persons	1	Persons	2730	3.224%
3.439%	142	Public Health	2	Skin and Connective Tissue Diseases	1830	2.161%
3.391%	140	Bacterial Infections and Mycoses	3	Public Health	1808	2.135%
2.882%	119	Therapeutics	4	Health Services	1466	1.731%
2.737%	113	Female Urogenital Diseases and Pregnancy Complications	5	Bacterial Infections and Mycoses	1332	1.573%
2.446%	101	Health Services	6	<i>Therapeutics</i>	1173	1.385%
2.155%	89	Reproductive Physiological Phenomena	7	<i>Health Personnel</i>	1122	1.325%
2.131%	88	Skin and Connective Tissue Diseases	8	<i>Investigative Techniques</i>	1048	1.237%
1.986%	82	Investigative Techniques	9	Chemical Actions and Uses	1040	1.228%
1.962%	81	Diagnosis	10	<i>Diagnosis</i>	1035	1.222%

Table 13: Categories that occurred in the top 10 for certain provider types only PE Doctors versus PE Nurses, categories that occurred in respective provider type only are displayed in bold. MG Doctors versus MG Nurses, categories that occurred in respective provider type only are displayed in *italic*.

As shown in Table 14, rural MG doctors need information about medication, treatment and diagnosis of disease like their urban counterparts, while rural PE doctors need information about medication and treatment (demonstrated by the “Yes”). According to Spath and Buttlar (Spath & Buttlar, 1996), nurses in urban areas need information regarding diagnosis of disease, medication, techniques, and equipment. Like their urban counterparts, the nurses from rural Minas Gerais and Pernambuco need information about diagnosis of diseases and medication. Although nurses do need information about equipment, this category was not as prominently discussed (i.e., frequency in the second quartile). As noted in Table 16, there is no corresponding category for "techniques". This is due to the fact that MeSH has no dedicated tree for this category. Thus, the status of rural nurses' information needs regarding techniques is not known at the current time.

Categories representing information needs	Urban Doctors (see Clarke (M. A. Clarke et al., 2013))	PE Doctors	MG Doctors	Urban Nurses (see Spath & Buttlar, 1996))	PE Nurses	MG Nurses
Treatment	Yes	“Therapeutics” 1 st Quartile 16/64	“Therapeutics” 1 st Quartile 18/107	Not mentioned	N/A	N/A
Diagnosis	Yes	“Diagnosis” 2 nd Quartile 23/64	“Diagnosis” 1 st Quartile 13/107	Yes	“Diagnosis” 1 st Quartile 10/83	“Diagnosis” 1 st Quartile 10/107
Medication	Yes	“Chemical Actions and uses” 1 st Quartile 5/64	“Chemical Actions and uses” 1 st Quartile 5/107	Yes	“Chemical Actions and uses” 1 st Quartile 13/83	“Chemical Actions and uses” 1 st Quartile 9/107
Techniques	Not mentioned	N/A	N/A	Yes	No appropriate corresponding category	No appropriate corresponding category
Equipment	Not mentioned	N/A	N/A	Yes	“Equipment and Supplies” 2 nd Quartile 28/83	“Equipment and Supplies” 2 nd Quartile 28/107

Table 14: Categories representing information needs of Urban versus Rural healthcare providers

Urban needs are derived from discussion in (M. A. Clarke et al., 2013) and (Spath & Buttlar, 1996). Rural needs indicate any category that was assigned with a frequency above the 3rd quartile.

4.5 Information Needs Theory

Like many professionals, nurses, doctors and CHWs express information needs. According to the "Anomalous State Knowledge" (ASK) theory, these professionals recognize a gap in their knowledge regarding a particular subject (Belkin, 2005; Case & Given, 2016). This, in turn, can lead the health professional to seek information to satisfy this gap in knowledge.

The information seeking behavior is often complex and involving many factors. These factors include: (1) context or environment in which the information seeker is located; (2) expertise or training and education; (3) the information recipient's needs, wants, and goals; (4) the information provider's needs, wants, and goals; (5) motivating and inhibiting factors; (6) characteristics and utility of the information channels selected and used by seekers; and (7) characteristics of information sources (Case & Given, 2016). In this case the healthcare with the information need is located in the basic health unit. The doctors and nurses have medical training. On the other hand, the CHWs have very basic training in disease prevention and health promotion (Bornstein & Stotz, 2008). The doctors and nurses, information recipients, need and information in order to provide care to their patients (goals). The CHWS need and want information to mentor, monitor, and educate the community to improve health and well-being (Galavote et al., 2011). The needs, wants, and goals of the information provider, doctors and nurses in the telehealth centers, involve giving health information to assist the rural healthcare providers in providing care to their patients. Motivating factors may include resolution of their doubts. Inhibiting factors may involve lack of Internet connectivity (de Souza et al., 2017; Joshi et al., 2011), or difficulties in using the telehealth system.

There are the characteristics and utility of the information channels selected and used by seekers. The channels of communication include written, electronic, face-to-face, etc. (Case & Given, 2016). In this case the channel of communication is electronic, through the telehealth system. In addition, there are characteristics of information sources. These characteristics include utility (e.g., relevance, timeliness, ease-of-use) and credibility (e.g.,

authority, reliability, lack of bias) (Case & Given, 2016). The information is relevant as the healthcare providers in the telehealth centers provide information specific to each case. The ease-of-use applies to the telehealth system (e.g. filling in the online forms for each case by the rural healthcare provider). Timeliness involves the time for the doctors and nurses in the telehealth centers to respond to the rural healthcare providers. For credibility, the doctors and nurses in the telehealth centers are an authoritative and reliable source of information relatively without bias as they provide additional citations from the literature along with their answers.

4.6 Conclusion

Although healthcare providers in Pernambuco and Minas Gerais have the same information needs, analysis of the top ten categories demonstrates that there are some differences. Healthcare providers need information regarding "Therapeutics" (treatment), "Female Urogenital Diseases and Pregnancy Complications" and "Reproductive Physiological Phenomena", evidenced by the requests for second opinions through the telehealth service about such topics as menopause, contraception, and sexually transmitted diseases. In Minas Gerais, healthcare providers need information about "Neoplasms", "Diagnosis" and "Cardiovascular Diseases", reflecting the need for EKG referrals (Alkmim et al., 2012), and the high risk of mortality from breast cancer in this region (Gonzaga et al., 2015).

Doctors, nurses, and CHWs in Pernambuco have the same information needs as well. However, the top ten categories show that there are some differences. Doctors preferentially expressed a need for information about "Chemical Actions and Uses"

(medication) and "Pathologic Processes"; nurses preferentially need information regarding "Investigative Techniques" (lab tests) and "Diagnosis"; and CHWs preferentially express a need for information about "Health Personnel", and "Nervous System Diseases.

In contrast, analysis of the top ten categories for healthcare providers in Minas Gerais indicate the similarities and differences in their information needs. Information needs preferentially expressed by doctors involve "Nervous System Diseases", "Female Urogenital Diseases and Pregnancy Complications", "Cardiovascular Diseases" and "Neoplasms", while those for nurses involve "Therapeutics", "Health Personnel", "Investigative Techniques", and "Diagnosis". There was a small number of category assignments made to Minas Gerais CHWs, making it difficult to draw inferences about their expressing preferentially information needs. Anecdotally I have been told that the policies governing access to telehealth facilities vary across regions. The information needs of healthcare providers from Pernambuco and Minas Gerais reflect their roles and health concerns of these regions.

It has been shown that urban doctors need information about treatment, diagnosis of disease, and medications (Cheng, 2004; C. Cimino & Barnett, 1991; M. A. Clarke et al., 2013; Kourouthanassis et al., 2015; Osheroff et al., 1991; Smith, 1996). Like their urban counterparts, doctors in rural Minas Gerais also need information about treatment, diagnosis of disease, and medication. However, doctors from rural Pernambuco expressed a need for information about medication and treatment. According to Spath and Buttlar (Spath & Buttlar, 1996), nurses in urban areas need information regarding diagnosis of

disease, medication, techniques, and equipment. Like their urban counterparts, the nurses from rural Pernambuco and rural Minas Gerais need information about diagnosis of diseases and medication. My research demonstrates that nurses in both regions have information needs concerning “Investigative Techniques”, which correspond to screening tests, demonstrating their role in health promotion.

Chapter 5: Analysis of Temporal Dynamics of Healthcare Providers' Information Needs

Much of the prior research has characterized information needs at a single point in time, due to the qualitative methods employed. These methods use questionnaires (Covell et al., 1985; Huth, 1989; Kourouthanassis et al., 2015), observations (C. Cimino & Barnett, 1991; Covell et al., 1985), and interviews (J. W. Ely et al., 1999; Osheroff et al., 1991). In this chapter I describe the methods used to evaluate the messages from the telehealth center in Pernambuco in order to characterize the temporal dynamics of the healthcare providers' information needs.

5.1 Materials

5.1.1 Data Set

The PE answers with the 90 unique categories assigned by RI at k of 5 and 3 top categories (the best-performing configuration aside from $k=1$) were collected and prepared for analysis. The metadata, representing years and months for 2010-2012, were gathered for each PE answer. Each of the three years contained four seasons, defined as follows: (1) summer with December through February "(e.g., December 2010, January 2011, February 2011), (2) fall with March through May, (3) winter with June through August, and (4) spring with September through November. The answers for each category were then separated into their corresponding season and year and quantified, giving 12 data points (one sum per season per year) for each category.

Due to the fact that the total number of answers was widely different for each year, the PE data set was normalized before analysis. To normalize the data, proportions for each of the 12 year/season data points were calculated. First, for each year/season combination, the number of answers for each of the 90 categories for that year/season were added together, giving 12 totals. For example, for summer 2010, the number of answers for each of the 90 categories in summer 2010 was summed, giving a total for summer 2010. In the next step, the number of answers at each year/season slot for each category was divided by the corresponding year/season total. For example, for the category "Virus Diseases", the number of answers for summer 2010 is divided by the total for summer 2010, and so on. Thus, the resulting data point reflects the proportion of answers to which "Virus Diseases" was assigned, rather than the raw number of assignments, providing a basis for comparison across time periods in the context of variation in the total number of reports. These discrepancies are an artifact of the data collection process. Sometimes, basic health units experience difficulties with Internet connectivity (de Souza et al., 2017; Joshi et al., 2011), affecting the number of questions asked. This, in turn, affects the amount of data collected.

5.1.2 Surveillance System in Brazil

Many countries have a surveillance system for reporting cases of infectious diseases. In Brazil, the national notifiable diseases information system (Brazilian Ministry of Health & Sistema Único de Saúde, n.d.) is used for this purpose (Galvão et al., 2008). In fact, the law that established SINAN also regulates the compulsory notification of infectious diseases, such as leprosy, cholera, meningitis, measles, tuberculosis and many others (Galvão et al., 2008). Surveillance happens at three levels: municipality, state and federal

(Galvão et al., 2008). Disease notification forms are completed at the basic health units, sent to the municipality, where the data is transferred to a SINAN file (Galvão et al., 2008). From here, the SINAN files are sent to the health region (composed of several municipalities) then on to the state (Galvão et al., 2008). The data from all of the municipalities are consolidated by the state and sent to the federal level (Galvão et al., 2008).

5.2 Methods

5.2.1 Outlier Detection

There are many definitions for outliers (Kannan, Manoj, & Arumugam, 2015). However, an outlier is often defined as: an observation that deviates markedly from other observations in the sample (Kannan et al., 2015; National Institute of Standards and Technology, 2013a). Outliers can be considered bad data and are often deleted from the data set (Kannan et al., 2015). However, outliers may also be due to random variation, and represent an interesting finding (National Institute of Standards and Technology, 2013a), thus, it is for this reason that I selected six methods for outlier detection: modified Z-Score (Kannan et al., 2015; Manoj, 2015; Seo, 2006), “adjusted Z-Score” (IBM, n.d.), Tukey inner fence (Seo, 2006; Tukey, 1977), Tukey outer fence (Seo, 2006; Tukey, 1977), Grubbs test (Grubbs, 1950, 1969; Grubbs & Beck, 1972; National Institute of Standards and Technology, 2013b), and Dixon Q test (Dixon, 1950; National Institute of Standards and Technology, 2015). As with many statistical tests, underlying assumptions have to be met, such as distribution of the data, and sample size. Many tests for outlier detection require a normal distribution (Kannan et al., 2015; Seo, 2006). Although the data does not follow a

normal distribution, these outlier detection methods can be used, as long as they are performed for exploratory purposes and not for data deletion (Seo, 2006). In addition, some tests can only be used on larger data sets, while others can be used on small data sets. The tests I use do work on smaller data sets. For additional reasons, I reduced the number of methods for outlier detection from six to four excluding: (1) adjusted Z-Score, as it can be too sensitive in detecting outliers; and (2) Tukey inner fence, as this test detects "possible" outliers (Seo, 2006). Thus, the four selected tests include: modified Z-Score, Tukey outer fence, Grubbs test, and Dixon Q test. Of the different Z-Score tests, the Modified Z-Score was chosen, as Iglewicz and Hoaglin (Iglewicz & Hoaglin, 1993) recommend this test for small sample sizes. In addition, these authors also note that the Z-Score can be affected by extreme values as it uses the mean and standard deviation (Iglewicz & Hoaglin, 1993). Although the Grubbs test uses the mean and standard deviation, this test was used for two reasons: (1) it can be used with small data sets, and (2) this test is being used for exploratory purposes, not data deletion. In the case of the Tukey tests, the outer fences have been selected, as they represent "probable far" outliers, whereas the Tukey inner fences represent "possible" outliers (Seo, 2006). These methods of outlier detection will provide a means to identify the year and season in which there is a greater need for information as represented by the categories.

Each of these tests uses different parameters for detecting outliers. The modified Z-Score is based upon the median and the median of the absolute deviation of the median (MAD) (Kannan et al., 2015; Manoj, 2015; Seo, 2006). The Tukey test measures outlier values based upon the median and calculated boundaries in the data (Seo, 2006; Tukey, 1977).

The Grubbs test is based upon the number of standard deviations that the extreme observations are from the mean (Grubbs, 1969). Finally, the Dixon Q test is based upon a value being too large (or small) compared to its nearest neighboring value (National Institute of Standards and Technology, 2015). These tests were chosen, as the different parameters provide different ways in which to explore the data.

The first test to detect category outliers was the modified Z-Score (Kannan et al., 2015; Manoj, 2015; Seo, 2006). The modified Z-Score was calculated using the following equation: modified $Z_i = 0.6745 * (X_i - X_m) / MAD$, where x_i is the observation, x_m the median of the data, and MAD is the median of absolute deviation about the median. $MAD = median \{|x_i - x_m|\}$ or the median of the (absolute value of (observation - median)). In this case, the observation (x_i) is the value representing the number of answers in a category's year/season slot. The median (x_m) was calculated by first ordering the values in the year/season slots in ascending order, summing the sixth and seventh values, and dividing by 2. For the data the MAD was calculated as: the median of (the absolute value of (the value for a category's year/season slot minus the median of the 12 year/season slots for that category)). A value for a year/season slot is considered an outlier when the modified Z-Score was greater than 3.5 (Iglewicz & Hoaglin, 1993).

Another test used to detect outliers for each category was the Tukey test (Tukey, 1977). Before calculating the values for this method of analysis, the data were sorted in ascending order. Then, the three values needed for this test were calculated: (1) Q1 (the upper boundary of the bottom quartile) which is the median of the lower half of the data set (i.e. median of the first six values); (2) Q3 (the lower boundary of the top quartile), which is the

median of the upper half of the data set (i.e. median of values seven through 12); and (3) IQR (interquartile range), which is the difference between Q3 and Q1. These values were then used for calculating: (1) *upper outer fence*, and (2) the *lower outer fence*. The equations were as follows: (1) $(Q3 + (3 * IQR))$ for the upper outer fence, and (2) $(Q1 - (3 * IQR))$ for the lower outer fence (Tukey, 1977). As suggested by Tukey (Tukey, 1977), values outside of the outer fences are considered a "probable far" outlier. In this case, any value in the data below the lower outer fence and above the upper outer fence was considered an outlier, and thus, was used for analysis.

Grubbs' test (Grubbs, 1950, 1969; Grubbs & Beck, 1972) was also used to detect outliers for each category. For this method of analysis, the following equation was utilized: $G = |(suspected\ outlier - mean) / SD|$. In this case, the mean represents the mean of the 12 values for the year/season slots for a category. *SD* is the standard deviation of the 12 values for the year/season slots for the same category. According to the statistical reference table provided in Grubbs and Beck (Grubbs & Beck, 1972), the critical value for the G statistic with a sample size of 12 is 2.285, leading to an outlier definition of $G > 2.285$.

The final method of analysis used for outlier detection was the Dixon's Q test (Dixon, 1950). Two equations were used to detect outliers, one for detecting high outliers, and one for detecting low outliers. The equation used for detecting high-value outliers was as follows: $Q = (y(n) - y(n-2)) / (y(n) - y2)$ (National Institute of Standards and Technology, 2015). The equation used for detecting low-value outliers was as follows: $Q = (y3 - y1) / (y(n-1) - y1)$ (National Institute of Standards and Technology, 2015). Before the calculations were performed, the values for each category were sorted from lowest to

highest. Then, for these equations: $y(n)$ was the highest (or 12th) value for the category; y_1 was the lowest (or first) value for the category; y_2 was the second lowest (or second) value for the category; y_3 was the third lowest (or third) value for the category; $y(n-1)$ was the second highest (or eleventh) value for the category; $y(n-2)$ was the third highest (or tenth) value for the category; $y(n-3)$ was the fourth highest (or ninth) value for the category. The Dixon Q critical value table (Rorabacher, 1991) was consulted. It was determined that the critical value for Q at a 0.05 level of significance with a sample size of 12 was 0.426. Thus, any value of Q greater than 0.426 was considered an outlier.

5.2.2 Qualitative Analysis

Grounded theory (Corbin & Strauss, 2007; Glaser & Strauss, 1967) involves construction of a hypothesis or discovery of concepts through data analysis (Faggiolani, 2011; P. Y. Martin & Turner, 1986). In this method, the researcher works through a set of overlapping steps: Note-taking, coding, memoing, and sorting (Dick, 2014). As with most research, the process begins with data collection. The sources of data are varied, but can include interviews, lectures, seminars, and expert group meetings (Ralph, Birks, & Chapman, 2014). In the next step, note-taking is performed. This involves examining a chunk of data, such as a sentence, an abstract, (in my case, an answer) and writing down the key concepts (Dick, 2014) represented by words or phrases (Woods, Gapp, & King, 2016). Then, coding begins by constant comparison of data (sentence or answer) to data (sentence or answer). The information from this comparison is written in the margins of the note-taking as identified themes or variables (categories) and properties (subcategories) (Dick, 2014;

Woods et al., 2016). As the coding progresses, links between categories may appear, which are noted in memos (Dick, 2014). Once there are no additional variables and properties, the researcher begins the process of sorting. Here, like memos are grouped together and are ordered in a sequence that makes the most sense (Dick, 2014). These steps are performed until saturation is reached and no new concepts or categories appear (Woods et al., 2016).

A modified form of grounded theory is deductive content analysis. In this type of analysis, the researcher begins with pre-set aims (Blackstone, 2012; Elo & Kyngäs, 2008; Pope, Ziebland, & Mays, 2000) or categories (Bradley, Curry, & Devers, 2007). According to Bradley and colleagues (Bradley et al., 2007), the initial step is to define the categories. Then, the data is reviewed, looking for patterns that fit into the pre-determined categories, in other words, moving from general to specific (Blackstone, 2012), or using a top-down approach (Soiferman, 2010). However, one must take care to avoid forcing data into these categories just because they exist (Bradley et al., 2007). Although forcing the data is a possibility, these categories do allow us to benefit from and build on insights from the field.

The modified grounded theory approach was chosen to analyze the answers for the outlier year/season slot of the selected categories. This method will assist in the discovery of concepts during these outlier seasons, providing a more complete picture of the change in healthcare providers' information needs during these times. These changes in information needs will then show important concepts pertinent to public health in the region.

Categories for which outliers were detected by the modified Z-Score, the Tukey outer fence, the Grubbs test, and the Dixon Q test were collected for qualitative analysis. There

is no consensus or widely accepted rationale for choosing between these four tests. Thus, I considered cases in which the four tests agree with one another, as having the strongest support for outlier status, and for using three of four tests in special cases. In addition, some categories in which outliers were detected by three tests were also chosen, using the following criteria: (1) instances where the category name contained the term “diseases”, as these categories may reflect diseases in the area; and (2) instances in which the category represented “Women’s issues”, as these categories may provide topics representing information needs important to the area. In addition, three categories, "respiratory Tract Diseases", "Bacterial Infections and Mycoses" and "Virus Diseases", were selected for qualitative analysis, as they may provide a view of the infectious diseases discussed by healthcare providers in Pernambuco, thus, confirming outbreaks of these diseases. As healthcare providers often need information regarding obstetrics and gynecology (Sanches et al., 2012), "Female Urogenital and Pregnancy Complications" was selected for qualitative analysis.

In the next step, the answers corresponding to the outlier year/season slots for the selected categories were obtained. For those categories for which there were no detected outliers, the answers from the year/season slots with the greatest number of answers were chosen for analysis. All of the answers within the selected year/season slot were analyzed using the modified grounded theory approach to determine the major and minor topics discussed (see Table 15). During the qualitative analysis two criteria were set: (1) major topics discussed were present in half or more than half of the answers, and (2) minor topics discussed were present in less than half of the answers.

Category	Number (out of 4) of outlier tests passed	Time period (year/season)	Number of Answers
Bacterial Infections & Mycoses	1	2012_Summer	11
Behavior and Behavior Mechanisms	3	2012_Spring	5
Biological Phenomena	4	2012_Spring	5
Endocrine System Diseases	4	2010_Spring	15
Equipment and Supplies	4	2012_Spring	6
Female Urogenital Diseases and Pregnancy Complications	1	2010_Spring	32
Immune System Diseases	4	2012_Fall	8
Investigative Techniques	4	2011_Fall	16
Neoplasms	4	2012_Spring	5
Organic Chemicals	3	2010_Summer	7
Polycyclic Compounds	3	2012_Summer	2
Respiratory Tract Diseases	3	2012_Fall	6
Virus Diseases	0	2010_Winter	15

Table 15: Thirteen categories selected for qualitative analysis - number of outlier tests passed, time period, and number of answers

5.2.3 Quantitative Analysis for Infectious Diseases

The notifiable infectious diseases from the qualitative analysis of "Respiratory Tract Diseases", "Bacterial Infections and Mycoses" and "Virus Diseases" were collected for further analysis. The data for these diseases were downloaded from the SINAN website (Brazilian Ministry of Health & Sistema Único de Saúde, n.d.) for Pernambuco, corresponding to the years 2010-2012. The number of answers assigned to each of the three categories for each month for 2010-2012 were obtained from the raw data. The monthly values representing the number of reported cases for each disease, during 2010-2012, was compared to the number of answers/month of its corresponding category, during the same

time period, using the Spearman's correlation coefficient. For example, the number of reported cases for dengue for 2010-2012 were compared to the number of answers assigned to "Virus Diseases". This was performed to determine the presence or absence of a relationship between the categories, representing healthcare providers' information needs, and the increase or decrease in incidence of the corresponding infectious disease. In other words, can the incidence of an infectious disease be confirmed by healthcare providers' information needs.

5.3 Results and Discussion

5.3.1 Outlier Detection

5.3.1.1 Results

Of the 90 unique categories in the PE data set, 38 unique categories did not have any outlier seasons (seasons in which the proportion of methods assigned to that category were outliers as compared with the other seasons in the set) according to all four methods of analysis (i.e. none of the four methods suggested an outlier - see Appendix G). As shown in Appendix G, 11 unique categories had an outlier season according to one of the four methods. Twelve unique categories have outlier seasons according to two of the four methods of analysis (see Appendix G). Of the 90 unique categories of the PE data set, 23 unique categories have an outlier season according to three of the four methods of analysis (see Appendix G). All 23 of these categories have a single maximum value season as an outlier (i.e. the category was assigned to a higher proportion of answers than in other seasons). Six unique categories have an outlier season according to all four methods of analysis (see Appendix G).

5.3.1.2 Discussion

It is believed that the outliers in many of these categories represent greater information needs during their corresponding year/season time periods. However, it is not clear why these information needs occur during a particular time period. For categories representing diseases, such as "Endocrine system Diseases", it is possible that these information needs may reflect an increase in incidence of diseases and conditions in Pernambuco, during these time periods. For other categories, such as "Biological Phenomena", it is difficult to discern the topic of the information need, thus, making it difficult to draw conclusions about these information needs during these particular times. The results from the qualitative analysis, as discussed in the next sections, may shed light on the reasons for greater information needs during the specified time periods.

As the answers were being prepared for qualitative analysis, we noticed that some of the categories, such as "Eye Diseases", had answers assigned to only two year/season temporal slots, leaving 10 year/season slots with zero assignments. The distribution of the assignment of answers to the 23 categories with outliers detected by three tests was further examined. It was determined that there were 19 categories with nine, ten, or 11 year/season temporal slots without answer assignments, and four categories with 0 or 1 year/season temporal slots without any answer assignments. Those four categories include: "Respiratory Tract Diseases", "Behavior and Behavior Mechanisms", "Organic Chemicals", and "Polycyclic Compounds".

One possible reason exists for this occurrence. With large numbers of zero counts, any representation at all for a category may appear as an outlier. This, in turn, may have led to incorrect detection of outliers for these categories. For this reason, it is difficult to draw meaningful conclusions from these 19 categories, thus, they were excluded from further analysis. Therefore, the above mentioned four categories were used for qualitative analysis.

5.3.2 Qualitative Analysis

5.3.2.1 Topics of Category Outliers Detected by Three Tests

Table 16 shows the results from the qualitative analysis discussed below. Of the categories with outliers detected by three methods of analysis, answers for "Behavior and Behavior Mechanisms" discussed breast feeding (3 of 5 answers) as a major topic and suicide (2 of 5 answers) as a minor topic. Answers for "Polycyclic Compounds" discussed crack (2 of 2 answers) as the major topic, and no minor topics. Answers for "Organic Chemicals" discussed Nystatin (4 of 7 answers) as a major topic and formocresol (2 of 7 answers) and Permethrin (1 of 7 answers) as minor topics.

Group	Category	Major/Minor Topics	Number answers out of total	Year/Season
Reportable Diseases	Respiratory Tract Diseases	Major Topic: Tuberculosis Minor Topic: Allergic Rhinitis	4 out of 6 2 out of 6	2012_Fall 2012_Fall
	Bacterial Infections and Mycoses	Major Topic: Tuberculosis Minor Topic: Bacterial Meningitis	7 out of 11 4 out of 11	2012_Summer 2012_Summer
	Virus Diseases	Major Topic: Hepatitis Minor Topic: Dengue Minor Topic: HIV	8 out of 15 4 out of 15 3 out of 15	2010_Winter 2010_Winter 2010_Winter
Categories with outliers detected by 3 of 4 tests	Behavior and Behavior Mechanisms	Major Topic: Breast Feeding Minor Topic: Suicide	3 out of 5 2 out of 5	2012_Spring 2012_Spring
	Polycyclic Compounds	Major Topic: Crack	2 out of 2	2012_Summer
	Organic Chemicals	Major Topic: Nystatin Minor Topic: Formocresol Minor Topic: Permethrin	4 out of 7 2 out of 7 1 out of 7	2010_Summer 2010_Summer 2010_Summer
Categories with outliers detected by 4 of 4 tests	Biological Phenomena	Major Topic: Wound Healing	5 out of 5	2012_Spring
	Endocrine System Diseases	Major Topic: Diabetes Minor Topic: Ovarian Cysts	12 out of 15 3 out of 15	2010_Spring 2010_Spring
	Equipment and Supplies	Major Topic: Bandages Major Topic: IUD	3 out of 6 3 out of 6	2012_Spring 2012_Spring
	Immune System Diseases	Major Topic: HIV Minor Topic: Allergic Rhinitis	6 out of 8 2 out of 8	2012_Fall 2012_Fall
	Investigative Techniques	Major Topic: Papanicolaou Test	16 out of 16	2011_Fall
	Neoplasms	Major Topic: Breast Cancer Prevention Minor Topic: Thyroid Cancer	4 out of 5 1 out of 5	2012_Spring 2012_Spring
Females Issues	Female Urogenital Diseases and Pregnancy Complications	Major Topic: Preventive Exams Minor Topic: Candidiasis Minor Topic: Contraception Minor Topic: Genital Herpes	20 out of 32 5 out of 32 5 out of 32 2 out of 32	2010_Spring 2010_Spring 2010_Spring 2010_Spring

Table 16: Major / Minor Topics for Outlier Seasons of Categories

Through qualitative analysis of the answers for each category, topics appear, providing a view of the categories that characterize healthcare providers' information needs, and by extension information about healthcare concerns in the region. It is through these topics that a picture of the healthcare concerns in the region can be seen. Thus, the messages provide an important source for understanding these information needs.

A topic of concern has been suicide. It has been noted that in the past few years suicide numbers have been increasing in Brazil (Stefanello, Cais, Mauro, Freitas, & Botega, 2008). One of the contributing factors is alcohol abuse (Stefanello et al., 2008). In fact, alcohol abuse has been on the rise in Brazil (Schmidt et al., 2011), possibly leading to the observed increase in suicide, and the need for information regarding this topic. Through previous research, it has been shown that suicides peak in spring with a secondary peak in autumn (Christodoulou C. et al., 2011). Discussion of suicide in the answers appears in the spring of 2012, corresponding to the spring peak found by these authors.

Medications have also been important topics. Nystatin is often discussed as a treatment for candida infections of the female urogenital system. Formocresol has also been discussed in the answers. This substance is used in a dental pulpotomy or root canal procedure. Previous research has shown that, in addition to obstetrics and gynecology, dentistry is another topic covered by teleconsultations at NUTES in Pernambuco (Brixey & Brixey, 2017; Sanches et al., 2012). Even though dentistry is a major topic for these healthcare providers, there is no apparent seasonality for the use of formocresol in root

canal procedures. Another topic of discussion in the answers, although a minor one, has been Permethrin. Permethrin is an insecticide and medication. As a medication, it is used to treat lice infestations and scabies. Lice appears in the warmer summer months, while there is no variation in the prevalence of scabies (Heukelbach J., Wilcke T., Winter B., & Feldmeier H., 2005). In fact, the need for information regarding these medications appears during summer of 2010, possibly relating to increase in lice infestation.

Pregnancy issues also represent other important information needs. One such topic includes breast feeding. One issue of concern is the use of illicit drugs, such as Crack, by breast feeding mothers. In this case, there does not seem to be any relationship between breast feeding and seasons. Mothers breast feed all year round. Additional topics related to women's issues were also discovered.

5.3.2.2 Women's Issues

Answers for "Female Urogenital Diseases and Pregnancy Complications" discussed preventive exams (20 of 32 answers) as a major topic and candidiasis (5 of 32 answers), contraception (5 of 32 answers) and genital herpes (2 of 32 answers) as minor topics. Answers for "Investigative Techniques" discussed Papanicolaou tests (16 of 16 answers) with no minor topics.

As demonstrated by many of these topics, the healthcare providers have concerns regarding female issues.

Other topics also include various infections such as genital herpes and candidiasis. In discussing most urogenital infections, the answers provide a description of signs and symptoms along with current methods of treatment. The literature does not provide any discussion regarding seasonality of vaginal candidiasis and genital herpes. Therefore, it is not clear why healthcare providers had a greater need for information about these diseases in the spring of 2010.

Preventive exams and Papanicolaou (Pap) tests are also frequently discussed topics, as they are often used for detecting the presence or absence of these various urogenital infections along with any cervical cancer (or premalignant lesions). Since these tests are used all year round, it is likely that there is no seasonality in their use. Therefore, it is not possible to determine why there is a greater need for information about the Pap test during fall of 2011.

A regularly discussed issue related to pregnancy is contraception, both during breast feeding, and in general, as discussed in spring of 2010. The answers provide information about different types of birth control methods, including birth control pills and IUDs. Although there was a spike in births in Brazil between 2007 and 2008, birth rates have been decreasing since then (Barrientos, 2018). In light of this information, there is no apparent reason for the need for information regarding birth control during spring of 2010.

5.3.2.3 Topics of Category Outliers Detected by Four Tests

Answers for "Biological Phenomena" discussed wound healing as a major topic (5 of 5 answers), with no minor topics. Answers for "Endocrine System Diseases" discussed diabetes as a major topic (12 of 15 answers) and ovarian cysts (3 of 15 answers) as a minor

topic. Answers for "Equipment and Supplies" discussed bandages (3 of 6 answers) and IUDs (3 of 6 answers) as major topics and no minor topic. Answers for "Immune System Diseases" discussed HIV (6 of 8 answers) as a major topic and allergic rhinitis (2 of 8 answers) as a minor topic. Answers for "Neoplasms" discussed breast cancer prevention (4 of 5 answers) as a major topic and thyroid cancer (1 of 5 answers) as a minor topic.

Issues related to diabetes are also important. It is known that there is a higher burden of disease from diabetes in the northeast region of Brazil (Leite et al., 2015). Despite the concerns about diabetes expressed healthcare providers in spring of 2010, there is no specific data showing a particularly high increase in diabetes during this time. Thus, the reason for the greater need of information regarding diabetes during 2010 is indeterminable.

Wound healing is another topic of discussion seen in the answers in spring of 2012. In analyzing these answers, healthcare providers needed information about wound dressings (bandages) in order to facilitate wound healing. Although one answer discussed diabetic wound care, the remaining answers discussed wound healing and bandages in general. Here again, there is no apparent reason for the greater need for information regarding wound healing and bandages during this time.

Although ovarian cysts are women's issues, they are also classified as an endocrine system disease. Ovarian cysts were discussed as a minor topic. In these answers, advice about their treatment was given. Ovarian cysts do not have seasonality, thus, the greater need for this information during spring of 2010 is not known.

Allergic rhinitis is briefly discussed in the answers. Information about its treatment and diagnosis are provided. There are both seasonal and perennial allergic rhinitis. Seasonal allergic rhinitis occurs during spring, summer, and early fall, while perennial allergic rhinitis occurs year-round (American College of Allergy, Asthma & Immunology, 2014). In fact, the discussion of allergic rhinitis corresponds to the fall, during which time healthcare providers may need information about both forms of allergic rhinitis.

Another topic of discussion is breast cancer prevention. The annual deaths from breast cancer increased in the Northeast region of Brazil from 1994 to 2011 (Freitas-Junior et al., 2012; Gonzaga et al., 2015). This is evidenced by the fact that one of the major topics of discussion is breast cancer prevention, through self-breast exams. Although this research only shows annual increases in deaths from breast cancer to 2011, it is entirely possible that deaths continued to increase through spring of 2012, and that healthcare providers' concerns may reflect a particularly high incidence of breast cancer during this time.

Intrauterine devices (IUDs), in particular, are often discussed, particularly in in spring of 2012. In one instance, a healthcare provider was concerned about conception during the use of an IUD. Barrientos (Barrientos, 2018) noted that birth rates in Brazil have been decreasing since 2008. Therefore, the reason for the greater need for information about contraception and IUDs is not clear.

Although not an issue exclusively experienced by women, thyroid diseases and cancer are also discussed in the answers. According to Duncan and colleagues (Duncan et al., 2012), cancer accounts for a vast majority of deaths. Although discussed as a minor topic, thyroid

cancer is a concern. In fact, the highest incidence of thyroid cancer in Central and South America has been observed in Brazil, Ecuador, Costa Rica and Colombia (Sierra, Soerjomataram, & Forman, 2016). This may explain a need for information regarding thyroid cancer in general, but not particularly in 2012 in Pernambuco.

5.3.2.4 Infectious Diseases Results and Discussion

Of the three categories representing infectious diseases, answers categorized as "Respiratory Tract Diseases", during fall of 2012, discussed tuberculosis as a major topic (4 of 6 answers), and allergic rhinitis as a minor topic. Answers for "Bacterial Infections and Mycoses", during summer of 2012, discussed tuberculosis as a major topic (7 of 11 answers), and bacterial meningitis as a minor topic (4 of 11 answers). Answers for "Virus Diseases" discussed hepatitis (8 of 15 answers) as a major topic and dengue (4 of 15 answers) and HIV (3 of 15 answers) as minor topics.

HIV/AIDS has been an important topic examined in the answers. In Brazil, the prevalence of HIV/AIDS infection is 1% in the general population and around 6% in high-risk groups such as drug users, gay males and female sex workers (Cavalcanti et al., 2012). HIV/AIDS also has particular regional incidence rates. Cavalcanti and colleagues (Cavalcanti et al., 2012) state that HIV/AIDS incidence rates are beginning to stabilize, decreasing in the southeastern and central-western regions and increasing in the northern, northeastern and southern regions. This increase in incidence explains the need for information about this topic. As an immune disease, healthcare providers had a greater need for information about HIV/AIDS during fall of 2012, while as a virus disease, they needed the information about

HIV/AIDS during winter of 2010. Therefore, these particular increases in the information needs may reflect time periods with greater increases in HIV/AIDS infection.

Dengue is a viral disease transmitted by the *Aedes* mosquito (Barcellos & Lowe, 2014). In Brazil, outbreaks of dengue are clustered in the northeast (Barcellos & Lowe, 2014). Barcellos and Lowe (Barcellos & Lowe, 2014) also note that the presence of outbreaks of dengue are not hampered by long dry seasons. It is thought that this is due to the fact that the local population resorts to storing water in improvised tanks, providing breeding sites for mosquitoes that spread the dengue virus (Caprara et al., 2009). Barcellos and Lowe (Barcellos & Lowe, 2014) note the dengue outbreaks in Recife were more severe during 2008 and 2010. This may explain healthcare providers' need for information about dengue during winter of 2010.

In addition to the previously discussed infections, hepatitis is also notifiable infectious disease. Although notification is necessary, it is difficult to estimate the hepatitis infection rates in the general population since a large number of viral infections are asymptomatic and the etiology of notified symptomatic cases cannot always be confirmed (Ximenes et al., 2010). Thus, the prevalence in the general population is unknown (Ximenes et al., 2010). Despite these difficulties, Azevedo and colleagues (Azevedo, Santos, Jerez Roig, & Souza, 2015) show that the incidence of hepatitis B and C in the northeast region of Brazil have increased from 1997-2010. Healthcare providers expressed a need for information during winter of 2010, coinciding with the higher incidence rate shown by these authors.

Meningitis is another notifiable infectious disease. It is a swelling of the protective membranes that cover the brain and spinal cord, and is caused by bacteria, viruses, and fungi (Centers for Disease Control and Prevention, 2017). Bacterial meningitis is the most severe form (WebMD, n.d.). Thus, it has become a public health problem for both developed and developing countries (Duarte, Amorim, Cuevas, Cabral-Filho, & Correia, 2005). In Recife, Pernambuco, Brazil, children are especially affected by bacterial meningitis (Duarte et al., 2005). Research has also shown the meningitis in Brazil shows a seasonal peak from May to October (Paireau, Chen, Broutin, Grenfell, & Basta, 2016), corresponding to late fall to mid-spring. Healthcare providers expressed a need for information during the summer of 2012, which does not correspond with this seasonal peak reported in the literature. This could possibly be explained by: (1) the fact that the category "Bacterial Infections and Mycoses" also covers tuberculosis; and (2) a possible increase in the number of cases of bacterial meningitis in Pernambuco during this time.

Tuberculosis is another notifiable infectious disease. It is caused by the *Mycobacterium* (Soares, Amaral, Zacarias, & Ribeiro, 2017). According to Soares and colleagues (Soares et al., 2017) there were more than one million new cases of tuberculosis with 70'000 deaths reported in Brazil from 2001 to 2014. There were 57,015 new cases of tuberculosis in the state of Pernambuco, during the same time period, with many of the occurrences spread out over time (Soares et al., 2017). According to Gaspar and colleagues (Gaspar, Nunes, Nunes, & Rodrigues, 2016), the incidence of tuberculosis in the northeast has been decreasing during 2002 to 2012. There was an expressed need for information about tuberculosis in the summer of 2012, at the time when the incidence of tuberculosis was

lower. Although not evident in the answers, healthcare providers may have been seeing cases of treatment nonadherence, which was still above Ministry of Health requirements (Soares et al., 2017).

5.3.3 Reportable Infectious Disease Analysis

5.3.3.1 Results

For each of these diseases, tuberculosis, meningitis, hepatitis, and dengue, the reported number of cases for each month was obtained from SINAN. In examining the correlation between each disease and its corresponding category, I found that there is little to weak correlation between the categories and their corresponding categories, with the exception of tuberculosis and "Respiratory Tract Diseases", which has a low moderate correlation (Spearman's $r = 0.5484$, see Table 17). Tuberculosis and "Bacterial Infections and Mycoses" are weakly and positively correlated with a correlation coefficient of 0.4109. On the other hand, tuberculosis and "Respiratory Tract Diseases" have a low moderate positive correlated with a correlation coefficient of 0.5484. Given that this is an estimate of correlation between two very different data sources, it suggests that the distribution of automatically-assigned categories does, to some extent and in selected cases, correspond to incidence rates of diseases in the region. However, further research is required to determine the constraints under which these data might reflect changes in the incidence of diseases of interest. Two diseases are negatively correlated with their corresponding categories: (1) Meningitis and "Bacterial Infections and Mycoses" with a correlation coefficient of -0.3784, and (2) Hepatitis and "Virus Diseases" with a correlation coefficient

of -0.0833. Dengue and "Virus Diseases" have a weak positive correlation coefficient of 0.4102.

Disease	Category	Spearman's <i>r</i> Coefficient
Dengue	Virus diseases	0.4102
Hepatitis	Virus diseases	-0.0833
Meningitis	Bacterial Infections and Mycoses	-0.3784
Tuberculosis	Bacterial Infections and Mycoses	0.4109
Tuberculosis	Respiratory Tract Diseases	0.5484

Table 17 Spearman's *r* coefficient

5.3.3.2 Discussion

As noted in the results, the Spearman's correlation coefficient was used for comparing the infectious diseases with their corresponding categories. The Spearman's correlation coefficient is used to assess the association of two variables in a population (Hanneman, Kposowa, & Riddle, 2012) in cases where the data has a nonnormal distribution and outliers (Zaiontz, 2018). According to Hanneman and colleagues (Hanneman et al., 2012), correlation coefficients with magnitude between 0.9 and 1.0 are considered to be very highly correlated; those between 0.7 and 0.9 are considered to be highly correlated; those between 0.5 and 0.7 are considered to be moderately correlated; those between 0.3 and 0.5 are considered to be weakly correlated; and those that are less than 0.3 have little if any correlation. In the cases of negative correlations, an increase in one variable (the number of disease reports) is associated with a decrease in the second variable (number of answers for the category). As noted in the results, the infectious diseases had little to weak correlations with their corresponding categories, except for tuberculosis with a low

moderate correlation. This low moderate correlation could possibly be explained by the fact that both the number of tuberculosis cases and the need for information changed very little over time. In the case of the information needs, it is not necessarily due to questions only about tuberculosis. These needs also covered other "Respiratory Tract Diseases" such as allergic rhinitis.

There are several possible reasons for weak correlations between the remaining infectious diseases and their corresponding categories. First, I am using a small data set, both in number of answers being assigned to the categories (1,377) and the number of years (3) being examined for changes in diseases and information needs. Next, it is also likely that the categories are not granular enough to detect changes for a specific infectious disease. Although there are changes in the notification of new cases for hepatitis, meningitis, and dengue, there is little to no change in information needs across the same time period, except for the noted outlier time period.

These results also suggest that this could be an area of future research. This would involve the use of larger data sets and more granular classifiers.

5.4 Conclusion

For many of the categories, no outlier was detected, reflecting little change in information need represented by these categories. However, there were outliers detected by four tests for six categories: "Biological Phenomena", "Endocrine System Diseases", "Equipment and Supplies", "Immune System Diseases", "Investigative Techniques", and "Neoplasms", possibly indicating an increase in a need for information for that season. Although 23 of

the categories had outliers in three of four methods of analysis, only three categories ("Behavior and Behavior Mechanisms", "Organic Chemicals", and "Polycyclic Compounds") show some change across time, again, possibly reflecting a need for information. The above categories, along with "Respiratory Tract Diseases", Bacterial Infections and Mycoses", "Virus Diseases", and "Female Urogenital Infections and Pregnancy Complications" were chosen for qualitative analysis, as the first three categories possibly represent reportable infectious diseases and the last category represents women's issues.

In qualitative analysis of these categories, various topics appeared, providing a view of the health concerns in the region. It is difficult to determine the reason that some topics appear in specific seasons. For instance, there is no apparent reason that healthcare providers need information about breast feeding during spring of 2012, as mothers breast feed all year-round. However, other topics have outliers in seasons that are consistent with the literature. For example, Christodoulou and colleagues (Christodoulou C. et al., 2011) demonstrate that suicides peak in spring with a secondary peak in autumn. Healthcare express a need for information about suicide in the spring of 2012, coinciding with the spring peak noted by these authors.

The relationship between the reported cases of these infectious diseases in SINAN was explored. It was determined that little to weak correlation exists between most of the infectious diseases and their corresponding categories. This is probably due to the fact that there is a small sample size, and that these categories are broad, including other diseases besides the infectious diseases. Despite these limitations, the weak correlation suggests that

syndromic surveillance using question-answer pairs from the Brazilian telehealth system may be an area for future research with larger data sets and more granular classifiers.

Chapter 6: Key Findings, Innovation, Contributions, Future Work and Conclusions

6.1 Summary

In the past, studies of healthcare providers' information needs have been limited to qualitative studies using small sample sizes. For instance, in one study physicians filled out 303 surveys (Kourouthanassis et al., 2015), and in another study they participated in 47 interviews and filled out 47 surveys (Covell et al., 1985). In observational studies, 103 physicians participated in one study (Ely et al., 1999), while 24 physicians participated in another study (Osheroff et al., 1991). Automated methods, such as k-NN, provide the means to use large sample sizes to categorize these information needs. For this research, I studied five distributional semantic models (RI, RI-RI, RI-RRI, RRI-RRI, and RI-NWE) in an effort to determine the best model for automatic categorization of healthcare providers' information needs. I discovered that k-NN with RI had performance comparable to that of categorization of MEDLINE abstracts with MeSH using MTI (Huang, Névéol, Lu, 2011; National Library of Medicine, Indexing Initiative, (2007)). These results provided confidence for using RI to categorize the information needs expressed in the messages of the unannotated data from Pernambuco, and Minas Gerais, Brazil.

The categorization of information needs showed similarities and differences across these two regions of Brazil. In Pernambuco, healthcare providers expressed concerns about pregnancy and female issues. In Minas Gerais, they expressed information needs regarding "Cardiovascular Diseases" and "Neoplasms" reflecting health concerns in this region.

Information needs across provider types (nurses and doctors) were also examined. Differences and similarities were found between doctors and nurses within and across these regions as demonstrated in Table 13.

Studies of health-related information needs in developed nations have shown that providers most frequently asked questions regarding treatment methods, diagnoses, and medications (Kourouthanassis et al., 2015; Clarke et al., 2013; Cheng, 2004; Smith, 1996, Cimino & Barnett, 1991; Osheroff et al., 1991). However, little is known about the information needs of rural healthcare providers in developing nations, such as Brazil. In this research, the characterized information needs of these rural doctors and nurses were compared to their urban counterparts. It was discovered that rural MG doctors need information about medication, treatment and diagnosis of disease like their urban counterparts, while rural PE doctors need information about medication and treatment. According to Spath and Buttlar (1996), nurses in urban areas need information regarding diagnosis of disease, medication, techniques, and equipment. Like their urban counterparts, the nurses from the rural Northeast (PE) and Southeast (MG) need information about diagnosis of diseases, while MG nurses also need information about medication. However, of note, nurses in both regions are distinguished by information needs concerning “Investigative Techniques”, which correspond to screening tests, demonstrating their role in health promotion.

In the last study, I examined the information needs for changes across time. The timestamped data from Pernambuco was evaluated for greatest change in information need using four tests for outlier detection: modified Z-Score, Tukey outer fence, Grubbs test, and Dixon Q test. As shown in Table 15, 13 categories were chosen for further analysis:

(1) 6 categories with outlier year/season time periods detected by all four tests, (2) 3 categories with outliers detected by three of four tests, (3) 3 categories possibly discussing infectious diseases, and (4) 1 category representing female issues. Next, the modified grounded theory approach (Blackstone, 2012) was used to analyze the answers from the outlier year/season time period, providing a picture of the information needs in the region. It was discovered that topics such as allergic rhinitis, suicidality, and scabies, conformed to expected seasonal occurrences corresponding to fall, spring, and summer respectively. In the final step, the reported incidences of the infectious diseases discussed in the answers (dengue, hepatitis, meningitis, and tuberculosis) were obtained from SINAN and compared to number of answers in the corresponding categories. Given the differences in the nature of the data concerned and the high level categories assigned by our system, even weak to moderate correlation across multiple categories suggests that with larger data sets and specific classifiers, the messages from the Brazilian telehealth system may provide a basis for a syndromic surveillance tool.

6.2 Implications

With this research, I discovered that the distributional semantic with k-NN is a broadly-applicable method for identification of information needs from inter-provider communication. This method provides new insight into the distribution of information needs across regions and provider types. This could be of value in a number of ways. First, it can inform the design of information resources (e.g. webinars) for dissemination by the telehealth centers. Second, it can inform the distribution of expertise, e.g. specialists. Given that this work occurs in the context of a large national health system, there is the potential

for central decision making to influence the distribution of provider expertise both to telehealth centers (e.g. centers with more queries concerning cancer may need an oncologist on site), and to practice in the regions. While it was difficult to draw conclusive interpretations from the small amount of data available for analysis across time, the results of the outlier test and accompanying qualitative analysis do suggest that these data may be of value for syndromic surveillance, given the development of customized categorization methods for this purpose.

6.3 Support for motivating hypotheses

These key findings provide support for my motivating hypotheses. My first hypothesis was that semi-automated analysis of healthcare providers' queries can assist in characterizing their information needs. The first hypothesis states that semi-automated text categorization of health care providers' communication in a telehealth system could provide insight into their information needs. The system assigned categories with performance similar to published performance of automated categorization systems in practical use, such a assignment of MeSH terms to abstracts in PubMed. Subsequent analysis of a large unannotated set of communication showed differences in information received that corresponded to known differences in expertise and healthcare concerns across regions. In addition, this analysis revealed differences in information needs across provider types. My second hypothesis was that the resulting characterized information needs can then be analyzed across provider types, as well as for seasonal and other variations. In the second hypothesis, changes in distribution of assigned categories over time can reveal changes in

information needs and may correspond to changes in disease distribution over time, suggesting possible utility for syndromic surveillance. Timestamped data constituted a relatively small proportion of the data available for analysis, as such it was not possible to draw firm conclusions in this regard. (2) However, using methods outlier detection and qualitative methods it was possible to identify seasons (within years) in which changes in information received corresponded to conditions with known seasonal distribution. In addition, there was a moderate positive correlation between annotation of the category “Respiratory Diseases” to answers over time and reported incidence of Tuberculosis in Pernambuco. Although these results should be interpreted with caution on account of the small number of data points available and the high-level nature of the categories assigned by the system (by design), they do provide some indication that these data sets could be of use for syndromic surveillance with further adaptation including, at a minimum, the development of customized disease-specific classifiers.

6.4 Limitations and Future Work

One limitation of this research is that the parameters of the distributional semantic models were not optimized. For instance, optimal dimensionality may likely influence categorization performance. In future work, various parameters, such as dimensionality, may be explored to best optimize the distributional semantic models. This, in turn, may lead to better performance of the algorithms used to characterize healthcare providers' information needs.

There were two limitations in the analysis of information needs. First, there was poor representation of CHWs in the MG data set. It has been stated that the policies governing access to telehealth facilities vary across regions. This may be mitigated by obtaining data sets from those regions in which CHWs have access to the telehealth system, providing additional data for analysis. Second, categories were used to characterize healthcare providers' information needs. It is likely the more granular DeCS terms will provide a more complete picture of these information needs. Future work would include characterizing these information needs using DeCS terms. In turn, these DeCS terms may provide a better classifier for syndromic surveillance.

There were limitations in outlier detection. The PE data set is small, which affects, in particular, the analysis of temporal dynamics. Although the outlier detection tests used in this research are used for small data sets, one must take care to use them for data exploration and not data deletion to determine any interesting findings (Seo, 2006). In addition, the PE data has a nonnormal distribution. This affects some outlier detection tests. Although a larger data set was available, it did not have temporal metadata. For future work, larger data sets with temporal timestamps could be obtained for analysis for temporal dynamics and syndromic surveillance. Finally, these larger data sets with timestamps, in combination with classifiers, may be used to build a surveillance tool.

In the current work, only one multi-label categorization technique, k-NN, was utilized to characterize healthcare providers' information needs. Accuracy of categorization may improve with other machine learning techniques. For future work, other multi-label classification techniques for categorization of healthcare providers' information needs may

be explored. For instance, Cronin and colleagues (Cronin, Fabbri, Denny, Rosenbloom, & Jackson, 2017) demonstrate that ensembles of binary classifiers can be used to label multi-label patient portal messages with promising performance, and a range of techniques have been developed to adapt binary classification approaches to multi-label categorization tasks (Gibaja & Ventura, 2014). Exploration of these and other methods will assist in determining the best method for improving the accuracy of categorization of information needs. Addressing these limitations would bridge the gap between this research-oriented system and a production system that could inform the allocation of resources and expertise within a geographically diffuse regional healthcare system.

6.5 Innovation

Information needs of healthcare providers are categorized, using messages from a telehealth system. This is innovative because prior work on information needs has relied on qualitative methods, that scale poorly to large data sets.

In addition, this research provides analysis of information needs in a rural setting in general, and in rural Brazil in particular. Little was known of the information needs in these regions prior to this analysis, and as such, this thesis makes a unique contribution to the literature on information needs.

This research looks for change of information needs across time. In previous research with web log data, specific known infectious disease outbreaks were compared to web searches to determine if these web searches could detect the known outbreak (Christaki, 2015; Hill

et al., 2013; Hulth et al., 2009; Polgreen et al., 2008). In this research, I was not looking for pre-determined syndromes, but rather for information needs as an area of investigation in their own right. Secondly, I analysed inter-provider communication which has not been investigated as thoroughly as consumer web searches (Eysenbach, 2006; Ginsberg et al., 2009; Scott-Wright et al., 2006; Seifter et al., 2010; Wilson & Brownstein, 2009).

Preliminary evaluation of utility of inter-provider communication within a telehealth system in a developing country as a data source for syndromic surveillance. The utility of such data for this purpose had not been evaluated previously.

6.6 Theory

My contribution was to apply the ASK theory (Belkin, 2005; Case & Given, 2016), mainly applied to library and information studies (Case & Given, 2016) to guide my work. In addition, this had not been applied in the context of inter-provider communication in a rural tele-health system previously. In doing so, I have extended the range of application of this framework.

6.7 Contribution to Informatics

This research shows that electronic messages from the telehealth system can be analyzed using automated methods, previously used for MeSH assignment to abstracts. This analysis can be used for characterizing information needs. It has been shown that methods of automated text categorization can reveal a previously unexploited utility of inter-provider communication in the context of a rural telehealth network. These methods provide a

means to study provider information needs at a scale beyond that possible with the qualitative methods that predominate in prior work. These electronic messages may also have some utility as a data source for syndromic surveillance, the potential for which was not recognized prior to this work.

6.8 Conclusion

This study demonstrates that when using k-NN, RI performs comparable to that published in the literature, and better than models incorporating background information at characterizing information needs expressed electronically.

It has been demonstrated that differences in healthcare providers' information needs across regions correspond to known health concerns in these regions. Doctors' information needs differ across regions, as well as from their urban counterparts. The information needs expressed by nurses are different than those identified for urban nurses.

Outlier detection demonstrates the change in information needs over time. These changes provide a picture of the concerns in the region, possibly indicating the use of the messages from the Brazilian telehealth system as syndromic surveillance. Thus, this research shows that there is the potential for inter-provider communication to be used as a basis for both large-scale analysis of information needs, and monitoring of changes in regional health concerns.

References

- AAPC Client Services. (2013). ICD-10: The History, the Impact, and the Keys to Success. Retrieved March 21, 2018, from <http://www.aapcps.com/resources/icd10-white-paper.aspx>
- Alkmim, M. B., Figueira, R. M., Marcolino, M. S., Cardoso, C. S., Pena de Abreu, M., Cunha, L. R., ... Ribeiro, A. L. P. (2012). Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil. *Bulletin of the World Health Organization*, 90(5), 373–378. <https://doi.org/10.2471/BLT.11.099408>
- Althouse, B. M., Ng, Y. Y., & Cummings, D. A. T. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Neglected Tropical Diseases*, 5(8), e1258. <https://doi.org/10.1371/journal.pntd.0001258>
- American College of Allergy, Asthma & Immunology. (2014). Allergic Rhinitis. Retrieved April 2, 2018, from <https://acaai.org/allergies/types/hay-fever-rhinitis>
- Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, 17–21.
- Aronson, Alan R., Bodenreider, O., Demner-Fushman, D., Fung, K. W., Lee, V. K., Mork, J. G., ... Rogers, W. J. (2007). From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. In *Proceedings of the ACL'2007 Workshop "BioNLP"* (pp. 105–112). Prague, Czech Republic. Retrieved from <https://mor.nlm.nih.gov/pubs/pdf/2007-bionlp-ara.pdf>
- Aronson, Alan R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association: JAMIA*, 17(3), 229–236. <https://doi.org/10.1136/jamia.2009.002733>

- Aronson, Alan R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM Indexing Initiative's Medical Text Indexer. *Studies in Health Technology and Informatics*, 107(Pt 1), 268–272. <https://doi.org/10.3233/978-1-60750-949-3-268>
- Assembléia Nacional Constituinte. (1988). Constituição [.gov]. Retrieved September 23, 2017, from http://www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm
- Atkinson, S., Fernandes, L., Caprara, A., & Gideon, J. (2005). Prevention and promotion in decentralized rural health systems: a comparative study from northeast Brazil. *Health Policy and Planning*, 20(2), 69–79. <https://doi.org/10.1093/heapol/czi009>
- Avila, M. M. M. (2011). [A case study of the Community Health Agents Program in Uruburetama, Ceará (Brazil)]. *Ciencia & Saude Coletiva*, 16(1), 349–360.
- Ayankogbe, O. O., Oyediran, M. A., Oke, D. A., Arigbabu, S. O., & Osibogun, A. A. (2009). ICPC-2 defined pattern of illnesses in a practice-based research network in an urban city in West Africa. *African Journal of Primary Health Care & Family Medicine*, 1(1), 4 pages. <https://doi.org/10.4102/phcfm.v1i1.3>
- Azevedo, A., Santos, M., Jerez Roig, J., & Souza, D. (2015). Incidence of viral hepatitis in Brazil from 1997 to 2010. *Journal Nursing UFPE on Line., Recife*, 9, 7375–7382.
- Bachrach, C. A., & Charen, T. (1978). Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Medical Informatics = Medecine Et Informatique*, 3(3), 237–254.

- Baker, L., Wagner, T. H., Singer, S., & Bundorf, M. K. (2003). Use of the Internet and e-mail for health care information: results from a national survey. *JAMA*, 289(18), 2400–2406. <https://doi.org/10.1001/jama.289.18.2400>
- Barbosa, A. K. P., de A Novaes, M., & de Vasconcelos, A. M. L. (2003). A web application to support telemedicine services in Brazil. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 56–60.
- Barcellos, C., & Lowe, R. (2014). Expansion of the dengue transmission area in Brazil: the role of climate and cities. *Tropical Medicine & International Health: TM & IH*, 19(2), 159–168. <https://doi.org/10.1111/tmi.12227>
- Barrientos, M. (2018, January 20). Brazil - Birth rate - Historical Data Graphs per Year. Retrieved April 1, 2018, from <https://www.indexmundi.com/g/g.aspx?c=br&v=25>
- Belkin, N. (2005). Anomalous state of knowledge. In K. E. Fisher, S. Erdelez, & L. McKechnie (Eds.), *Theories of Information Behavior* (pp. 44–48). Information Today, Inc. Retrieved from https://books.google.com/books/about/Theories_of_Information_Behavior.html?id=ll6qzqhIj8wC
- Bernardo, T. M., Rajic, A., Young, I., Robiadek, K., Pham, M. T., & Funk, J. A. (2013). Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of Medical Internet Research*, 15(7), e147. <https://doi.org/10.2196/jmir.2740>
- Biblioteca Virtual em Saúde. (n.d.-a). DeCS - Health Sciences Descriptors. Retrieved March 11, 2018, from <http://decs.bvs.br/I/homepagei.htm>

- Biblioteca Virtual em Saúde. (n.d.-b). Formative Second Opinion (SOF). Retrieved September 23, 2017, from http://aps.bvs.br/formative-second-opinion-sof/?l=en_US
- BIREME. (2017, August 23). Health Sciences Descriptors - Wiki [wiki]. Retrieved September 23, 2017, from http://wiki.bireme.org/en/index.php/Health_Sciences_Descriptors
- BIREME. (n.d.-a). LILACS - Database search. Retrieved September 23, 2017, from <http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&base=LILACS&lang=i&form=A>
- BIREME. (n.d.-b). LILACS - Wiki [wiki]. Retrieved September 23, 2017, from <http://wiki.bireme.org/en/index.php/LILACS>
- Blackstone, A. (2012). *Sociological Inquiry Principles Qualitative and Quantitative Methods v. 1.0*. Boston, MA: Flat World Knowledge. Retrieved from <https://2012books.lardbucket.org/pdfs/sociological-inquiry-principles-qualitative-and-quantitative-methods.pdf>
- Bonham, M. D. (1990). BIREME: Latin American and Caribbean Health Sciences Information Center. *Bulletin of the Medical Library Association*, 78(2), 119–123.
- Borchardt, S. M., Ritger, K. A., & Dworkin, M. S. (2006). Categorization, prioritization, and surveillance of potential bioterrorism agents. *Infectious Disease Clinics of North America*, 20(2), 213–225, vii–viii. <https://doi.org/10.1016/j.idc.2006.02.005>

- Bornstein, V. J., & Stotz, E. N. (2008). [Concepts involved in the training and work processes of community healthcare agents: a bibliographical review]. *Ciencia & Saude Coletiva*, *13*(1), 259–268.
- Boulton, R. (2014, September). A Portuguese stop word list [archive]. Retrieved September 23, 2017, from <http://snowball.tartarus.org/algorithms/portuguese/stop.txt>
- Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative Data Analysis for Health Services Research: Developing Taxonomy, Themes, and Theory. *Health Services Research*, *42*(4), 1758–1772. <https://doi.org/10.1111/j.1475-6773.2006.00684.x>
- Brauer, G. W. (1992). Telehealth: the delayed revolution in health care. *Medical Progress Through Technology*, *18*(3), 151–163.
- Bravata, D. M., McDonald, K. M., Smith, W. M., Rydzak, C., Szeto, H., Buckeridge, D. L., ... Owens, D. K. (2004). Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Annals of Internal Medicine*, *140*(11), 910–922.
- Brazilian Ministry of Health, & Sistema Único de Saúde. (n.d.). SINANWEB - Página inicial. Retrieved March 24, 2018, from <http://portalsinan.saude.gov.br/>
- Breyer, B. N., Sen, S., Aaronson, D. S., Stoller, M. L., Erickson, B. A., & Eisenberg, M. L. (2011). Use of Google Insights for Search to track seasonal and geographic kidney stone incidence in the United States. *Urology*, *78*(2), 267–271. <https://doi.org/10.1016/j.urology.2011.01.010>

- Brixey, J., & Brixey, J. J. (2017). An Exploratory Analysis of Questions Submitted to a Brazilian Telemedicine System. *Studies in Health Technology and Informatics*, 245, 1253. <https://doi.org/10.3233/978-1-61499-830-3-1253>
- Callahan, A., Pernek, I., Stiglic, G., Leskovec, J., Strasberg, H. R., & Shah, N. H. (2015). Analyzing Information Seeking and Drug-Safety Alert Response by Health Care Professionals as New Methods for Surveillance. *Journal of Medical Internet Research*, 17(8), e204. <https://doi.org/10.2196/jmir.4427>
- Campos, F. E., Haddad, A. E., Wen, C. L., Alkmin, M. B. M., & Cury, P. M. (2009). The National Telehealth Program in Brazil: an instrument of support for primary health care. *Latin American Journal of Telehealth*, 1(1), 39–66.
- Caprara, A., Lima, J. W. de O., Marinho, A. C. P., Calvasina, P. G., Landim, L. P., & Sommerfeld, J. (2009). Irregular water supply, household usage and dengue: a bio-social study in the Brazilian Northeast. *Cadernos De Saude Publica*, 25 Suppl 1, S125-136.
- Cardoso, A. dos S., & Nascimento, M. C. do. (2010). [Communication in the Family Health Program: the health agent as an integrating link between the team and the community]. *Ciencia & Saude Coletiva*, 15 Suppl 1, 1509–1520.
- Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 49(10), 1557–1564. <https://doi.org/10.1086/630200>

- Case, D. O., & Given, L. M. (2016). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior: 4th Edition*. (J.-E. Mai, Ed.) (4th edition). Bingley, UK: Emerald Group Publishing Limited. Retrieved from <https://www.amazon.com/Looking-Information-Research-Seeking-Behavior/dp/1785609688>
- Castilla, A. C., Furuie, S. S., & Mendonça, E. A. (2007). Multilingual information retrieval in thoracic radiology: feasibility study. *Studies in Health Technology and Informatics*, 129(Pt 1), 387–391.
- Cavalcanti, A. M. S., Brito, A. M. de, Salustiano, D. M., Lima, K. O. de, Silva, S. P. da, & Lacerda, H. R. (2012). Recent HIV infection rates among HIV positive patients seeking voluntary counseling and testing centers in the metropolitan region of Recife - PE, Brazil. *The Brazilian Journal of Infectious Diseases: An Official Publication of the Brazilian Society of Infectious Diseases*, 16(2), 157–163.
- Centers for Disease Control and Prevention. (2017, November 29). Meningitis. Retrieved March 26, 2018, from <https://www.cdc.gov/meningitis/index.html>
- Chan, E. H., Sahai, V., Conrad, C., & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases*, 5(5), e1206.
<https://doi.org/10.1371/journal.pntd.0001206>
- Cheng, G. Y. T. (2004). A study of clinical questions posed by hospital clinicians. *Journal of the Medical Library Association: JMLA*, 92(4), 445–458.

- Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to Train good Word Embeddings for Biomedical NLP (pp. 166–174). Presented at the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany.
<https://doi.org/10.18653/v1/W16-2922>
- Choi, J., Cho, Y., Shim, E., & Woo, H. (2016). Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*, *16*(1), 1238. <https://doi.org/10.1186/s12889-016-3893-0>
- Christaki, E. (2015). New technologies in predicting, preventing and controlling emerging infectious diseases. *Virulence*, *6*(6), 558–565.
<https://doi.org/10.1080/21505594.2015.1040975>
- Christodoulou C., Douzenis A., Papadopoulos F. C., Papadopoulou A., Bouras G., Gournellis R., & Lykouras L. (2011). Suicide and seasonality. *Acta Psychiatrica Scandinavica*, *125*(2), 127–146. <https://doi.org/10.1111/j.1600-0447.2011.01750.x>
- Cimino, C., & Barnett, G. O. (1991). Analysis of physician questions in an ambulatory care setting. *Proceedings. Symposium on Computer Applications in Medical Care*, 995–999.
- Cimino, J. J. (1996). Review paper: coding systems in health care. *Methods of Information in Medicine*, *35*(4–5), 273–284.
- Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, *37*(4–5), 394–403.

- Clarke, C., Bouland, D., Reed-Rowe, H., Friedman, L., & Meyer, B. (2011). Technology Heals—Bringing Telemedicine to Palau. Retrieved September 23, 2017, from <http://healthspan.ucsd.edu/2011/02/Pages/hs-palau.aspx>
- Clarke, M. A., Belden, J. L., Koopman, R. J., Steege, L. M., Moore, J. L., Canfield, S. M., & Kim, M. S. (2013). Information needs and information-seeking behaviour analysis of primary care physicians and nurses: a literature review. *Health Information and Libraries Journal*, 30(3), 178–190. <https://doi.org/10.1111/hir.12036>
- Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2), 240–256. <https://doi.org/10.1016/j.jbi.2009.09.003>
- Cohen, T., & Widdows, D. (2009). Empirical distributional semantics: methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2), 390–405. <https://doi.org/10.1016/j.jbi.2009.02.002>
- Coletti, M. H., & Bleich, H. L. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association: JAMIA*, 8(4), 317–323.
- Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS One*, 6(8), e23610. <https://doi.org/10.1371/journal.pone.0023610>

- Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., & Peipins, L. A. (2005). Cancer Internet search activity on a major search engine, United States 2001-2003. *Journal of Medical Internet Research*, 7(3), e36.
<https://doi.org/10.2196/jmir.7.3.e36>
- Corbin, J. M., & Strauss, A. (2007). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd edition). Los Angeles, Calif: SAGE Publications, Inc.
- Coumou, H. C. H., & Meijman, F. J. (2006). How do primary care physicians seek answers to clinical questions? A literature review. *Journal of the Medical Library Association: JMLA*, 94(1), 55–60.
- Covell, D. G., Uman, G. C., & Manning, P. R. (1985). Information needs in office practice: are they being met? *Annals of Internal Medicine*, 103(4), 596–599.
- Cronin, R. M., Fabbri, D., Denny, J. C., Rosenbloom, S. T., & Jackson, G. P. (2017). A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*, 105, 110–120.
<https://doi.org/10.1016/j.ijmedinf.2017.06.004>
- Cufino Svitone, E., Garfield, R., Vasconcelos, M. I., & Araujo Craveiro, V. (2000). Primary health care lessons from the northeast of Brazil: the Agentes de Saúde Program. *Revista Panamericana De Salud Publica = Pan American Journal of Public Health*, 7(5), 293–302.
- Darmoni, S. J., Soualmia, L. F., Griffon, N., Grosjean, J., Kerdelhué, G., Kergourlay, I., & Dahamna, B. (2013). Multi-lingual search engine to access PubMed

- monolingual subsets: a feasibility study. *Studies in Health Technology and Informatics*, 192, 966. <https://doi.org/10.3233/978-1-61499-289-9-966>
- de Araújo Novaes, M., Pinto Barbosa, A. K., Soares de Araújo, K., Lacerda de A Couto, J. M., Araújo, G., & Sarmiento, L. (2005). Experiences on the use of a second opinion software for the primary care. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 889.
- de Holanda, A. L. F., Barbosa, A. A. de A., & Brito, E. W. G. (2009). [Reflections around the performance of community health agents in oral health strategies]. *Ciencia & Saude Coletiva*, 14 Suppl 1, 1507–1512.
- de Melo, M. do C. B., Nunes, M. V., Resende, R. F., Figueiredo, R. R., Ruas, S. S. M., Dos Santos, A. de F., ... de Aguiar, R. A. T. (2017). Belo Horizonte-Telehealth: Incorporation of Teleconsultations in a Health Primary Care System. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*. <https://doi.org/10.1089/tmj.2017.0165>
- de Souza, C. H. A., Morbeck, R. A., Steinman, M., Hors, C. P., Bracco, M. M., Kozasa, E. H., & Leão, E. R. (2017). Barriers and Benefits in Telemedicine Arising Between a High-Technology Hospital Service Provider and Remote Public Healthcare Units: A Qualitative Study in Brazil. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 23(6), 527–532. <https://doi.org/10.1089/tmj.2016.0158>
- Dee, C., & Blazek, R. (1993). Information needs of the rural physician: a descriptive study. *Bulletin of the Medical Library Association*, 81(3), 259–264.

- Departamento de Informática do SUS. (2008). Portal da saúde - Informações de Saúde (TABNET) - Epidemiológicas e Morbidade. Retrieved September 23, 2017, from <http://www2.datasus.gov.br/DATASUS/index.php?area=0203>
- Dias, R. D. S., Marques, A. D. F. H., Diniz, P. R. B., Silva, T. A. B. D., Cofiel, L., Mariani, M. M. D. C., ... Tavares, H. (2015). Telemental health in Brazil: past, present and integration into primary care. *Archives of Clinical Psychiatry (São Paulo)*, 42(2), 41–44. <https://doi.org/10.1590/0101-608300000000046>
- Dick, B. (2014, May 28). Grounded theory: a thumbnail sketch. Retrieved March 30, 2018, from <http://www.aral.com.au/resources/grounded.html>
- Dickerson, S., Reinhart, A. M., Feeley, T. H., Bidani, R., Rich, E., Garg, V. K., & Hershey, C. O. (2004). Patient Internet use for health information at three urban primary care clinics. *Journal of the American Medical Informatics Association: JAMIA*, 11(6), 499–504. <https://doi.org/10.1197/jamia.M1460>
- Diniz, P. R. B., Ribeiro Sales, F. J., & de Araújo Novaes, M. (2016). Providing Telehealth Services to a Public Primary Care Network: The Experience of RedeNUTES in Pernambuco, Brazil. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 22(8), 694–698. <https://doi.org/10.1089/tmj.2015.0209>
- Dixon, W. J. (1950). Analysis of Extreme Values. *The Annals of Mathematical Statistics*, 21(4), 488–506.
- Dorsch, J. L. (2000). Information needs of rural health professionals: a review of the literature. *Bulletin of the Medical Library Association*, 88(4), 346–354.

- dos Santos, K. T., Saliba, N. A., Moimaz, S. A. S., Arcieri, R. M., & Carvalho, M. de L. (2011). [Community Health Agent: status adapted with Family Health Program reality?]. *Ciencia & Saude Coletiva*, *16 Suppl 1*, 1023–1028.
- Duarte, M. C. M. B., Amorim, M. R., Cuevas, L. E., Cabral-Filho, J. E., & Correia, J. B. (2005). Risk factors for death from meningococcal infection in Recife, Brazil. *Journal of Tropical Pediatrics*, *51*(4), 227–231.
<https://doi.org/10.1093/tropej/fmi006>
- Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). Influenza forecasting with Google Flu Trends. *PloS One*, *8*(2), e56176.
<https://doi.org/10.1371/journal.pone.0056176>
- Duncan, B. B., Chor, D., Aquino, E. M. L., Bensenor, I. M., Mill, J. G., Schmidt, M. I., ... Barreto, S. M. (2012). Chronic non-communicable diseases in Brazil: priorities for disease management and research. *Revista De Saude Publica*, *46 Suppl 1*, 126–134.
- Ellis, I. (2004). Is telehealth the right tool for remote communities? Improving health status in rural Australia. *Contemporary Nurse*, *16*(3), 163–168.
- Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. <https://doi.org/10.1111/j.1365-2648.2007.04569.x>
- Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., & Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ (Clinical Research Ed.)*, *319*(7206), 358–361.

- Ely, John W., Osheroff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D. C., Stevermer, J. J., & Pifer, E. A. (2002). Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ (Clinical Research Ed.)*, *324*(7339), 710.
- Ely, John W., Osheroff, J. A., Maviglia, S. M., & Rosenbaum, M. E. (2007). Patient-care questions that physicians are unable to answer. *Journal of the American Medical Informatics Association: JAMIA*, *14*(4), 407–414.
<https://doi.org/10.1197/jamia.M2398>
- Eysenbach, G. (2006). Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 244–248.
- Faggiolani, C. (2011). Perceived Identity: applying Grounded Theory in Libraries. *JLIS.it*, *2*(1). <https://doi.org/10.4403/jlis.it-4592>
- Finn, N. B., & Bria, W. F. (Eds.). (2009). *Digital Communication in Medical Practice* (2009th ed.). Springer.
- Freitas-Junior, R., Gonzaga, C. M. R., Freitas, N. M. A., Martins, E., & de Cássia de Maio Dardes, R. (2012). Disparities in female breast cancer mortality rates in Brazil between 1980 and 2009. *Clinics*, *67*(7), 731–737.
[https://doi.org/10.6061/clinics/2012\(07\)05](https://doi.org/10.6061/clinics/2012(07)05)
- Friedman, C. (1997). Towards a comprehensive medical language processing system: methods and issues. *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 595–599.

- Galavote, H. S., do Prado, T. N., Maciel, E. L. N., & de Cássia Duarte Lima, R. (2011). Disclosing the work processes of the community health agents on the Family Health Strategy in Vitória (ES, Brazil). *Ciencia & Saude Coletiva*, 16(1), 231–240.
- Galvão, P. R. S., Ferreira, A. T., Maciel, M. D. G. G., De Almeida, R. P., Hinders, D., Schreuder, P. A. M., & Kerr-Pontes, L. R. S. (2008). An evaluation of the Sinan health information system as used by the Hansen's disease control programme, Pernambuco State, Brazil. *Leprosy Review*, 79(2), 171–182.
- Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4), 42–47.
- Gao, W., Yang, B.-B., & Zhou, Z.-H. (2016). On the Robustness of Nearest Neighbor with Noisy Data. *ArXiv:1607.07526 [Cs]*. Retrieved from <http://arxiv.org/abs/1607.07526>
- Garg, A., & Turtle, K. M. (2003). Effectiveness of training health professionals in literature search skills using electronic health databases--a critical appraisal. *Health Information and Libraries Journal*, 20(1), 33–41.
- Gaspar, R. S., Nunes, N., Nunes, M., & Rodrigues, V. P. (2016). Temporal analysis of reported cases of tuberculosis and of tuberculosis-HIV co-infection in Brazil between 2002 and 2012. *Jornal Brasileiro De Pneumologia: Publicacao Oficial Da Sociedade Brasileira De Pneumologia E Tisiologia*, 42(6), 416–422. <https://doi.org/10.1590/S1806-37562016000000054>

- Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology*, *10*(11), e1003892. <https://doi.org/10.1371/journal.pcbi.1003892>
- Gibaja, E., & Ventura, S. (2014). Multi-label learning: a review of the state of the art and ongoing research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(6), 411–444. <https://doi.org/10.1002/widm.1139>
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. <https://doi.org/10.1038/nature07634>
- Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research* (1st edition). New York: Aldine Pub. Co.
- Gonzaga, C. M. R., Freitas-Junior, R., Curado, M.-P., Sousa, A.-L. L., Souza-Neto, J.-A., & Souza, M. R. (2015). Temporal trends in female breast cancer mortality in Brazil and correlations with social inequalities: ecological time-series study. *BMC Public Health*, *15*, 96. <https://doi.org/10.1186/s12889-015-1445-7>
- Green, M. L., & Ruff, T. R. (2005). Why do residents fail to answer their clinical questions? A qualitative study of barriers to practicing evidence-based medicine. *Academic Medicine: Journal of the Association of American Medical Colleges*, *80*(2), 176–182.
- Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, *21*(1), 27–58.

- Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21. <https://doi.org/10.2307/1266761>
- Grubbs, F. E., & Beck, G. (1972). Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations: *Technometrics*: Vol 14, No 4. *Technometrics*, *14*(4), 847–854. <https://doi.org/10.1080/00401706.1972.10488981>
- Haddad, A. E., Skelton-Macedo, M. C., Abdala, V., Bavaresco, C., Mengehel, D., Abdala, C. G., & Harzheim, E. (2015). Formative second opinion: qualifying health professionals for the unified health system through the Brazilian Telehealth Program. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, *21*(2), 138–142. <https://doi.org/10.1089/tmj.2014.0001>
- Hagihara, A., Miyazaki, S., & Abe, T. (2012). Internet suicide searches and the incidence of suicide in young people in Japan. *European Archives of Psychiatry and Clinical Neuroscience*, *262*(1), 39–46. <https://doi.org/10.1007/s00406-011-0212-8>
- Hammond, W. E., Jaffe, C., Cimino, J. J., & Huff, S. M. (2012). Chapter 7: Standards in Biomedical Informatics. In E. H. Shortliffe & J. J. Cimino (Eds.), *Biomedical Informatics: Computer Applications in Health Care and Biomedicine* (3rd ed., pp. 265–311). Springer.
- Hanneman, R. A., Kposowa, A. J., & Riddle, M. D. (2012). *Basic Statistics for Social Research* (1 edition). San Francisco, CA: Jossey-Bass.
- Harzheim, E., Duncan, B. B., Stein, A. T., Cunha, C. R. H., Goncalves, M. R., Trindade, T. G., ... Pinto, M. E. B. (2006). Quality and effectiveness of different approaches

- to primary care delivery in Brazil. *BMC Health Services Research*, 6, 156.
<https://doi.org/10.1186/1472-6963-6-156>
- Hastie, T., Friedman, J., & Tibshirani, R. (2001). Additive Models, Trees, and Related Methods. In *The Elements of Statistical Learning* (pp. 257–298). Springer, New York, NY. https://doi.org/10.1007/978-0-387-21606-5_9
- Henry, S., Cuffy, C., & McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of Biomedical Informatics*, 77, 111–119.
<https://doi.org/10.1016/j.jbi.2017.12.006>
- Heukelbach J., Wilcke T., Winter B., & Feldmeier H. (2005). Epidemiology and morbidity of scabies and pediculosis capitis in resource-poor communities in Brazil. *British Journal of Dermatology*, 153(1), 150–156.
<https://doi.org/10.1111/j.1365-2133.2005.06591.x>
- Hill, S., Merchant, R., & Ungar, L. (2013). LESSONS LEARNED ABOUT PUBLIC HEALTH FROM ONLINE CROWD SURVEILLANCE. *Big Data*, 1(3), 160–167. <https://doi.org/10.1089/big.2013.0020>
- Huang, M., Névéol, A., & Lu, Z. (2011). Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association: JAMIA*, 18(5), 660–667. <https://doi.org/10.1136/amiajnl-2010-000055>
- Hulth, A., Rydevik, G., & Linde, A. (2009). Web queries as a source for syndromic surveillance. *PloS One*, 4(2), e4378. <https://doi.org/10.1371/journal.pone.0004378>
- Huth, E. J. (1989). The underused medical literature. *Annals of Internal Medicine*, 110(2), 99–100.

- Hutwagner, L., Thompson, W., Seeman, G. M., & Treadwell, T. (2003). The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 80(2 Suppl 1), i89-96.
- IBM. (n.d.). Modified z score. Retrieved March 30, 2018, from https://www.ibm.com/support/knowledgecenter/SSWLKY_1.0.0/com.ibm.spss.analyticcatalyst.help/analytic_catalyst/modified_z.html
- Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.
- Jain, N. L., & Friedman, C. (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 829–833.
- Johnson, A. K., Mikati, T., & Mehta, S. D. (2016). Examining the themes of STD-related Internet searches to increase specificity of disease forecasting using Internet search terms. *Scientific Reports*, 6, 36503. <https://doi.org/10.1038/srep36503>
- Johnson, C. D., Noyes, J., Haines, A., Thomas, K., Stockport, C., Ribas, A. N., & Harris, M. (2013). Learning from the Brazilian community health worker model in North Wales. *Globalization and Health*, 9, 25. <https://doi.org/10.1186/1744-8603-9-25>
- Joshi, A., Novaes, M. A., Iyengar, S., Machiavelli, J. L., Zhang, J., Vogler, R., & Hsu, C. E. (2011). Evaluation of a tele-education programme in Brazil. *Journal of*

Telemedicine and Telecare, 17(7), 341–345.

<https://doi.org/10.1258/jtt.2011.101209>

Kanerva, P. (1988). *Sparse Distributed Memory*. MIT Press. Retrieved from

<https://mitpress.mit.edu/books/sparse-distributed-memory>

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 103–6). Erlbaum.

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling Methods for Identifying Outliers. *International Journal of Statistics and Systems*, 10(2), 231–238.

Kluthcovsky, A. C. G. C., & Takayanagui, A. M. M. (2006). Community health agent: a literature review. *Revista Latino-Americana De Enfermagem*, 14(6), 957–963.

Kman, N. E., & Bachmann, D. J. (2012). Biosurveillance: a review and update. *Advances in Preventive Medicine*, 2012, 301408. <https://doi.org/10.1155/2012/301408>

Kourouthanassis, P. E., Mikalef, P., Ioannidou, M., & Pateli, A. (2015). Exploring the online satisfaction gap of medical doctors: an expectation-confirmation investigation of information needs. *Advances in Experimental Medicine and Biology*, 820, 217–228. https://doi.org/10.1007/978-3-319-09012-2_15

Kumar, S., & Krupinski, E. (Eds.). (2008). *Teleradiology* (2008 edition). Berlin: Springer.

Leite, I. da C., Valente, J. G., Schramm, J. M. de A., Daumas, R. P., Rodrigues, R. do N., Santos, M. de F., ... Mota, J. C. da. (2015). Burden of disease in Brazil and its

- regions, 2008. *Cadernos De Saude Publica*, 31(7), 1551–1564.
<https://doi.org/10.1590/0102-311X00111614>
- Liang, B., & Scammon, D. L. (2013). Incidence of online health information search: a useful proxy for public health risk perception. *Journal of Medical Internet Research*, 15(6), e114. <https://doi.org/10.2196/jmir.2401>
- Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), 265–266.
- Liu, C., Cao, L., & Yu, P. S. (2014). A hybrid coupled k-nearest neighbor algorithm on imbalance data. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 2011–2018). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/IJCNN.2014.6889798>
- Lowe, H. J., & Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, 271(14), 1103–1108.
- Lussier, Y. A., Shagina, L., & Friedman, C. (2001). Automating SNOMED coding using medical language understanding: a feasibility study. *Proceedings. AMIA Symposium*, 418–422.
- Maldonado, J. M. S. de V., Marques, A. B., & Cruz, A. (2016). Telemedicine: challenges to dissemination in Brazil. *Cadernos De Saude Publica*, 32Suppl 2(Suppl 2), e00155615. <https://doi.org/10.1590/0102-311X00155615>
- Mancini, F., Sousa, F. S., Teixeira, F. O., Falcão, A. E. J., Hummel, A. D., da Costa, T. M., ... Pisa, I. T. (2011). Use of Medical Subject Headings (MeSH) in Portuguese

- for categorizing web-based healthcare content. *Journal of Biomedical Informatics*, 44(2), 299–309. <https://doi.org/10.1016/j.jbi.2010.12.002>
- Manoj, K. (2015, July). *Outlier detection in datamining*. Manonmaniam Sundaranar University, Tamil Nadu, India. Retrieved from <http://hdl.handle.net/10603/131060>
- Mao, Y., & Lu, Z. (2017). MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of Biomedical Semantics*, 8(1), 15. <https://doi.org/10.1186/s13326-017-0123-3>
- MapasBlog. (n.d.). Mapas de Pernambuco. Retrieved March 21, 2018, from <https://mapasblog.blogspot.com/2011/11/mapas-de-pernambuco.html>
- Martin, D. I., & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (1st ed.). Routledge Handbooks Online. <https://doi.org/10.4324/9780203936399.ch2>
- Martin, P. Y., & Turner, B. A. (1986). Grounded Theory and Organizational Research. *The Journal of Applied Behavioral Science*, 22(2), 141–157. <https://doi.org/10.1177/002188638602200207>
- Marzari, C. K., Junges, J. R., & Selli, L. (2011). [Community health agents: profile and education]. *Ciencia & Saude Coletiva*, 16 Suppl 1, 873–880.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. Retrieved from <http://arxiv.org/abs/1301.3781>

- Milnovich, G. J., Williams, G. M., Clements, A. C. A., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet. Infectious Diseases*, 14(2), 160–168. [https://doi.org/10.1016/S1473-3099\(13\)70244-5](https://doi.org/10.1016/S1473-3099(13)70244-5)
- Miller, N., Lacroix, E. M., & Backus, J. E. (2000). MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. *Bulletin of the Medical Library Association*, 88(1), 11–17.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Cambridge, MA: The MIT Press.
- Mork, J. G., Yepes, A. J. J., & Aronson, A. R. (2013). *The NLM Medical Text Indexer System for Indexing Biomedical Literature*. Retrieved from https://ii.nlm.nih.gov/Publications/Papers/MTI_System_Description_Expanded_2013_Accessible.pdf
- Musen, M. A. (1992). Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research, an International Journal*, 25(5), 435–467.
- National Institute of Standards and Technology. (2013a, October 30). Detection of Outliers. Retrieved March 24, 2018, from <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>
- National Institute of Standards and Technology. (2013b, October 30). Grubbs' Test for Outliers. Retrieved March 24, 2018, from <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm>

- National Institute of Standards and Technology. (2015, October 13). Dixon Test.
Retrieved March 24, 2018, from
<https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/dixon.htm>
- National Library of Medicine. (2007). Indexing Initiative. Retrieved September 23, 2017,
from http://wayback.archive-it.org/org-350/20130703100840/http://ii.nlm.nih.gov/Eval_Analysis/Eval_2007/summary.shtml
- Nazareth, J. V., de Souza, K. V., Beinler, M. A., Barra, J. S., Brüggemann, O. M., & Pimenta, A. M. (2017). Special attention to women experiencing high-risk pregnancy: Delivery, care assistance and neonatal outcomes in two Brazilian maternity wards. *Midwifery*, 53, 42–48.
<https://doi.org/10.1016/j.midw.2017.07.009>
- Neghme, A. (1975). Operations of the Biblioteca Regional de Medicina (BIREME). *Bulletin of the Medical Library Association*, 63(2), 173–179.
- Ocampo, A. J., Chunara, R., & Brownstein, J. S. (2013). Using search queries for malaria surveillance, Thailand. *Malaria Journal*, 12, 390. <https://doi.org/10.1186/1475-2875-12-390>
- Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10), e1003256.
<https://doi.org/10.1371/journal.pcbi.1003256>

- Osheroff, J. A., Forsythe, D. E., Buchanan, B. G., Bankowitz, R. A., Blumenfeld, B. H., & Miller, R. A. (1991). Physicians' information needs: analysis of questions posed during clinical teaching. *Annals of Internal Medicine*, *114*(7), 576–581.
- Paim, J., Travassos, C., Almeida, C., Bahia, L., & Macinko, J. (2011). The Brazilian health system: history, advances, and challenges. *Lancet (London, England)*, *377*(9779), 1778–1797. [https://doi.org/10.1016/S0140-6736\(11\)60054-8](https://doi.org/10.1016/S0140-6736(11)60054-8)
- Paireau, J., Chen, A., Broutin, H., Grenfell, B., & Basta, N. E. (2016). Seasonal dynamics of bacterial meningitis: a time-series analysis. *The Lancet Global Health*, *4*(6), e370–e377. [https://doi.org/10.1016/S2214-109X\(16\)30064-X](https://doi.org/10.1016/S2214-109X(16)30064-X)
- Pan, E., Cusack, C., Hook, J., Vincent, A., Kaelber, D. C., Bates, D. W., & Middleton, B. (2008). The value of provider-to-provider telehealth. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, *14*(5), 446–453. <https://doi.org/10.1089/tmj.2008.0017>
- Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A.-J. (2009). More diseases tracked by using Google Trends. *Emerging Infectious Diseases*, *15*(8), 1327–1328. <https://doi.org/10.3201/eid1508.090299>
- Pereira, T. A., & Montero, E. F. de S. (2012). DeCS terminology and the new rules on orthography of Portuguese language: guidelines for an update. *Acta Cirurgica Brasileira*, *27*(7), 509–514.
- Peres, C. R. F. B., Caldas Júnior, A. L., da Silva, R. F., & Marin, M. J. S. (2011). [The community health agent and working as a team: the easy and difficult aspects]. *Revista Da Escola De Enfermagem Da U S P*, *45*(4), 905–911.

- Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 47(11), 1443–1448. <https://doi.org/10.1086/593098>
- Pope, C., Ziebland, S., & Mays, N. (2000). Analysing qualitative data. *BMJ*, 320(7227), 114–116. <https://doi.org/10.1136/bmj.320.7227.114>
- Ralph, N., Birks, M., & Chapman, Y. (2014). Contextual Positioning: Using Documents as Extant Data in Grounded Theory Research. *SAGE Open*, 4(3). <https://doi.org/10.1177/2158244014552425>
- Resnick, M. P., Santana, F., de Araujo Novaes, M., Shamenek, F. S., Frieden, L., & Iyengar, M. S. (2013). Representing second opinion requests from primary care within the Brazilian tele-health program: international classification of primary care, second edition. *Studies in Health Technology and Informatics*, 192, 1190.
- Rezende, E. J. C., Tavares, E. C., Alves, H. J., dos Santos, A. de F., & de Melo, M. do C. B. (2013). Teleconsultations in public primary care units of the city of belo horizonte, Brazil: profile of patients and physicians. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 19(8), 613–618. <https://doi.org/10.1089/tmj.2012.0179>
- Ricci, M. A., Caputo, M. P., Callas, P. W., & Gagne, M. (2005). The use of telemedicine for delivering continuing medical education in rural communities. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 11(2), 124–129. <https://doi.org/10.1089/tmj.2005.11.124>

- Rinde, E., Nordrum, I., & Nymo, B. J. (1993). Telemedicine in rural Norway. *World Health Forum*, 14(1), 71–77.
- Rodrigues Netto, N., Mitre, A. I., Lima, S. V. C., Fugita, O. E., Lima, M. L., Stoianovici, D., ... Kavoussi, L. R. (2003). Telementoring between Brazil and the United States: initial experience. *Journal of Endourology*, 17(4), 217–220.
<https://doi.org/10.1089/089277903765444339>
- Rorabacher, D. B. (1991). Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level. *Analytical Chemistry*, 63(2), 139–146.
<https://doi.org/10.1021/ac00002a010>
- Rossignol, L., Pelat, C., Lambert, B., Flahault, A., Chartier-Kastler, E., & Hanslik, T. (2013). A method to assess seasonality of urinary tract infections based on medication sales and google trends. *PloS One*, 8(10), e76020.
<https://doi.org/10.1371/journal.pone.0076020>
- Ruas, S. S. M., & Assunção, A. Á. (2013). Teleconsultations by primary care physicians of Belo Horizonte: challenges in the diffusion of innovation. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 19(5), 409–414. <https://doi.org/10.1089/tmj.2012.0095>
- Sachpazidis, I., Ohl, R., Polanczyk, C., Torres, M., Messina, L., Sales, A., & Sakas, G. (2005). Applying Telemedicine to Remote and Rural Underserved Regions in Brazil using eMedical Consulting Tool. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology*

- Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2, 2191–2195. <https://doi.org/10.1109/IEMBS.2005.1616897>
- Sahlgren, M. (2005). An introduction to random indexing. In *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.
- Sanches, L. M. P., Alves, D. S., Lopes, M. H. B. M., & Novaes, M. A. (2012). The practice of telehealth by nurses: an experience in primary healthcare in Brazil. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 18(9), 679–683. <https://doi.org/10.1089/tmj.2012.0011>
- Sandin, F., Emruli, B., & Sahlgren, M. (2017). Random Indexing of Multidimensional Data. *Knowl. Inf. Syst.*, 52(1), 267–290. <https://doi.org/10.1007/s10115-016-1012-2>
- Santillana, M., Nsoesie, E. O., Mekar, S. R., Scales, D., & Brownstein, J. S. (2014). Using clinicians' search query data to monitor influenza epidemics. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 59(10), 1446–1450. <https://doi.org/10.1093/cid/ciu647>
- Schmidt, M. I., Duncan, B. B., Azevedo e Silva, G., Menezes, A. M., Monteiro, C. A., Barreto, S. M., ... Menezes, P. R. (2011). Chronic non-communicable diseases in Brazil: burden and current challenges. *Lancet (London, England)*, 377(9781), 1949–1961. [https://doi.org/10.1016/S0140-6736\(11\)60135-9](https://doi.org/10.1016/S0140-6736(11)60135-9)

- Scott-Wright, A., Crowell, J., Zeng, Q., Bates, D., & Greenes, R. (2006). Analysis of information needs of users of MEDLINEplus, 2002 - 2003. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 699–703.
- Seifter, A., Schwarzwald, A., Geis, K., & Aucott, J. (2010). The utility of “Google Trends” for epidemiological research: Lyme disease as an example. *Geospatial Health*, 4(2), 135–137. <https://doi.org/10.4081/gh.2010.195>
- Seo, S. (2006, August 9). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. University of Pittsburgh ETD. Retrieved from <http://d-scholarship.pitt.edu/7948/>
- Sierra, M. S., Soerjomataram, I., & Forman, D. (2016). Thyroid cancer burden in Central and South America. *Cancer Epidemiology*, 44 Suppl 1, S150–S157. <https://doi.org/10.1016/j.canep.2016.07.017>
- Sinha, A. (2000). An overview of telemedicine: the virtual gaze of health care in the next century. *Medical Anthropology Quarterly*, 14(3), 291–309.
- Sitbon, L., Bruza, P. D., & Prokopp, C. (2012). Empirical analysis of the effect of dimension reduction and word order on semantic vectors. *International Journal of Semantic Computing*, 06(03), 329–351. <https://doi.org/10.1142/S1793351X12500055>
- Smith, R. (1996). What clinical information do doctors need? *BMJ (Clinical Research Ed.)*, 313(7064), 1062–1068.
- Soares, M. L. M., Amaral, N. A. C. do, Zacarias, A. C. P., & Ribeiro, L. K. de N. P. (2017). Sociodemographic, clinical and epidemiological aspects of Tuberculosis

- treatment abandonment in Pernambuco, Brazil, 2001-2014. *Epidemiologia E Servicos De Saude: Revista Do Sistema Unico De Saude Do Brasil*, 26(2), 369–378. <https://doi.org/10.5123/S1679-49742017000200014>
- Soiferman, L. K. (2010). *Compare and Contrast Inductive and Deductive Research Approaches*. Retrieved from <https://eric.ed.gov/?id=ED542066>
- Spath, M., & Buttlar, L. (1996). Information and research needs of acute-care clinical nurses. *Bulletin of the Medical Library Association*, 84(1), 112–116.
- Stefanello, S., Cais, C. F. da S., Mauro, M. L. F., Freitas, G. V. S. de, & Botega, N. J. (2008). Gender differences in suicide attempts: preliminary results of the multisite intervention study on suicidal behavior (SUPRE-MISS) from Campinas, Brazil. *Revista Brasileira De Psiquiatria (Sao Paulo, Brazil: 1999)*, 30(2), 139–143.
- Thacker, S. B., Qualters, J. R., Lee, L. M., & Centers for Disease Control and Prevention. (2012). Public health surveillance in the United States: evolution and challenges. *MMWR Supplements*, 61(3), 3–9.
- The Apache Software Foundation. (n.d.). Lucene. Retrieved March 7, 2018, from <https://lucene.apache.org/>
- Tran, N., Luong, T., & Krauthammer, M. (2007). Mapping terms to UMLS concepts of the same semantic type. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 1136.
- Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document

- retrieval. *Bioinformatics (Oxford, England)*, 25(11), 1412–1418.
<https://doi.org/10.1093/bioinformatics/btp249>
- Tukey, J. W. (1977). *Exploratory Data Analysis* (1 edition). Reading, Mass: Pearson.
- Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188.
<https://doi.org/10.1613/jair.2934>
- Vahedi, K., & Amarenco, P. (2000). Cardiac Causes of Stroke. *Current Treatment Options in Neurology*, 2(4), 305–318.
- Vasuki, V., & Cohen, T. (2010). Reflective random indexing for semi-automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, 43(5), 694–700.
<https://doi.org/10.1016/j.jbi.2010.04.001>
- Virtual-Brazil.com. (n.d.). Map of Minas Gerais, Brazil. Retrieved March 21, 2018, from <http://www.v-brazil.com/tourism/minas-gerais/map-minas-gerais.html>
- Wahle, M., Widdows, D., Herskovic, J. R., Bernstam, E. V., & Cohen, T. (2012). Deterministic binary vectors for efficient automated indexing of MEDLINE/PubMed abstracts. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2012*, 940–949.
- WebMD. (n.d.). What Is Meningitis? Retrieved March 26, 2018, from <https://www.webmd.com/children/understanding-meningitis-basics>
- Westbrook, J. I., Coiera, E. W., & Gosling, A. S. (2005). Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the*

American Medical Informatics Association: JAMIA, 12(3), 315–321.

<https://doi.org/10.1197/jamia.M1717>

White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B., & Horvitz, E. (2013). Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association: JAMIA*, 20(3), 404–408.

<https://doi.org/10.1136/amiajnl-2012-001482>

Widdows, D., & Cohen, T. (2010). The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. *Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010)*, 9–15.

Widdows, D., & Cohen, T. (2015). Reasoning with vectors: A continuous model for fast robust inference. *Logic Journal of the IGPL*, 23(2), 141–173.

<https://doi.org/10.1093/jigpal/jzu028>

Widdows, D., & Ferraro, K. (2008). Semantic vectors: A scalable open source package and online technology management application. In *Sixth international conference on Language Resources and Evaluation (LREC)*.

Wikipedia. (2018, March 5). States of Brazil. In *Wikipedia*. Retrieved from

https://en.wikipedia.org/w/index.php?title=States_of_Brazil&oldid=828989203

Wilson, K., & Brownstein, J. S. (2009). Early detection of disease outbreaks using the Internet. *CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, 180(8), 829–831.

<https://doi.org/10.1503/cmaj.090215>

- Woods, P., Gapp, R., & King, M. A. (2016). Generating or developing grounded theory: methods to understand health and illness. *International Journal of Clinical Pharmacy*, 38(3), 663–670. <https://doi.org/10.1007/s11096-016-0260-2>
- Woolf, S. H., & Benson, D. A. (1989). The medical information needs of internists and pediatricians at an academic medical center. *Bulletin of the Medical Library Association*, 77(4), 372–380.
- World Health Organization. (2018). History of ICD. Retrieved March 21, 2018, from www.who.int/classifications/icd/en/?HistoryOfICD.pdf
- Ximenes, R. A. de A., Pereira, L. M. B., Martelli, C. M. T., Merchán-Hamann, E., Stein, A. T., Figueiredo, G. M., ... Cardoso, M. R. A. (2010). Methodology of a nationwide cross-sectional survey of prevalence and epidemiological patterns of hepatitis A, B and C infection in Brazil. *Cadernos De Saude Publica*, 26(9), 1693–1704.
- Yang, Y., & Liu, X. (1999). A Re-examination of Text Categorization Methods. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42–49). New York, NY, USA: ACM. <https://doi.org/10.1145/312624.312647>
- Zaiontz, C. (2018). Spearman's Rank Correlation [Blog]. Retrieved April 26, 2018, from <http://www.real-statistics.com/correlation/spearmans-rank-correlation/>
- Zhou, X., Ye, J., & Feng, Y. (2011). Tuberculosis surveillance by analyzing Google trends. *IEEE Transactions on Bio-Medical Engineering*, 58(8). <https://doi.org/10.1109/TBME.2011.2132132>

Zollo, S. A., Kienzle, M. G., Henshaw, Z., Crist, L. G., & Wakefield, D. S. (1999). Tele-education in a telemedicine environment: implications for rural health care and academic medical centers. *Journal of Medical Systems*, 23(2), 107–122.

Appendix A: Map – States and Regions in Brazil (Wikipedia, 2018)



Appendix B: Map – State of Pernambuco (MapasBlog, n.d.)



Appendix C: Map – State of Minas Gerais
(Virtual-Brazil.com, n.d.)



Appendix D: Committee for the Protection of Human Subjects



Committee for the Protection of Human Subjects

3410 Texas Street, Suite 1006
Houston, Texas 77030

Dr. Trevor Cohen
UT-H - SBMI - Health Informatics

October 28, 2013

HSC-SBMI-13-0716 - *The Development and Evaluation of Automated Methods of Text Analysis that Can Be Used to Interpret the Health-Related Queries of Primary Care Practitioners*

The above named project is determined to qualify for exempt status according to 45 CFR 46.101(b)

CATEGORY #4 : *Research, involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified directly or through identifiers linked to the subjects.*



Health Insurance Portability and Accountability Act:
Exempt from HIPAA

CHANGES: Should you choose to make any changes to the protocol that would involve the inclusion of human subjects or identified data from humans, please submit the change via iRIS to the Committee for the Protection of Human Subjects for review.

STUDY CLOSURES: Upon completion of your project, submission of a study closure report is required. The study closure report should be submitted once all data has been collected and analyzed.

Should you have any questions, please contact the Office of Research Support Committees at 713-500-7943.

Appendix E: Letter of Agreement – Statement of Authorization for Research

Letter of Agreement - Statement of Authorization for Research

Project Title: The development and evaluation of automated methods of text analysis that can be used to interpret the health-related queries of primary care practitioners.
Author: Melissa P. Resnick
Adviser: Trevor Cohen, MBChB, PhD (Chair)
Co-Adviser: Chiehwen Hsu, MS, PhD

To: NUTES

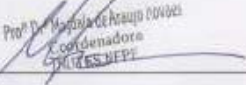
Dear All,

I, Melissa P. Resnick, Ph.D. student at the School of Biomedical Informatics, University of Texas Health Science Center of Houston, request authorization to perform doctoral research in the Telehealth Center, NUTES, institution under your responsibility.

This study will be part of the work of doctoral research. The aim of this project is to Evaluate the sensitivity of terminological systems in order to support Telehealth Primary Care Services delivery. The terminologies taken into account to perform this study will be the International Classification for Primary Care v.2, the International Classification of Diseases v.10, the Medical Subject Headings (MeSH) and its English translation correspondent, Health Sciences Descriptors (MeSH).

This study does not imply burden for the institution and incur no risks involved. Nor will any payment be made for the institution in question. At any time during the study you may request clarification or settle any doubts that may exist.

I, Magdala de Araújo Novães, after reading and understanding the text above, I give permission for the implementation of data collection in this institution.


Prof. Dr. Magdala de Araújo Novães
Coordenadora
NUTES/UFPE

Recife, 2nd of August of 2013.

NUTES | Hospital das Clínicas | 21 andar | Av. Prof. Moraes Rego s/n | Cidade Universitária | Recife | PE | CEP 50.670-420
Fone: +55 81 2126.0000 | Fax: +55 81 2126.0001 | www.nutes.ufpe.br | www.ufpe.br | contato@nutes.ufpe.br

1 de 1

Appendix F: Alphabetical list of 109 unique categories from training data set

[Bold Highlight = Denotes 19 unique categories not found in PE unannotated data set]

4	Amino Acids, Peptides, and Proteins	Anesthesia and Analgesia	Animal Diseases	Bacteria
Bacterial Infections and Mycoses	Behavior and Behavior Mechanisms	Behavioral Disciplines and Activities	Biological Factors	Biological Phenomena
Biomedical and Dental Materials	Body Regions	Carbohydrates	Cardiovascular Diseases	Cell Physiological Phenomena
Cells	Chemical Actions and Uses	Chemical Phenomena	Chemically-Induced Disorders	Circulatory and Respiratory Phenomena
Circulatory and Respiratory Physiological Phenomena	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Data Collection	Dentistry	Diagnosis
Digestive System	Digestive System and Oral Physiological Phenomena	Digestive System Diseases	Education, Nonprofessional	Endocrine System Diseases
Environment and Public Health	Enzymes	Enzymes and Coenzymes	Equipment and Supplies	Eukaryota
Eye Diseases	Female Urogenital Diseases and Pregnancy Complications	Fluids and Secretions	Food and Beverages	Genitalia, Female
Geographic Locations	Health Care Economics and Organizations	Health Care Quality, Access, and Evaluation	Health Facilities	Health Occupations
Health Personnel	Health Promotion	Health Services	Health Services Administration	Health Services Management
Health Surveillance	Hemic and Immune Systems	Hemic and Lymphatic Diseases	Heterocyclic Compounds	Homeopathy
Hormones, Hormone Substitutes, and Hormone Antagonists	Human Activities	Immune System Diseases	Inorganic Chemicals	Integumentary System
Investigative Techniques	Lipids	Macromolecular Substances	Male Urogenital Diseases	Mathematical Concepts
Menotropins	Mental Disorders	Metabolic Phenomena	Musculoskeletal and Neural	Musculoskeletal Diseases

			Physiological Phenomena	
Musculoskeletal System	Neoplasms	Nervous System	Nervous System Diseases	Nutritional and Metabolic Diseases
Ocular Physiological Phenomena	Organic Chemicals	Otorhinolaryngologic Diseases	Parasitic Diseases	Particulate Matter
Pathologic Processes	Pathological Conditions, Anatomical	Persons	Pharmaceutical Preparations	Physiological Phenomena
Plant Preparations	Plant Structure	Polycyclic Compounds	Population Characteristics	Psychological Phenomena and Processes
Public Health	Reproductive Physiological Phenomena	Respiratory Tract Diseases	Science and Health	Signs and Symptoms
Skin and Connective Tissue Diseases	Social Sciences	Stomatognathic Diseases	Stomatognathic System	Surgical Procedures, Operative
Technology, Industry, and Agriculture	Therapeutics	Tissues	Urogenital System	Vaccines
Virus Diseases	Viruses	Waste Products	Wounds and Injuries	

Appendix G: Outlier Detection – 90 categories x 4 tests

# tests (out of 4) outliers detected in	Category	Number of Assignments by Model	Number of year/season slots (out of 12) with zero assignments	Modified Z-Score Test	Grubbs Test	Tukey Outer Fence Test	Dixon Q-Test
3	Amino Acids, Peptides, and Proteins	11	6	2011_Spring	2011_Spring		2011_Spring
3	Anesthesia and Analgesia	1	11		2010_Spring	2010_Spring	2010_Spring
0	Bacteria	5	8				
1	Bacterial Infections and Mycoses	186	0				(Max) 2012_Summer (Min) 2012_Spring
3	Behavior and Behavior Mechanisms	81	0	2012_Spring	2012_Spring		2012_Spring
3	Behavioral Disciplines and Activities	2	10		2012_Spring	2012_Spring	2012_Spring
0	Biological Factors	8	7				
4	Biological Phenomena	29	2	2012_Spring	2012_Spring	2012_Spring	2012_Spring
0	Biomedical and Dental Materials	17	3				
0	Body Regions	7	8				
2	Cardiovascular Diseases	74	0	2012_Spring			2012_Spring
3	Cell Physiological Phenomena	1	11		2010_Fall	2010_Fall	2010_Fall
2	Chemical Actions and Uses	123	0		(Min) 2012_Spring		(Max) 2012_Summer (Min) 2012_Spring
3	Chemical Phenomena	1	11		2010_Winter	2010_Winter	2010_Winter
3	Chemically-Induced Disorders	28	1				
2	Circulatory and Respiratory Physiological Phenomena	6	7		2011_Spring		2011_Spring

1	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	26	4				2012_Spring
0	Data Collection	12	5				
0	Dentistry	30	3				
0	Diagnosis	101	0				
3	Digestive System	2	10		2011_Fall	2011_Fall	2011_Fall
0	Digestive System Diseases	37	4				
0	Education, Nonprofessional	12	5				
4	Endocrine System Diseases	65	0	2010_Spring 2012_Spring	2012_Spring	2012_Spring	2012_Spring
0	Environment and Public Health	76	0				
0	Enzymes	4	8				
4	Equipment and Supplies	40	1	2012_Spring	2012_Spring	2012_Spring	2012_Spring
2	Eukaryota	3	9			2011_Spring	2011_Spring
3	Eye Diseases	2	10		2010_Fall	2010_Fall	2010_Fall
1	Female Urogenital Diseases and Pregnancy Complications	154	0	2010_Spring 2011_Spring			
1	Fluids and Secretions	11	5				2012_Fall
3	Genitalia, Female	3	9		2010_Summer	2010_Summer	2010_Summer
3	Health Care Economics and Organizations	3	9		2012_Summer	2012_Summer	2012_Summer
0	Health Care Quality, Access, and Evaluation	36	1				
3	Health Facilities	2	11		2010_Winter	2010_Winter	2010_Winter
0	Health Occupations	29	3				
1	Health Personnel	127	0				(Max) 2011_Summer (Min) 2011_Spring
1	Health Promotion	3	9			2011_Fall	
2	Health Services	164	0		(Min) 2011_Spring		Min (2011_Spring)
0	Health Services Administration	4	8				
0	Health Services Management	58	0				

0	Health Surveillance	42	0				
0	Hemic and Lymphatic Diseases	34	2				
0	Heterocyclic Compounds	50	2				
3	Homeopathy	1	11		2010_Winter	2010_Winter	2010_Winter
2	Hormones, Hormone Substitutes, and Hormone Antagonists	9	7		2012_Fall		2012_Fall
3	Human Activities	2	10		2011_Summer	2011_Summer	2011_Summer
4	Immune System Diseases	44	2	2012_Fall	2012_Fall	2012_Fall	2012_Fall
3	Inorganic Chemicals	2	10		2011_Winter	2011_Winter	2011_Winter
3	Integumentary System	1	11		2010_Fall	2010_Fall	2010_Fall
4	Investigative Techniques	106	1	2011_Fall	(Min) 2012_Spring	(Min) 2012_Spring	(Max) 2011_Fall (Min) 2012_Spring
0	Male Urogenital Diseases	63	2				
3	Menotropins	1	11		2011_Winter	2011_Winter	2011_Winter
0	Mental Disorders	58	0				
2	Musculoskeletal and Neural Physiological Phenomena	8	8		2011_Fall		2011_Fall
0	Musculoskeletal Diseases	31	1				
4	Neoplasms	104	0	2012_Spring	2012_Spring	2012_Spring	(Min) 2010_Fall (Max) 2012_Spring
0	Nervous System Diseases	72	0				
0	Nutritional and Metabolic Diseases	108	0				
2	Ocular Physiological Phenomena	2	10			2011_Spring	2011_Spring
3	Organic Chemicals	58	0	2010_Summer	2010_Summer		2010_Summer
0	Otorhinolaryngologic Diseases	13	4				
0	Parasitic Diseases	34	1				
2	Particulate Matter	2	10			2010_Spring	2010_Spring
0	Pathologic Processes	67	0				
2	Pathological Conditions, Anatomical	36	2		2012_Fall		2012_Fall
0	Persons	349	0				
2	Pharmaceutical Preparations	13	4		2010_Summer		2010_Summer

1	Physiological Phenomena	70	0	2012_Fall			
3	Polycyclic Compounds	24	1	2012_Summer	2012_Summer		2012_Summer
1	Population Characteristics	21	3				2011_Summer
3	Psychological Phenomena and Processes	13	4	2010_Summer	2010_Summer		2010_Summer
0	Public Health	221	0				
0	Reproductive Physiological Phenomena	132	0				
3	Respiratory Tract Diseases	41	0	2012_Fall	2012_Fall		2012_Fall
3	Science and Health	3	11		2010_Winter	2010_Winter	2010_Winter
0	Signs and Symptoms	67	0				
0	Skin and Connective Tissue Diseases	134	0				
2	Social Sciences	20	1	2012_Summer			2012_Summer
1	Stomatognathic Diseases	79	1				2011_Winter
1	Stomatognathic System	21	3				2011_Winter
0	Surgical Procedures, Operative	73	1				
0	Technology, Industry, and Agriculture	17	3				
0	Therapeutics	163	0				
3	Tissues	2	10		2011_Fall	2011_Fall	2011_Fall
3	Urogenital System	1	11		2010_Fall	2010_Fall	2010_Fall
1	Vaccines	24	2				(Min) 2012_Spring
0	Virus Diseases	67	1				
0	Viruses	9	7				
0	Wounds and Injuries	30	2				

