UNIVERSITY OF CALIFORNIA
RIVERSIDE

Ontology-Based Analysis of Online Healthcare Data

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Matthew Thomas Wiley

June 2016

Dissertation Committee:

Dr. Vagelis Hristidis (Evangelos Christidis), Chairperson
Dr. Eamonn Keogh
Dr. Stefano Lonardi
Dr. Vassillis Tsotras

The Dissertation of Matthew Thomas Wiley is approved:

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I would like to express my sincerest appreciation and gratitude to my advisor Professor Vagelis Hristidis. Thank you for the continuous support of my Ph.D. study and allowing me to pursue various research projects without objection. Your mentorship has been vital to my education and growth as a research scientist, and your advice on both research as well as my career has been priceless. Thank you for making my time at UCR an enjoyable endeavor! I would also like to thank my committee members, professor Eamonn Keogh, professor Stefano Lonardi, and professor Vassillis Tsotras for serving as my committee members and providing general advice and guidance through my examinations and course work.

A special thanks to my coauthors and colleagues, your collective guidance is the only reason this dissertation is possible! Chapter 2 of this dissertation, in full, is a reprint of the material as is appearing in the 2014 EDBT conference proceedings, titled Efficient Concept-based Document Ranking. Thank you to the co-authors Vagelis Hristidis, who supervised and directed the research in that publication, and Anastsios Arvanitis for providing technical expertise. Thank you Nhat Le and professor Robert El-Kareh (UCSD) for your technical and medical expertise on the methods and results presented in Chapter 3, Predicting Future Medical Concepts in Electronic Health Records. Chapter 4 of this dissertation, in full, is a reprint of the material as is appearing in the Journal of Biomedical Informatics, June 2014, titled Pharmaceutical Drugs Chatter on Online Social Networks. Thank you to the co-authors Vagelis Hristidis and Kevin Esterling, who both supervised and directed the research in that publication, and Canghong Jin for your technical expertise.

Chapter 5 of this dissertation, in full, is a reprint of the material as is appearing in the Journal of Health Services Research, March 2016, titled Providers Attributes Correlation Analysis to Their Referral Frequency and Awards. Thank you to the co-authors Vagelis Hristidis, who supervised and directed the research in that publication, and Ryan Rivas for your technical expertise. I would also like to thank my colleagues for supporting me at the lab and offering your wisdom when I got stuck. Thank you Shiwen Cheng, Eduardo Ruiz, Abhijith Kashyap, Moloud Shahbazi, and Shouq Sadah. Also, I would like to thank the reviewers and editors of Extending Database Technology (EDBT), Journal of Biomedical Informatics, and Journal of Health Services Research. Your feedback improved the strength of each publication and in turn improved the strength of this dissertation.

I would like to thank the grants that have supported me during my Ph.D. study, including National Science Foundation grants IIS-1216032, IIS-1216007, IIS-1447826, and IIP-1448848. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

And the best for last: I was to thank my family for their unconditional support and love during my graduate studies. I especially want to thank my wife Shu and my parents!

ABSTRACT OF THE DISSERTATION

Ontology-Based Analysis of Online Healthcare Data

by

Matthew Thomas Wiley

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, June 2016
Dr. Vagelis Hristidis (Evangelos Christidis), Chairperson

The wide-spread adoption of electronic health records combined with the surge of online healthcare data have created unique data analysis challenges at the intersection of computing and healthcare. These challenges include extracting meaningful concepts from clinical notes and online social networks, as well as defining scalable algorithms and knowledge discovery techniques that utilize domain-specific knowledge representations, such as biomedical ontologies. Several ontologies have been built for the healthcare domains, which include information on diseases, procedures, drugs, and relationships between them.

As a first research contribution, we study how to efficiently find medical documents semantically similar to a given document. An application of this is finding patients similar to a current patient.

We define a novel algorithm for computing similarity between two sets of documents, where each document is a set of medical concepts represented by an ontology. We evaluate the scalability and performance of our methods using a real dataset of electronic health records.

Our second research contribution studies how predict medical concepts in a patient's health record. For that, we then consider the sequence of notes in the current patient's healthcare record, and use the records of similar patients to predict the current patient's future diagnoses.

Our third contribution is the analysis of the relationship between a health online social network's characteristics, such as moderation or anonymity, and its content – we focus on pharmaceutical drug discussions. The proposed techniques include novel methods for extracting and analyzing medical concepts from social media posts. We evaluate these techniques with several online social networks, and show how each type of online social network influences its pharmaceutical-related discussions.

Lastly, we propose a data-driven analysis to discover how the quality indicators of individual healthcare providers, such as peer awards, are associated with a rich set of attributes, such as years of experience, found in publicly available datasets. Our proposed analysis pipeline includes novel methods for mapping entities across multiple sources, building classifiers of provider quality, and identifying localized attributes of provider quality.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1. Introduction

The increased prevalence of both online and private healthcare data has catalyzed unique data analysis challenges at the intersection of computing and healthcare. This prevalence is driven by two factors:

(1) the wide-spread adoption of Electronic Health Records (EHRs);

(2) the ubiquity of Online Social Networks (OSNs) in the healthcare domain.

In 2014, 97% of U.S. non-federal acute care hospitals had a certified EHR system, and 75% had adopted basic usage of their EHR system – a 15% increase from 2013 [1]. The rapid adoption of EHRs has enabled novel datasets to be published by both public and private entities, such as anonymized patient records, quality ratings of hospitals, outcomes of physician groups, patient ratings of healthcare services, and prescription and procedure data for individual healthcare providers [2-6].

Moreover, several biomedical ontologies have been developed specifically for healthcare domains, including information on diseases, procedures, drugs, and the relationships between them [7-10]. The combination of novel datasets and domain-specific knowledge representations have engendered new research directions with potential for high impact, such as measuring similarity between two patients or predicting temporal trends of patients based on historical EHR data [11-15]. However, the computational challenges of defining scalable algorithms and effective knowledge discovery techniques remains an open problem – as discussed in Chapters 2 and 3.

Further, the ubiquity of OSNs has created new sources of healthcare information, including discussions of pharmaceutical drugs and search portals for healthcare providers. These

OSNs include several forums that focus on a specific drug or disease, along with websites that allow patients to rate healthcare providers [16-19]. These OSNs and rating websites have also engendered new research directions with potential for high impact, such as detecting adverse drug reactions in OSNs, the quality and safety of content generated on OSNs, crowdsourcing drug efficacy for a specific disease, or defining signals of quality based on patient reviews [20-24]. However, a systemic method for analyzing online healthcare data using effective knowledge discovery techniques and domain-specific algorithms remains an open problem – as discussed in Chapters 4 and 5.

As a first research contribution, we study how to efficiently compute semantic similarity between medical documents, with an application of finding similar patients to a given patient. We define each document as a patient's EHR, where the EHR data is represented as a set of concepts derived from an ontology; medical researchers have found this approach to be effective for searching and finding similar EHRs [11, 14, 25]. But, the related literature has yet to define scalable algorithms that focus on efficiency and performance. Our methods consider both query relevance and query similarity. As an example of a relevance query, consider a clinical researcher searching an EHR database for patients that qualify to participate in a clinical trial for a new breast cancer treatment; information such as past treatments and specific symptoms may quality a patient for the trial, and thus the researcher wishes to find the most relevant patient records with respect to a set of medical concepts. As an example of a similarity query consider a physician who wishes to find the most similar medical case from an EHR database for a specific patient, using similar clinical indicators.

In Chapter 2, we formally define these problems and show that they pose unique computational challenges due to the semantic nature of the application – the similarity between two EHRs is a function of the minimum semantic distances from each concept of one document to a concept of the other document and vice versa. We then present efficient data structures and algorithms for both of these problems (relevance queries and similarity queries) and evaluate the scalability and performance of our methods using a real dataset of EHRs.

Our second research contribution studies how to predict medical concepts in a patient's health record. For that, we evaluate the hypothesis that we can leverage similar EHRs to predict possible future medical concepts, such as disorders, in a patient's EHR. Enabling users to find similar patients based on a given EHR has potential to improve the quality of care and create novel applications.

In Chapter 3, we define novel methods to represent EHRs as time-based prefixes and suffixes, where each prefix and suffix is represented as a set of concepts from a medical ontology. Next, we define methods to extend semantic similarity functions from the literature to our time-based prefix and suffix representation. We then evaluate each of the similarity methods using a real dataset of EHRs.

Moreover, patients are not the only stakeholders who stand to benefit from these predictions. Clinicians and clinical researchers can also benefit from a what-if analysis based on similar patients. For example, when a doctor is answering questions for a patient or the patient's family, such an analysis may be helpful as supporting evidence. Further,

such predictions can be used in load-balancing emergency departments and optimizing the number of available staff based on demand.

Our third contribution is the analysis of the relationships between a OSN's characteristics and its content – in the context of pharmaceutical drug mentions. We analyze the impact of a given OSN's characteristic along four distinguishing dimensions: (1) general vs health-specific OSNs; (2) OSN moderation rules; (3) OSN registration requirements; and (4) OSNs with a question and answer format. Healthcare providers may use our analysis to pick the right OSNs or to advise patients regarding their information needs. Our analysis is also useful to future researchers of OSNs who may find our results informative while choosing OSNs as data sources.

In Chapter 4, we present a novel pipeline for mining this OSN data, where we extend existing data mining techniques towards health-specific OSNs. We evaluate our pipeline and present our results on 10 separate OSNs. We synthesize our results into actionable items for both healthcare providers and future researchers of healthcare discussions on OSNs.

Our last contribution is a data-driven analysis to discover how the quality indicators of individual healthcare providers are associated with a rich set of attributes found in publicly available datasets. We consider referral frequency and peer awards as quality indicators of healthcare providers, and a plethora of attributes, including years of experience, medical school, patient reviews, hospital affiliations, and technology usage. We carried out this analysis using a combination of publicly available datasets and provider rating websites. We present our data-driven analysis and results in Chapter 5.

Several studies have performed qualitative analyses of provider quality, yet none have performed a data-driven, quantitative analysis of provider attributes. Hence research is lacking on the association between information from provider rating websites and publicly available data. This leaves several data-driven questions unanswered, such as which attributes determine a peer-nominated award, and do these attributes also correlate with attributes that determine a provider's referral frequency?

Chapter 6 summarizes and integrates the main findings presented in this dissertation dissertation.

Chapter 2. Efficient Concept-Based Document Ranking

Summary: Recently, there is increased interest in searching and computing the similarity between Electronic Medical Records (EMRs). A unique characteristic of EMRs is that they consist of ontological concepts derived from biomedical ontologies such as UMLS or SNOMEDCT. Medical researchers have found that it is effective to search and find similar EMRs using their concepts, and have proposed sophisticated similarity measures. However, they have not addressed the performance and scalability challenges to support searching and computing similar EMRs using ontological concepts. In this chapter, we formally define these important problems and show that they pose unique algorithmic challenges due to the nature of the search and similarity semantics and the multi-level relationships between the concepts. In particular, the similarity between two EMRs is a function of the minimum semantic distance from each concept of one document to a concept of the other and vice versa. We present an efficient algorithm to compute the similarity between two EMRs. Then, we propose an early-termination algorithm to search for the top-k most relevant EMRs to a set of concepts, and to find the top-k most similar EMRs to a given EMR. We experimentally evaluate the performance and scalability of our methods on a large real EMR data set.

2.1. Introduction

Adoption and usage of Electronic Medical Records (EMRs) has become commonplace in healthcare organizations. An EMR contains systematic documentation of health care delivered to a patient over a period of time. Each medical record includes a variety of information recorded by health care providers, such as progress notes, lab results, discharge

summaries, medication, problem lists etc. Figure 2.1 shows an excerpt of a clinical note describing a patient visit. A large part of an EMR is free text that contains numerous medical terms. In an effort to standardize EMRs many ontologies have been developed that describe medical concepts and their associations, like MeSH, RxNorm and SNOMED-CT. Links between a term that appears in an EMR and ontological concepts can be created using structured data entry tools [26] or by parsing the text of clinical notes using NLP tools like cTAKES [27] or MetaMap [28].

"Patient here for follow up ***diabetes*** *care. Computer print out of* ***blood sugar*** *shows average of 201 with 1.7 tests. There is* ***hypoglycemia*** *about 2-3 times a week. Current Medications:* ***- CELLCEPT 500MG po twice daily - FROSEMIDE 80MG po daily"***

Figure 2.1. Excerpt of a clinical note

Several types of relationships exist between these terms that are captured in the ontology structure. For example, in Figure 2.2 representing a small part of the SNOMED-CT ontology, a "heart valve finding" is a type of "cardiac finding". Some medical terms can also be synonymous, e.g. "heart attack" and "myocardial infraction" represent the same ontology concept. Previous studies [14, 29] have shown that leveraging these concept associations can significantly improve the effectiveness of free-text search on EMRs. For instance, consider the query "aortic valve stenosis". Intuitively, documents that do not contain the actual query terms, but contain similar concepts such as "thrombosis", "embolus" or slightly more general ones such as "heart disease" or "heart valve finding" can be considered as relevant to the query. Thereby, documents are routinely viewed in medical literature as sets of concepts [11, 14, 29-32] and several sophisticated measures have been proposed to quantify concept-concept similarity [25, 33-35].

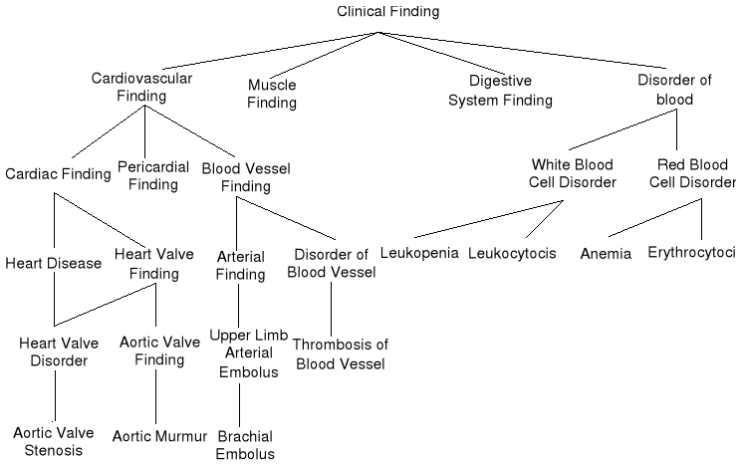


Figure 2.2. A subgraph of the SNOMED-CT ontology

Concept-based similarity search has proved to be beneficial in other domains as well. Lord et al. [7] compare genes and proteins based on the similarity of their functions that is captured in the Gene Ontology (GO), rather than based on their sequence similarity. The similarity between two gene representations can be used in order to predict gene functions or protein interactions [34].

Therefore, in this chapter we adopt the view of a document as a set of ontological concepts, as proposed in the biomedical literature, although we do recognize that also considering the free text that is not associated with concepts has the potential to further improve the retrieval quality. We study two query types: *relevance* and *similarity queries*, which are the most frequent in practice. As an example of a relevance query, consider a clinical researcher searching an EMR database for patients that qualify to participate in a clinical trial for a new breast cancer treatment. Specific symptoms and past treatments for breast cancer, which can be represented as a set of concepts, may qualify the patient for the trial. Thus, the researcher wishes to find the most relevant patient records with respect to a set of medical concepts. As an example of a similarity query, a physician who wishes to be

assisted in finding the right medical treatment for a patient can search a database of EMRs for patients with similar clinical indicators such as vital signs or medical history. Patient similarity assessment is also a very important task in the context of patient cohort identification for comparative effectiveness studies [15]. The key difference between the two aforementioned query types is that for relevance queries, we are not concerned with the concepts of the returned EMRs that are not related to the query concepts. In contrast, for similarity queries we care about the two-directional similarity between the query EMR and result EMRs.

In a relevance query (hereafter termed RDS for Relevant Document Search), the user specifies a query that consists of a set of concepts with the goal to retrieve the k most relevant documents (e.g., EMRs). In a similarity query (termed SDS for Similar Document Search), the user inputs a query document d_q with the goal to retrieve the k most similar ones. In order to evaluate each query type, we derive a ranking of documents that depends on the similarity between the query concepts (or query document respectively) with documents in the collection. This distance similarity is a function of the similarities of individual concepts. According to previous studies [11, 34], complex distance metrics do not clearly improve the correlation with the results provided by domain experts, whereas they affect efficiency [36]. Therefore, we adopt a simple distance metric represented as the shortest path connecting two concepts [37]. Further, for measuring the distance between two EMRs we use the document-document similarity measure proposed by Melton et al. [11], where the similarity between two documents is a function of the minimum semantic distance from each concept of the one document to a concept of the other and vice versa.

Each ontology may contain thousands of concepts that can be associated with several paths. For instance, SNOMED-CT ontology has a size of 300K concepts with up to 29 paths per concept; the UMLS metathesaurus contains over 2.9 million concepts. Thus, efficiency and scalability challenges arise. However, according to a recent work [36], and to the best of our knowledge, previous works on similarity metrics have serious limitations in terms of performance. Motivated by this, in this chapter we take a first step towards improving the efficiency and scalability of concept-based retrieval.

A baseline method is to precompute the distance of each concept with all documents in the collection and build an inverted file with a distance-based sorting for each postings list. After that, a top-k algorithm (e.g., Threshold Algorithm (TA) [38]) can be applied to find the k documents with the minimum distances. Although building such an index offline is feasible, applying TA for SDS queries is very inefficient. Due to the dual nature of the document-document distance, whenever TA examines a document, the postings lists for each concept in that document also need to be accessed and the distances from the query document determined. Since each posting list provides sequential access, in the worst case we would have to access all documents in each list (refer to Section 2.4.1 for details).

In an effort to address these shortcomings, we propose a uniform methodology to evaluate both RDS and SDS queries. Our query evaluation technique consists of two parts: (i) calculating query-document distances using the distances of the concepts they contain, and (ii) ranking documents based on their distance from the query. For the first part we propose an algorithm termed DRC that reduces the cost of query-document distance calculation from $O(n^2)$ to $O(nlogn)$ where n is the number of concepts in the query or examined

document. Using a variation of the Radix Tree, our algorithm constructs a subgraph of the ontology that contains only the query and document concepts and uses this index to efficiently calculate the query-document distance.

For ranking documents, we present an algorithm for retrieving the k most relevant (resp. similar) documents, following a parallel branch and bound traversal of the ontology starting from the query concept nodes. Our processing strategy balances the costs associated with distance calculation and graph traversal by probing the DRC algorithm for a document only if it is highly likely that it will belong to the query results. For this purpose, we propose an error estimation function for the semantic distances. Essentially, the error estimation measures how close the currently calculated distance of a document is to its actual distance based on the subset of query nodes already "covered" by the document. Thus, we avoid examining documents with large error estimate, such as those that "cover" only a few query concepts (recall that each document might contain hundreds or thousands of concepts in total).

An additional advantage of our method is that it does not require any distances precomputation. Our algorithm can integrate new documents into its computation on-the-fly; i.e., when a new patient arrives at the point-of-care, we can instantly add his or her EMR to our database. In contrast, TA would have to update every concept inverted index with the distance from the newly added EMR.

Contributions: The contributions of chapter are as follows:

- We define two important and challenging types of queries on concept-rich document corpuses, i.e., relevance and similarity.

- Based on a variation of a Radix tree, we propose an algorithm to reduce the cost of evaluating document-document distances from $O(n^2)$ to $O(nlogn)$.
- We present a threshold-based algorithm for efficiently identifying the k most relevant/similar documents.
- We provide a thorough experimental evaluation of our methods on a real EMR database. Our results show that our algorithms significantly outperform baseline strategies in terms of performance and scalability.

Outline: The rest of this chapter is structured as follows. Section 2.2 gives a review of the related work and Section 2.3 provides some technical background and defines the semantic distances and semantic similarity queries. In Section 2.4, we present an algorithm for efficiently calculating document-query and document-document distances. We employ these distance methods for our algorithm presented in Section 2.5, which is used to evaluate both query types. Section 2.6 reports an experimental evaluation of our methods and Section 2.7 concludes the paper and discusses future work.

2.2. Related Work

Fernandez et al. [36] provide a comprehensive survey of semantic based searching approaches that have been proposed in the past. One classification of these approaches is based on the query model followed; some approaches utilize structured ontology query languages such as SPARQL, whereas others assume a keyword searching paradigm. In an effort to combine the flexibility of keyword search with the expressiveness of structured queries, Pound et al. [39] propose a hybrid approach where keyword queries are disambiguated to structured queries based on the vocabulary of the knowledge base. Our

approach falls into the keyword search category. Additionally, we also address the case of semantic similarity queries where the input is a document instead of a set of keywords.

For keyword-based searching, ontology-based query expansion techniques have proved very beneficial for improving the retrieval quality [40]. For instance, Matos et al. [41] follow a concept- oriented query expansion methodology to search biomedical publications by expanding gene concepts related to the query with related concepts such as protein and pathway names. Likewise, query expansion techniques have been applied by Lu et al. [42] on the PubMed database to significantly improve the results' precision.

In order to address some of the arising performance challenges, [43] and [44] propose to index together terms that appear frequently in common in user queries. Their approach requires additional space and does not consider the semantic distance between concepts, thus it cannot be used to rank documents based on their distances from the query terms, which is very useful if an ontology is available for the domain. Ding et al. [45] studied index optimization by grouping terms that appear in the subtree of a taxonomy. Concept-instance relationships were used to apply query substitutions, e.g., the query term "pet" may be replaced by "cat" or "dog". Compared to this chapter our focus is on query evaluation, rather than index maintenance. Further, our methods are not limited to concept-instance taxonomies but can be used in DAGs in general.

XOntoRank [46] considers keyword search against a corpus of XML documents with ontological references. XOntoRank returns subtrees that (i) either contain or (ii) are associated with the query terms through the ontological references. XOntoRank will not return any partial matches and it cannot be used on "bi-directional" distance functions such

as the one proposed by Melton et al. [11]. Tao et al. address the problem of finding nearest neighbors in XML trees [47]. Given a query node q and a keyword w, a nearest keyword (NK) query returns the node that is the nearest to q among all nodes associated with w. The authors present an indexing scheme that allows answering NK queries efficiently. However, in our scenario the query keywords are not known apriori. Further, the proposed method cannot be applied for document-document similarity queries where bidirectional distance metrics apply.

In order to measure the semantic distance between ontology concepts, several metrics have been proposed; these metrics have been reviewed thoroughly [25, 33-35]. In [25] semantic measures are generally categorized as either: (i) *structured-based* or (ii) *information content-based*. Structured-based metrics exploit the geometrical structure of the ontology, such as the length of the shortest path connecting two concepts [37], or the depth of the concepts in the hierarchy [48], etc. Information content-based approaches capture the amount of information content shared by two concepts. Information content depends on the probability of occurrence of any descendant node of c [49]; i.e., it is proportional to the size of c's subtree including c. Resnik [49] and Lin [50] proposed different distances that measure the information content of the least common ancestor (LCA) of two nodes compared with the information content fully associated with the individual concepts. According to previous user studies with domain experts [11, 34], complicated distance metrics do not clearly improve the retrieval effectiveness, therefore in this chapter we adopt the shortest path distance metric as proposed by [37] for measuring concept-concept distance and the similarity metric proposed by [11] as a measure of similarity between

documents that contain ontological concepts, since it has been shown to be effective for medical records.

2.3. Preliminaries

2.3.1 Ontologies and Radix Trees

Concept Ontology. Let $\mathcal{D}$ be a document corpus, where each document consists of terms derived from a vocabulary V. Let $\mathcal{C} \subseteq \mathcal{V}$ be the set of terms that are mapped to concepts derived from an ontology, where each $c_i \in \mathcal{C}$ is associated either with a single term or with several terms (synonyms) from $\mathcal{V}$.

In this chapter we will focus on domain ontologies that describe concept hierarchies, which is the type of ontology typically found in the medical domain. For instance, MeSH descriptors are organized in a hierarchical structure that allows searching at various levels of specificity, whereas the Gene Ontology is a Directed Acyclic Graph (DAG). In general, a concept hierarchy is represented as a Directed Acyclic Graph (DAG) $G = \{C, E\}$, where C is the set of nodes representing concepts and E is a set of edges between concepts representing relationships such as *is-a*, *part-of*, etc.

In Figure 2.3 every path from the root to a concept $c_i \in \mathcal{C}$ is encoded using the Dewey Decimal Coding. Dewey is a prefix-based scheme where if a node c_j is a child of c_i and $l\{c_i\}$ is the label of a path from the root to c_i, then the path label of c_j is $l\{c_i\}.j$, where $j \in \{1, 2, \dots, |children(c_i)|\}$.

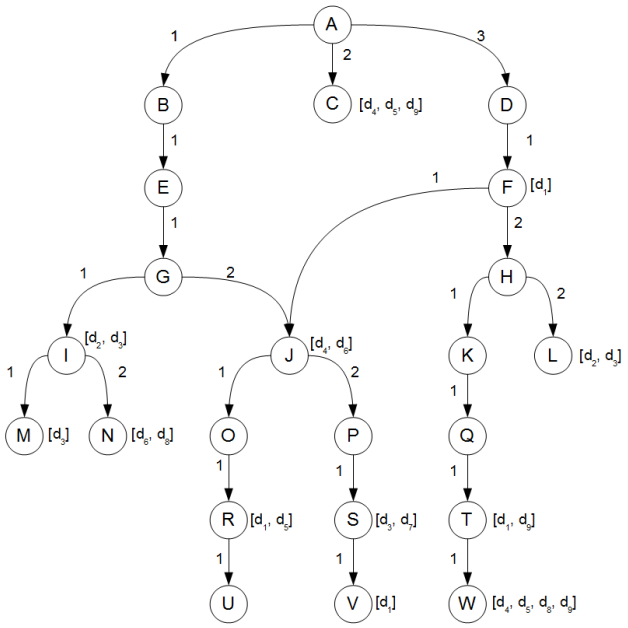


Figure 2.3. A labeled DAG representing an ontology

Radix Trees. A trie index is a data structure used to store strings, where each path represents a unique string. In order to reduce the space consumption of tries, various techniques have been proposed including path compression or adaptive indexing of the internal nodes of the trie [51]. In case of path compression, nodes with only one child can be merged with their child, yielding a space-optimized index known as a Patricia or Radix Tree. In this chapter we use a Radix index to represent a document as a set of concepts. Since our ontology is a DAG, each concept can be associated with several paths, therefore our index is not a tree but a DAG. The Radix DAG maintains the set of path labels to each concept in the document. Note that we only merge children that represent a concept in the document with parents that do not represent any concept in the document. Figure 4 shows $,T,V\}$ using the ontology from Figure 3. The concepts contained in the document are denoted with squares. Nodes B, E, G, and J have been merged into one node with edge

label 1.1.1.2. In Section 4, we describe a variation of the Radix DAG to speed up the calculation of distances between nodes in the ontology.

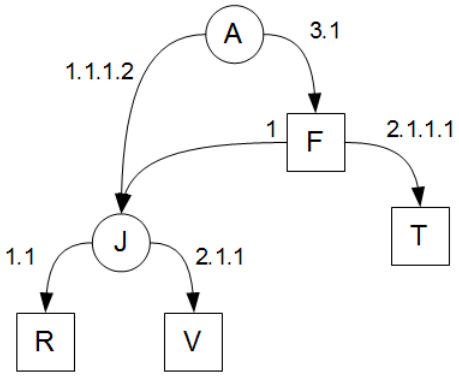


Figure 2.4. Indexing $d = \{F, R, T, V\}$ using a Radix DAG

2.3.2. Semantic Distances

Let the semantic distance between $c_i, c_j \in \mathcal{C}$ be defined as $D(c_i, c_j)$. In this chapter we focus on the case where the semantic distance between two concepts is their shortest path distance, as proposed in [37] and evaluated on medical records in [25] . Note that we consider a path as valid, only if it passes through a common ancestor of c_i, c_j. For instance, the shortest path distance $D(G, F)$ is not 2 but 5 because it has to pass through one of their common ancestors, A.

Next, we build on the concept-concept distance definition and define document-concept, document-query and document-document distances. First, we define the distance between a document $d \in \mathcal{D}$ and a concept $c \in \mathcal{C}$ as $D_{dc}(d, c)$[1]. $D_{dc}(d, c)$ is equal to the distance of c from the nearest concept in C that is associated with d:

[1] 1 dc is used to denote that function D_{dc} measures document-concept distance as opposed to concept-concept distance for D, and is not related to variables d and c.

Equation 2.1

$$D_{dc}(d, c) = \min_{ci \in d} D(c_i, c)$$

Given a query consisting of a set of concepts $q = \{q_1, \ldots, q_n\}$, define the distance of a document d from the query q as $D_{dq}(d, q)$[23]:

Equation 2.2

$$D_{dq}(d, q) = \sum_{i=1}^{n} D_{dc}(d, q_i)$$

We define the semantic distance between two documents as $D_{dd}(d_1, d_2)$. For this purpose we adopted the symmetric interpatient distance function as proposed by Melton et al. [11], where we assumed that all concepts have equal weights. Thus, computing $D_{dd}(d_1, d_2)$ requires two calculations: one for deriving d_2 starting from d_1, and another for deriving $d_1$starting from d_2; i.e., we calculate the distance of any concept in $d_1$from the nearest concept found in $d_2$and vice-versa, while normalizing by the number of concepts in the document:

Equation 2.3

$$D_{dd}(d1, d_2) = \frac{\sum_{c_i \in d_1} D_{dc}(d_2, c_i)}{|C_1|} + \frac{\sum_{c_j \in d_2} D_{dc}(d_1, c_j)}{|C_2|}$$

where $|C_1|$, $|C_2|$ represent the number of concepts in documents d_1 and $d_2$respectively. Note that unlike the document-query distance (Equation 2.2), Equation 2.3 is symmetric.

2.3.3. Similarity Queries

Now we introduce two important queries that arise when searching on a collection of documents that contain concepts derived from a domain ontology:

---

[2] dq denotes document-query distance. 3

[3] When merging the distances (scores) of documents produced by multiple queries (i.e. in query expansion) $D_{dq}(d, q_i)$ needs to be normalized with the size of q_i.

Definition 2.1 (Relevant Document Search - RDS). *Given a set of query concepts* $q = \{q_1, \dots, q_n\}$*, a document collection* $\mathcal{D}$ *and a positive integer* k*, determine the set* $\mathcal{D}' \subset \mathcal{D}$*, such that* $|\mathcal{D}'| = k$ *and* $\forall d' \in \mathcal{D}', d \in \mathcal{D} - \mathcal{D}', D_{dq}(d', q) \leq D_{dq}(d, q)$.

Definition 2.2 (Similar Document Search - SDS). *Given a query document* d_q*, a document collection* $\mathcal{D}$ *and a positive integer* k*, determine the set* $\mathcal{D}' \subset \mathcal{D}$*, such that* $|\mathcal{D}'| = k$ *and* $\forall d' \in \mathcal{D}', d \in \mathcal{D} - \mathcal{D}', D_{dd}(d', d_q) \leq D_{dd}(d, d_q)$.

As mentioned in the introduction, RDS are suitable for exploratory queries, where the user is looking for documents relevant to a set of concepts. Recall the clinical researcher seeking qualifying candidates for a clinical trial. In this case, it is not important if a patient's record contains additional concepts not specified in the query, as long as the patient record is associated with some of the query concepts. On the other hand, SDS are appropriate for patient similarity queries, which have an inherent symmetric property.

2.4. Distance Calculation Algorithm

We now discuss document-query and document-document distance calculation. In Section 2.4.1 we describe the limitations of the baseline methods; in Section 2.4.2 we present a data structure, termed D-Radix, which we use in Section 2.4.3, where we propose our algorithm for calculating distances between documents efficiently.

2.4.1. Baseline Strategies

One approach for calculating document-query and document-document distances is to precompute all pairwise concept-concept distances. The space required to maintain these distances would be $O(|C|^2)$. Even if it were possible to build this index, at query time, for each examined document we have to select the concepts with the minimum distances and

calculate the distances based on Equations 2.2 or 2.3. Assuming n_q, n_d concepts in the query and the document, we must calculate $O(n_q n_d)$ distances for each examined document. Unfortunately, a typical EMR may contain thousands of concepts; in this case the naïve approach is not an option.

Another baseline method is to calculate offline the minimum distance of each concept from all documents in the collection based on Equation 2.1, which would require $O(|\mathcal{D}||\mathcal{C}|)$ space, where $|\mathcal{D}|$is the size of the collection; $|\mathcal{D}|$ can be in the millions and $|\mathcal{C}|$ is 2.9 million for the UMLS metathesaurus. Then we could build a postings list for each concept by sorting the (doc_id, distance) pairs in ascending order. After that, we could apply the threshold algorithm [38] to find the k documents with the minimum distances for the RDS query type. However, applying a threshold algorithm for SDS queries pose several challenges. First, due to the dual nature of the document-document distance, whenever the threshold algorithm examines a document, the postings lists for each concept contained in that document also need to be accessed, and the distances from the query document determined. Since the postings lists provide sequential access, in the worst case for each list we should access $O(\mathcal{D})$ elements (documents). Further, the query document itself may contain thousands of concepts, thus we would have to traverse thousands of lists in parallel and maintain intermediate results in memory. Even worse, the lower bound threshold used by TA would assume that a partially examined document does not contain any concept other than those found so far, which does not allow for effective pruning in practice for the SDS query case.

2.4.2. The D-Radix Index

In order to address the scalability shortcomings of the baseline methods, in Section 2.4.3 we propose a more efficient algorithm for computing document-query and document-document distances. In contrast with baseline methods, our method does not require any precomputation of distances. Distance calculation is conducted at query time by utilizing a variation of the Radix that we introduce, termed D-Radix DAG (Distance-Radix DAG). Given a document d and a query q, a D-Radix DAG indexes all concepts that exist in either d or q. Additionally, each node contains the node's distance from the nearest node in d and q respectively. More formally:

Definition 2.3. *Given two sets of concepts* d *and* q*, a D-Radix DAG* $T_{d,q}$ *is a DAG* $G(C[D_{dc}(d, c_i), D_{dc}(q, c_i)], E)$ *where there is a node* $c_i \in C$ *for every common prefix found in* $c \in d \cup q$ *and if* $\exists! e\{c_j, c_k\}$ *and* $c_j, c_k \notin d \cup q$ *then* c_j, c_k *are merged into* c_i. $D_{dc}(d, c_i)$ *and* $Ddc(q, ci)$ *represent the distances of node ci from the nearest* $c_d \in d$ *and* $c_q \in q$ *respectively, as given in Equation 2.1*.

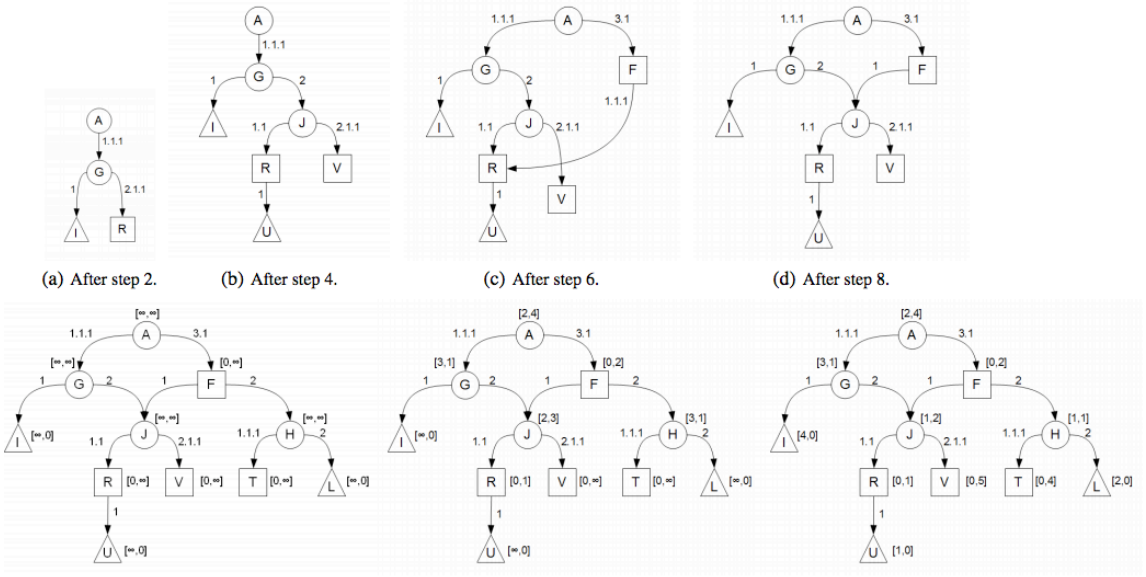


Figure 2.5. Running example of the DRC algorithm

Example 2.1. *Figure 5(g) shows an example of a D-Radix DAG for a document* $d = \{F, R, T, V\}$. *Document and query concepts are represented with squares and triangles respectively. Each node is associated with two numbers: the first number is the distance from the nearest document concept, and the second one is the distance from the nearest query concept.* ■

Assuming that we have such an index structure available, then in the case of an RDS query, in order to calculate $D_{dq}(d, q)$ we can apply Equation 2.2 using the nearest document distance attached to each of the query nodes. Distances from the nearest query nodes are ignored. Hence, we get $D_{dq}(d, q) = D_{dc}(d, I) + D_{dc}(d, L) + D_{dc}(d, U) = 4 + 2 + 1 = 7$. Similarly, in the case of a SDS query with $d_q = \{I, L, U\}$, we can calculate $D_{dd}(d, dq)$ based on Equation 2.3 where we use the distances from the nearest document node attached to each of the query nodes and the distances from the nearest query node attached to each of the document nodes.

Apart from having two distances associated with each node, a significant difference of the D-Radix index compared to the Radix Tree is that in a D-Radix two concept-nodes are not merged, even if there is no branch in any of the two nodes. In particular, we only merge children that represent a concept in the document or query with parents that do not represent any concept in the document or query. For instance, in a Radix Tree nodes R and U would have been merged; in the D-Radix they are kept separate.

2.4.3. The DRC Algorithm

DRC Overview. The DRC (D-Radix Construction) algorithm consists of a construction and a tuning phase. The construction phase builds a D-Radix DAG for indexing query and document concepts. All shortest distances are initially set to ∞ with one exception; if the inserted node is a document concept then the shortest distance from the document is set to 0, whereas if it is a query concept then the shortest distance from the query is set to 0. Once the index has been constructed, DRC propagates the shortest distance information by executing a bottom-up traversal followed by a top-down traversal. The distance information for a node is updated based on the minimum of (i) its distance, (ii) and the distance from its children or parents plus the length of the edge. We show that DRC calculates query-document and document-document distances in $O((|P_q| + |P_d|)\, log(|P_q| + |P_d|))$ time, where P_q and P_d represent the number of paths leading to concepts from the query and the document respectively.

D-Radix DAG Construction. Constructing a D-Radix DAG is quite more complex compared to the construction of a Radix tree. The main reason is that, since we have to build a DAG rather than a tree, each step involves the *insertion of both a node and a path*

to that node. Each inserted path has to be matched with edges that already appear in the index. Further, each partial match (path address) has to be checked against the set of nodes already inserted to the index, since it may define an alternative path to such a node. In that case, the insertion algorithm has to avoid adding a path twice, such that duplicate paths will not be propagated to the subtree. For the same reason, an already inserted edge may be split. We examine some of these cases based on a running example presented next. The details of inserting a path address are explained in the next paragraph. Algorithm 1 shows the complete pseudocode of DRC for the RDS query case. The SDS case is similar except that (i) we use the distances from both document and query and (ii) the distance is calculated based on Equation 2.3.

Path Insertion. Insertion for the D-Radix DAG is similar to that of Radix trees, except that a path may define a node already contained in$T_{d,q}$. $T_{d,q}$ is a hash of nodes, where each node contains zero or more pointers to other nodes in the hash; these pointers represent child edges. $T_{d,q}$ also contains a pointer to the root node, which is created during initialization of Algorithm 1. Thus, the insertion algorithm starts at the root and traverses $T_{d,q}$ until all pointers have been updated correctly. Pseudocode for path insertion is given in Function InsertPath.

Algorithm 2.1. DRC Algorithm for RDS Queries

Input: d: a document, q: a query
Output: D_{dq}: document-query distance
Variables: $T_{d,q}$: a D-Radix DAG on d, q,
P_d: lexicographically sorted list of Dewey addresses for each $c_j \in d$,
P_q: lexicographically sorted list of Dewey addresses for each $q_i \in q$,
$l\{n_d\}$: next Dewey address from P_d, $l\{n_q\}$: next Dewey address from P_q

```
1  begin
2      //Index Construction Phase;
3      retrieve P_d; retrieve and sort P_q;
4      insert(root) into T_{d,q};
5      l{n_d} := P_d.first; l{n_q} := P_q.first;
6      while (P_d.hasNext or P_q.hasNext) do
7          if (l{n_d} ≤ l{n_q}) then
8              c_n := n_d;
9              InsertPath(l{n_d}, T_{d,q});
10             l{n_d} := P_d.next;
11         else
12             c_n = n_q;
13             InsertPath(l{n_q}, T_{d,q});
14             l{n_q} := P_q.next;
15         if c_n ∈ q then
16             D_q(q, c_n) := 0;
17         else
18             D_q(q, c_n) := ∞;
19     //Tuning Phase;
20     //Traverse T_{d,q} bottom-up;
21     foreach c_j ∈ T_{d,q} do
22         D_q(d, c_j) := min{D_q(d, c_j), min_{c_k}{D_q(d, c_k)+D(c_j, c_k)}};
23         for all c_k where c_k is a child of c_j;
24     //Traverse T_{d,q} top-down;
25     foreach c_j ∈ T_{d,q} do
26         D_q(d, c_j) := min{D_q(d, c_j), min_{c_k}{D_q(d, c_k)+D(c_k, c_j)}};
27         for all c_k where c_k is a parent of c_j;
28         if (c_j ∈ q) then
29             D_dq(d, q) := D_dq(d, q) + D_query(d, c_j);
30     return D_dq(d, q);
```

Let n_i be the node to be inserted with path address $l\{n_i\}$. Execution begins by initializing the variables: $u = \epsilon, v = l\{n_i\}$, and $c_n = T_{d,q}.root$ (lines 2-4). Function InsertPath maintains the invariant that u is a common prefix of $l\{n_i\}$, and v is the suffix of $l\{n_i\}$ not matched by u. c_n is used to keep track of the current node in the traversal (line 6); u defines a path to node c_n in $T_{d,q}$.

While variable v is not equal to ϵ, $l\{n_i\}$ has not been fully inserted into $T_{d,q}$ (line 5). Hence, we examine each child edge of c_n, seeking an edge that shares a common prefix with v (lines 5-10). Only one such edge may exist. If no such edge exists, then n_i is a child of c_n with edge label v (lines 11-13). Otherwise, v contains a prefix exactly equal to m, or v shares a prefix with m that is not equal to ϵ or m. If v contains a prefix exactly equal to m, then u, v, and c_n are updated to reflect traversal from c_n to n (lines 14-17). This is accomplished by concatenating m to u, removing the prefix of m from v, and setting $c_n = n$.

If v shares a prefix with m that is not equal to ϵ or m, then the edge between c_n and n must be modified to include the LCA of $l\{n_i\}$ and $T_{d,q}$. Thus, the child edge from c_n to m is removed (line 19). Let variable lcp be equal to the Longest Common Prefix (LCP) of v and m (line 20). The path defining the LCA is u concatenated with lcp (line 21). Once we have the Dewey address of the LCA, we look up its corresponding node identifier (line 22). Next, we add a child edge from c_n to the LCA with lcp as the edge label (line 23). Then an edge is added from the LCA to node n with edge label $m.substring(lcp.length + 1)$ (line 24); plus one removes the "." trailing the lcp string. If the LCA is not equal to n_i,

then we also add a child edge from the LCA to ni with edge label $v.substring(lcp.length + 1)$ (lines 25-26); again, plus one removes the "." trailing the lcp string.

Example 2.2. *Consider* $d = \{F, R, T, V\}$ *from Figure 2.3 and* $q = \{I, L, U\}$. P_d *and* P_q *are listed in Table 2.1. First, DRC retrieves the lists and inserts a root node into* $T_{d,q}$. *In the first step, DRC processes I with address 1.1.1.1. A node for I is created along with an edge from A to I. Next, DRC processes R with address 1.1.1.2.1.1. After matching 1.1.1.2.1.1 with 1.1.1.1, DRC splits edge 1.1.1.1 into 1.1.1 (the common prefix) and 1. Thus, DRC will also insert node G with address 1.1.1, and insert the remaining path address to R as an edge 2.1.1. The resulting D-Radix DAG is shown in Figure 2.5(a). In the third step, DRC processes U with 1.1.1.2.1.1.1, which subsumes 1.1.1.2.1.1 (node R), thus it inserts an extra edge from R to U as well as node U.*

In the fourth step, DRC has to process node V with 1.1.1.2.2.1.1. This step splits the edge between G and R with 1.1.1.2 (node J) as the LCA; V (2.1.1) is added with J as the parent (Figure 2.5(b)). In the fifth step, node F with address 3.1 is added with the root as the parent. In the sixth step, DRC processes node R with address 3.1.1.1.1. Node R already exists in $T_{d,q}$, *but an edge between F and R is missing. Thus, DRC adds this edge as shown in Figure 2.5(c). The seventh step processes node U with address 3.1.1.1.1.1. This address is completely matched in* $T_{d,q}$, *thus* $T_{d,q}$ *is not modified. The eighth step processes node V with address 3.1.1.2.1.1. By matching this address, the DRC algorithm decides that the edge between F and R has to be split into addresses 3.1.1, 1.1 and 2.1.1. DRC performs a*

lookup and finds out that 3.1.1 corresponds to node J which already appears in $T_{d,q}$. Therefore, the edge between F and R is modified to be between F and J but no new node will be created. Also, DRC finds out that address 2.1.1 (node V) already appears in $T_{d,q}$. The result of this step is illustrated in Figure 2.5(d). The ninth step adds node T as a child of node F. Finally, in the tenth step DRC processes node L with address 3.1.2.2. By matching this address, DRC has to split the edge between F and T and insert 3.1.2 (node H) as a parent of T and L. The result of the tenth step with the initial distances assigned to each node is illustrated in Figure 2.5(e). ■

Table 2.1. Dewey path address lists for DRC

Node	Labels	Step #
P_d		
R	1.1.1.2.1.1	2
V	1.1.1.2.2.1.1	4
F	3.1	5
R	3.1.1.1.1	6
V	3.1.1.2.1.1	8
T	3.1.2.1.1.1	9
P_q		
I	1.1.1.1	1
U	1.1.1.2.1.1.1	3
U	3.1.1.1.1.1	7
L	3.1.2.2	10

Distances Tuning. Obtaining the shortest distance for each node requires a bottom-up traversal followed by a top-down traversal. Let D_q symbolize the distance from the nearest query node; then the distance at each node is recursively updated as:

Equation 2.4.

$$D_q(d, c_j) = min\{D_q(d, c_j), \min_{c_k}\{D_{q(d,c_k)} + D(c_k, c_j)\}\}$$

$\forall c_k$ where c_k is a child of c_j for the bottom up traversal, and a parent of c_j for the top-down traversal. A similar formula is used for computing distances from document nodes for an SDS query.

Correctness. Let u, v be two nodes of $T_{d,q}$. One of the following conditions holds: (i) u is a descendant of v, or (ii) u is an ancestor of v, or (iii) u and v share a common ancestor. Executing a bottom-up and a top-down traversal always propagates the correct distance information for u and v in the first two cases. Further, recall from Section 3.1 that valid paths must pass through a common ancestor. Based on the order of the traversals all distance information is only propagated along valid paths that contain the common ancestor of two nodes. Since $T_{d,q}$ has only one root, the common ancestor of any u, v is always visited. Therefore, in the third case u and v will always have the correct distance information.

Figure 2.5(e) shows the D-Radix DAG from Example 2.2 after the completion of the construction phase of DRC. The bottom-up traversal propagates these distances up to the root, as shown in Figure 2.5(f). The top-down traversal propagates these distances down to the leaves of the D-Radix DAG, as illustrated in Figure 2.5(g).

After finishing both the construction and tuning steps, the final distance is computed using Equation 2.2 or 2.3, depending on the query type. Distances are progressively calculated during the top-down traversal as each document and query node is visited.

Complexity Analysis. Let P_q and P_d be the sets of path addresses to concepts of the query and the document respectively. The D-Radix constructed by DRC will contain $O(|P_q| + |P_d|)$ nodes. The construction phase loops over each path address. Since the height of the

D-Radix index is $log(|P_q| + |P_d|)$, the construction phase takes $O((|P_q| + |P_d|)\, log(|P_q| + |P_d|))$ time. The traversals required for the tuning phase are completed in $O(|P_q| + |P_d|)$. Hence, the total complexity of DRC is $O((|P_q| + |P_d|)\, log(|P_q| + |P_d|))$.

2.5. K-Nearest Document Search Algorithm

This section presents our algorithm for evaluating RDS and SDS queries, termed kNDS (k-Nearest Document Search Algorithm). We first present the challenges that our problem poses, and a general overview of the proposed algorithm. Next, we provide details regarding the algorithm's execution.

2.5.1. Baseline Methods

A naïve approach to evaluate RDS or SDS queries is to calculate the distances of all documents in the collection from the query (or query document); and then select k with the minimum distances. Clearly, this is prohibitively expensive and inefficient. Ideally, we would prefer to maintain a sorted list of documents ordered by their semantic distances from the query, such that unexamined documents would always have a larger distance, thus we could prune those documents. However, as we discussed in Section 2.4.1, this threshold-based approach would require precomputing the distance of each document in the collection from any concept in the ontology, i.e., $O(|\mathcal{D}||\mathcal{C}|)$ space, and it would not be useful for the SDS query due to the dual nature of the semantic distance of Equation 2.3.

2.5.2. Challenges and Tradeoffs

In order to overcome this problem, we propose a solution that does not require any distance precomputation but exploits a threshold-based technique to prune irrelevant documents.

Our algorithm, termed kNDS, is based on the idea of query expansion. In particular, we start our search by considering documents that contain the exact query terms and then we follow a breadth-first traversal of the ontology graph to retrieve documents that contain similar concepts. Our goal is the following: if at some point during the graph traversal we already have found k documents with final distances (i.e., we have covered all query nodes and the total distance of each document from the query has been determined), then we can prune all documents for which we have not calculated their exact distances, as long as their lower bound is greater than of the k already examined ones. Before delving into the details of kNDS we first explain how we calculate partial and lower bound distances.

Iteration $l + 1$, $l \geq 0$ examines concepts having distance l from a query concept. Assume that during iteration $l + 1$ for the breadth-first search starting from query node q_i we traverse a concept node c_j , such that c_j is contained in document d and c_j is the first concept for document d seen for query node q_i. Then, we know that $D_{dc}(d, q_i) = D(c_j , q_i) = l$. If no concept for document d is found, then the lower bound for the distance $D_{dc}(d, q_i)$, termed $D^- d_c(d, q_i)$ is equal to $l + 1$.

Example 2.3. *Consider a query* $q = \{I, L, U\}$ *and document* $d = \{F, R, T, V\}$. *Then, starting a parallel breadth-first search from each query concept in Figure 2.3, in the second iteration we examine nodes:* G, M, N, R, H. *Only* R *is contained in* d, *thus the actual distance* $D_{dc}\{d, U\}$ *is 1. For the rest of the query nodes it holds that:* $D_{dc}^-\{d, I\} = D_{dc}^-\{d, L\} = 2$. ■

Let $M_d(q_i, d)$ be a hash that maps a node q_i to a distance value l if during the breadth-first search starting from query node q_i a concept that belongs to document d has been found

with distance l from q_i. For instance in our previous example, during the second step of the traversal, $M_d(q_i, d)$ would contain the element $\{U, 1\}$. Note that values for each key in M_d are only set once so that M_d maintains the minimum distance from each q_i. Then, we define the partial (currently calculated) $D_{dq}^{partial}(d, q)$ and the lower bound distance $D_{dq}^{-}(d, q)$ between a document and a query (for RDS) as:

Equation 2.5.

$$D_{dq}^{partial}(d, q) = \sum_{q_i \in M_d} M_d(q_i, d)$$

Equation 2.6.

$$D_{dq}^{-}(d, q) = \sum_{q_i \in M_d} M_d(q_i, d) + \sum_{q_i \notin M_d} (l + 1)$$

Let $M'_d(c_i, q)$ be a hash for document d with a value for c_i if and only if a concept for document d has been found during any of the breadth first searches for any query node in q; values for each key in M'_d are only set once, hence M'_d contains the minimum distances. Then the partial and lower bound distances for SDS are:

Equation 2.7.

$$D_{dd}^{partial}(d_1, d_2) = \frac{D_{dq}^{partial}(d_2, d_1)}{|C_1|} + \frac{\sum_{c_i \in M'_{d_2}} M'_{d_2}(c_i, d_1)}{|C_2|}$$

Equation 2.8.

$$D_{dd}^{-}(d_1, d_2) = \frac{D_{dq}^{-}(d_2, d_1)}{|C_1|} + \frac{\sum_{c_i \in M'_{d_2}} M'_{d_2}(c_i, d_1) \; + \; \sum_{c_i \notin M'_{d_2}} (l + 1)}{|C_2|}$$

kNDS proceeds in a branch-and-bound fashion. Starting from the query nodes, it performs a breadth-first traversal of the ontology, retrieves documents that contain the visited concept nodes and iteratively updates their partial distances using Equations 2.5 or 2.7, depending on the query type. Similarly, it calculates a lower bound distance based on

Equations 2.6 and 2.8. Then, we can check whether some of the documents can be pruned by comparing their lower bound distance with the partial distances of already examined ones.

The challenge is that during graph traversal, it is highly unlikely to discover documents that would cover all query nodes early during the algorithm execution, especially if the query (or query document) contains many terms. Moreover, in general we would like to avoid calling the DRC algorithm to calculate actual (final) distances because this is an expensive operation. In fact, there is a tradeoff between the distance calculation cost (DRC execution) and graph traversal. If we execute DRC too soon we may waste time to compute the distance of irrelevant documents. On the other hand, if we wait until finding several concepts of a document before running DRC, this may explode the ontology traversal cost, since non-visited nodes are kept in a priority queue in memory.

Then when would it be preferable to calculate the actual distance of a document from the query in order to prune some documents? To answer this question, kNDS algorithm maintains an error estimate that compares the partial distance of the document with the document's lower bound distance based on the following formula:

Equation 2.9.

$$\epsilon_d = 1 - \frac{D_{dq}^{partial}(d,q)}{\bar{D_{dq}}(d,q)}$$

kNDS compares the calculated error with an error threshold ϵ_θ. If the error estimate is lower (i.e., the partial distance yields quite a good estimate of the actual distance), then kNDS probes DRC in order to compute the actual distance. Otherwise, kNDS continues the graph traversal until having a better distance estimation

Note that determining a good error threshold ϵ_θ generally depends on several factors such as: (i) the query type (RDS or SDS), (ii) the query size, (iii) the ontology characteristics (fan-out, average number of paths to each concept node, etc.), and (iv) the document collection statistics (e.g., if a document contains concepts that are close to each other in the ontology, the average number of concepts per document, etc.). Thereby, we use the error threshold as an input parameter to the algorithm. We include a detailed sensitivity analysis on this parameter in the experimental section (Section 6).

2.5.3 The kNDS Algorithm

We first describe the data structures used followed by the details of the algorithm execution.

Data Structures. kNDS maintains the following data structures:

- A queue, denoted as E_c, used to perform breadth-first traversal of the ontology, where each element contains a concept node and the corresponding query node from which the traversal originated, denoted as $\{c_j , q_i\}$.
- A list of documents, denoted as L_d, where each element contains a document d and its partial and lower distances.
- A binary heap H_k of the top-k most similar documents found so far and their respective distances from the query. This heap contains documents for which their distances have been determined; it is ordered in reverse $D_{dq}(d, q)$.
- A hashset S_d of documents that have been examined.

We also assume the availability of an index that allows us to traverse the ontology efficiently (this would typically fit in memory) as well as an inverted and a forward index that map concepts to documents and vice-versa (memory or disk-based).

Algorithm Execution. The algorithm execution consists of two steps: breadth-first expansion and distance calculation. In the following we provide details for each step. The complete pseudocode of the kNDS algorithm is given in Algorithm 2.2. Additional engineering optimizations are described at the end of this section. Breadth-first Expansion. Initially all data structures are empty (line 2). E_c is initiated by inserting each $qi \in q$ into E_c (line 4). kNDS performs multiple breadth-first traversals of the ontology starting from each query node. For each node c_j in the queue, we maintain its distance to the query concept that was the source of c_j . We use this distance to compute the two document distances described above. In each iteration the following operations are conducted:

- The breadth-first traversal proceeds to the next depth level, e.g., at iteration l kNDS processes all nodes with distance (depth) l from any of the query nodes. Note that we enforce the traversal to follow only valid paths in the ontology (passing through a common ancestor), as we discussed in Section 2.3.1.
- For each traversed node c_j , the node's neighbors are inserted to E_c (lines 9-10). Ec maintains a natural ordering of elements via insertion. In order to distinguish elements that have different depths, we include a null insertion $\{\emptyset, \emptyset\}$ after finishing each iteration (lines 5 and 14). Note that a node can be visited several times during the ontology traversal. Labeling a visited node is more expensive, since it would require to maintain a large structure with all (c_j , q_i) already visited.

Algorithm 2.2. kNDS Algorithm

Input: $\mathcal{D}$: a document collection, q: a query, G: a concept ontology,
k: a positive integer, ϵ_θ: a distance error threshold, $D(c_j)$: inverted index on c_j
Output: the k most similar documents to q
Variables: E_c: nodes' queue, L_d: a list of documents,
H_k: a heap of the k most similar documents to q,
S_d: a hash of documents that have been examined,
D_k^+: the distance of the k-th element in H_k from q,
D^-: the lower bound of the distance from q of the first element in L_d

```
1  begin
2      L_d := ∅; H_k := ∅; S_d := ∅; D^- := 0; D_k^+ := ∞;
3      foreach q_i ∈ q do
4          E_c.push(q_i, q_i);
5      E_c.push(∅, ∅);
6      while (D^- < D_k^+ and E_c ≠ ∅) do
7          while (E_c.head() ≠ {∅, ∅}) do
8              E_c.pop() → {c_j, q_i};
9              foreach c_l: ∃E(c_l, c_j) ∈ G or ∃E(c_j, c_l) ∈ G do
10                 E_c.push(c_l, q_i);
11             foreach d ∈ D(c_j) and d ∉ S_d do
12                 calculate(D^-_dq(d, q));
13                 L_d.push(d, D^-_dq(d, q));
14         E_c.pop(); E_c.push(∅, ∅);
15         sort(L_d);
16         calculateError(L_d.first) → ε_d;
17         while (ε_d ≤ ε_θ and L_d ≠ ∅) do
18             L_d.removeFirst() → d;
19             calculate(D_dq(d, q));
20             S_d.push(d);
21             if |H_k| < k then
22                 H_k.push(d, D_dq(d, q));
23                 H_k.find-min() → D_k^+;
24             else if D_dq(d, q) < D_k^+ then
25                 H_k.delete-min();
26                 H_k.push(d, D_dq(d, q));
27                 H_k.find-min() → D_k^+;
28             L_d.first() → D^-;
29             calculateError(L_d.first) → ε_d;
30             foreach d_i ∈ H_k do
31                 if D_dq(d_i, q) ≤ D^- then
32                     output d_i;
33     foreach d_i ∈ H_k do
34         output d_i;
```

- For each traversed node c_j , all documents that contain c_j and have not been examined before (i.e., they are not found inS_d) are inserted to L_d (lines 11-13). If the document already exists in L_d, its lower bound distance as well as the current distance are updated (line 12). For each document, we also maintain the query nodes from which the search originated, so that we do not increase a distance if the document is associated with a second concept that originated from a covered query node.

Distance Calculation. After completing a breadth-first expansion, kNDS proceeds to analyze collected documents. First, it sorts L_d by increasing $D_{dq}^{-}(d, q)$ (line 15). Then, it calculates the estimation error (ϵ_d) for the first element (line 16). If$\epsilon_d \leq \epsilon_\theta$, where ϵ_θ is the error threshold, then the document must be analyzed, i.e., the document is removed from L_d, added to S_d and the actual distance is calculated by calling upon DRC (lines 17-20). Otherwise, kNDS proceeds to the next breadth-first iteration. Each document for which the actual distance has been determined is compared with the documents contained in a min-heap H_k, where $H_k$contains the k documents with the currently lowest actual distances. If the new document's distance is lower than the distance of the k-th element of H_k (or $|H_k| < k$), then the new document replaces the last element of H_k (or it is inserted into H_k respectively) (lines 22- 26). Documents from L_d are examined iteratively until either $L_d$is empty or $\epsilon_d < \epsilon_\theta$ (line 17) or D^- is higher than the distance of the k-th element in H_k (line 6); in the last case kNDS terminates and the contents of H_k are returned as the query results (lines 33-34).

Example 2.4. *Following Example 2.2, assume an RDS query with* $q = \{F, I\}, \theta = 1,$ *and* $k = 2$ *on the document collection and the ontology depicted in Figure 2.3. Table 2.2 shows the contents of various data structures during the execution of kNDS. Every two rows represent one iteration of the main while loop. The first row shows the contents at the start of the respective iteration and the second row shows the contents after retrieving the neighbors for each node in* E_c *and updating* L_d*. kNDS begins by adding the query nodes to* E_c*, and initializing* $D^- = 0$*,* $D_k^+ = \infty$ *(row 1). The algorithm then pushes each neighbor of F and I into* E_c*, and initializes* L_d *(row 2). The top-2 documents (*d_1 *and* d_2*) are then analyzed and added to* H_k *and* S_d*;* D^- *is set to 1 using the lower bound distance of* d_3*, and* D_k^+ *is set to 4 using the actual distance of d1 (row 3). Since* $D^- < D_k^+$*, kNDS continues to the next iteration. Next, kNDS processes* E_c *adding the respective neighbors and updates* L_d *(row 4). Note, node J has now been added twice to* E_c*; once for F and once for I. Also note that although G is a parent of J, the BFS for query node F did not push* $\{G, F\}$ *to* E_c*; this is due to the valid path rules discussed in Section 2.3.1. kNDS then examines* L_d*and sets* D^- *to 3 using the lower bound of* d_4*, and* D_k^+ *to 2 using the final distance of* d_3*. Since* $D^- \geq D_k^+$ *, kNDS terminates and outputs the contents of* H_k*as the top-2 results.* ■

Table 2.2. Running example of the kNDS algorithm

Iteration	S_d	L_d	E_c	H_k	D^-	D_k^+
0	$\emptyset$	$\emptyset$	$\{F,F\}\{I,I\}\{\emptyset,\emptyset\}$	$\emptyset$	0	∞
0	$\emptyset$	$\{d_1,1\}\{d_2,1\}\{d_3,1\}$	$\{D,F\}\{\mathrm{H},F\}\{J,F\}\{G,I\}$ $\{M,I\}\{N,I\}\{\emptyset,\emptyset\}$	$\emptyset$	0	∞
1	$\{d_1,d_2\}$	$\{d_3,1\}$	$\{D,F\}\{H,F\}\{J,F\}\{G,I\}$ $\{M,I\}\{N,I\}\{\emptyset,\emptyset\}$	$\{d_2,2\}\{d_1,4\}$	1	4
1	$\{d_1,d_2\}$	$\{d_3,2\}\{d_6,2\}\{d_4,3\}$	$\{A,F\}\{K,F\}\{L,F\}\{O,F\}$ $\{P,F\}\{E,I\}\{J,I\}\{\emptyset,\emptyset\}$	$\{d_2,2\}\{d_1,4\}$	1	4
END	$\{d_1,d_2,d_3,d_6\}$	$\{d_1,3\}$	$\{A,F\}\{K,F\}\{L,F\}\{O,F\}$ $\{P,F\}\{E,I\}\{J,I\}\{\emptyset,\emptyset\}$	$\{d_2,2\}\{d_3,2\}$	3	2

Correctness. We will show that kNDS algorithm always outputs the top-k documents with the lowest distances from the query. Any document $d \in \mathcal{D}$ can be in one of the following 3 states: (i) already examined, i.e. contained in S_d, (ii) partially visited, i.e. contained in L_d, or (iii) not visited yet. Recall that kNDS maintains a min-heap H_k with the documents found so far that have the lowest distances. Whenever the final distance of a new document is calculated (line 19), i.e. the documents moves from state (ii) to (i), if $D_{dq}(d,q) < D_k^+$ then the new document replaces the old one in H_k (lines 24-26). This step ensures that $\nexists d \in S_d - H_k : D_{dq}(d,q) < D_k^+$. Now recall that partially visited documents are kept in L_d sorted on their lower distance (lines 11-13). kNDS continues as long as the first document in L_d has $D^- < D_k^+$ (line 6). When kNDS terminates, since L_d is sorted in ascending lower distance, all documents in L_d will have $D^- > D_k^+$ so they can be safely discarded. Finally, let l be the distance of the concepts examined in the breadth-first traversal at the current iteration from q. Then for RDS it holds that $\forall d \in L_d, d' \in \mathcal{D} - \{S_d \cup \mathrm{L_d}\}, D^- \leq D_{dq}^-(d,q) \leq |Q|(l+1) \leq D_{dq}^-(d',q) \leq D_{dq}(d',q)$. A similar inequality holds also for the SDS query based on Equations 2.7 and 2.8. Therefore, when kNDS

terminates it holds $D^- \leq D_{dq}(d', q)$. In other words, any not visited document will always have a greater distance than those already examined.

Complexity Analysis. The worst case for the cost of kNDS happens when the number of iterations (line 6) is maximized or all documents in the corpus have to be examined (each document's distance is computed). Each iteration performs a breadth-first step, so the maximum number of iterations is equal to the longest path in the ontology L. Normally $|\mathcal{D}| > L$. Further, based on our analysis in Section 2.4.3, each distance calculation has a $O((|P_q| + |P_d|)\, log(|P_q| + |P_d|))$ cost where P_q and P_d represent the sets of path addresses to concepts of the query and the document respectively. Therefore, assuming $|\mathcal{D}|$ iterations in the worst case the complexity of kNDS will be $O(|\mathcal{D}|(|P_q| + |P_d|)\, log(|P_q| + |P_d|))$. Note that the cost for the heap reorganization in each iteration (line 15) is dominated by the cost of the distance calculation, since in practice the number of documents kept in the heap is $|\mathcal{D}'| \ll |\mathcal{D}|$.

Optimizations. In order to speed up the algorithm execution we also apply the following optimizations:

- When updating the distances of a document in L_d, if the calculated lower distance grows larger than that of the k-th element in H_k, then the document is removed from L_d.
- Since the size of L_d might grow large, instead of sorting L_d after each iteration we build a partial sorted heap H_d that contains $n \geq k + 1$ documents ordered by $D^-_{dq}(d, q)$. The reason for enforcing $n \geq k + 1$ is that in the most favorable scenario, the first k elements in the heap will be the final query results. In that case,

we need to know the lower bound distance of the next element in order to check the termination condition.

- As we discussed before, for each document d we maintain the number of distinct query concepts or their neighbors for which we have found that d is associated with. If all query nodes are found already then we can use the current distance instead of applying the DRC algorithm.
- kNDS can progressively output results from H_k during the algorithm execution. If the distance of a document d in H_k is lower than or equal to the lower bound distance of the first element in L_d (or H_d), then d must be in the top-k most similar documents and can be reported as a query result (lines 30-32).

2.6. Experimental Evaluation

2.6.1. Experimental Setting

Dataset. Experiments were conducted using a subset of the MIMIC II clinical database [6]. This subset consists of 42,144 clinical notes over 983 patients. There are four different types of notes available for each patient: (i) MD Notes (816 documents), (ii) Nursing Notes (28,133 documents), (iii) Radiology Reports (12,373 documents), and (iv) Discharge Summaries (822 documents).

For our experiments, we used two different document collections. Our purpose was to examine the performance of our methods on data sets with different characteristics in terms of size, average number of concepts contained per document, total number of distinct concepts in the collection, etc. The first document collection that we used consists of the Radiology Reports documents; we refer to this corpus as RADIO. For the second collection

we constructed a patient records corpus. For this purpose, we treated all clinical notes associated with a patient as a single document. Since the new document includes all different types of notes, it contains more concepts and these concepts are more densely distributed in the ontology. On the other hand, RADIO contains fewer concepts per document and it less cohesive. Table 2.3 reports some statistics for the two document collections used in the experiments.

Table 2.3. Document Corpus Statistics

	Patient	**Radio**
Total Documents	983	12,373
Total Concepts	16,811	8,629
Avg. Tokens / Document	8,148	273.7
Avg. Concepts/Document	706.6	125.3

We used the SNOMED-CT ontology where we considered only edges that represent *is-a* relationships. In total, there are 296,433 concepts. Each node has an average of 4.53 children. On average there are 9.78 path addresses per concept with length equal to 14.1.

In order to link the medical documents with the SNOMED-CT ontology we applied the following procedure. First, we analyzed each document in order to identify and expand abbreviations based on a public list of medical abbreviations. Next, we used the MetaMap tool [28] in order to identify UMLS concepts associated with terms in the clinical notes. We indexed only UMLS concepts that correspond to SNOMED-CT concepts. Negation of concepts was identified using MetaMap as well. According to domain experts, negated concepts are not relevant when measuring inter-patient similarity [14]. Therefore, we only consider concepts with positive polarity; e.g., we exclude concepts contained in phrases such as "absence of bradycardia". We have built an index of the ontology, an inverted index on concepts and a forward index to map documents to concepts. Depending on the

collection and ontology sizes and memory availability, the indexes can be memory or disk-based. In our experiments the inverted and forward indexes where loaded into a MySQL database for indexing, thus we will also include performance analysis that measures the database access times.

Experimental setup. All experiments were carried out on an Intel i3 2.1 GHz CPU with 6 GB RAM running Windows 7 and MySQL 5.2.4. All algorithms were implemented in Java 7 with a 4 GB heap and a 64-bit JVM. In order to avoid memory overflow when inserting too many elements into the nodes' queue during a breadth-first expansion step, we set a maximum queue size of 50K elements. Whenever the size of the queue reaches this limit, the graph traversal halts and kNDS is forced to examine the collected set of documents. In practice, the queue size limit can be eliminated by implementing kNDS as a MapReduce job. Each mapper would be responsible for one iteration of the BFS traversal starting from one query node; reducers would do the book-keeping and execute the distance calculation algorithm, if needed.

Parameters. Table 2.4 describes the parameters under investigation; default values are shown in bold. For each experiment, we vary each parameter while keeping the rest in their default values. Additionally, we set a depth and a collection frequency (cf) threshold such that we exclude generic or very common concepts (such as "disease" or "blood" respectively). For depth threshold we used a default value of 4, i.e., we excluded all concepts in a depth level that is lower than 4. This includes over 99% of the concepts. We found that the number of concepts filtered by the cf threshold depends on the distribution of the dataset. Therefore, we used $\mu + \sigma$ as the default cf threshold for each dataset, where

μ is the estimated mean and σ is the estimated standard deviation; $\mu + \sigma$ includes about 92% of the concepts. In order to examine the statistical significance of our results, we ran a two-tailed t-test for the times reported in Figure 2.9 with two sample variances and found out that the execution times measured are statistically significant with p-value < 0.001.

Table 2.4. Values for parameters; default values shown in bold

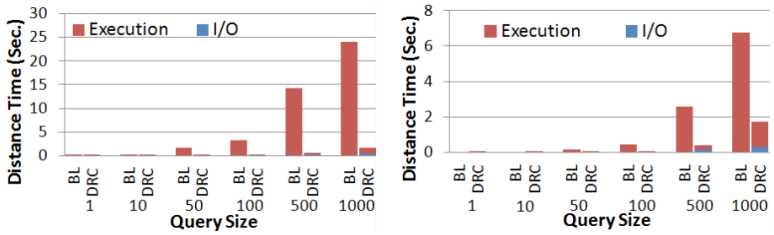
Parameter	Range
Number of Results (k)	3,5,**10**,50,100
Query Size (Q)	1,**3**,5,10

2.6.2. Experimental Results

Previous works [14] [11] have studied the effectiveness of the distance metrics that we have used, hence our experiments will focus on efficiency. Our goal is to examine the performance of distance calculation separately from the document search algorithm. Thus we conducted two experiments: (i) the first one evaluates the performance on different algorithms for calculating document-document distances, (ii) the second one measures the benefit from our pruning strategy on query evaluation. We discuss which algorithms we compare at the beginning of the respective experiment.

Distance Calculation Experiments. The goal of the first experiment is to measure the scalability of the distance calculation methods against the query size, i.e., the number of concepts in the query document. As we discussed in Section 2.4.1, building a matrix for all concept-concept distances would impose a large space requirement. Thus, in order to have a fair comparison, we compared two methods that do not require index maintenance, i.e., DRC against a baseline that calculates the document to document distances at the query time by computing the respective minimum concept distances. Our experiments examine the scalability of the two methods when varying the query size n_q over a workload of 5000

randomly generated query documents with n_q concepts each. Figure 2.6 shows the average time required by the baseline (BL) and the DRC algorithm for the two document collections that we examined. As expected, in all experiments when the query size grows larger, the time required by the baseline methods grows quadratically. In contrast, DRC algorithm takes less than two seconds in the worst case, and grows with $nlogn$ rate as shown in Section 2.4.3 (n_q is proportional to P_q).



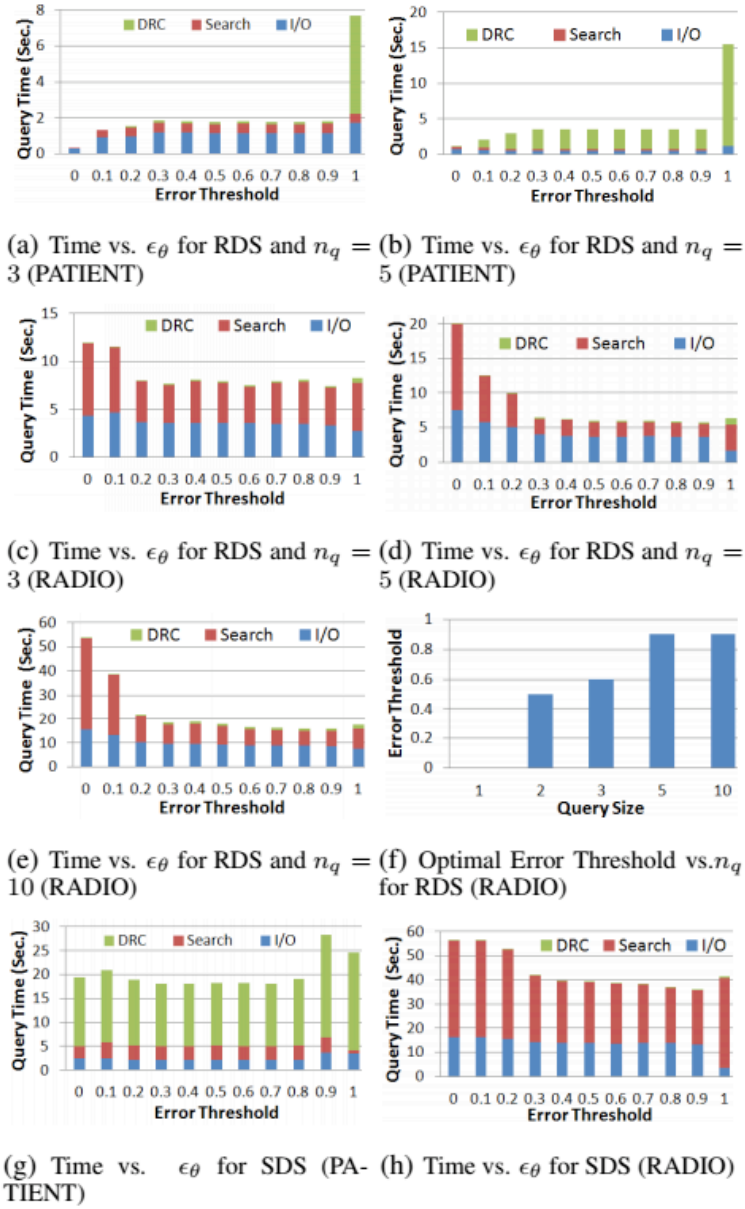
(a) Time vs. n_q for SDS (PATIENT) (b) Time vs. n_q for SDS (RADIO)

Figure 2.6. Distance calculation time vs. query size n_q

Document Ranking Experiments. This experiment compares kNDS against a baseline method that does not apply any pruning of documents. In order to isolate the performance gains achieved because of the documents pruning that kNDS applies, *we used the DRC algorithm as the distance calculation component for both kNDS and the baseline method*. Note that we did not consider a TA [38] variation as a competitor algorithm since it is impractical for the SDS query due to the problems that we discussed in Section 2.4.1. We conducted experiments for both RDS and SDS queries. All query experiments measure the average times taken over 100 randomly generated queries; in the case of SDS, documents were randomly picked from the corpus. Each experiment measures user time spent for distance calculations using DRC, ontology traversal time (applies only for kNDS) and the I/O time of each algorithm.

Sensitivity Analysis vs. Error Threshold. In the first set of experiments, we conduct a sensitivity analysis vs. the error threshold that is used an input parameter input of the kNDS algorithm. The examined range of values covers two extreme variations of kNDS; $\epsilon_\theta = 0$ represents a strategy where the algorithm waits until having visited all concepts of a document, i.e. it will calculate an exact distance for this document. On the other hand, when $\epsilon_\theta = 1$, then kNDS would directly calculate the actual distance of a document the first time it visits any concept node linked with the document.

We first examine the performance of kNDS for different values of ϵ_θ when varying the query size for the RDS query type. Plots 2.7(a)-2.7(b) show the measured times for the PATIENT collection. An interesting observation is that in this setting the optimal value for ϵ_θ is always 0, i.e., the best strategy is to find all query nodes before examining a document. The reason is that the PATIENT collection contains many concepts that are very close to each other. Thereby, it is highly likely that another document that contains a neighbor node may belong to the query results instead. Thus, in most of the queries, it is more efficient to wait until finding all query nodes in a document. Another important factor is that because of the large number of concepts contained in each document, the DRC calculation part is considerably expensive and dominates the total time for larger query sizes. This is another reason to avoid redundant distance calculations as much as possible.



(a) Time vs. ϵ_θ for RDS and $n_q = 3$ (PATIENT)

(b) Time vs. ϵ_θ for RDS and $n_q = 5$ (PATIENT)

(c) Time vs. ϵ_θ for RDS and $n_q = 3$ (RADIO)

(d) Time vs. ϵ_θ for RDS and $n_q = 5$ (RADIO)

(e) Time vs. ϵ_θ for RDS and $n_q = 10$ (RADIO)

(f) Optimal Error Threshold vs.n_q for RDS (RADIO)

(g) Time vs. ϵ_θ for SDS (PATIENT)

(h) Time vs. ϵ_θ for SDS (RADIO)

Figure 2.7. Query time vs distance error threshold ϵ_θ

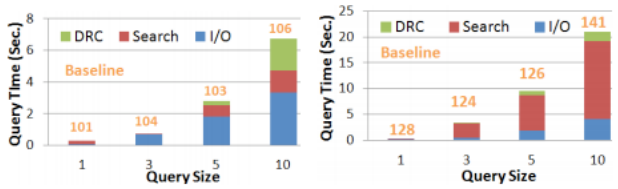
Plots 2.7(c)-2.7(e) show the results for the RADIO collection. In contrast with PATIENT documents, we notice that in this case the query times are highly dependent on the error threshold and they are generally lower for larger thresholds. Further, the distance calculation cost is rather small. The reason is that RADIO documents contain fewer concepts. These concepts are generally sparsely distributed in the ontology graph. Thus, it

is sufficient to find some documents that contain only a small subset of the query concepts in order to probe the distance calculation. Since the distance calculation is not expensive, making a false judgment does not affect the performance. As expected, the best error threshold is larger for larger query sizes (plot 2.7(f)) requiring less query nodes to be found before calculating distances. Plots 2.7(g) and 2.7(h) plot the query times measured for various error thresholds for SDS query.

Regardless of the error threshold used, kNDS outperforms the baseline algorithm, where the baseline times are shown in plots 2.9(a)-2.9(d). However, the results of the above analysis allow us to find a good setting for the error threshold. In the following experiments we set the default error thresholds for the PATIENT and RADIO collections to 0.5 and 0.9 respectively. The percentage of examined documents (i.e. documents for which DRC was probed) that were eventually part of the top-k query results justifies our settings for the error threshold parameter. Specifically, for RDS in the PATIENT dataset, 99% of the documents for which the actual distance was calculated were returned in the top-k results. For SDS queries over 60% of the examined documents were reported as results; this percentage could be improved by increasing the node queue limit that may cause excessive calls to DRC.

Scalability vs. Query Size. Next we varied the query size nq and measured the execution time needed by the baseline and kNDS on a workload of RDS queries. Results are depicted in plots 2.8(a)-2.8(b). As expected, processing times increase with the size of n_q with rate roughly nlogn, which supports the complexity analysis in Section 2.5.3 (n_q is proportional to P_q). Note that lower query sizes cause fewer calls to DRC so kNDS can often terminate

before exceeding the queue limit. In all settings kNDS is the most efficient algorithm with a large performance gain over the baseline.

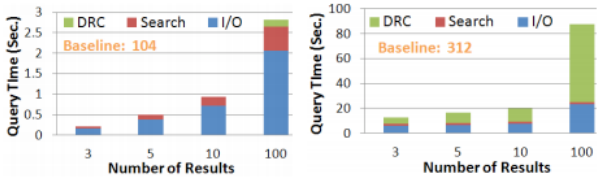


(a)Time vs. n_q for RDS (PATIENT) (b) Time vs. n_q for RDS (RADIO)

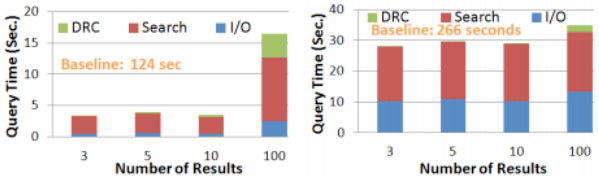
Figure 2.8. Query time vs. query size n_q

Performance Analysis vs. Number of Results. Finally, we examined the behavior of the algorithms for evaluating RDS and SDS queries when varying the number of results k. Plots 2.9(a)-2.9(d) show the results for the two document collections used. The baseline algorithm has to calculate the distances for all documents in the collection; thus its performance is independent from k whereas kNDS uses a termination condition in order to prune some documents. In all experiments, kNDS outperforms the baseline method with a broad margin. For example, for the default setting where $k = 10$ in the PATIENT collection, kNDS takes less than 1 sec to run, whereas the baseline method takes 104 secs. The performance gains of kNDS are more significant in SDS, e.g., for $k = 10$, kNDS is 99% faster. Again notice that for the PATIENT collection, most of the processing time is used for distance calculation; this is due to the large number of concepts contained in each patient record.

Finally, as shown in the plots, the performance of kNDS is not affected significantly by k. For instance, for $k = 100$ and a SDS query, the kNDS algorithm is 89% faster than the baseline.



(a) Time vs. k for RDS (PATIENT) (b) Time vs. k for SDS (PATIENT)



(c) Time vs. k for RDS (RADIO) (d) Time vs. k for SDS (RADIO)

Figure 2.9. Query time vs. number of results k

2.7. Conclusion

In this chapter we studied two important and challenging types of queries arising when searching over concept-rich document collections, i.e., relevance and similarity queries. Such queries are frequently encountered in Electronic Medical Record (EMR) systems. We proposed an algorithm that reduces the cost of query evaluation from $O(n^2)$ to $O(nlogn)$ by using a variation of the Radix Tree. We presented an efficient early-termination algorithm to search for the top-k most relevant/similar documents that avoids redundant distance calculations following a branch and bound approach. We experimentally evaluated our algorithms against baseline strategies on real clinical data and we showcased the advantages of our methods in terms of efficiency and scalability

Chapter 3. Predicting Future Medical Concepts in Electronic Health Records

Summary: Medicine 2.0 creates the need for applications that find similar patients based on a patient's electronic health record (EHR). We evaluate the hypothesis that we can leverage similar EHRs to predict possible future medical concepts (e.g. disorders) in a patient's EHR. We represent patients' EHRs using time-based prefixes and suffixes, where each prefix or suffix is a set of medical concepts from a medical ontology. We compare the prefixes of other patients in the collection with the state of the current patient using various inter-patient distance measures. The set of similar prefixes yields a set of suffixes, which we use to determine probable future concepts for the current patient's EHR. We evaluated our methods on the MIMIC II dataset of patients, where we achieved precision up to 63% and recall up to 62%. Our results show that this is a promising direction of research.

3.1. Introduction

Medicine 2.0 – the intersection of Web 2.0 and healthcare services, applications, and tools – brings new opportunities for patients to actively contribute to their own care [52, 53]. Allowing users to find similar patients based on their electronic health record (EHR) has the potential to improve the quality of care and expand options for healthcare solutions [54]. This approach may lead to novel applications for patients such as self-management recommendations based on big data aggregation across cohorts [55]. Applications that allow patients to find, discuss, and share health data and information can improve patient outcomes while raising meaningful discussions in disease management [23, 56]. Therefore, finding patients with similar experiences and conditions is a critical step for

patients to contribute to their own care. This capability is becoming more important as more patient records are available, for instance through health social networks that aim to connect patients, which drive the need for patient-centered health informatics [57].

We evaluate the *hypothesis* that we can predict possible future medical concepts in a patient's EHR, by leveraging the EHRs of other patients in the collection. For that, we first use various inter-patient similarity measures to locate other EHRs that have time-based prefixes similar to the current patient's EHR. Then, we process the time-based suffixes of the matched EHRs to determine which medical concepts are probable for the future of the current patient's EHR.

We represent each patient as a set of medical concepts from SNOMED-CT (Systemized Nomenclature Of MEDical Clinical Terms) [9]. We extract medical concepts using the MetaMap library [28]. Then, to identify similar patients, we adopt various distance functions studied in the literature [11, 14]. We show how to extend these distance functions to predict future medical concepts given a query patient. We demonstrate and evaluate these methods on the MIMIC II Clinical Database, which contains patient data from visits to an intensive care unit (ICU) [6]. We present a detailed evaluation of the accuracy of our techniques for various confidence parameters. Our results show that this is a promising approach to predict possible future concepts in a patient's EHR.

While we use the MIMIC II database to evaluate our methods, our methods are applicable to any database of EHRs where a rich set of medical concepts can be extracted for various time instances (e.g., hospital visits) during a patient's care. For example, we can apply our framework to BioMed Central's Cases Database based on medical case reports [58]. The

cases database records the evaluation and progression of a patient's medical history, and is also rich in textual information that describes medical concepts associated with each case. Patients are not the only stakeholders who stand to benefit from the prediction of future medical concepts in an EHR; clinicians and clinical researchers can also benefit from a what-if analysis based on similar patients. For example, when a doctor is answering questions for a patient or the patient's family, such analysis may be helpful as supporting evidence. Moreover, the clinician may view the changes in the probable future EHR of a patient if a specific therapy is undertaken. From a research standpoint, clinical researchers may be interested in finding patients with similar predicted concepts when performing non-randomized studies, e.g., for matching cases and controls.

3.2. Background

We leverage previous work, which has studied several distance functions for finding similar patients. Methods include a bag-of-words approach, information content, path length in ontology, common ancestors, and combinations thereof. To the best of our knowledge, none of these methods has been applied on prefixes of EHRs or for the purpose of predicting future medical concepts in EHRs.

Cao et al. used case-based reasoning to find similar patients based on clinical text [12]. They found that medical concepts are superior features versus a bag-of-words approach, which is an approach we also adopt. Similar to this chapter, they restricted medical concepts to a specific subset of semantic types. However, they did not consider semantic similarity between concepts – e.g., two concepts may be neighbors in the SNOMED-CT ontology – when comparing patients.

Plaza et al. looked at concept graphs for measuring inter-patient similarity [14]. Given a set of concepts for a patient, all ancestors of each concept are retrieved and assigned a weight based on their depth, where deeper concepts have higher weights. This method is investigated and explained in greater detail in our methods described below.

Melton et al. investigated approaches including a bag-of-concepts and average path length [11]. They explored weighted and un-weighted path lengths. Paths were weighted based on information content, common descendants, and information content of descendants. Both the bag-of-concepts and the un-weighted average path lengths are investigated and described in greater detail in our methods below.

Works on aggregating patient data for analytics employ a patient database in order to provide recommendations, analysis, and/or predictions. Gotz and others at IBM have developed an interactive system to aid domain experts in retrospective patient cohort analysis [59-61]. Similar to our work, their system finds a cohort of similar patients based on the EHR of the physician's current patient via symptoms. Statistics for the cohort are aggregated and visualized using a variety of techniques, including an outflow graph that models the evolution of symptoms over time and the respective outcomes. Unlike our work, they do not predict future medical concepts nor do they use ontologies when measuring similar patients. However, their work complements our work in that the user can use predicted symptoms to explore possible outcomes in the outflow graph.

Roitman et al. at IBM Research developed a system that allows users to perform an exploratory search over social-medical data [62]. Data includes EHRs and social data such as treating physicians and family members. Users build a query based on attributes (e.g.

Abilify 20 mg), and/or relations (e.g. relType:PatientMed). Facets are provided to refine the results. Unlike our work, IBM's system does not predict future medical concepts nor does it find similar patients via ontologies.

PatientsLikeMe has also examined the effects of aggregating patient data [56, 63]. An online survey found that users reported several benefits from having access to aggregated patient statistics. Further, they found a correlation between perceived benefit and the number of features used by a user. Our work aims at increasing the value of such aggregated data by predicting possible concepts.

3.3. Methods

We apply our framework to the MIMIC II clinical database, which contains patient data collected from multiple ICUs from a medical center in Boston over a seven-year period [6]. Several types of data are collected during a visit, including radiology reports, and nursing and physician notes. We parse each note to extract medical concepts from the text, as discussed below. Each note is associated with a timestamp that represents its creation time. We use these timestamps to map notes to events – i.e., patient transfers, as explained below – generating a list of concepts for each event. An overview of this process is shown in Figure 3.1.

First we parsed medical concepts from each type of note using the MetaMap library [28]. This library maps free text to biomedical concepts in the Unified Medical Language System (UMLS) [8]. Each concept in the UMLS corresponds to one or more semantic types.[64] Previous work has shown that diseases, symptoms, procedures, body parts, and medications are the most important UMLS semantic types for the purpose of measuring

patients' similarity[14]. Negated concepts are identified and ignored as previous work has shown that absent concepts are not relevant to patient similarity [14]. We adopt this approach and discard all concepts not belonging to any of these types. After obtaining a list of relevant concepts, each concept from the UMLS is converted to a concept from SNOMED-CT using the MRCONSO table [65].

A single patient visit may consist of several transfers between wards, ICUs, radiology, and other care units. Each of these transfers is considered to be a *census event* in the MIMIC II database. An example visit is shown in Figure 3.1. The rationale for this definition of an event, which we also adopt in this chapter, is that each time a patient enters a new care unit there may be a significant change in the patient's status, e.g., the patient's condition worsened and he was transferred to the surgical ICU. For instance, if Bob was admitted to the surgical ICU, then transferred to the regular ward, and then transferred back to the surgical ICU again, his visit would consist of three events e_1, e_2, and e_3. Each of these events would be associated with a set of concepts.

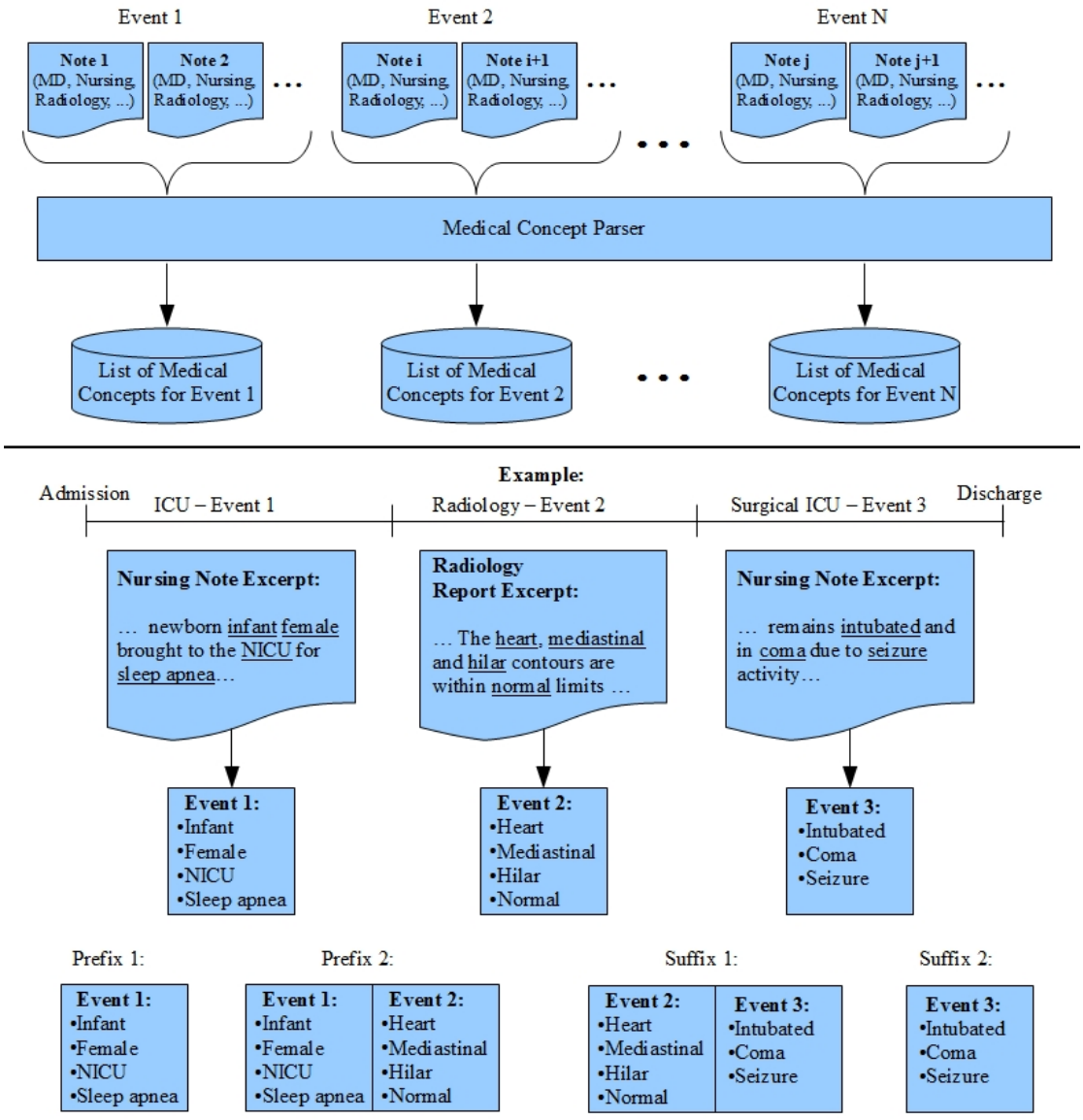
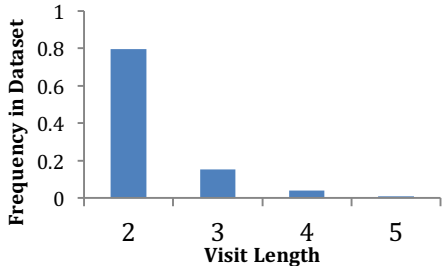


Figure 3.1. An overview of our system, where a patient visit is split to a list of events and each event is associated with a set of concepts. Time is represented along the horizontal axis. The notes are parsed using MetaMap to generate sets of medical concepts that are associated with each event. These sets are then used to generate prefixes and suffixes for each visit. In the example, the patient was admitted to the ICU, transferred to radiology, and then sent to the surgical ICU. Since this example contains three events, there are two possible prefixes and suffixes.

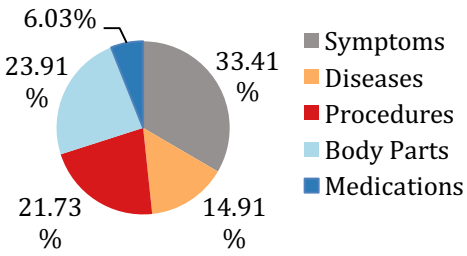
If a patient visits a hospital multiple times, then each visit is treated independently, that is, viewed as a different patient for the purpose of our similarity matching algorithm. This decision is not critical for the MIMIC II dataset, since most patients only have one visit as

discussed below. It has been shown that the above concept of census events provides an effective coarse timeline of a patient's EHR, where concepts within an event are semantically associated to each [66].

We generated some descriptive statistics of the MIMIC II database using the event sequences. We only consider admissions (visits) with more than one event, because we cannot define prefixes or suffixes for visits with one event. In total there are 1418 visits for 1369 unique patients. Figure 3.2(a) shows the distribution of the number of events per visit. The average visit contained 2.3 events, with a standard deviation of 0.007; a majority of visits contain two events. Figure 3.2(b) shows the distribution of semantic types in this dataset. Symptoms, body parts, and procedures dominate the dataset. Figure 3.3(a) shows the average number of concepts at each event position, and Figure 3.3(b) shows the distribution of the total number of unique concepts in a visit by visit length (i.e. number of events in visit).

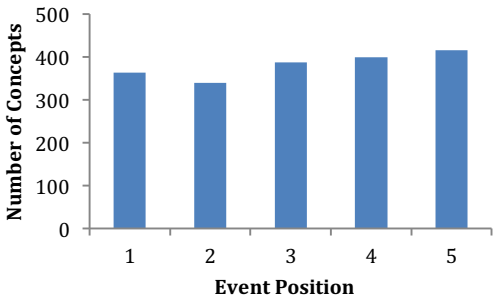


(a) Distribution of visits by the number of events. The average length is 2.3 events with a standard deviation of 0.007; About 80% of the visits contain two events.

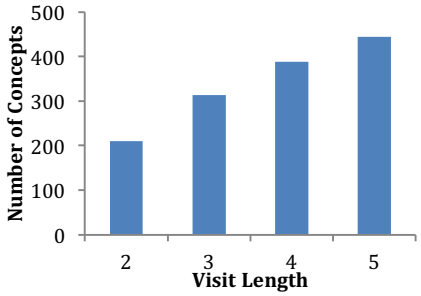


(b) Distribution of the semantic types for concepts from the MIMIC II database. Symptoms, body parts, and procedures are the semantic type for over 75% of concepts.

Figure 3.2. General statistics for the MIMIC II database. These statistics are calculated over 1418 visits for 1369 unique patients. (a) shows the distribution of the number of events per visit and (b) shows the distributino of semantic types across all visists. A majority of visits contain 2 events and concepts are dominated by symptoms, body parts, and procedures.



(a) Number of concepts per event position. The number of concepts usually grows as the event position increases.



(b) Number of unique concepts by visit length. The number of unique concepts contained in a visit grow as the length of the visit increases.

Figure 3.3. Statistics for the number of concepts by event position and visit length. (a) shows the average number of concepts at each event position and (b) shows the average number of unique concepts by visit length. An interesting observation is that longer visits contain more concepts.

Given a query EHR, we compare each prefix EHR in our dataset to find the most similar ones. We consider the following types of dissimilarity functions:

1. Bag-of-Concepts (*BoC*)
2. Common Ancestors (*CA*)
3. Average Path Length (*APL*)

Let A and B be the sets of concepts representing two EHRs that we wish to compare.

In the *BoC* approach, the dissimilarity between A and B is defined as the sum of the number of concepts that appear in A but not in B or in B but not in A, divided by the size of their union[11]; note the union of A and B is also a set, and therefore the size of the union only considers each concept once:

Equation 3.1.

$$BoC(A,B) = \frac{|A \setminus B| + |B \setminus A|}{|A \cup B|}$$

This dissimilarity function takes values between 0 and 1, where 0 represents maximum similarity and 1 represents minimum similarity. Note that *BoC* is symmetric, that is, $BoC(A,B) = BoC(B,A)$.

In the *CA* approach, for each concept in A we retrieve all ancestor concepts in the SNOMED-CT concept hierarchy (we only consider *is-a* links) and assign to each concept and its ancestors a weight as described next [14]. For each concept $c_a \in A$, let α be the set of all ancestors of c_a including c_a. For each $c_b \in \alpha$, define β as the set of all of c_b's ancestors along with c_b. The weight of c_b is defined as the number of ancestors in common with c_a, divided by the total number of distinct ancestors for c_a and c_b, i.e., $w(c_b)_{w.r.t.\, c_a} = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$. The intuition is that c_a is assigned a weight of 1, and each ancestor is assigned a weight relative to its distance from c_a. For each α derived from A, $w(c_b)_{w.r.t.\, c_a}$ is computed for every concept in α. Weights are averaged if a node obtains more than one weight, defined as $w(c_b)$.

Let A^α be the set of all concepts that are ancestors of at least one concept in A, and B^α be the set of all concepts that are ancestors of at least one concept in B. When computing the dissimilarity from A to B, we examine each concept in A^α and check if it exists in B^α. If it exists, then the concept in A^α is assigned a value equal to its own weight, and zero otherwise. Once the weights are updated, they are then summed and normalized by the maximum similarity possible for A [14]:

Equation 3.2.

$$CA(A,B) = 1 - \frac{\sum_{c_i \in A \cup A^\alpha} c_i \cdot \begin{cases} w(c_i) & if\ c_i \in B^\alpha \\ 0 & otherwise \end{cases}}{\sum_{c_i \in A \cup A^\alpha} w(c_i)}$$

In order to make *CA* a measure of dissimilarity, we subtract the score from 1, yielding a score where 0 represents maximum similarity and 1 represents minimum similarity. Note that by this definition *CA* is not symmetric, but we could extend this dissimilarity function to be symmetric by summing both A to B and from B to A, and use $CA(A,B) + CA(B,A)$. However, we only investigate non-symmetric *CA* in this chapter.

The *APL* measure uses the SNOMED-CT ontology to find the most semantically similar concept in B for each concept in A. In particular, the semantic dissimilarity between two concepts in the SNOMED-CT ontology is the minimum number of *is-a* links that we must traverse to reach the one from the other. For instance, the path length between *Malignant tumor of breast* and *Malignant tumor of prostate* is four because they both share the grandparent *Malignant neoplastic disease*, under the SNOMED-CT hierarchy. We sum the distances across all concepts in A to obtain the dissimilarity of A to B [11]:

Equation 3.3.

$$APL(A,B) = \frac{\sum_{a \in A} minPath(a,B)}{|A|}$$

where $|A|$ is the number of concepts in A. A score of zero implies maximum similarity.

APL by definition is not symmetric; symmetric *APL* (referred as *APL_SYM*) is the sum of A to B and B to A. That is:

Equation 3.4.

$$APL\_SYM = APL(A,B) + APL(B,A)$$

Given a query EHR, we find similar prefixes of other EHRs in the database, and then aggregate the suffixes of these matched EHRs to generate the predicted concepts. First we compute the dissimilarity between the query EHR and EHR prefixes (prefixes of the events sequence of an EHR). Let Q_k^p be a query EHR (simply referred as query) represented as a

set of concepts obtained from the first k events (a prefix as denoted by the p superscript) of a patient visit. Note that in a clinical setting, we would use the whole patient EHR as Q_k^p, since we want to predict future concepts given the current state of the patient. Let D be the database of patient visit EHRs. The objective is to find a set of EHRs in D whose dissimilarity with respect to Q_k^p is less than some *dissimilarity threshold* τ:

Equation 3.5.

$$DisSim_{\{P_i \in D\}}(P_i, Q_k^p) < \tau$$

where P_i is a prefix of events for a single visit and $DisSim$ is a dissimilarity function, which can be any of the aforementioned dissimilarity functions.

Each visit in the database can be split to a prefix P_i and a suffix S_i in different ways for each choice of prefix length. If multiple prefixes match for the same visit, we only consider the one with the lowest dissimilarity. Given a query Q_k^p, the suffix space $S(Q_k^p)$, or simply S, is the union of the suffixes of the (visits with prefixes) similar to Q_k^p prefixes:

Equation 3.6.

$$S = \cup_{\{S_i \in D\}} S_i \mid DisSim(P_i, Q_k^p) < \tau$$

For each query prefix, Q_k^p, we define the query suffix Q_{k+1}^s, which represents the visit events starting from the *k+1st* until the last, to be the ground truth of the query. Of course, when patient events are evaluated prospectively, Q_{k+1}^s is unknown. Then precision is equal to the size of the intersection of Q_{k+1}^s with S divided by the size of S, whereas recall is the same intersection divided by the size of the query:

Equation 3.7.

$$precision = \frac{|Q_{k+1}^s \cap S|}{|S|} \qquad recall = \frac{|Q_{k+1}^s \cap S|}{|Q_{k+1}^s|}$$

Unique precision and recall are defined as the precision and recall of concepts that do not exist in the prefix Q_k^p. The rationale is that we may be less interested to discover possible future concepts that already in the query EHR.

As mentioned above, a first parameter is the dissimilarity threshold τ. A second parameter is needed to exclude from S concepts with low confidence, that is, concepts that appear in few suffixes. We introduce parameter P which is the probability threshold for concepts in S. The probability of a concept c is $p(c) = \frac{Number\ of\ matched\ suffixes\ that\ contain\ c}{Number\ of\ matched\ suffixes}$. We only include concepts in S with $p(c) > P$. We study the role of these parameters in the precision and recall in our experiments.

3.4. Results

3.4.1. Anecdotal Example of Predicting Future Medical Concepts

We start with a short real anonymized example from the MIMIC II dataset, to demonstrate the potential utility of our approach. Bob was involved in a motor vehicle collision where he struck his head and lost consciousness. He arrived at the ICU with a chief complaint of severe shoulder pain and bleeding from his nostrils. After arriving at the ICU (event 1), Bob is transferred to the SICU for further tests (event 2). During his stay at the SICU, the staff observes symptoms of pneumonia and pulmonary aspirations. Radiology tests reveal that Bob indeed has both pneumonia and pulmonary aspirations.

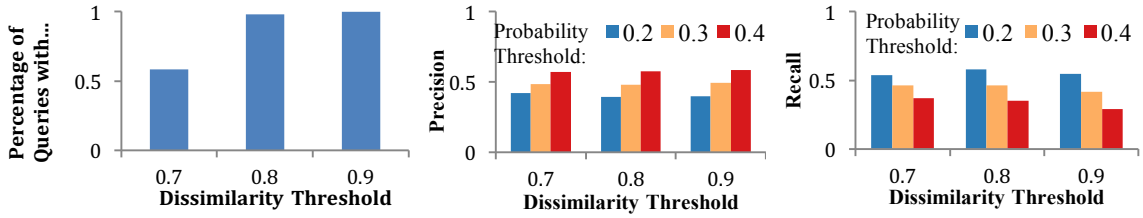
We execute our prediction method using event 1 as query. In particular, we use CA with $\tau = 0.5$ and $P = 0.3$. Out of the similar patient suffixes, 50% contain concepts pneumonia and pulmonary aspiration; these percentages are substantially higher than the prior probabilities (probabilities in whole database), which are 31% for pneumonia and 24% for

pulmonary aspiration. Another interesting concept not found in Bob's prefix, but found in his suffix is diabetes mellitus type 2; 75% of the similar suffixes contained this concept versus a prior of 59%.

3.4.2. Detailed Results

Next, we present our results for each of the aforementioned dissimilarity functions. Our query set consists of 114 visits out of the 1418 total visits with at least two events, which were selected randomly. From these 114 visits, we generate 152 queries, where each query is a prefix of a visit such that the suffix only contains a single event.

For each dissimilarity function, we first measure the percentage of queries (EHRs) for which our method predicts at least one medical concept against the threshold τ; then we measure precision and recall against both τ and P, shown in Figures 3.4, 3.5, 3.6, and 3.7 for *BoC*, *CA*, *APL*, and *APL_SYM* respectively. In each of these cases, there is naturally a trade-off between precision and recall – a higher precision degrades the quality of the recall. We observe that the precision is around 0.5, which we believe is still useful, especially given the relatively small size of the database, as we explain in the Discussion section below.

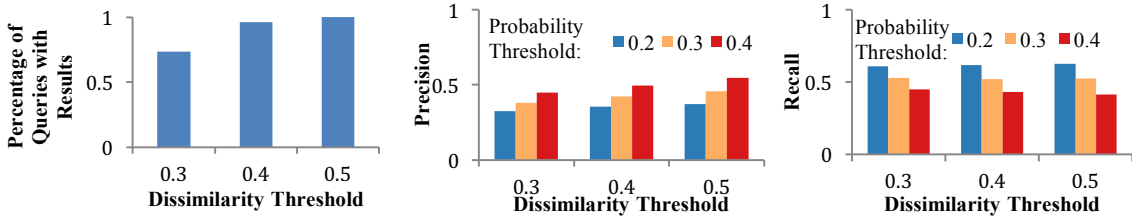


(a) The percentage of queries that contain a result versus the dissimilarity threshold τ.

(b) The precision of BoC versus the dissimilarity and probability thresholds.

(c) The recall of BoC versus the dissimilarity and probability thresholds.

Figure 3.4. Results for BoC. The percentage of queries with a result is shown in (a), the precision in (b), and the recall in (c).

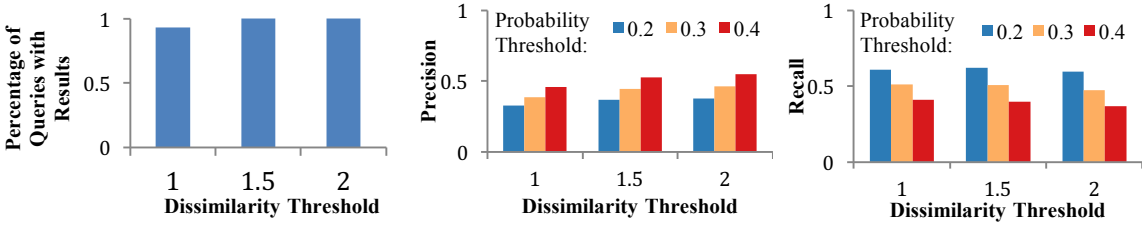


(a) The percentage of queries that contain a result for CA.

(b) The precision of CA versus the dissimilarity and probability thresholds.

(c) The recall of CA versus the dissimilarity and probability thresholds.

Figure 3.5. Results for CA. The percentage of queries with a result is shown in (a), the precision in (b), and the recall in (c).

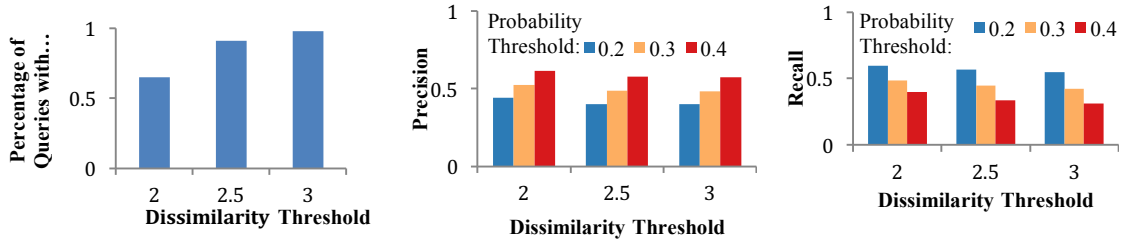


(a) Percentage of queries that contain a result for APL.

(b) Precision of APL versus the dissimilarity and probability thresholds.

(c) Recall of APL versus the dissimilarity and probability thresholds.

Figure 3.6. Results for APL. The percentage of queries with a result is shown in (a), the precision in (b), and the recall in (c).



(a) Percentage of queries that contain a result for APL_SYM.

(b) Precision of APL_SYM versus the dissimilarity and probability thresholds.

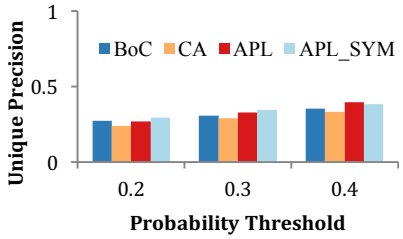
(c) Recall of APL_SYM versus the dissimilarity and probability thresholds.

Figure 3.7. Results for APL_SYM. The percentage of queries with a result is shown in (a), the precision in (b), and the recall in (c).

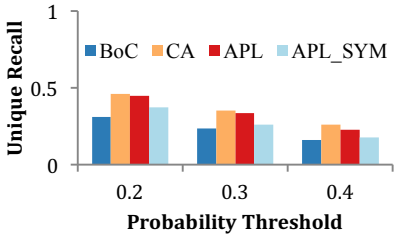
The above graphs do not account for concepts that are already in the query. This distinction may be useful for some applications like searching for undetected disorders. Due to space constraints, we cannot report the unique precision and recall results for all combinations of τ and P. For that, for each dissimilarity function we pick a single value of τ shown in Table 3.1. Note that as shown in Figures 3.4-3.7, the selection of τ does not have a big influence of the precision and recall. We then freeze the parameter τ for each dissimilarity function and examine their unique precision and unique recall, as shown in Figure 3.8. We observe that the precision drops compared to the non-unique precision, as expected.

Table 3.1. Values of dissimilarity threshold τ used in Figures 3.8 and 3.9.

Dissimilarity Function	τ
BoC	0.7
CA	0.5
APL	1.5
APL_SYM	2.0



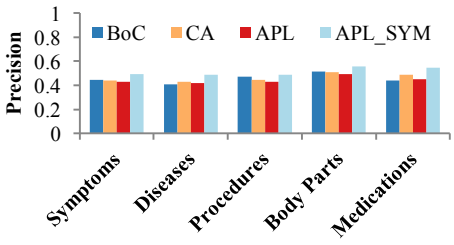
(a) The unique precision of each dissimilarity function with the dissimilarity threshold from Table 3.1.



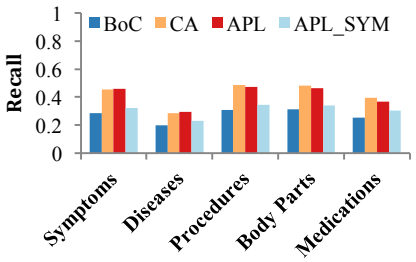
(b) The unique recall of each dissimilarity function with the dissimilarity threshold from Table 3.1.

Figure 3.8. Unique precision shown in (a) and unique recall shown in (b). Color represents different dissimilarity functions clustered around a specific value for *P*.

Figure 3.9 shows the precision and recall for each type of medical concept for each dissimilarity function with $P = 0.3$ and the τ values from Table 3.1. We see that *APL_SYM*, which has been shown to be an accurate dissimilarity function [11], maximizes the precision while *APL* and *CA* maximizes the recall, since it is less strict than *APL_SYM* – compare Figures 3.5(a), 3.6(a), and 3.7(a) – and hence more concepts are returned.



(a) The precision of each semantic type with the dissimilarity threshold from Table 3.1 and P = 0.3.



(b) The recall of each semantic type with the dissimilarity threshold from Table 3.1 and P = 0.3.

Figure 3.9. Precision shown in (a) and recall shown in (b) for each semantic type using each dissimilarity function.

3.5. Discussion and Limitations

We have presented a promising, as supported by our results, technique for predicting future medical concepts in a patient's EHR. This technique could have useful applications for patients, clinicians and clinical researchers. Our results show that we can achieve good

precision and recall for many of the query EHRs, even for the relatively small dataset of 1418 visits (with more than one event). However, due to the small size of the dataset, there are query EHRs for which we cannot find enough closely similar visits to make a reliable prediction. We expect that this limitation would be alleviated if we used a much larger dataset, which is realistic given the recent trend to merge and integrate EHR systems [67-69].

We recognize that in its current form, our system is not accurate enough for deployment. However, we measured accuracy across all concepts identified in the EHR; future efforts will try to identify specific concepts that perform best and limit the scope of the system to those concepts. For example, concern arises when giving patients or their families access to our proposed methods – incorrectly predicting an undesired concept may incur unneeded stress and anxiety. For that, we may calibrate the confidence parameters to achieve higher precision and have an expert manually select the set of concepts that are appropriate to present to patients. As one example of a potential application, such a controlled prediction module could be deployed in a patient portal of a health insurance company, where a patient can already view his or her EHR.

Other modifications of this technique might improve its accuracy. In this chapter we split the EHR of a patient based on the census events, which correspond to transfers of the patient from one unit to another. If a more detailed timeline was available – for instance we could assign a timestamp to each concept in the EHR – our methods could be modified to consider such finer-granularity prefixes and suffixes. Another decision we made is to treat two visits of a patient independently instead of concatenating. This decision is

reasonable for MIMIC II, where most patients only have one visit, but for other longer-care datasets it may be useful to reexamine this decision.

Our work has some limitations. One limitation is that we did not weigh concepts by their clinical importance. For example, the concept *cardiac arrest* is more important in terms of similarity and predictive value than the concept *coughing*. Moreover, the importance of a clinical concept depends on the application and domain. Another limitation is that we only considered patients from a single site and only ICU patients. It is not clear how well our methods would translate to other levels of acuity, such as general hospital floor, outpatient setting, etc.; this limitation motivates evaluation of datasets other than MIMIC II. Furthermore, it is not clear how accurate our system needs to be in order to be useful to patients, clinicians, and researchers. This could be addressed via user evaluations of our methods.

3.6. Conclusion

We have presented a promising technique for predicting future medical concepts in a patient's EHR by leveraging a database of similar EHRs. Using a database of real patients, we evaluated three types of inter-patient similarity measures and identified two important parameters that influence the accuracy of our technique. This evaluation revealed limitations to our approach and identified key steps for further research. These steps include increasing the size of the EHR database, identifying important clinical concepts, and evaluating datasets other than the MIMIC II database. Each step is critical for practical implementation of our technique.

Chapter 4. Pharmaceutical Drugs Chatter on Online Social Networks

Summary: The ubiquity of Online Social Networks (OSNs) is creating new sources for healthcare information, particularly in the context of pharmaceutical drugs. We aimed to examine the impact of a given OSN's characteristics on the content of pharmaceutical drug discussions from that OSN. We compared the effect of four distinguishing characteristics from ten different OSNs on the content of their pharmaceutical drug discussions: (1) General vs. Health OSN; (2) OSN moderation; (3) OSN registration requirements; and (4) OSNs with a question and answer format. The effects of these characteristics were measured both quantitatively and qualitatively. Our results show that an OSN's characteristics indeed affect the content of its discussions. Based on their information needs, healthcare providers may use our findings to pick the right OSNs or to advise patients regarding their needs. Our results may also guide the creation of new and more effective domain-specific health OSNs. Further, future researchers of online healthcare content in OSNs may find our results informative while choosing OSNs as data sources. We reported several findings about the impact of OSN characteristics on the content of pharmaceutical drug discussion, and synthesized these findings into actionable items for both healthcare providers and future researchers of healthcare discussions on OSNs. Future research on the impact of OSN characteristics could include user demographics, quality and safety of information, and efficacy of OSN usage.

4.1. Introduction

Numerous Online Social Networks (OSNs)[4] host Medicine 2.0 applications that focus specifically on user reviews of drugs [18, 19, 70-74]. Previous work has analyzed these discussions and confirmed that online drug reviews serve their purpose – i.e. users discuss medications and their effect on a disease or physical condition [75]. However, research is lacking on the impact of a given OSN's characteristics on the content of that OSN's discussions; e.g., if an OSN requires registration (e.g., providing an email address), does that affect the types of drugs users are willing to discuss?

Medicine 2.0 applications foster online communities where patients discuss their own healthcare decisions and experiences [52, 53]. These applications allow clinical researchers and citizen scientists to conduct crowdsourced health studies that complement traditional clinical trials in the public health research ecosystem [55, 76]. Such studies benefit other forms of knowledge generation, such as consumers' opinions of pharmaceutical drugs [77]. This knowledge is important: 24% of adults that use the Internet have read online reviews of a particular drug or medical treatment [78].

Moreover, there is increased interest from the research community in analyzing health-related content of OSNs. Previous work includes analyzing the content of health-related OSN discussions in terms of safety and quality, and detecting adverse drug reactions and events in OSN discussions; yet, previous work has not covered the impact of an OSN's characteristics on its discussions.

---

[4] We use the term Online Social Networks (OSNs) to define social media platforms where users share content through messages; we further define these messages as posts. Examples of OSNs include Twitter and WebMD.

Therefore, we analyzed the effect of four distinguishing characteristics of OSNs on a given OSN's content. These characteristics include: (1) OSN type – general (e.g. Twitter) versus health (e.g. WebMD); (2) if a given OSN moderates its posts; (3) if a given OSN requires registration; and (4) if a given OSN's discussions are in a Question and Answer (Q&A) format. We analyzed these characteristics both quantitatively (e.g., distribution of posts by drug type) and qualitatively (e.g., examining posts with the most frequent co-occurring medical concepts). Our results show that these OSN characteristics indeed affect the content of discussions related to pharmaceutical drugs. These effects include the type of discussions, the type of drugs discussed, the subjectivity of discussions, and the medical concept content.

4.2. Related Work

Recently, there is increased interest in analyzing the content of health-related discussions in OSNs. Related work has chronicled the utility and potential benefit/harm of health-related discussions in OSNs; related work has focused on specific aspects of the information found in OSN discussions, but none focus on the impact of OSN characteristics; we demonstrate through our results that the characteristics of the OSNs adversely affect the type of content contained within each OSN. Coupling our findings with this related work provides possible (further) explanations of the findings from the related work. Another research area of recent interest at the intersection of healthcare and OSNs is detecting adverse drug events in OSN posts; the overreaching goal is real-time pharmacovigilance via the Internet. Our work complements this related work by giving

further insight into the impact of OSN characteristics on discussions related to pharmaceutical drugs.

4.2.1. Analyzing Health Content of OSNs

Denecke and Nejdl [75] analyzed various Medicine 2.0 content and found that patient-authored postings contain more drug-related concepts than any other post. Further, they showed that drug reviews contain many disease related concepts and concluded that users searching for drugs or disorders will find results in patient-authored posts [75]. Lu *et al*. [79] studied the content of three discussion boards, from an online health community; they used one discussion board on diabetes and two on cancer. They found that drug-related postings accounted for a larger fraction of topics discussed on the diabetes board than the cancer boards [79].

Several works have looked at diabetes-related OSNs. Weitzman *et al*. [80] analyzed the quality and safety of diabetes-related OSNs and found that the quality/safety of information was variable across the ten sites under analysis. Shrank *et al*. [81] also qualitatively analyzed 15 diabetes-related OSNs – all of which feature a discussion or question forum – and they found a wide range in the number of members (from 3,000 to 300,000), one-third of the OSNs provided physicians answering questions, and two-thirds had site administrators reviewing posts. Zhang *et al*. [22] analyzed posts from a Facebook diabetes group and found that over 60% of posts were providing information, followed by emotional support (17%) and eliciting information (12%).

Greene *et al*. [82] qualitatively analyzed the communications of Facebook communities dedicated to diabetes. They found many benefits for patients participating in these

communities, such as community support and access to specialized knowledge, with little evidence of these communities supporting risky behaviors; however, one quarter of posts were explicit advertisements, some of which advertise non-FDA (Food and Drug Administration) approved products [82]. Two-thirds of posts were descriptions of personal experiences in diabetes management and a quarter of posts contained sensitive information unlikely to be revealed in doctor-patient interactions [82].

Goeuriot *et al*. [83] built and evaluated sentiment lexicons using drug reviews from a health social network. They built a general lexicon based on existing lexicons from the literature, and a domain lexicon based on drug reviews from the health social network. They showed that opinion mining of health social networks is possible, and using a combination of the general and domain lexicons achieves the best results [83].

4.2.2. Detecting Adverse Events in OSNs

Bian *et al*. [21] built two classifiers based on Twitter posts; one classifier to predict if a user (or someone they know) has used a particular drug, and a second classifier to classify if a post describes an adverse drug event. They obtain reasonable accuracy, but cite the noise in Twitter posts as one limitation to their approach [21]. Chee *et al*. [84] looked at predicting whether a drug will be withdrawn by the FDA using posts in Yahoo! Groups. While their classifier predicted many false positives (in the sense that a false positive is still on the market), a majority of the false positives with the greatest scores have been withdrawn from some market for a period of time [84].

Yang *et al*. [85] used association rule mining to detect adverse drug events in a health social network. Using data from the FDA, they confirmed correlations between drugs and

adverse reactions in the posts [85]. Leaman *et al*. [20] validated that user comments from a health social network can be mined for adverse drug events. They built a lexicon based on manual annotations of users' posts and achieve reasonable accuracy using lexical matching [20].

4.3. Methods

4.3.1. Datasets

Our analysis used the ten OSNs listed in Table 4.1. Each of these OSNs was categorized as either a *general OSN* or a *health OSN*. General OSNs include Twitter, Google+, and Pinterest, which were chosen due to their popularity and various methods of sharing messages. We chose health OSNs based on their popularity and various methods of reviewing drugs; we only considered posts in health OSNs that originate from specific forums for reviewing drugs. Hence, posts from general forums or "Ask an Expert" forums were not collected from the health OSNs. Table A.1 of Appendix A lists the dates for which posts were collected and URLs for each OSN.

Table 4.1. Various categorizations of each OSN. An OSN is moderated if a message is reviewed before becoming public. If registration is required, users must create an account before contributing content. An OSN is a Q&A format if reviews are formulated as comments/questions and replies/answers.

Dataset	Health (H) or General (G)?	Moderated?	Registration Required?	Q&A Format?
Twitter	G	N	Y	N
Google+	G	N	Y	N
Pinterest	G	N	Y	N
DailyStrength	H	N	Y	Y
Drugs.com	H	Y	N	Y
DrugLib.com	H	Y	N	Y
everydayHealth	H	N	N	Y
MediGuard	H	Y	Y	N
medications	H	Y	Y	Y
WebMD	H	N	N	Y

Each OSN was categorized further based on its moderation, registration requirements, and review format, as listed in Table 1; these categorizations are similar to related work that studies diabetes-related OSNs [80, 81]. We consider an OSN to be moderated if a message is reviewed before becoming public. An OSN requires registration if it is necessary to create an account before publishing content. An OSN has a Q&A format if posts are formatted as comments/questions with replies/answers. We ignored categorizing each OSN based on whether users can posts anonymously, as this categorization is the same as the health versus general OSN category. Even if a health OSN requires registration, users have the option to post anonymously.

4.3.2. Data Collection

First we obtained a list of the 200 most popular drugs by prescriptions dispensed from RxList.com [86]. We then removed variants of the same drug (e.g., different milligram dosages) resulting in 122 unique drug names. This list was used as a filter for finding relevant posts. Posts from general OSNs were only considered relevant if one of the drug names was found in the post's text, whereas drug reviews from health OSNs were only collected for each of the 122 drugs. The full list of drugs is given in Tables A.2, A.3, and A.4 of Appendix A.

For each OSN, we analyzed the layout of the website and built a crawler using Apache HttpComponents [87] – a library that enables web applications to obtain HTML content as if a web browser had downloaded and displayed the webpage; Twitter was handled separately using the Twitter API with the drug name list as a filter to collect tweets during the dates specified in Table 1. Data for the rest of the OSNs was gathered by

programmatically employing the search feature located on the respective OSN's website, where each drug name was specified as a query; e.g., we used Apache HttpComponents to search for Abilify on Google+. In the case of Pinterest and Google+, we collected all posts associated with the query; whereas the crawlers for health OSNs used the top search result that links to drug reviews (determining valid link patterns was done manually for each health OSN). The result is a series of HTML pages associated with a query for each OSN. Next, we extracted knowledge from each of the HTML pages using unique wrappers such as element id, location, or style. The wrappers and their content were extracted using jsoup, a Java HTML parser [88]. All pages for a given OSN follow the same HTML format, thus each of the wrappers were only defined once per OSN.

Posts in health social networks contain metadata such as gender, age, length of membership, username, etc. However, we limited our data collection to the post text and date (if available), to respect users' privacy. We collected all data in accordance with each OSN's terms of use, and therefore an OSN's data will not be made publicly available without first obtaining permission from the respective OSN.

Relevant posts obtained from the crawlers were further processed before the data analysis, as illustrated in Figure 4.1(A). First, non-English posts are removed from the general OSNs (health OSNs only contained English posts); we used a Bayesian filter based on language profiles generated from Wikipedia [89]. Next, we removed all hyperlinks and we corrected spelling mistakes in each of the posts; we corrected spelling errors using the first suggestion from HunSpell [90], an open source spell checker employed by several software packages. Lastly, we marked or removed duplicate posts for reasons described in the following

subsection. The result is a database of user posts that are relevant to the input list of prescription drug names for each OSN.

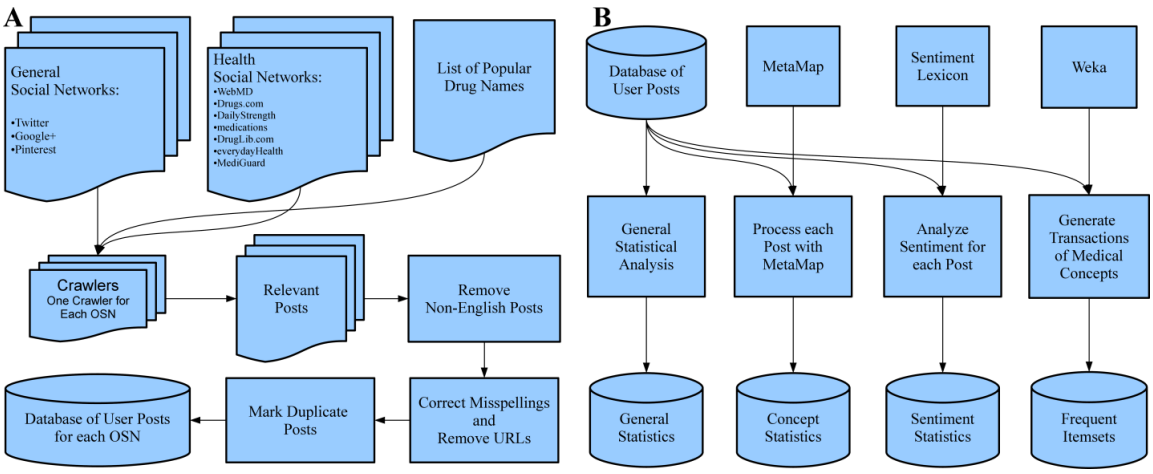


Figure 4.1. (A) an illustration of the data collection and preprocessing. Each crawler obtains a list of relevant posts using the OSNs as a seed and the list of drug names as a filter. These posts are then processed generating a database of English-only posts that have their spelling corrected. Lastly, duplicate posts are marked. (B) An overview of the data analysis performed on the database of user posts. Four different types of results are generated by the data analysis: general statistics, concept statistics, sentiment statistics, and frequent itemsets.

Some OSNs contain duplicate posts for various reasons. For example, a retweet on Twitter is reposting a tweet from another user; this is a common method to share interesting tweets with other users, and thus generates many duplicate tweets. Therefore, posts from general OSNs containing duplicated text are marked as duplicates and ignored in our analysis unless noted otherwise. Health OSNs contain duplicate posts for two reasons: (1) variants of a particular drug name (e.g., generic versus brand name) may lead to the same webpage using the OSN's search feature; and (2) users may accidently hit the submit button repeatedly, which duplicates their review on the website. All duplicate posts in the health OSNs are ignored in each analysis.

4.3.3. Methods for Data Analysis

The database created by the data collection process is then analyzed with four separate analyses: general statistics, medical concept statistics, sentiment statistics, and association rule mining. This process is illustrated in Figure 4.1(B). Since some OSNs have many more posts than others, we computed the average between each network when combining multiple OSNs into one result, rather than computing the average over all posts; otherwise, the results from Twitter or DailyStrength would decimate the results from each of the other OSNs.

4.3.3.1. Methods for General Statistics

One general statistic is the frequency of drugs based on their category. Drugs.com has a publicly available taxonomy of all drugs listed on its website [91], where one drug may be classified into multiple categories. We mapped our list of drug names to each of its top level categories as listed in the Drugs.com taxonomy; the distribution of these categories for our drug list is visualized in Figure A.1 of Appendix A. The full list of drug names along with their respective category or categories is given in Tables A.2, A.3, and A.4 of Appendix A.

For each OSN, we computed the frequency of each drug category and normalized this frequency by the total number of posts. For each OSN in a given category, we averaged the percentages of each drug category separately, and divided the sum of these percentages by the number of OSNs in the given category. Thus each OSN's distribution is weighted equally when presenting the distribution for the category. Otherwise, an OSN with many

posts would dominate the category's distribution. Duplicate posts from general OSNs were included.

We analyzed OSN similarity by ranking the most frequent drugs. We measured similarity between each pair of ranked lists by using Spearman's footrule [92]. This measure of similarity considers the distance of each item (in terms of its rank) between two ranked lists. If the lists are identical, the value will be equal to zero, whereas a value of one denotes the maximum measure of disarray between the two lists. Other general statistics are presented in Appendix B.

4.3.3.2. Methods for Medical Concept Statistics

The MetaMap tool [28] was employed to annotate each post with medical concepts from the Unified Medical Language System (UMLS). The UMLS [8] is a compendium of several medical-focused ontologies. Thus MetaMap effectively represents each post as a set of medical concepts from the UMLS.

MetaMap was originally intended to annotate text for academic publications in the biomedical field, such as those available in PubMed. Related work has shown that MetaMap is not perfect for processing social media posts [93]. Thus, we manually inspected the annotations produced by MetaMap, and we removed annotations where MetaMap consistently misclassified UMLS concepts. A majority of mistakes were words that were misinterpreted as abbreviations in the social media posts. Other common mistakes included colloquial phrases not common to academic literature in the biomedical field. Some common mistakes include:

- the first-person narrative “I” was mapped to the UMLS concept for “Iodine” (C0021968)
- “so” was mapped to “Somalia” (C0021968)
- “fed” was mapped to “fish eye disease” (C0342895)
- “lol”, “LOL” were mapped to “LOXL1 gene” (C1416898)
- “OMG” and “omg” were mapped to “OMG gene” (C1417949)
- “said” was mapped to “Simian Acquired Immunodeficiency Syndrome” (C0080151)

Mistakes similar to the ones given above were deleted from the MetaMap annotation results. We systematically analyzed each OSN by ordering every concept by its frequency and analyzing distinct phrases that were mapped for each concept.

Every concept in the UMLS is associated with one or more semantic types [64] (e.g., *Disease or Syndrome*). Each semantic type belongs to one of fifteen semantic groups [94], also defined by the UMLS. We analyzed the distribution of five semantic groups that relate to medical concepts, which include *Procedures*, *Disorders*, *Physiology*, *Chemicals and Drugs*, and *Anatomy*.

We considered the similarity of medical concept content between each OSN by ranking the most frequent semantic types. Again, we only considered semantic types that relate to medical concepts using the same five aforementioned semantic groups. We measured the similarity between each pair of ranked lists using Spearman's footrule; this is analogous to using Spearman’s footrule for measuring OSN similarity with the most frequent drugs. Other medical concept statistics are presented in Appendix B.

4.3.3.3. Methods for Sentiment Statistics

The goal of sentiment analysis is to measure the average polarity and emotion of each post. Both are achieved by mapping phrases in each post to phrases from a sentiment lexicon. We use SentiWordNet [95], which contains a dictionary of phrases where each phrase is associated with a positive, negative, and objective score. Every term in SentiWordNet is subject to the constraint that the sum of the positive, negative, and objective score must equal one.

SentiWordNet distinguishes phrases based on their sense and part of speech. Therefore we tagged each word with its part of speech using the Stanford Core NLP tagger [96]. In order to remove variants of words, we stemmed both the posts and the terms in SentiWordNet; this was done to normalize words, e.g., rain, rains, and raining all become rain. Phrases form the posts are then mapped to phrases from SentiWordNet using the longest possible match first. In the case where one term has multiple senses, we averaged the score of all senses for the given term. We then computed the positive, negative, and objective scores of each post by averaging the scores from every mapped term. The sentiment of a given OSN is measured by averaging the sentiment of all posts within that OSN. In the appendix we also present results from the NRC word-emotion lexicon [97] for analyzing the emotion of each OSN: negative–positive, anger–fear, trust–disgust, and anticipation–surprise.

4.3.3.4. Methods for Frequent Itemsets

Association rule mining is a data mining technique that learns relations between items given a database of transactions by first discovering frequent itemsets [98]. We applied this technique using UMLS concepts as items, where we considered each post to be a single

transaction. Items were restricted based on their semantic groups; we analyzed frequent itemsets for medical concepts only and all UMLS concepts. Further, frequent itemsets were discovered separately for the health and general OSNs. For implementation we used the Weka machine learning toolkit [99]. Due to the large number of items and transactions, we employed the FP-growth algorithm [100] for discovering frequent itemsets. We removed trivial itemsets and only report itemsets that show interesting trends between categorizations of OSNs. Duplicate posts from general OSNs were included.

4.4. Results

Appendix B reports general statistics and medical concept statistics for each OSN. Next, we compare the ten OSNs to each other using two measures of similarity. These measures include similarity between the most frequent drugs and the most frequent semantic types using Spearman's footrule. The first measure shows which OSNs are similar based on the frequency of discussions about particular drugs, whereas the second measure shows which OSNs are similar based on the medical content (defined by the semantic types of the extracted concepts) in the discussions. Figure 4.2 illustrates these measures for each of the ten OSNs using metric multidimensional scaling [101].

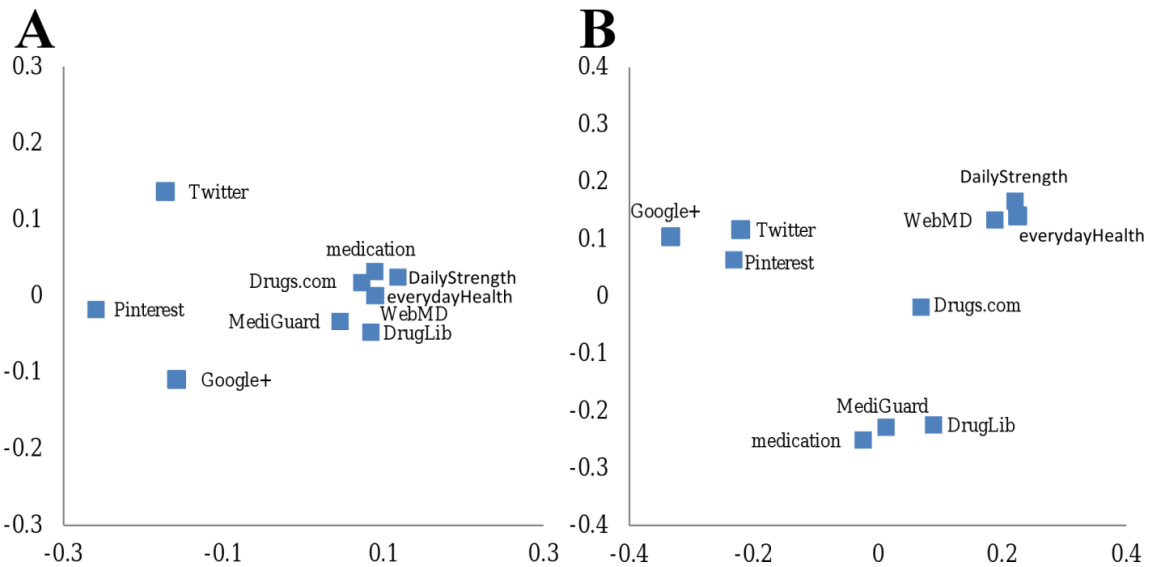


Figure 4.2. (A) multidimensional scaling of OSN similarity using Spearman's footrule with the top 25 most frequent drugs for each OSN. (B) multidimensional scaling of OSN similarity using Spearman's footrule with the top 30 semantic types for each OSN.

As shown in Figure 4.2(A), there are three primary clusters of OSNs, with the general ONSs belonging to the top-left cluster, the non-moderated health OSNs belonging to the top-right cluster and the moderated health OSNs belonging to the bottom cluster. The reason for this clustering, also discussed in Sections 4.4.1 and 4.4.2, is that *these three groups mention different types of drugs*. The only OSN left out of these clusters is Drugs.com, which is a moderated health OSN; Drugs.com is separated from the other moderated health OSNs due to a higher number of psychotherapeutics in its top 25 drugs. Figure 4.2(B) shows one cluster, which contains the health OSNs, and the three general OSNs separated from that cluster and each other. This figure suggests that *the medical content, in terms of UMLS semantic types, of health OSNs is similar, and differs from the medical content found in general OSNs*; further, this figure also suggests that the medical content in general OSNs varies across each OSN. For example, over 50% of the concepts in Twitter relate to Chemicals and Drugs, whereas Google+ and Pinterest have 36% and

44% respectively. Therefore, Twitter is more likely to contain semantic types relating to Chemicals and Drugs in its top 25 semantic types.

The remainder of our results section examines each categorization of OSNs, and it is divided into four parts: (1) general versus health OSNs; (2) health OSNs that are non-moderated versus moderated; (3) health OSNs with registration versus no registration; and (4) health OSNs with a Q&A format versus health OSNs with a review format. We omitted general OSNs from the last three categorizations of OSNs, since they all belong to the same categories (e.g., all are non-moderated).

4.4.1. General versus Health OSNs

Figure 4.3 compares the distributions of drug category frequency, polarity, and semantic groups of the health and general OSNs with the distribution of a uniform baseline. In Figure 4.3(A), this baseline is the distribution of the drug categories reported in Figure A.1. The baselines for Figures 4.3(B)-(C) assume a uniform distribution for all items matched in the database; e.g. the baseline in Figure4. 3(B) assumes a uniform distribution for all terms matched from SentiWordNet.

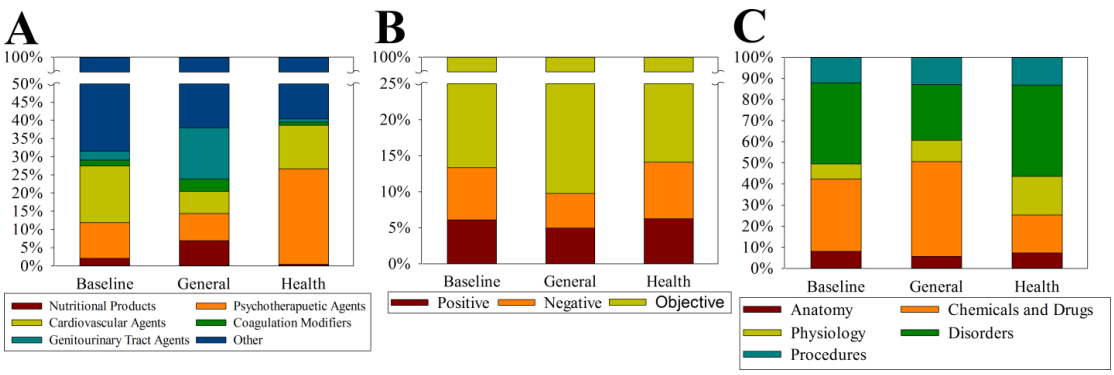


Figure 4.3. An overview of the analysis for general OSNs versus health OSNs: (A) the distribution of drug category frequencies; (B) the distribution of polarity; and (C) the distribution of semantic groups. Each baseline represents a uniform distribution: (A) assumes each drug from the drug list will appear with equal probability; (B) assumes each term mapped from SentiWordNet will appear with equal probability; and (C) assumes each UMLS concept extracted from the posts will appear with equal probability.

Table 4.2 illustrates the major differences visualized in Figure 4.3. This table reports the highest absolute relative change of each item when compared to the baseline distributions. For example, there is a 473% increase in the number of posts related to genitourinary tract agents in general OSNs compared to the assumption that each drug would appear with equal probability. General OSNs have a decrease in both negative and positive polarity because their posts are more objective.

Table 4.2. Highest absolute relative changes of each item compared with the baselines shown in Figure 4.3.

Drug Category		
Genitourinary Tract Agents	General	+473%
Nutritional Products	General	+237%
Psychotherapeutic Agents	Health	+167%
Coagulation Modifiers	General	+107%
Nutritional Products	Health	-79%
Genitourinary Tract Agents	Health	-63%
Cardiovascular Agents	General	-61%
Coagulation Modifiers	Health	-47%
Psychotherapeutic Agents	General	-24%
Cardiovascular Agents	Health	-23%
Polarity		
Negative	General	-33%
Positive	General	-18%
Semantic Group		
Physiology	Health	+158%
Chemical and Drugs	Health	-48%
Physiology	General	+41%
Disorders	General	-31%
Chemical and Drugs	General	+31%

Figure 4.3(A) shows some interesting trends between the types of drugs discussed in general and health OSNs. Firstly, both general and health OSNs have a smaller number of posts about cardiovascular agents compared to the baseline, and therefore *users of any OSN are less likely to post about cardiovascular agents such as Digoxin or Flomax*. The other drug categories show opposing trends between health and general OSNs – *drugs such as Viagra, Niaspan, and Warfarin are more common in general OSNs than drugs such as Cymbalta or Abilify, whereas the opposite is true for health OSNs*.

Figure 4.3(B) illustrates the differences in polarity between the health and general OSNs. *General OSNs use more objective terms; whereas health OSNs use more subjective terms*. We speculate that this is because of the anonymity in health OSNs, where users often use name aliases, and hence can discuss more personal and subjective topics. Results for

emotion, which are reported in Appendix C, show no significant differences between general and health OSNs.

Figure 4.3(C) illustrates the type of medical concepts discussed for general and health OSNs compared to a baseline that assumes each UMLS concept appears with equal probability. There is a large increase in the number of concepts relating to physiology in health OSNs, but a decrease in the number of concepts relating to chemicals and drugs. General OSNs have more concepts relating to chemicals and drugs, and fewer concepts related to disorders. Further, these results suggest that *users of health OSNs are concerned with the effects of drugs on physiology, whereas users of general OSNs are either using drug names as slang or drug names in advertisements*.

4.4.1.1 A Qualitative Analysis of General and Health OSNs

Table 4.3 reports the most frequent itemsets of size 1 of medical concepts for health and general OSNs; itemsets of larger sizes are reported in Appendix C. *Health OSNs contain medical conditions, drug names and symptoms where the concept for sleep dominates* with a frequency of over 10%. *General OSNs contain many specific drugs names, where Viagra and Ibuprofen dominate* with frequencies over 15% and 10% respectively. Larger itemsets show that *general OSNs contain frequent itemsets of drugs that serve a similar purpose*; e.g., Ibuprofen, Tylenol, and Advil. *In general ONSs, drugs are often used as slang or in jokes*; e.g., "Viagra for women has been around for centuries. It's called money". *Funny news items are popular in general OSNs*; for example, Appendix C illustrates a series of frequent itemsets referring to Viagra and heart attack, which references a news story about a man that took a bottle of Viagra and died from a heart attack after having sex for 12 hours.

Table 4.3. Frequent itemsets of size 1 for medical concepts.

Health OSNs		**General OSNs**	
Sleep	10.20%	Viagra	15.35%
Depression	4.81%	Ibuprofen	10.55%
Headache	4.11%	Penicillins	3.36%
Tired	4.02%	Ambien	2.65%
Weight Gain	3.86%	Oxycodone	2.19%
Anxiety	3.62%	Sleep	1.90%
Eating	3.48%	Cialis	1.68%
Mental Suffering	3.22%	Eating	1.23%
Dizziness	3.17%	Acids	1.17%
Lisinopril	3.15%	Tramadol	1.16%

Table 4.4 reports frequent itemsets of size 1 of all concepts for health and general OSNs; itemsets of larger sizes are reported in Appendix C. Concepts for help, physician, milligram and started dominate health OSNs with frequencies greater than 12%, revealing that *users of health OSNs are discussing their experiences with their medications, and the differing strategies employed by their physicians*; e.g., "Because of my sleep troubles from Lexapro, [My doctor] started me on a new drug, Ambien to help me sleep with a dosage of 5 mg". *General ONSs contain posts from online pharmacies that advertise drugs for the best price with no prescription needed*; e.g., "[URL] with best price naprelan 250mg in internet rx overnight South Dakota". *Breaking news items about pharmaceutical drugs are popular in general OSNs*; as illustrated in Appendix C, the United States Food and Drug Administration recommended lower dosages of Ambien for patients during our Twitter data collection.

Table 4.4. Frequent itemsets of size 1 for all UMLS concepts.

Health OSNs		**General OSNs**	
Help	16.78%	milligram	8.30%
Physicians	15.23%	Internet	5.72%
milligram	13.75%	Dosage	3.64%
Started	12.24%	Tablet Dosing Unit	3.04%
Sleep	8.65%	Order	2.97%
Dosage	7.77%	Physicians	2.14%
Better	7.57%	Prices	2.11%
To be stopped	5.50%	Scripts	1.93%
Etiology aspects	5.39%	Buying	1.89%
Life	5.33%	Fast	1.74%

4.4.2. Moderated versus Non-moderated Health OSNs

Figure 4.4 compares distributions of drug category frequency, polarity, and semantic groups of moderated and non-moderated health OSNs with the distribution of all health OSNs as a baseline; Table 4.5 illustrates the major differences visualized in Figure 4.4, analogous to Table 4.2 and Figure 4.3. Appendix D reports the general statistics and medical concept statistics for moderated and non-moderated health OSNs.

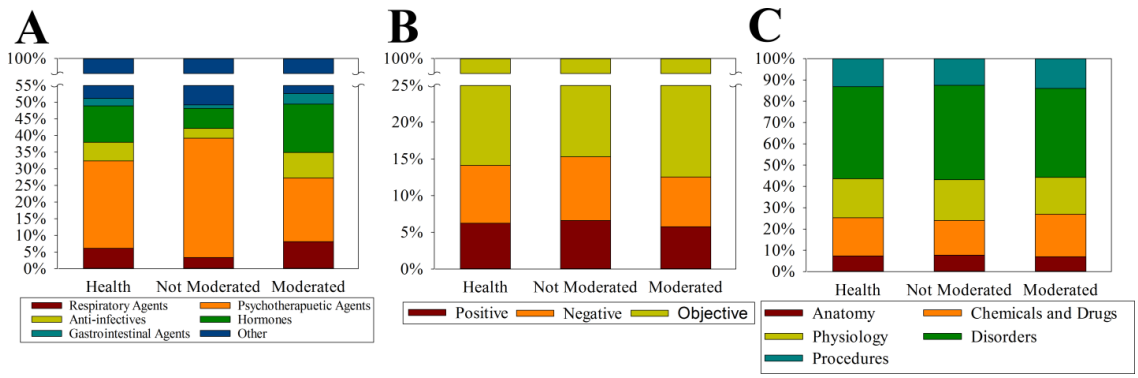


Figure 4.4. An overview of the analysis for moderated and not moderated OSNs. (A) The distribution of drug category frequencies; (B) the distribution of polarity; and (C) the distribution of semantic groups. The original distribution of all health OSNs is used as the baselines.

Table 4.5. Highest absolute relative changes of each item compared with the health OSN baseline shown in Figure 4.4.

Drug Category		
Gastrointestinal Agents	Non-moderated	-53%
Anti-infectives	Non-moderated	-48%
Respiratory Agents	Non-moderated	-45%
Hormones	Non-moderated	-45%
Gastrointestinal Agents	Moderated	+40%
Psychotherapeutic Agents	Non-moderated	+36%
Anti-infectives	Moderated	+36%
Respiratory Agents	Moderated	+34%
Hormones	Moderated	+34%
Psychotherapeutic Agents	Moderated	-27%
Polarity		
Negative	Moderated	-14%
Negative	Non-moderated	+10%
Semantic Groups		
Chemicals and Drugs	Moderated	+12%
Chemical and Drugs	Non-moderated	-9%

Figure 4.4(A) compares the distribution of drug categories between health OSNs, non-moderated health OSNs, and moderated health OSNs. As noted in Table 4.5, *moderation affects the types of drugs users are willing to discuss; psychotherapeutic agents observed a 36% increase in frequency amongst non-moderated health OSNs and a 27% decrease in moderated health OSNs*. Conversely, gastrointestinal agents, hormones, anti-infectives, and respiratory agents all observed an increase for moderated health OSNs, and a decrease for health OSNs that are not moderated.

Figure 4.4(B) compares the distribution of polarity between health OSNs, non-moderated health OSNs, and moderated health OSNs. Also noted in Table 4.5, *moderation decreases the overall subjectivity, whereas non-moderated health OSNs increases subjectivity*. Thus, introducing moderation adds a level of objectivity to health OSNs.

Figure 4.4(C) reports the effect of moderation on semantic groups, and Appendix D reports the effect of moderation on emotion. *Overall, moderation has little effect on the medical*

concept content and emotional terms in health OSNs. However, moderated health OSNs did have a slight increase on the number of terms relating to trust, whereas non-moderated health OSNs decreased the number of terms relating to trust. Further, moderated health OSNs increased the number of concepts relating to Chemicals and Drugs by 12%, whereas lack of moderation decreased these concepts by 9%. Appendix D reports frequent itemsets for health OSNs with and without moderation. These itemsets show that *users prefer non-moderated health OSNs when discussing psychotherapeutics and psychological conditions*.

4.4.3. Registration versus no Registration in Health OSNs

Figure 4.5 compares distributions of drug category frequency, polarity, and semantic groups of health OSNs that do or do not require registration with the distribution of all health OSNs as a baseline; Table 4.6 illustrates the major differences visualized in Figure 4.5. Appendix E reports the general statistics and medical concept statistics for health OSNs that do or do not require registration.

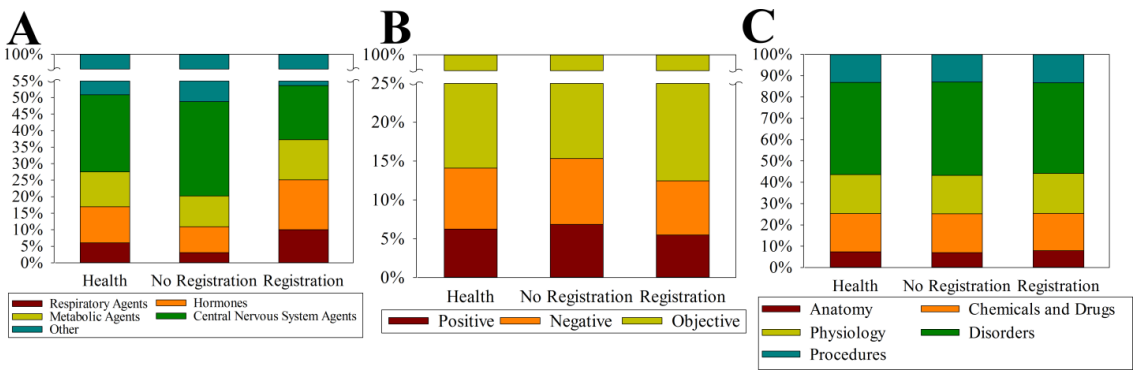


Figure 4.5. An overview of the analysis for health OSNs that do or do not require registration. (A) The distribution of drug category frequencies; (B) the distribution of polarity; and (C) the distribution of semantic groups. The original distribution of the health OSNs is used as a baseline.

Table 4.6. Highest absolute relative changes of each item compared with the health OSN baseline shown in Figure 4.4.

Drug Category		
Respiratory Agents	Registration	+65%
Respiratory Agents	No Registration	-49%
Hormones	Registration	+37%
Central Nervous System Agents	Registration	-30%
Hormones	No Registration	-28%
Central Nervous System Agents	No Registration	+22%
Metabolic Agents	Registration	+16%
Metabolic Agents	No Registration	-12%
Polarity		
Positive	Registration	-12%
Negative	Registration	-11%

Figure 4.5(A) compares the distribution of drug categories for health OSNs that do or do not require registration against all health OSNs as a baseline. As noted in Table 4.6, *registration affects the types of drugs users are willing to discuss; central nervous system agents observed a 30% decrease in frequency amongst health OSNs that require registration and a 22% increase in health OSNs that do not require registration*. Conversely, health OSNs that require registration have a 65% increase in posts about respiratory agents, but health OSNs that do not require registration report a 49% decrease in posts about respiratory agents.

Figure 4.5(B) compares the distribution of polarity for health OSNs that do or do not require registration against all health OSNs as a baseline. *Similar to moderated health OSNs, requiring registration reduces the amount of subjectivity in health OSNs*.

Figure 4.5(C) reports the effect of registration on semantic groups, and Appendix E reports the effect of registration on emotion. *Overall, registration has little effect on the medical concept content and emotional terms in health OSNs*. Appendix E reports frequent itemsets

for health OSNs that do or do not require registration. Similar to moderation, these itemsets show that *users prefer health OSNs that do not require registration when discussing psychotherapeutics and psychological conditions*.

4.4.4. Review versus Q&A format

Figure 4.6 compares distributions of drug category frequency, polarity, and semantic groups of health OSNs that have a review format with health OSNs that have a Q&A format with the distribution of all health OSNs as a baseline; Table 4.7 illustrates the major differences visualized in Figure 4.6. Appendix F reports the general statistics and medical concept statistics for health OSNs with a review or Q&A format.

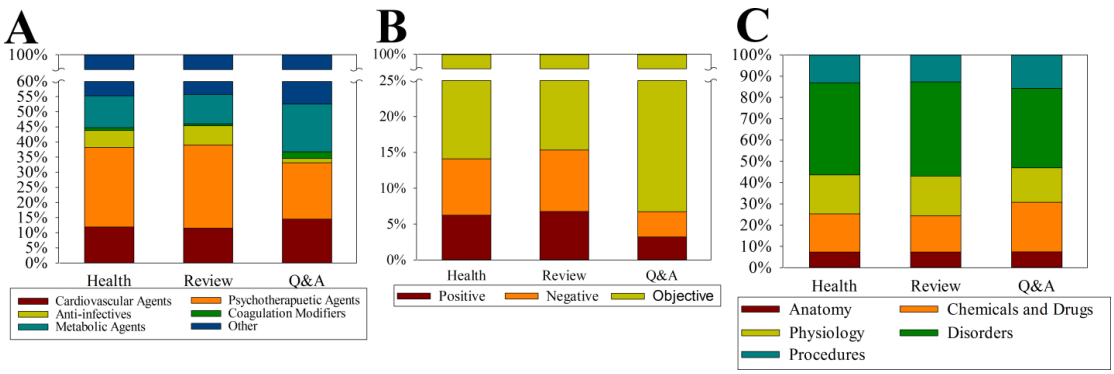


Figure 4.6. An overview of the analysis for health OSNs with a review or Q&A format. (A) The distribution of drug category frequencies; (B) the distribution of polarity; (C) the distribution of semantic groups. The original distribution of the health OSNs is used as a baseline.

Table 4.7. Highest absolute relative changes of each item compared with the health OSN baseline shown in Figure 4.6.

Drug Category		
Coagulation Modifiers	Q&A	+155%
Anti-infectives	Q&A	-75%
Metabolic Agents	Q&A	+50%
Psychotherapeutic Agents	Q&A	-29%
Coagulation Modifiers	Review	-26%
Cardiovascular Agents	Q&A	+21%
Polarity		
Negative	Q&A	-55%
Positive	Q&A	-49%
Semantic Groups		
Chemicals and Drugs	Q&A	+30%
Procedures	Q&A	+19%
Disorders	Q&A	-13%

Figure 4.6(A) compares the distribution of drug categories for health OSNs that have a review format or Q&A format. Health OSNs that have a Q&A format have a 155%, 50%, and 21% increase in posts related to coagulation modifiers, metabolic agents, and cardiovascular agents respectively. Posts about psychotherapeutic agents and anti-infectives observed a decrease of 29% and 75% in health OSNs with a Q&A format. This suggests that *users are less likely to ask questions about Abilify or Penicillin, but users are more likely to ask questions about Warfarin, Advair, or Lipitor*.

Figure 4.6(B) compares the distribution of polarity for health OSNs that have a review format or Q&A format. Health OSNs with a Q&A format are much more objective than health OSNs with a review format, with decreases of 55% and 49% to negativity and positivity respectively. Thus, *users of health OSNs with a Q&A format tend to post in an objective manner, rather than subjective opinions regarding a particular drug*.

Figure 4.6(C) compares the distribution of semantic groups for health OSNs that have a review format or Q&A format. Health OSNs with a Q&A format observed an increase of

30% and 19% for Chemicals and Drugs and Procedures respectively; whereas Disorders and Physiology observed decreases of 13% and 11% respectively. This suggests *users ask questions that focus on drugs and procedures rather than questions about specific disorders or effects on their physiology*.

4.5. Discussion

Our results section has demonstrated the similarities and differences of OSNs in the context of pharmaceutical chatter in OSNs. Together, these data may help inform patients and healthcare providers about the type of content related to pharmaceutical drugs on OSNs. As pointed out by Eysenbach, OSNs (including health OSNs) are essentially an apomediated environment [52], where users take over the role of intermediary and guide other users to relevant and accurate information.

Based on our findings, healthcare providers could advise patients on the use of OSNs. Examples include: the prevalence and legitimacy of online pharmacies due to the high number of advertisements from online pharmacies in general OSNs; general OSNs are good sources of breaking news, particularly if that news was reported by a trusted source such as United States Food and Drug Administration; thousands of other patients are discussing health conditions and their treatments on health OSNs, yet these discussions may be subjective or biased; health OSNs that require registration, have moderation, or a Q&A format tend to be more objective, and thus information is less opinionated.

Our results may also guide the creation of new and more effective domain-specific health OSNs. Furthermore, these data may help future researchers that study OSNs make informed decisions about the social networks chosen for study when consider health

content in OSNs. In the context of pharmaceutical drug chatter in OSNs: general OSNs are sources of jokes, news, and advertisements; health OSNs are sources of user experiences' with pharmaceutical drugs and strategies employed by their physicians for a particular medical condition or set of medical conditions; also, sleep and sleep related problems are a common theme throughout health OSNs. Drugs and diseases relating to the brain or central nervous system are more frequently discussed on health OSNs that are non-moderation and do not require registration respectively. In contrast, more prevalent diseases, such as asthma, hypertension, or high cholesterol are more frequently discussed on health OSNs that have moderation or require registration. Lastly, users are more likely to ask questions in public spaces about respiratory agents and hormones.

4.5.1 Limitations

We did not consider demographics of users in this study as this information was not present in every source. Therefore, we cannot generalize our results to the general population. However, given that nearly 1 in 4 adults in 2011 that used the Internet, also looked for reviews on drugs or medical treatments [78], we argue that our results are still consequential to a substantial portion of the general population.

Another limitation of our work is that we did not remove messages that would be considered spam. The definition of spam is subjective – health social networks would remove pharmaceutical advertisements, whereas general social networks would not; mind that general social networks are successful for detecting and removing spam. Albeit modern social networks have become very good at eliminating spam [102-104], therefore we believe that this is not a serious problem.

There are also technical limitations with our approach. Due to the volume of Twitter posts, we only selected a two-week sample of posts, whereas we collected as many posts as possible for each of the other datasets. Ideally, we would examine all posts from Twitter since Twitter's beginning. Due to crawling constraints, we did not consider every social network where users post messages with respect to pharmaceutical drugs. MetaMap is not perfect for annotating social media posts, but we did clean up its output by removing annotations that are obviously incorrect. While the UMLS is a compendium of several medically focused ontologies, an ideal ontology for OSN posts about pharmaceutical drugs would be built using a specialized lexicon for health-related posts in social media; such a lexicon would also apply to the sentiment lexicons, where terms such as “omg” and “lol” are not mapped to any word in each of the sentiment lexicons used in this chapter.

4.6. Conclusion

With the objective to analyze the impact of OSN characteristics on the content of pharmaceutical drug discussions, we have reported several patterns of information from ten different OSNs. We demonstrated that an OSN’s characteristics affect the type of discussions, the type of drugs discussed, the subjectivity of discussions, and the medical concept content. We synthesized these findings and proposed actionable items for both healthcare providers and future researchers of healthcare discussions on OSNs. Future research on the effect of OSN characteristics in healthcare discussions could include user demographics, quality and safety of information, and efficacy of OSN usage.

Chapter 5. Provider Attributes Correlation Analysis to their Referral Frequency and Awards

Summary: There has been a recent growth in health provider search portals, where patients specify filters—such as specialty or insurance—and providers are ranked by patient ratings or other attributes. Previous work has identified attributes associated with a provider's quality through user surveys. Other work supports that intuitive quality-indicating attributes are associated with a provider's quality.

We adopt a data-driven approach to study how quality indicators of providers are associated with a rich set of attributes including medical school, graduation year, procedures, fellowships, patient reviews, location, and technology usage. In this chapter, we only consider providers as individuals (e.g., general practitioners) and not organizations (e.g., hospitals). As quality indicators, we consider the referral frequency of a provider and a peer-nominated quality designation. We combined data from the Centers for Medicare and Medicaid Services (CMS) and several provider rating web sites to perform our analysis.

Our data-driven analysis identified several attributes that correlate with and discriminate against referral volume and peer-nominated awards. In particular, our results consistently demonstrate that these attributes vary by locality and that the frequency of an attribute is more important than its value (e.g., the number of patient reviews or hospital affiliations are more important than the average review rating or the ranking of the hospital affiliations, respectively). We demonstrate that it is possible to build accurate classifiers for referral frequency and quality designation, with accuracies over 85%.

Our findings show that a one-size-fits-all approach to ranking providers is inadequate and that provider search portals should calibrate their ranking function based on location and specialty. Further, traditional filters of provider search portals should be reconsidered, and patients should be aware of existing pitfalls with these filters and educated on local factors that affect quality. These findings enable provider search portals to empower patients and to "load balance" patients between younger and older providers.

List of Abbreviations

- CMS – **C**enters for **M**edicare and **M**edicaid Services
- PQRS – **P**hysician **Q**uality **R**eporting **S**ystem
- HMO – **H**ealthcare **M**aintenance **O**rganization
- NPI – **N**ational **P**rovider **I**dentifier
- EHR – **E**lectronic **H**ealth **R**ecords
- eRx – Electronic Prescriptions

5.1. Background

Recently, there has been an increased interest in provider search portals such as Vitals.com and Healthgrades.com [16, 17]. A key challenge for these portals is to identify attributes that determine the quality of a provider, and to make these attributes available to their users. Provider search portals typically allow users to rank providers by location, patient rating, or last name, and users may filter providers by medical school or affiliated hospital rankings. However, ranking based on patient reviews may be ineffective as the wide majority of patient ratings are positive, and previous research has shown that patients mostly rate providers on office wait times and visit durations [105-109]. Further, better medical schools do not necessarily create better providers, as a provider's residency has a stronger impact on that provider's clinical style [110].

Other studies have assessed the qualitative attributes of provider quality via surveys [111-114]. These studies show that accurate diagnosis and treatment, probity, good communication and listening skills, sensitivity towards feelings, and tailoring treatment options are the qualitative attributes of provider quality. Unfortunately, measuring these qualitative attributes for all providers is impossible given the available information on providers and provider search portals. CMS may publish performance data for individual providers in the future, such as medical procedure outcomes, but more subjective attributes such as listening skills may still be largely unavailable.

Given the lack of data on qualitative attributes and the sparsity and bias of patient reviews of provider quality, we focus on quantitative attributes of providers in this study. There is a rich set of data available for each provider, however, a key challenge in using a data-driven approach is finding the ground truth–i.e., a set of "good" providers–to guide our analysis of important attributes for provider quality. The Centers for Medicare and Medicaid Services (CMS) has defined quality measures, such as the Physician Quality Reporting System (PQRS), but PQRS data is only publicly available for group practices with more than 25 providers and hence is not applicable to individuals [4].

In our approach we view referral frequency and peer-nominated quality designations as indicators for provider quality, although we understand that these measures have their own flaws and limitations as discussed in the limitations section. We view both peer-nominated awards and referral frequency as a peer-validated quality measures—i.e., a provider would not receive many referrals or nominations if he or she has not garnered the trust of their peers, which implies high-quality ratings from the local community. We adopt a data-

driven approach to discover the provider attributes that are associated with these quality indicators. Our focus is to study the correlations among a wide range of provider attributes and indicators of quality, keeping in mind that correlation is not equal to causation, nor are our quality measures comprehensive (unfortunately there are no comprehensive quality indicators for individual providers that are publicly available).

5.2. Related Work

The related work can be split into two categories: provider search sites and attributes associated with provider quality. Previous work shows that providers are being rated online, as one out of every six physicians has been rated online [24]. Moreover, provider rating websites have observed increases in usage from less than 1% to over 30% for specific specialties from 2005 to 2010 [24]. Further, several studies have attempted to identify attributes of provider quality, but these studies focus on qualitative aspects of medical practice (e.g., communication skills) rather than quantitative aspects (e.g., medical school rank).

5.2.1. Online Provider Search Sites

There has been increased interest in provider search portals with over 30 studies and reviews appearing in peer-reviewed journals [115, 116]. The previous related work has studied the topic of provider ratings online, but these studies are focused solely on user generated content and do not consider the rich set of provider data readily available. Ellimoottil et al. studied online reviews of 500 urologists from Vitals.com and found that each physician was rated 2.4 times on average and 86% of physicians had positive ratings [106]. Wan and Dimov analyzed online reviews of 300 allergists from three popular

provider review websites, and they also found that a majority of reviews were positive [117]. Further, they reported a statistical difference when categorizing reviews by the physician's graduation year, which showed that physicians who graduated more recently obtained more positive scores. Kadry et al. analyzed 4999 online provider ratings from the 10 most popular websites that rate providers, and they found that a majority of reviews are positive. Further, Kadry et al. suggest that a single overall rating to evaluate providers is sufficient to assess a patient's opinion of a provider [107].

Verhoef et al. published a review on provider rating websites as tools to understand quality of care, and they found that several studies indicate a relationship between ratings and quality of care [115]. However, Verhoef et al. point out that provider rating websites have some drawbacks, including anonymity of ratings and the fact that the population on social media is not representative of the actual patient population. Due to the anonymity of the ratings, the overall scores of each provider are susceptible to fraud [115]. Hence, provider ratings may not be reliable for assessing the quality of a provider. Segal et al. examined online surgeon reviews and whether those reviews are able to track surgeon volume [118]. They showed that high volume surgeons can be differentiated from lower volume surgeons by using the number of ratings, the number of text comments for a surgeon, and the ratio of positive and negative comments.

5.2.2. Attributes Associated with Provider Quality

Several surveys have examined the qualitative attributes of providers and, but none have focused on the quantitative attributes of providers. Lee et al. assessed the attributes that make a good provider by generating a list of characteristics and surveying medical students,

faculty, patients, and primary care providers [111]. Their survey showed that all participants regarded accurate diagnosis and treatment as the most important attribute and keeping up-to-date as the second most important attribute. Lambe and Bristow also surveyed a panel of experts from a wide range of medical specialties on the most important attributes of good providers [112]. They found that probity, recognition that patient care is the primary concern of a provider, good communication and listening skills, and recognition of one's own limits were among the top attributes. As with Lee et al., Labe and Bristow sought to identify qualitative attributes of top providers.

Schattner et al. surveyed 445 patients at hospitals and clinics, asking each patient to select the four most important attributes from a questionnaire of 21 arbitrary attributes [114]. The most essential attributes selected were professional expertise, patience and attentiveness, informing the patient, and representing the patient's interest. Further, Schattner et al. found that significantly more attributes were selected in the domain of patient's autonomy over the domain of professional expertise. Luthya et al. also examined attributes of good providers from the patient's perspective via a survey [113]. They found that sensitivity towards feelings and tailoring treatment options were the most important attributes for good providers. Similar to the other studies, Luthya et al. focused on the qualitative attributes of good providers.

None of the aforementioned studies—on both provider search sites and attributes of provider quality—have performed a data-driven, quantitative analysis of provider attributes. Hence, research is lacking on the association between information from provider rating websites and publicly available data, such as the patient's perspective via

user reviews, credentials of the provider (e.g., medical school), and professional attributes (e.g., accepted insurance plans). This leaves several data-driven questions unanswered. E.g., which attributes determine a peer-nominated award, and do these attributes also correlate with attributes that determine a provider's referral frequency? And, are reviews based on wait times useful for finding distinguished providers, or providers who receive many referrals?

5.3. Methods

We collected detailed data from a diverse set of sources including CMS data on providers and hospitals, U.S. News rankings of medical schools and hospitals, and additional provider information and patient reviews from Vitals.com and Healthgrades.com. We then mapped entities across sources, creating a database of 608,935 providers; this database is then used in each of our analyses. We converted each provider's information to a set of intuitive quantitative attributes. For instance, medical school, residency, and fellowship were converted to integers based on the U.S. News & World Report ("U.S. News") medical school rankings [119-121]. Affiliated hospitals were mapped to specialty-specific rankings as defined by U.S. News (e.g., cancer, gynecology, urology, etc.). Figure 5.1 presents an overview of our methods.

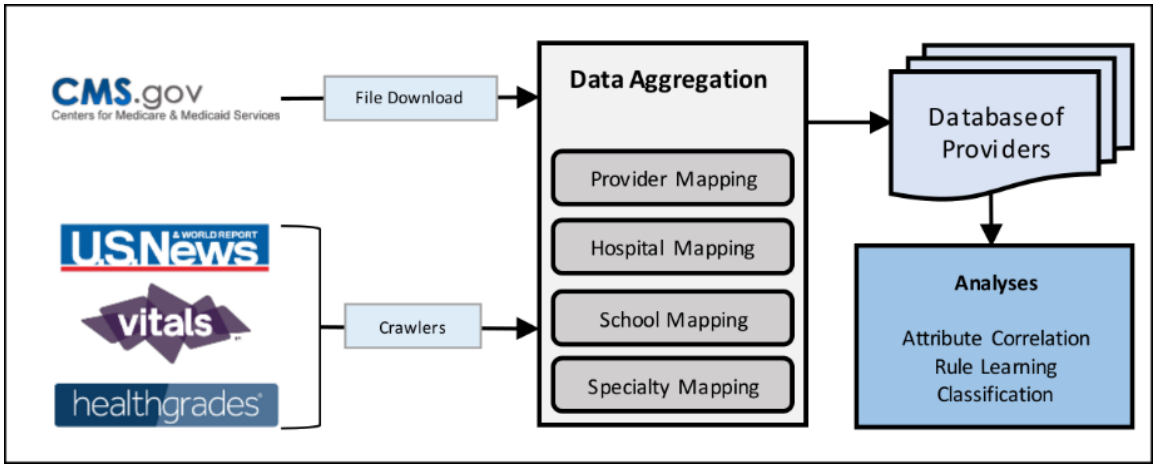


Figure 5.1. An overview of our methods from data collection to aggregation to analysis.

5.3.1. Quality Indicators

For referrals we selected CMS's 2012-2013 30 day interval public dataset of Medicare and Medicaid referral patterns [122]. In this data set, referrals are only considered when a provider services a patient 30 days after another provider serviced the same patient—given that the first provider is listed as a referring provider on the second provider's CMS claim. Medicare Part A and B beneficiaries, in most cases, do not need referrals to see specialists enrolled in Medicare; however Medicare Part C beneficiaries on Healthcare Maintenance Organization (HMO) plans are required to have a referral to see a specialist (certain exceptions exist, such as annual mammogram screenings) [123, 124]. In 2013, 9.3 of the 50 million Medicare beneficiaries were enrolled in a Part C HMO plan, up from 8.5 million in 2012; in both years these beneficiaries accounted for 65% of all Part C beneficiaries [125, 126]. Thus approximately 20% of all Medicare beneficiaries must obtain a referral to see a specialist; moreover, regardless of insurance plan, most radiological procedures require a physician referral. Further, primary care physician referrals are amongst the leading factors patients consider when choosing physicians [127].

For rule learning and classification purposes, Referral Frequency is converted into a nominal attribute with five distinct values based on the provider's referral frequency relative to other providers:

1. None (never referred, e.g., a general practitioner)
2. Very Low (normalized referrals greater than 0 and less than or equal to 0.25)
3. Low (normalized referrals greater than 0.25 and less than or equal to 0.5)
4. High (normalized referrals greater than 0.5, less than or equal to 0.75)
5. Very High (normalized referrals greater than 0.75).

For quality designation we selected the Castle Connolly designation; each year Castle Connolly distinguishes top providers both nationally and regionally through a peer nomination process that involves over 50,000 providers and hospital and healthcare executives [128]. Castle Connolly receives over 100,000 nominations each year, and a physician-led research team awards top providers from these nominations. Regional awardees are leaders in their communities and national awardees are physicians who attract patients from across the country [129]. Analogous to Castle Connolly, several organizations have internal peer-nominated awards (e.g., Kaiser Permanente Medical Group awards; the American Academy of Family Physicians awards Family Physician of the Year). However, unlike Castle Connolly, these types of awards are not as comprehensive nor do they consider a wide pool of physicians across several specialties. Hence we focus on Castle Connolly awards as other awards are limited by the number of awardees and their geographical and medical specialty diversity.

5.3.2. Data Collection

Insurance information and patient ratings were collected from both Vitals.com and Healthgrades.com [16, 17]. Medical school and hospital rankings were collected from U.S. News's reports [119, 121]. CMS has released several datasets for health providers (and hospitals) based in the U.S. This includes general information such as the provider's specialties, medical training, and hospital affiliations [130, 131]. Other provider information includes the Healthcare Common Procedure Coding System (HCPCS), physician referrals, and prescription data [2, 5, 122]. Note that all CMS datasets link providers using a National Provider Identifier (NPI). CMS hospital information includes name, location, and a unique identifier which is used to link each NPI to affiliated hospitals [3]. CMS data was downloaded directly from cms.gov [2, 3, 5, 122, 130, 131]. Separate crawlers were built using jsoup [88]—a Java library that obtains and parses HTML pages—for each of the other data sources: Vitals.com, Healthgrades.com, and U.S. News.

In total, we collected information on 3.2 million distinct providers from CMS, 4600 distinct hospitals from CMS, 1.9 million distinct providers from Healthgrades.com, 1 million distinct providers from Vitals.com, 1,956 hospitals from U.S. News, and 149 distinct medical schools from U.S. News. After appropriate data transformations and entity mappings, we generated the set of provider attributes listed in Tables 5.1 and 5.2.

Table 5.1. List of attributes used in our analysis based on the data collected (continued in Table 5.2).

Category	Attribute	Description	Source
Quality Indicators	Referral Frequency	Normalized number of referrals.	CMS
	Castle Connolly Award	Whether or not the provider is recognized by Castle Connolly as a distinguished provider.	Vitals.com
General Information	Gender	Male or female, as specified in the CMS data.	CMS
	Accepting New Patients	Whether or not the provider is accepting new patients.	Vitals.com and Healthgrades.com
	Specialties	A set of attributes, one for each specialty, e.g., cardiologist.	CMS
	Census Division	One of the nine regional divisions as defined by the U.S. Census Bureau, based on the provider's location [132].	CMS
	Number of Organization Members	Number of organization members, e.g., 1 for a private practice with 1 provider.	CMS
	Languages	A set of attributes that represent languages spoken by the provider.	Healthgrades.com
	Number of Spoken Languages	The number of spoken languages spoken by the provider.	Healthgrades.com
	Accepts Medicare Insurance	Whether or not the provider accepts Medicare assignments.	CMS
	PQRS	Whether or not the provider participates in the Physician Quality Reporting System (PQRS)[4].	CMS
	EHR	Whether or not the provider uses an Electronic Health Record (EHR) system.	CMS
	eRx	Whether or not the provider uses electronic prescriptions.	CMS

Table 5.2. List of attributes used in our analysis based on the data collected.

Category	Attribute	Description	Source
HCPCS Information	Procedure Types	A set of binary attributes, one for each type of procedure performed by the provider. HCPCS cover anything billable to Medicare, from new visits to transplants.	CMS
	Relative Cost of Procedures	The relative cost of the provider's procedures, normalized to [0,1] by all providers within a 30-mile radius.	CMS
	Relative Procedure Volume	The relative volume of the provider's procedures, normalized to [0,1].	CMS
	Number of HCPCS Beneficiaries	Number of beneficiaries for all HCPCSs for the provider.	CMS
Prescriber Information	Prescription Types	The types of drugs prescribed by the provider (brand and generic names handled separately).	CMS
	Number of Rx Beneficiaries	Number of Medicare beneficiaries from the prescriber dataset.	CMS
Hospital Affiliations	Affiliated Hospital Score	The maximum score from the provider's hospital affiliations, where the score of each hospital affiliation depends upon the provider's specialties and U.S. News scoring of hospitals.	CMS (to get hospitals) and U.S. News (for score of hospitals)
	Number of Affiliated Hospitals	Number of hospital affiliations for the provider.	
Insurance	Number of Accepted Insurances	Number of insurers accepted by the provider.	Vitals.com and Healthgrades.com
	Individual Insurers	A set of attributes, one for each insurer accepted by the provider, e.g., Humana.	Vitals.com and Healthgrades.com
Medical Experience	Medical School Rank	Ranking of the provider's medical school by primary care rating.	CMS and U.S. News
	Years of Experience	The difference between 2014 and the year the provider graduated medical school.	CMS
	Credentials	The provider's credentials, e.g., MD, DO, FACP, etc.	CMS
	Residency Rank	Ranking of the provider's residencies by primary care rating.	Healthgrades.com and U.S. News
	Fellowship Rank	Ranking of provider's fellowships by primary care rating.	Healthgrades.com and U.S. News
	Number of Residencies	Number of the provider's residencies.	Healthgrades.com
	Number of Fellowships	Number of the provider's fellowships.	Healthgrades.com
Disciplinary Information	Number of Malpractices	Number of malpractices of the provider.	Healthgrades.com
	Number of Sanctions	Number of sanctions of the provider.	Healthgrades.com
	Number of Board Actions	Number of disciplinary board actions of the provider.	Healthgrades.com
Average Ratings from Patient Reviews	Patient Review Ratings	A set of attributes based on user reviews: Overall Rating, Ease of Appointment, Follows Up After Visit, Promptness, Spends Time with Me, Courteous Staff, Bedside Manner, and Accurate Diagnosis.	Merge reviews from Vitals.com and Healthgrades.com
	Number of Patient Reviews	Number of patient reviews for the provider.	Vitals.com and Healthgrades.com

The Referral Frequency attribute is log transformed as its distribution is observed to be exponential; we then normalize Referral Frequency to the interval [0,1]. Analogous transformations are applied to the Relative Cost of Procedures and Relative Procedure Volume attributes. Years of Experience and all variables with the prefix "Number" are represented as numeric attributes. A few of the attributes are single binary variables, such as electronic prescriptions (eRx) and Accepting New Patients. Attributes that appear as combinations are represented as sets of binary attributes, including Credentials, Specialties, Languages, Procedure Types, Prescription Types, and Individual Insurers. Methods for computing values for Medical School Rank, Residency Rank, Fellowship Rank, and Affiliated Hospitals' Score are described in the next subsection.

5.3.2. Entity Mappings

The names of medical schools and hospitals listed by U.S. News differ from the names in the CMS data. E.g., "University of California, Riverside," "University of California — Riverside" and "UC Riverside" all refer to the same school. Therefore, we used a string edit distance metric—the minimum number of operations (insert and delete) to transform one string into another string—to map CMS names to U.S. News names for all medical schools and hospitals with more than 100 occurrences; each of these mappings were then manually reviewed as some results were incorrect or no mappings exist (as in cases where a medical school is located outside of the U.S. or a hospital is not listed by U.S. News). This generated 231 medical school mappings and 2029 hospital mappings. The medical school mappings were then used to assign values for each provider's Medical School Rank, Fellowship Rank, and Residency Rank, where null (unknown value) is used for providers

whose medical schools are missing from the mappings.

The hospital rankings listed by U.S. News scores hospitals across several specialties for adults and children; for each hospital listed, the hospital's score, name, location, and rankings were collected. Further, the hospital specialties reported by U.S. News do not always correspond to the specialties listed by CMS. In particular, CMS uses a taxonomy of medical specialties that consider subspecialties whereas U.S. News uses broad categories for specialties [133]. Note that this mapping is not necessarily one-to-one; e.g., a provider specializing in internal medicine may map to several categories listed by U.S. News. Therefore, we manually mapped all specialties with more than 100 occurrences to the specialties used by U.S. News. CMS specialties are self-selected by providers; 195 of the 653 specialties have less than 100 providers. These rare specialties included technicians (e.g., Biomedical Engineering), therapists (e.g., Poetry Therapist), Clinical Nurse Specialists (a majority of nurses are marked as practitioners instead of specialists), and Molecular Genetics. This generated 5651 mappings. We then used these mappings to assign scores to each of the affiliated hospitals. For each affiliated hospital, we compute the average score of the hospital with respect to the provider's specialties as a hospital's score varies by specialty. We then assign Hospital Affiliation Score to the hospital affiliation with the maximum score (i.e., the best affiliation), where null values are used for providers whose hospital affiliations are missing from the hospital mappings.

Several attributes were collected from our crawlers, including Castle Connolly Award, Accepting New Patients, language, fellowship, residency, disciplinary actions, and patient reviews information. Thus for each provider, we mapped their CMS data to Vitals.com

and Healthgrades.com provider profiles. In particular, we mapped 608,935 providers between CMS, Vitals.com, and Healthgrades.com; 25,514 of whom have received a Castle Connolly award. To map CMS providers to providers from other sources, we followed a hybrid automatic-manual data integration approach. First, we identified a promising set of attributes to use for mapping, specifically: first name, middle name, last name, address, medical school, graduation year, affiliated hospitals, and specialties. For each attribute we constructed a customized mapping algorithm. For example, the mapping between first names is computed using the Levenshtein distance between the two strings; medical schools and hospitals used their respective mappings. Then, we assigned weights to each attribute's matching score based on a large number of accuracy experiments, where the authors defined the ground truth mappings. We then computed a mapping threshold based on the mapping scores via more accuracy experiments. We obtained a precision of 100% and a recall of 94% for our Vitals.com mapping, and a precision of 98% and a recall of 93% for our Healthgrades.com mapping.

5.3.3. Attributes Analysis and Classification Methods

We examined the information gain and correlation of each of the attributes from Tables 1 and 2 with respect to Castle Connolly Award and Referral Frequency. Information gain is used to filter the set of attributes such that only discriminative attributes are correlated and employed for classification. We then mined rules using RIPPER, a rule learning algorithm, and classified Castle Connolly Award and Referral Frequency to validate the selected attributes [134]. Rule learning algorithms (e.g., RIPPER) are employed to discover relationships between attributes in large data sets; for example, given a dataset of

transactions at a supermarket, a rule learning algorithm discovers which items are commonly bought together. Weka, an open source set of tools for data mining, was employed in each of our analyses [99].

As expected, we found that the data is highly imbalanced for both Castle Connolly Award and Referral Frequency. Only 4% of all mapped providers have received a Castle Connolly award and 42% of all mapped providers have zero referrals; a majority of providers with zero referrals specialized in Internal Medicine, Family Medicine or Emergency Medicine. This imbalance poses computational challenges for rule learning and hinders trivial classifiers. Further, only analyzing the data at the national level will omit local trends, such as state-wide Electronic Health Record (EHR) and eRx incentive programs. Thus we stratified our original dataset by each provider's state and perform our rule learning and classification tasks at both the national and state levels. Intuitively, attributes that may be discriminative in California are not the same attributes that are discriminative in New York. Moreover, healthcare is regulated both at the state and federal levels. These regulations, along with demographics and population health, create localized trends in healthcare.

We investigated the classification task using random forests and 5-fold cross-validation. Random forests has been shown to work well on imbalanced datasets [135, 136]. We applied cost-sensitive training to each classifier, where each example is weighted based on its output label. Thus, the model treats errors from each class label equally. For example, given 100 training examples with two classes, an even split would have 50 positive examples and 50 negative examples; however, if only 4 examples are positive, then applying a weight of 50/4 = 12.5 for each positive example, and a weight of 50/96 = 0.52

for each negative example will yield a cost-sensitive dataset where both the positive and negative examples are treated equally. Further, cost-sensitive training allows each classifier to make meaningful classifications; otherwise a classifier could simply guess false for Castle Connolly Award and obtain a precision of 96% and a sensitivity of 0%.

Each experiment used a 5-fold cross validation for training and testing purposes. In all experiments we set the number of trees to 20, the maximum depth to 1 + (0.01 * n) and number of features to 1 + (0.025 * n), where n is the number of features. These parameters, which are modeled after the default parameters, were chosen using a validation phase, where we enumerated different combinations of all three parameters and validated the settings on three randomly selected states; we repeated the random selection of states ten times for each combination. As noted in the methods, we used cost-sensitive training datasets, that weigh each example based on its class label, to avoid trivial classifiers (e.g., always classifying Castle Connolly Award=false yields a classifier with 96% accuracy).

5.4. Results

In this section we report the results of our analyses for Referral Frequency=Very High and Castle Connolly Award=true. First we report some general statistics on Castle Connolly Award=true and Referral Frequency=Very High. Next we report correlations between Referral Frequency and Castle Connolly Award, along with correlations of attributes. Last, we present a summary of our classification results. Detailed rule learning results are reported in Appendix K.

5.4.1. General Statistics of Providers

First we analyzed some general statistics and demographics of providers at the national level; demographics of providers are presented in Appendix G. Figure 5.2(A-D) presents the distributions of Years of Experience, Number of Affiliated Hospitals, Number of Organization Members, and Number of Patient Reviews for all providers, Castle Connolly Award=true, and Referral Frequency=Very High. Several interesting observations may be made from Figure 2. Firstly, providers that receive many referrals are likely to have at least a decade of experience or they are likely to be affiliated with several hospitals; however, patient review frequency and organization size have less of an impact on referral frequency. On the other hand, a provider is more likely to receive a Castle Connolly award if she or he has over 10 years of experience, works at a larger organization, and receives at least 1 or more reviews online. Assuming the average age of a student entering medical school is 22, that medical school requires four years of training, a majority of providers with a Castle Connolly award are between the ages 46 and 66.

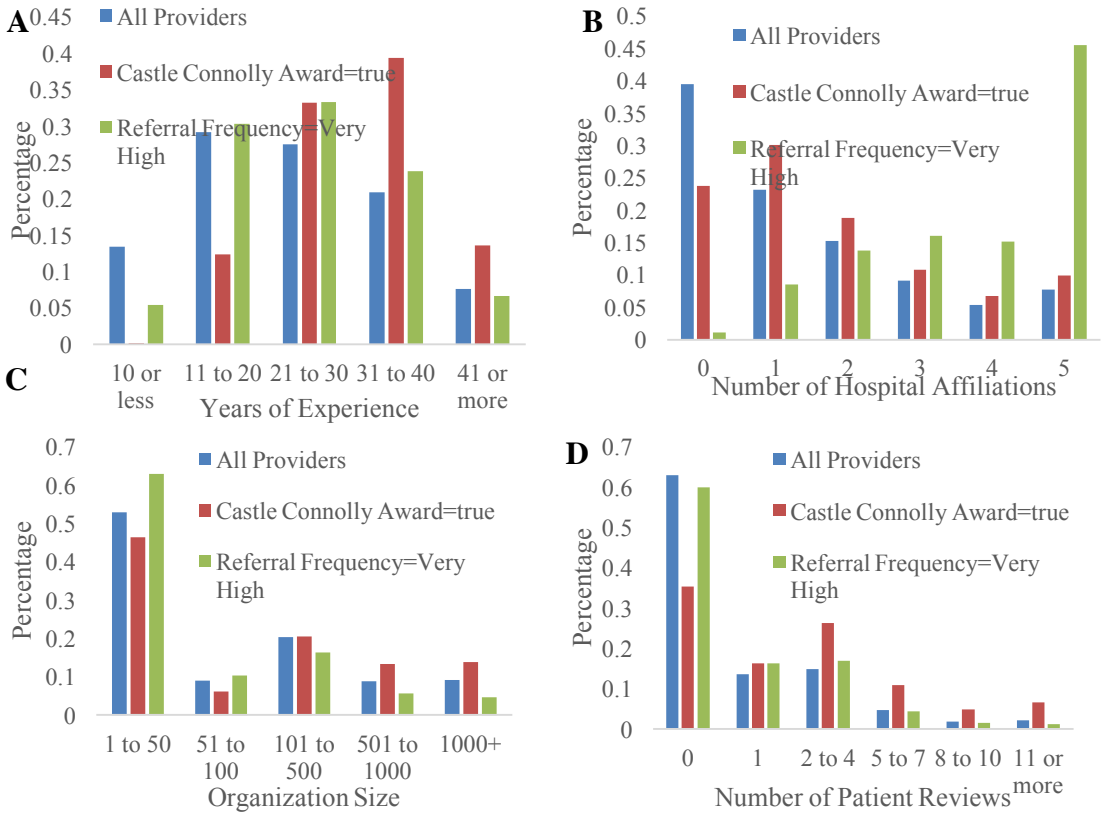


Figure 5.2. Distributions of YearsExp, NumHospitals, NumOrgMembers, and NumReviews for all providers, Castle Connolly Award=true, and Referral Frequency=Very High.

Table 5.3 lists the top 10 specialties ranked by the proportion of providers who have Referral Frequency=Very High; Wilcoxon signed-rank tests showed all differences to be significant with p less than 0.001. As expected, radiology and its subspecialties have a high concentration of providers who are referred frequently. Interventional cardiology and internal medicine is the only top 10 specialty not related to radiology; this is likely because heart disease is the leading cause of death for both men and women in the U.S. [137]. Further, interventional cardiology and internal medicine accounts for over 23% of providers with Referral Frequency=Very High.

Table 5.3. Top 10 specialties ranked by the proportion of providers who have Referral Frequency=Very High.

Specialty	Total Number of Providers for Given Specialty	Percentage of Referral Frequency=Very High within Given Specialty
Diagnostic Ultrasound	760	52.1%
Body Imaging	1076	51.9%
Neuroradiology	1725	50.7%
Diagnostic Neuroimaging	169	49.1%
Vascular and Interventional Radiology	1558	48.8%
Diagnostic Radiology	15,957	48.1%
Nuclear Radiology	1011	47.5%
Pediatric Radiology	629	45.3%
Nuclear Cardiology	7991	47.5%
Interventional Cardiology and Internal Medicine	3817	40.5%

Table 5.4 lists the top 10 specialties ranked by the proportion of Castle Connolly awards within the respective specialty; Wilcoxon signed-rank tests showed all differences to be significant with p less than 0.001. Pediatric and oncology specialists have higher rates of Castle Connolly awards than general specialties, such as internal medicine with a rate of 2% or family medicine with a rate of 1%. However, internal medicine has the highest number of Castle Connolly awards, accounting for 9.8% of all Castle Connolly awards.

Table 5.4. Top 10 specialties ranked by the proportion of Castle Connolly awards within the respective specialty.

Specialty	Total Number of Providers for Given Specialty	Percentage of Castle Connolly Award=true within Given Specialty
Gynecologic Oncology	980	29.6%
Pediatric Surgery	926	24.8%
Reproductive Endocrinology	429	12%
Pediatric Urology	97	23.7%
Oncology Surgery	1021	22.3%
Pediatric Nephrology	550	21.8%
Otology and Neurotology	205	20.9%
Colon and Rectal Surgery	1445	20.0%
Pediatric Pulmonology	1081	18.8%
Pediatric Endocrinology	1058	18.8%

5.4.2. Attribute Correlations and Discriminative Power

We computed the correlation of Referral Frequency and Castle Connolly Award=true, along with the average number of referrals for Castle Connolly Awards. We found that the Pearson correlation of Referral Frequency and Castle Connolly Award is positive, but very low, specifically 0.058. However, this low correlation is not surprising as Castle Connolly Award reflects peer recognition whereas Referral Frequency reflect patient volume. Further, a provider with high volume may not necessarily be recognized as an outstanding provider, or an outstanding provider may not necessarily have high volume. For example, a provider may receive a referral because he or she is prompt to perform a test and has an efficient office, and not necessarily because he or she is an outstanding provider. Hence, high referrals and peer awards can be viewed as just two of the possible quality indicators, describing different quality aspects.

Table 5.5 reports strong and negligible correlations of attributes with respect to referral frequency. Several of these correlations are due to the nature of referrals, thus we focus on nonobvious correlations. Unexpected correlations include:

(1) User ratings and number of reviews are negligibly correlated with referral frequency. Hence, referrals are more likely based on physician-to-physician trust, and establishing relationships with other physicians could be more important than being popular with patients.

(2) Referral Frequency is strongly correlated with the number of affiliated hospitals and the total number of affiliations is more important than the score of the respective affiliations.

(3) Years of experience and insurance information are negligibly correlated with referral frequency. That is, simply accepting more insurance plans or practicing medicine for a longer period of time is not sufficient to secure more referrals.

We also examined correlations of Referral Frequency=Very High at the state level with the aim to observe local trends in providers with frequent referrals, as reported in Appendix H.

Table 5.5. Selected correlations of attributes with respect to referral frequency. The p-value for all correlations is less than 0.01, except for the ones with an asterisk.

Strong Correlations	**Correlation**
HCPCS: Initial Hospital Care	0.46
Number of Hospital Affiliations	0.43
Number of HCPCS Beneficiaries	0.33
Relative Procedure Volume	0.26
HCPCS: X-ray Exam of Abdomen	0.24
Number of Rx Beneficiaries	0.23
Internal Medicine and Cardiovascular Disease	0.22
Diagnostic Radiology	0.21
Family Medicine	-0.20
Obstetrics and Gynecology	-0.16
Number of Fellowships	0.10
Negligible Correlations	
Hospital Score	-0.02
Patient Review Ratings	[-0.01,0.01]*
Number of Patient Reviews	-0.01
Number of Accepted Insurances	0.02
Years of Experience	0.01
Medical School Rank	0.03
Individual Insurers	[-0.4,0.5]*
Residency Rank	0.01*

A majority of attributes have negligible correlations (less than or equal to 0.05) with respect to Castle Connolly Award=true, except for those attributes listed in Table 5.6. This table suggests that providers with Castle Connolly awards have a diverse set of attributes; however, providers that see new patients or speak multiple languages are more likely to have a Castle Connolly award. We report state-level correlations of Castle Connolly Award=true in Appendix H, which reports a correlation for female gender in nine states.

Table 5.6. Attributes with a correlation greater than 0.05 with respect to Castle Connolly Award=true. The p-value for all correlations is less than 0.001.

Attributes with correlation greater than 0.05	Correlation
HCPCS: New Office/Outpatient Visit	0.13
Language=Spanish	0.08
Insurance=Aetna Health	0.06
Number of Spoken Languages	0.06

Table 5.7 reports the top 10 most discriminative attributes for Castle Connolly Award in terms of information gain. This table suggests that whether a provider has a Castle Connolly award may be discriminated by the quantity of an attribute rather than the value of the attribute. E.g., the number of patient reviews of a provider is more discriminative than the review scores; the number of fellowships and residencies is more discriminative than the institution rankings. The top 10 most discriminative attributes for Referral Frequency are reported in Appendix I.

Table 5.7. The top 10 most discriminative attributes for Castle Connolly Award in terms of information gain.

Most Discriminative Attributes for Castle Connolly Award
Number of Fellowships
Years of Experience
Number of Patient Reviews
HCPCS: New Office/Outpatient Visit
Number of Residencies
Accepting New Patients
Number of Organization Members
Number of Accepted Insurances
Family Medicine
Number of Spoken Languages

5.4.3. Classification Results

We evaluated classifiers at the national level and state level using the parameters from the methods for both Referral Frequency and Castle Connolly Award. In both cases, state-by-state classifiers outperformed national classifiers; state-level results are reported in Appendix J. Thus, finding discriminative attributes to classify Castle Connolly providers

or providers with high referral frequency is easier using attributes at the local level, and these local influencers should be modeled in each classifier separately.

Table 5.8 reports the confusion matrix for the discretized Referral Frequency classifier at the national level. For Referral Frequency=Very High, we observed an accuracy of 96%, sensitivity of 52%, specificity of 98%, and a positive predictive value of 78%. A majority of errors (Type I and Type II) were either classified as or labeled as Referral Frequency=High. Errors for other categories were similar, where a majority of errors occurred relative to the ordering of categories; compare Referral Frequency=Low with Referral Frequency=Very Low and Referral Frequency=High. Thus, provider referral frequency may be discretized and classified at the national level, with reasonable accuracies due to the correlations of attributes with referrals frequency.

Table 5.8. Confusion matrix of discretized Referral Frequency at the national level.

Classified as →	Referral Frequency= None	Referral Frequency= Very Low	Referral Frequency= Low	Referral Frequency= High	Referral Frequency= Very High
Referral Frequency= None	225,329	22,576	4652	3603	462
Referral Frequency= Very Low	7289	22,637	11,136	504	1
Referral Frequency= Low	5540	18,936	60,688	16,107	26
Referral Frequency= High	2187	2347	31,522	131,589	4916
Referral Frequency= Very High	219	92	350	16,789	19,260

Table 5.9 reports the confusion matrix for the Castle Connolly classifier at the national level. Based on this table we observed a balanced sensitivity, specificity, accuracy, and

precision, 77%. However due to the large number of false negatives, our positive predictive value is not as promising at 13%; although a trivial classifier would have a positive predictive value of 0%. Hence peer awards are difficult to predict based on the attributes of a provider. State-level classifiers observed more accurate results, as reported in Appendix J.

Table 5.9. Confusion matrix of Castle Connolly Award at the national level.

Classified as →	Castle Connolly Award=false	Castle Connolly Award=true
Castle Connolly Award=false	448,689	130,927
Castle Connolly Award=true	5791	19,623

5.5. Discussion

Our results have demonstrated and identified several attributes that are both correlated and discriminative for providers who are frequently referred. Further, we showed that most correlations are negligible with Castle Connolly awards at the national level, which suggests that a one-size-fits-all approach to ranking providers is inadequate. However, we demonstrated that these attributes are indeed discriminative for both referral frequency and Castle Connolly awards via rule learning and classification, and that these attributes are better discriminators at the state level due to local influencers. Hence, provider search portals should not use a global ranking formula across the whole country or across all specialties, but instead learn different weights for each attribute based on the user's location or provider's specialty.

Moreover, our findings have consistently demonstrated that the frequency of an attribute is more important than the value of an attribute—e.g., the number of reviews of a provider is more important than the individual review ratings. Thus, current filters for provider

search portals, such as medical school ranking, patient review rating, or hospital affiliation ranking, do not necessarily determine quality. Instead, emphasis should be placed on the number of reviews, fellowships, residencies, insurers, or hospital affiliations. The implication of these results is that quality of care is affected by providers who have a more diverse set of experiences and access to a larger set of services. Expanding services and increasing experience can be achieved through accepting more insurance plans and increasing hospital affiliations. Income is directly tied to rates of mortality, morbidity, and access to healthcare; thus accepting a wider range of insurance plans will expose the provider to a more diverse set of patients and episodes [138]. Further, hospital affiliations usually require an existing relationship—where leadership alignment promotes the collaboration. Thus, best practices are shared, along with an expansion of services in a cost-effective manner [139]. Lastly, providers who encourage patients to author reviews will have a more comprehensive picture of their skills online, even if they are a 3 or 4-star doctor. As the 5-star doctor with a handful of reviews may have solicited these reviews from family and friends, and thus the 5-star rating is inaccurate.

The locality of quality factors should also be captured when ranking providers, as pointed out in the Appendix. For example, states with higher rates of Castle Connolly awards suggest more nominations, and hence more providers seek peer-review processes such as accreditation programs, which have been shown as tools to increase quality of care [140]. Similarly, demographics and credentials affect referral rates and Castle Connolly awards. E.g., nine states report correlations between females and Castle Connolly awards whereas zero states report correlations for males, and 50 of 51 states (including Washington D.C.)

have correlations with pediatricians. Our rule learning results show that factors such as specific prescriptions or procedures affect referral frequency by locality, and varying years of experience and organization size affect Castle Connolly awards by locality.

Hence patients should be educated on the local factors that determine provider quality within their community, and patients should be made aware of the pitfalls of existing filters in provider search portals. For example, patients should compare the number of hospital affiliations of each provider with the average number of hospital affiliations of providers in the patient's community, and patients should be aware that a majority of patient reviews are scored based on wait times and visit durations. This education would allow provider search portals to highlight younger providers with less years of experience who have attributes in common with older providers who have high marks in quality. Hence our work enables provider search portals to empower patients and to "load balance" patients between younger and older providers without sacrificing quality of care.

The next stage of this research will include more performance measures and patient survey data as they are made available by CMS and other sources. We expect performance measures to correlate with quality, and hence these measures should improve the accuracy of our inferences and predictions. We also plan to integrate organizational attributes into our algorithms, such as payment data and performance measures of hospitals. For example, CMS has released surveys of patients' experience with hospitals, which reports hospital-level attributes such as doctor and nurse communication, cleanliness of hospital environment, and willingness to recommend the hospital [3, 141]. Integrating organizational data and performance measures will enable us to build a provider reputation

rating system, where, for each provider, we identify attributes that would improve the provider's reputation.

5.5.1. Limitations

A limitation of this chapter is that our results are tied to CMS, Vitals.com, and Healthgrades.com data. This analysis depends on successfully mapping between these data sources, and the accuracy of these data sources is not guaranteed; e.g., errors made by an optical character recognition program—a popular method for amassing data from PDF files—will create inaccurate data. Moreover, attributes change over time. Consider a provider who moves to a new office and updates his or her address with CMS, but Vitals.com has yet to process the update. Thus, these two sources become inconsistent and mappings are unsuccessful as location is a critical factor when mapping providers. Other attributes that become inconsistent over time include: last name, subspecialties, and hospital affiliations. Further, providers who do not participate in Medicare and Medicaid will have several missing attributes, and referrals outside of Medicare and Medicaid are omitted. However, we collected data on and successfully mapped 608,935 providers. Another limitation is that a majority of providers have zero reviews; this is likely due to the fact that only 4% of Internet users post online reviews for providers, and previous work has shown that most providers have zero reviews [24].

Another limitation is the usage of referral frequency and Castle Connolly awards as quality indicators. Firstly, these indicators are not comprehensive—CMS has defined measures for physician quality via PQRS, but this data is currently not publicly available at the provider level. Further, PQRS measures are condition specific, and while this information

is useful for a provider search portal, our analysis focused on a condition insensitive analysis of provider quality. We understand that the number of referrals greatly depends on the specialty; normalizing this number by the specialty could potential lead to another quality measure. Further, while the Castle Connolly award is prestigious and rigorously vetted, the award is biased towards providers who have more experience, because providers with more experience have had more time to build their reputation. However, our results show that several other attributes are also discriminative and years of experience alone does not determine a Castle Connolly designation.

5.6. Conclusion

We studied which attributes from a provider's profile correlate with and discriminate against referral volume and peer-nominated awards. Our findings have shown that a one-size-fits-all approach to provider ranking is inadequate, and that local influencers on provider quality must be considered when ranking providers. In turn, patients should be aware of the pitfalls of current provider search portals, and patients should be educated on the local factors influencing provider quality. Provider search portals that integrate these findings effectively will empower patients and enables these portals to "load balance" patients between younger and older providers without sacrificing quality of care.

Chapter 6. Conclusion

This dissertation has presented novel algorithms and knowledge discovery techniques that solve computational challenges at the intersection of healthcare and computing. Chapter 2 presented an efficient and scalable algorithm for computing semantic similarity between two documents, where each document is represented as a set of medical concepts from an ontology. Our algorithm is applicable to both relevance and similarity queries, which are frequently encountered when utilizing an EHR database. Our algorithm reduces the complexity of a naïve approach from $O(n^2)$ to $O(nlogn)$ by using a variation of the Radix Tree. Our early-termination algorithm to search for the top-k most relevant or similar documents avoids redundant calculations following a branch and bound approach. Our experimental evaluation on real clinical data showcased the advantages of our methods in terms of efficiency and scalability.

Chapter 3 presents a promising technique for predicting future medical concepts in a patient's EHR by leveraging a database of similar EHRs. Using a database of real patients, we evaluated three types of inter-patient similarity measures and identified two important parameters that influence the accuracy of our technique. This evaluation revealed limitations to our approach but also identified key steps for future research.

Chapter 4 presented a detailed analysis and presented a pipeline that extends existing techniques for OSNs with pharmaceutical drug discussions. We analyzed each OSN against four distinguishing characteristics, and we demonstrated that these characteristics affect the type of discussions, the type of drugs discussed, the subjectivity of discussions, and the medical concept content. We synthesized these findings and proposed actionable

items for both healthcare providers and future researchers of healthcare discussions on OSNs.

Lastly, Chapter 5 presented a data-driven analysis of which attributes from a provider's profile correlate with and discriminate against referral volume and peer-nominated awards. Our findings show that a one-size-fits-all approach to provider ranking is inadequate, and that local influencers on provider quality must be considered when ranking providers. In turn, patients should be aware of the pitfalls of current provider search portals, and patients should be educated on the local factors influencing provider quality. Provider search portals that integrate these findings effectively will empower patients and enable these portals to load balance patients between younger and older providers without sacrificing quality of care.

Appendices

Appendix A. Online Social Network and Drug Summary

A.1. Online Social Network Summary

Table A.1 lists each of the ten Online Social Networks (OSNs) investigated in this appendix with their respective website, and the start and end dates of posts collected from each OSN. Not every OSN marks posts with timestamps, therefore these networks were marked with the date they were crawled.

Table A.1. An overview of the OSNs analyzed in this appendix. The start and end dates represent the timestamp of the first and last post from each dataset. An asterisk denotes the date an OSN was crawled for OSNs that do not mark posts with an exact timestamp.

Dataset	URL	Start	End
Twitter	www.twitter.com	Dec. 29, 2012	Jan. 15, 2013
Google+	plus.google.com	Jan. 1, 2011	Jan. 31, 2013
Pinterest	www.pinterest.com	N/A	Feb. 11, 2013*
Daily Strength	www.dailystrength.org	N/A	Jan. 15, 2013*
Drugs.com	www.drugs.com	Apr. 2, 2007	Jan. 23, 2013
DrugLib.com	www.druglib.com	N/A	Feb. 11, 2013*
everydayHealth	www.everydayhealth.com	Jan. 2, 2001	Jan. 31, 2013
MediGuard	www.mediguard.org	Jan. 21, 2007	Jan. 31, 2013
medications	www.medications.com	N/A	Feb. 13, 2013*
WebMD	www.webmd.com	Sept. 18, 2007	Jan. 19, 2013

A.2. Drug Summary

Tables A.2, A.3 and A.4 list the most popular drugs by prescriptions dispensed, as given on RxList.com [142]. Each of these drugs was classified into one or more drug groups, according to the drug taxonomy available on Drugs.com [91]. Each drug is associated with one or more categories.

Table A.2. Listing of drugs that were classified as Gastrointestinal Agents, Genitourinary Tract Agents, Topical Agents, Alternative Medicines, Nutritional Products, and Coagulation Modifiers.

Gastrointestinal Agents	Genitourinary Tract Agents	Topical Agents	Alternative Medicines	Nutritional Products	Coagulation Modifiers
Famotidine	Cialis	Mupirocin	Lovaza	Folic	Plavix
Nexium	Detrol	Nasonex		Klor-Con	Warfarin
Omeprazole	Viagra	Premarin		Niaspan	
Pantoprazole		Xalatan			
Ranitidine					

Table A.3. Listing of drugs that were classified as Hormones, Anti-infectives, Psychotherapeutic Agents, and Respiratory Agents.

Hormones	Anti-infectives	Psychotherapeutic Agents	Respiratory Agents
Levothyroxine	Amoxicillin	Abilify	Advair
Levoxyl	Azithromycin	Amitriptyline	Albuterol
Loestrin	Cefdinir	Citalopram	Cheratussin
Methylprednisolone	Cephalexin	Cymbalta	Combivent
NuvaRing	Ciprofloxacin	Effexor	Fexofenadine
Ocella	Doxycycline	Fluoxetine	Flovent
Prednisone	Fluconazole	Lexapro	Fluticasone
Premarin	Levaquin	Paroxetine	Hydrocodone
Synthroid	Penicillin	Seroquel	Proair
TriNessa	Sulfamethoxazole	Sertraline	Promethazine
		Trazodone	Proventil
		Zyprexa	Singulair
			Spiriva
			Ventolin

Table A.4. Listing of drugs that were classified as Metabolic Agents, Cardiovascular Agents, and Central Nervous System Agents.

Metabolic Agents	Cardiovascular Agents	Central Nervous System Agents
Actonel	Amlodipine	Alprazolam
Actos	Atenolol	Ambien
Alendronate	Benazepril	Amphetamine
Allopurinol	Benicar	Aricept
Crestor	Carvedilol	Carisoprodol
Glyburide	Clonidine	Celebrex
Januvia	Digoxin	Clonazepam
Lantus	Diltiazem	Concerta
Lipitor	Diovan	Cyclobenzaprin
Lovastatin	Enalapril	Diazepam
Metformin	Flomax	Gabapentin
Niaspan	Furosemide	Hydrocodone
Pravastatin	Hydrochlorothiazide	Ibuprofen
Simvastatin	Isosorbide	Lorazepam
Tricor	Lisinopril	Lyrica
Vytorin	Metoprolol	Meloxicam
Zetia	Toprol	Namenda
	Triamterene	Naproxen
	Verapamil	Oxycodone
		Oxycontin
		Promethazine
		Propoxphyene
		Suboxone
		Tramadol
		Vyvanse
		Zolpidem

Figure A.1 visualizes the distribution of the drug categories listed in Tables A.1, A.2, and A.3. Central nervous system agents, cardiovascular agents, and metabolic agents make up roughly fifty percent of the drugs investigated in this appendix.

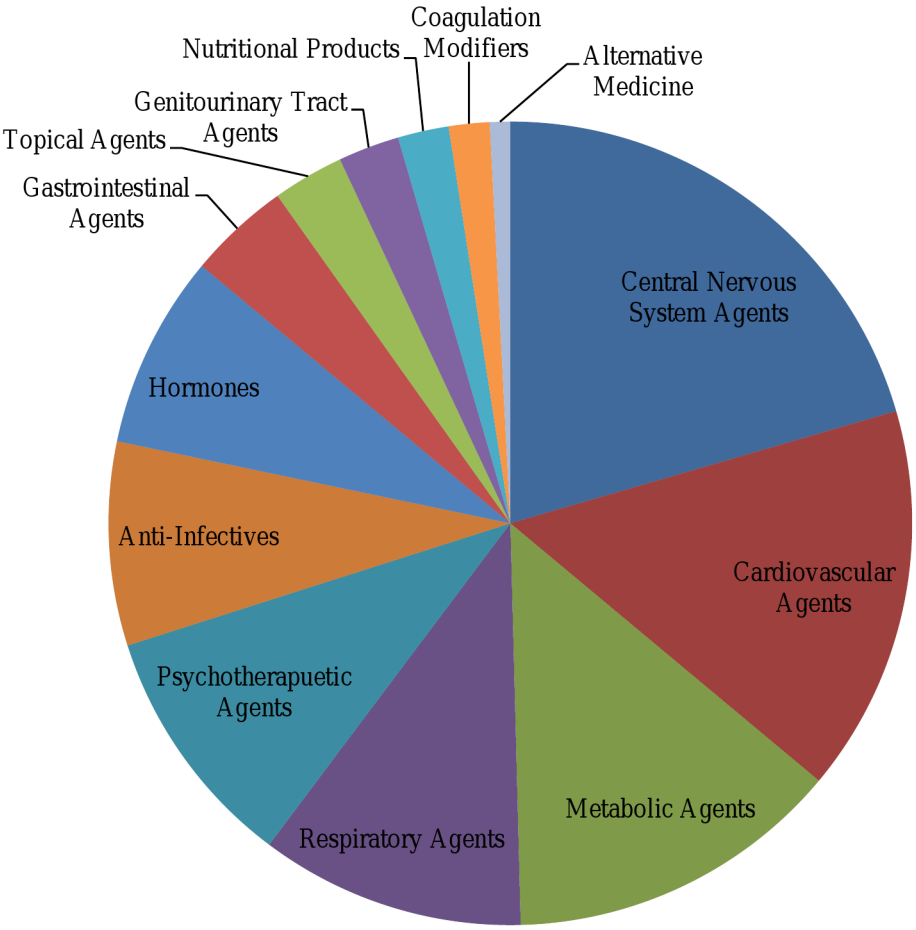


Figure A.1. Distribution of drug categories for the list of drug names, as classified by the Drugs.com taxonomy.

Appendix B. General Statistics and Medical Concept Statistics

Table B.1 reports the number of posts, number of unique posts, posts per day, and the average length of each post. General OSNs tend to have many more duplicate posts than health OSNs due to advertisements and reposting of content. MediGuard is an exception since all drug reviews for a particular drug are listed under the brand name, and its search feature has up-to-date information on generic to brand drug name mappings.

General OSNs such as Twitter, Pinterest, and Google+ contain many more posts over a shorter period of time than the health OSNs; compare 354,128 posts in Twitter over

eighteen days with 5,451 posts in WebMD over five years. This difference is also emphasized by the number of posts per day. However, the average length of a post from a general OSN is much smaller than the average length of a post in all health OSNs, with the exception of DailyStrength. This is due to the nature of drug reviews in DailyStrength – a majority of reviews are short phrases such as "works for me" or "doesn't work".

Table B.2 summarizes the medical concept content of each OSN in terms of the number of medical concepts per post and per word. These results were computed across all concepts in a given post and for medical concepts that are unique in a given post. Except for DailyStrength, each health OSN contains a higher number of concepts per post, but the concentration of medical concepts per word is higher in general OSNs than health OSNs. These results coupled with the observations from Table B.1 suggest that users are sharing stories about their experiences with a particular drug in health OSNs; whereas users in general OSNs are expressing shorter thoughts with more medical concepts, such as advertisements, news, educational material, or jokes. Again, the only exception is DailyStrength.

Table B.1. General statistics for each of the OSNs. The total number of posts, total number of unique posts, average posts per day, and average words per post are given.

Dataset	Total Posts	Unique Posts	Percent Unique	Avg. Posts per Day	Std. Dev.	Avg. Words per Post	Std. Dev.
Twitter	354,128	284,387	80.3%	19,673	8,138	13.7	5.0
Google+	11,803	8,706	73.7%	15.5	25.7	39.6	70.1
Pinterest	8,706	5,876	66.5%	N/A	N/A	24.9	33.2
DailyStrength	81,514	72,522	88.9%	N/A	N/A	15.7	13.4
Drugs.com	5,451	4,994	91.6%	2.4	2.3	64.5	41.8
DrugLib.com	974	959	98.4%	N/A	N/A	121.4	84.1
everydayHealth	852	820	96.2%	0.19	0.74	77.5	51.6
MediGuard	21,278	15,126	71.0%	6.9	47.6	72.8	65.4
medications	35,050	34,997	99.8%	N/A	N/A	133.9	135.6
WebMD	28,482	27,705	97.2%	14.2	7.0	61.2	60.9

Table B.2. Overview of medical concept content. The average number of concepts, total number of concepts, and the average number of concepts per word are shown; these results only consider concepts from semantic groups related to medicine.

Dataset	Avg. Concepts per Post	Std. Dev.	Avg. Unique Concepts per Post	Std. Dev.	Avg. Concepts per Word	Avg. Unique Concepts per Word	Total Concepts	Unique Concepts
Twitter	2.9	1.9	2.5	1.5	0.223	0.189	836,167	13,007
Google+	6.1	8.4	5.3	5.9	0.181	0.165	53,218	5,849
Pinterest	4.8	5.8	4.3	4.6	0.219	0.201	28,136	4,067
DailyStrength	1.9	2.2	1.8	2.0	0.130	0.127	158,669	4,820
Drugs.com	9.7	6.6	8.3	5.3	0.158	0.142	48,425	3,500
DrugLib.com	19.6	12.3	15.0	7.8	0.173	0.138	18,818	2,305
everydayHealth	11.4	8.0	9.5	6.3	0.156	0.137	9,339	1,577
MediGuard	7.6	8.5	6.1	6.2	0.110	0.095	160,660	6,308
Medications	18.7	19.0	12.7	10.8	0.150	0.119	654,340	9,169
WebMD	8.8	8.6	7.5	6.5	0.167	0.153	244,589	6,535

Appendix C. General versus Health OSNs

Table C.1 summarizes general and medical concept statistics for the two groupings of OSNs. General OSNs contain more posts with fewer words per post, but general OSNs have a smaller percentage of unique posts. Health OSNs contain more concepts per post due to their increased length, but these OSNs have fewer concepts per word. General OSNs cover more medical concepts than health OSNs.

Table C.1. Summary of general statistics and medical concept statistics for general and health OSNs.

Category	Total Posts	Unique Posts	Words Per Post	Average Concepts per Post	Avg. Concepts per Word	Unique Concepts
General	373,637	298,969 (80%)	26.1	4.6	.208	17,429
Health	173,601	157,123 (90%)	64.7	9.7	.149	13,130

Figure C.1 compares distributions of emotional pairs of the health and general OSNs with the distribution of a uniform baseline, where the baseline assumes a uniform distribution for every term mapped from the NRC word-emotion lexicon [97]. This lexicon contains over 14,000 words manually labeled by humans via crowdsourcing. Each term is assigned one or more emotional-pairs from the following set: (1) negative–positive; (2) joy–sadness; (3) anger–fear; (4) trust–disgust; and (5) anticipation–surprise. Since joy–sadness is similar to positive–negative, and we compute positive, negative, and objective scores using SentiWordNet [95], our analysis omits results for the emotional pairs joy–sadness and positive–negative from the NRC lexicon. Analogous to the SentiWordNet process, we stemmed both the posts and the terms in the NRC lexicon before computing the emotion scores. We then mapped phrases from the NRC lexicon to phrases in the posts using the longest possible match first. Next, we computed the score for each emotional-pair of each post by averaging the emotion scores from every mapped term. The final score for each emotional pair is then computed by averaging the emotion scores of all posts within a given OSN. We can see in Figure C.1(A) that 55% of the terms mapped from the NRC lexicon are related to fear, whereas 45% are related to anger. Table C.2 reports the highest absolute relative changes of each emotional pair shown in Figure C.1. Health and general OSNs

follow the same trends with respect to the baseline. Both groups observe an increase in fear, trust, and anticipation terms, and a decrease in anger, disgust, and surprise terms.

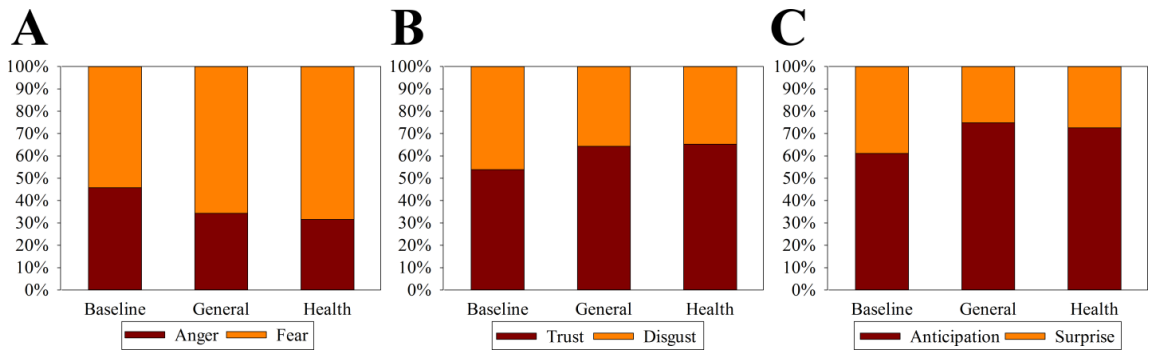


Figure C.1. An overview of the emotion analysis for general OSNs versus health OSNs. (A) The distribution of fear–anger; (B) the distribution of disgust–trust; and (C) the distribution of surprise–anticipation.

Table C.2. Highest absolute relative changes of the emotional pairs compared with the baselines shown in Figure C.1.

Emotion		
Surprise	General	-35%
Anger	Health	-31%
Surprise	Health	-29%
Fear	Health	+27%
Anger	General	-24%
Disgust	Health	-24%
Disgust	General	-22%
Anticipation	General	+24%
Trust	Health	+21%
Fear	General	+21%
Trust	General	+19%
Anticipation	Health	+18%

Frequent itemsets of medical concepts for health and general OSNs are reported in Tables C.3 and C.4 of Appendix C. General OSNs are dominated by specific drug names, where drugs with similar purposes often co-occur together. Examples form Table C.4 include**:** Ibuprofen, Tylenol, and Advil; Viagra, Cialis, Levitra, Promethazine, Xanax, and Percocet. Drugs co-occur in a single post for multiple reasons:

- Online pharmacies advertise multiple drugs that serve a single purpose; e.g., "[URL] order Viagra Cialis and Levitra in internet shop without script California !".
- Users will associate conditions with each drug from a single group; e.g., "WORSE HEADACHE EVER!!! #TYLENOL #IBUPROFEN #ADVIL"
- Popular culture references; e.g., "Xanax, Percocet, Promethazine" is a quote from a popular song.

Drugs such as Viagra and Oxycodone are often used in jokes; e.g., "oxymoron – a moron that loves OxyContin" and "Viagra for women has been around for centuries. It's called money". These drugs also appear in posts that share news articles; e.g., "Oxycontin abusers switching to heroin [URL]".

Another interesting itemset from Table C.4 is Viagra, watching, and awkward; this itemset refers to posts that discuss the awkwardness of watching Viagra commercials with one's family. Lastly, there are a series of itemsets referring to Viagra, death, sex, and myocardial infraction. These itemsets are referring to an odd news article where a man took a full bottle of Viagra and died from a heart attack after having sex for 12 hours. This news article was shared over 1300 times within our Twitter dataset.

Sleep dominates health OSNs with a frequency of over 10%. As shown in Table C.3, sleep occurs in six of the ten most frequent itemsets of size two. Several itemsets refer to drugs and the conditions they treat:

- Lisinopril and hypertension.
- Singular and allergies/asthma.

- Ibuprofen and headache.
- Sleep and Ambien.
- Cholesterol and Lipitor
- Lexapro/Cymbalta, anxiety, and depression

Itemsets of symptoms are also common to health OSNs, such as headache, dizziness, and nausea. Health OSNs also contain frequent itemsets of drugs and their side effects: Lisinopril, sleepiness, and coughing; NuvaRing and decreased Libido; Singulair and depression.

We further examined frequent itemsets of all possible concepts for both general and health OSNs, reported in Tables C.5 and C.6. These itemsets yield further insight into the types of conversations users have in each grouping of OSNs. Several itemsets for general OSNs are related to advertisements from online pharmacies; these itemsets include concepts such as Internet, mail, priority, prices, best, low, scripts, buying, fast, and milligram. An interesting itemset from Table C.5 is Viagra and herbal, which refers to advertisements such as "[URL] herbal remedy for Viagra. Overnight shipping, order today!". Lastly, Table C.6 reveals one breaking news item (at the time of data collection), where the United States Food and Drug Administration recommended lower dosages of Ambien.

Itemsets for health OSNs reveal that users are discussing their experiences with their medications, and the differing strategies employed by their physicians; the concept for physicians appears in over half of the frequent itemsets for both Tables C.5 and C.6. These posts typically discuss a problem and an action; e.g., "my doctor increased my dosage to

20mg"; and "[My doctor] put me on Lisinopril but stopped taking it after constantly coughing day and night".

Table C.3. Frequent itemsets of size 2 for medical concepts.

Health OSNs			**General OSNs**		
Depression	Anxiety	1.05%	Ibuprofen	Headache	0.61%
Lisinopril	Coughing	0.91%	Ibuprofen	Acetaminophen	0.45%
Singulair	Asthma	0.79%	Viagra	Male	0.44%
Sleep	Tired	0.75%	Ambien	Sleep	0.44%
Sleep	Sleeplessness	0.65%	Viagra	Cialis	0.41%
Sleep	Depression	0.64%	Viagra	Penile Erection	0.36%
Sleep	Eating	0.61%	Ibuprofen	Advil	0.36%
Sleep	Anxiety	0.60%	Ibuprofen	Tylenol	0.33%
Headache	Nausea	0.58%	Viagra	Sexual intercourse	0.33%
Sleep	Ambien	0.57%	Viagra	Female	0.32%

Table C.4. Frequent itemsets of size 3 for medical concepts.

Health OSNs				**General OSNs**			
Singulair	Hypersensitivity	Asthma	0.25%	Ibuprofen	Tylenol	Advil	0.13%
Sleep	Lisinopril	Coughing	0.16%	Viagra	Cialis	Levitra	0.12%
Lisinopril	Blood pressure	Coughing	0.16%	Viagra	Male	Sexual intercourse	0.07%
Sleep	Depression	Anxiety	0.15%	Viagra	Watching	Awkward	0.06%
Depression	Anxiety	Lexapro	0.14%	Promethazine	Xanax	Percocet	0.06%
Libido	NuvaRing	Sexual intercourse	0.14%	Viagra	Male	Died	0.06%
Headache	Dizziness	Nausea	0.13%	Viagra	Died	Myocardial Infarction	0.06%
Depression	Singulair	Asthma	0.12%	Viagra	Sexual intercourse	Died	0.06%
Libido	NuvaRing	Contraceptives	0.11%	Viagra	Male gender	Myocardial Infarction	0.05%
Lisinopril	Blood pressure	Hypertension	0.11%	Ibuprofen	Advil	Motrin	0.05%

Table C.5. Frequent itemsets of size 2 for all UMLS concepts.

Health OSNs			**General OSNs**		
Physicians	Started	3.83%	milligram	Internet	0.89%
Physicians	milligram	3.73%	Viagra	commercial	0.84%
milligram	Started	3.34%	Internet	Tablet Dosing Unit	0.82%
Help	Physicians	3.02%	Viagra	Hardness	0.75%
milligram	Dosage	3.01%	Internet	Scripts	0.73%
Help	Sleep	2.69%	milligram	Tablet Dosing Unit	0.67%
Help	milligram	2.63%	Prices	best (quality)	0.54%
Physicians	Dosage	2.55%	Priority	Mail	0.50%
Physicians	Better	2.25%	Viagra	Herbal	0.49%
Help	Started	2.14%	Ibuprofen	milligram	0.48%

Table C.6. Frequent itemsets of size 3 for all UMLS concepts.

Health OSNs				**General OSNs**			
Physicians	Milligram	Started	1.31%	Dosage	Ambien	US FDA	0.33%
Physicians	Milligram	Dosage	1.17%	Internet	Priority	Mail	0.20%
milligram	Started	Dosage	0.92%	Dosage	Ambien	Cut	0.18%
Help	Physicians	milligram	0.89%	Internet	Prices	best (quality)	0.17%
Help	Physicians	Started	0.88%	Viagra	Hardness	Physical findings	0.16%
Physicians	Started	Better	0.84%	Viagra	commercial	Watching	0.15%
Physicians	Started	Dosage	0.78%	Internet	Financial cost	low	0.14%
Physicians	milligram	Better	0.72%	Viagra	commercial	Awkward	0.14%
Help	milligram	Started	0.70%	milligram	Internet	Tablet Dosing Unit	0.13%
Physicians	Started	Last	0.69%	Dosage	US FDA	Recommendation	0.13%

Appendix D. Non-moderated versus Moderated Health OSNs

Table D.1 summarizes general and medical concept statistics for moderated and non-moderated health OSNs. Moderated OSNs contain many more words per post, due to their inclusion of medications and DrugLib.com, both of which contain over 120 words per post. Thus, moderated health OSNs also contain more concepts per post and cover more concepts than non-moderated health OSNs.

Table D.1. Summary of general statistics and medical concept statistics for not moderated and moderated OSNs.

Category	**Total Posts**	**Unique Posts**	**Words Per Post**	**Average Concepts per Post**	**Avg. Concepts per Word**	**Unique Concepts**
Non-moderated	110,848	101,047 (91%)	51.5	7.4	.151	7,875
Moderated	62,753	56,076 (89%)	99.3	13.9	.148	11,651

Figure D.1 reports the effect of moderation on the emotional pairs. Moderated OSNs decreased the number of disgusting terms and increased the number of trusting terms, whereas lack of moderation had the opposite effect. Otherwise, moderation had little or no effect on the emotional and medical content of drug reviews in health OSNs.

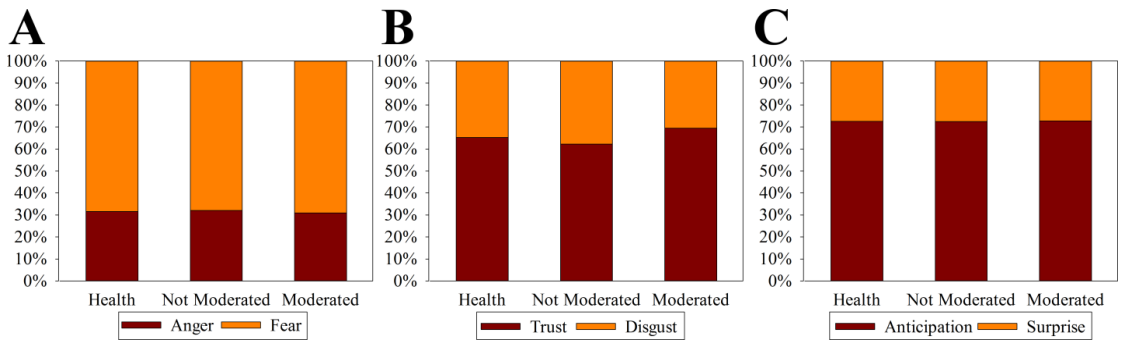


Figure D.1. An overview of the semantic group and emotion analysis for moderated and non-moderated OSNs. (A) The distribution of semantic groups; (B)-(C) the distribution of the emotional pairs fear–anger, disgust–trust, and surprise–anticipation.

Table D.2. Relatives changes of each item compared with the health OSN baseline shown in Figure D.1.

Emotion		
Disgust	Moderated	-12%
Disgust	Non-moderated	+9%

Tables D.3-5 report frequent itemsets for health OSNs with and without moderation. Sleep is common to both groupings of OSNs, but sleep is more frequent for non-moderated health OSNs. Frequent itemsets from non-moderated health OSNs concur with Figure 4(A) from Section 4.2, in that psychotherapeutic agents (Lexapro and Cymbalta), along with psychological conditions (panic attacks, mental suffering, depression, and anxiety) are frequent; whereas these drugs are not found in the frequent itemsets of moderated health OSNs, and these conditions are not as frequent. Moderated health OSNs contain concepts relating to the respiratory and cardiovascular systems, including Lisinopril, Singulair, Lipitor, asthma, coughing, hypertension, blood pressure, and cholesterol. Further, moderated health OSNs also discuss the contraceptive NuvaRing and its side effect of decreased libido.

Table D.3. Frequent itemsets of size 1 for medical concepts.

Non-moderated Health OSNs		Moderated Health OSNs	
Sleep	10.93%	Sleep	9.19%
Depression	5.11%	Lisinopril	6.44%
Weight Gain	3.70%	Singulair	6.25%
Tired	3.70%	Depression	6.01%
Anxiety	3.63%	Headache	5.66%
Headache	3.00%	Mental Suffering	5.61%
Dizziness	2.68%	Eating	5.26%
Drowsiness	2.58%	Prednisone	4.91%
Nausea	2.52%	Levaquin	4.61%
Lexapro	2.33%	Tired	4.47%

Table D.4. Frequent itemsets of size 2 for medical concepts.

Non-moderated Health OSNs			Moderated Health OSNs		
Depression	Anxiety	1.10%	Lisinopril	Coughing	1.82%
Sleep	Anxiety	0.57%	Singulair	Asthma	1.81%
Sleep	Depression	0.57%	Lisinopril	Listerine	1.51%
Sleep	Ambien	0.56%	Singulair	Hypersensitivity	1.25%
Depression	Lexapro	0.55%	Lipitor	cholesterol	1.12%
Sleep	Tired	0.52%	Sleep	Tired	1.06%
Depression	Cymbalta	0.51%	Lisinopril	Blood pressure	1.04%
Sleep	Sleeplessness	0.50%	Sleep	Depression	1.00%
Sleep	Eating	0.46%	Depression	Anxiety	0.98%
Sleep	Drowsiness	0.41%	Asthma	Hypersensitivity	0.91%

Table D.5. Frequent itemsets of size 3 for medical concepts.

Non-moderated Health OSNs				Moderated Health OSNs			
Depression	Anxiety	Lexapro	0.19%	Singulair	Asthma	Hypersensitivity	0.58%
Sleep	Depression	Anxiety	0.14%	Sleep	Lisinopril	Coughing	0.34%
Depression	Anxiety	Cymbalta	0.12%	NuvaRing	Libido	Sexual intercourse	0.32%
Headache	Dizziness	Nausea	0.09%	Lisinopril	Coughing	Blood pressure	0.31%
Sleep	Depression	Sleeplessness	0.08%	Singulair	Depression	Asthma	0.27%
Sleep	Sleeplessness	Ambien	0.08%	Singulair	Happiness	Asthma	0.25%
Depression	Anxiety	Panic Attacks	0.08%	Singulair	Mental Suffering	Asthma	0.24%
Depression	Anxiety	Mental Suffering	0.07%	NuvaRing	Libido	Contraceptives	0.23%
Depression	Weight Gain	Anxiety	0.07%	Sleep	Singulair	Asthma	0.23%
Sleep	remembering	Ambien	0.07%	Lisinopril	Blood pressure	Hypertension	0.23%

Appendix E. Registration versus No Registration for Health OSNs

Table E.1 summarizes general and medical concept statistics for health OSNs that do or do not require registration. Registration has little effect on these statistics, with the average number of words and medical concepts being roughly equal.

Table E.1. Summary of general statistics and medical concept statistics for health OSNs that do and do not require registration.

Category	Total Posts	Unique Posts	Words Per Post	Average Concepts per Post	Avg. Concepts per Word	Unique Concepts
No Registration	35,759	34,478 (96%)	81.2	12.4	.164	7,567
Registration	137,842	122,645 (89%)	74.3	9.4	.13	11,839

Figure E.1 reports the effect of registration on the emotional pairs. Overall, registration had little or no effect on the emotional and medical content of drug reviews in health OSNs.

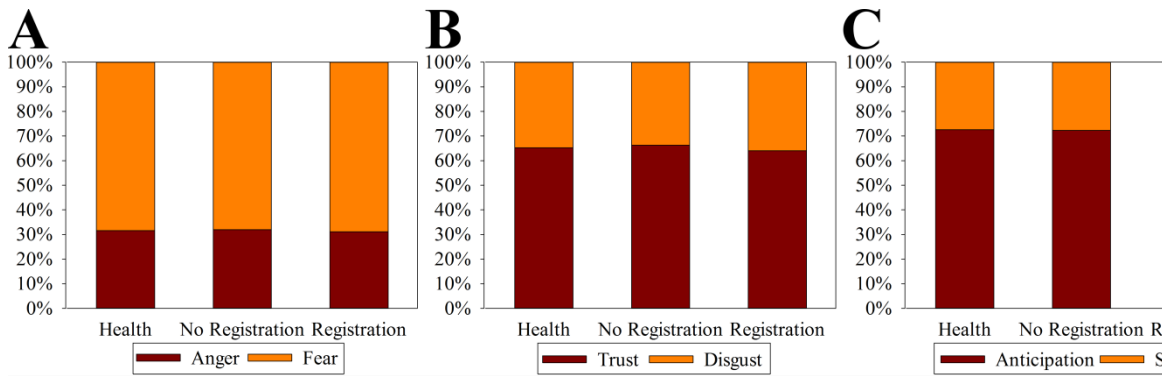


Figure E.1. An overview of the semantic group and emotional analysis for health OSNs that do and do not require registration. (A)-(C) The distribution of the emotional pairs fear–anger, disgust–trust, and surprise–anticipation.

Table E.2. Relatives changes of each item compared with the health OSN baseline shown in Figure E.1.

Emotion		
Disgust	Registration	+3%
Disgust	No Registration	+3%

Tables E.3-5 report frequent itemsets for health OSNs that do or do not require registration. Similar to moderation, sleep is common to both groupings of OSNs, but is more prevalent

in health OSNs that do not require registration. Further, concepts relating to psychotherapeutics and psychological conditions are common in health OSNs that do not require registration; analogous to the frequent itemsets for non-moderated health OSNs. Health OSNs that require registration have similar frequent itemsets to that of health OSNs with moderation, which focus on respiratory and cardiovascular drugs and conditions, such as Lisinopril, Lipitor, Singulair, asthma, and allergies. Also similar to health OSNs with moderation, health OSNs with registration have NuvaRing and its side effect libido as frequent itemsets.

Table E.3. Frequent itemsets of size 1 for medical concepts.

No Registration Health OSNs		**Registration Health OSNs**	
Sleep	12.26%	Sleep	9.51%
Depression	6.51%	Depression	4.61%
Headache	5.59%	Vision	3.93%
Tired	5.36%	Weight Gain	3.76%
Anxiety	4.92%	Headache	3.62%
Dizziness	4.83%	Personal appearance	3.61%
Happiness	4.49%	Lisinopril	3.58%
Nausea	4.42%	Tired	3.58%
Blood pressure finding	4.28%	Singulair	3.48%
Weight Gain	4.18%	Eating	3.29%

Table E.4. Frequent itemsets of size 2 for medical concepts.

No Registration Health OSNs			**Registration Health OSNs**		
Depression	Anxiety	1.78%	Lisinopril	Coughing	1.01%
Sleep	Ambien	1.19%	Singulair	Asthma	0.99%
Sleep	Depression	1.16%	Depression	Anxiety	0.80%
Depression	Lexapro	1.13%	Singulair	Hypersensitivity	0.69%
Sleep	Sleeplessness	1.13%	Sleep	Tired	0.65%
Sleep	Tired	1.05%	Lipitor	cholesterol	0.60%
Sleep	Anxiety	1.01%	Lisinopril	Blood pressure	0.57%
Depression	Cymbalta	0.99%	Sleep	Depression	0.57%
Sleep	Eating	0.94%	Hypersensitivity	Asthma	0.53%
Dizziness	Nausea	0.85%	Headache	Nausea	0.50%

Table E.5. Frequent itemsets of size 3 for medical concepts.

No Registration Health OSNs				Registration Health OSNs			
Depression	Anxiety	Lexapro	0.41%	Singulair	Hypersensitivity	Asthma	0.32%
Sleep	Depression	Anxiety	0.38%	Sleep	Lisinopril	Coughing	0.19%
Depression	Anxiety	Cymbalta	0.25%	NuvaRing	Libido	Sexual intercourse	0.18%
Sleep	Depression	Sleeplessness	0.24%	Lisinopril	Coughing	Blood pressure	0.17%
Depression	Anxiety	Happiness	0.23%	Depression	Singulair	Asthma	0.15%
Headache	Dizziness	Nausea	0.22%	Singulair	Happiness	Asthma	0.14%
Depression	Anxiety	Panic Attacks	0.21%	Singulair	Suffering	Asthma	0.13%
Sleep	Sleeplessness	Ambien	0.20%	NuvaRing	Libido	Contraceptives	0.13%
Depression	Anxiety	Weight Gain	0.19%	Sleep	Singulair	Asthma	0.12%
Depression	Anxiety	Zoloft	0.19%	NuvaRing	Sexual intercourse	Contraceptives	0.12%

Appendix F. Review vs Q&A OSNs

Table F.1 summarizes general and medical concept statistics for health OSNs with a review format and with a Q&A format. The format has little effect on these statistics, with the average number of words and medical concepts being roughly equal.

Table F.1. Summary of general statistics and medical concept statistics for health OSNs with a review format and Q&A format.

Category	Total Posts	Unique Posts	Words Per Post	Average Concepts per Post	Avg. Concepts per Word	Unique Concepts
Review	152,323	141,997 (93%)	62.4	11.6	.156	12,152
Q&A	21,278	15,126 (71%)	72.8	7.6	.110	6,308

Figure F.1 reports the effect of health OSN format on the emotional pairs. Health OSNs with a Q&A format observed a 26% decrease in disgusting terms and a 20% increase in trusting terms. Further, the format has little effect on the emotional pair surprise–anticipation; however, health OSNs with a Q&A format observed a decrease of 10% in anger terms, and an increase of 5% in fear terms.

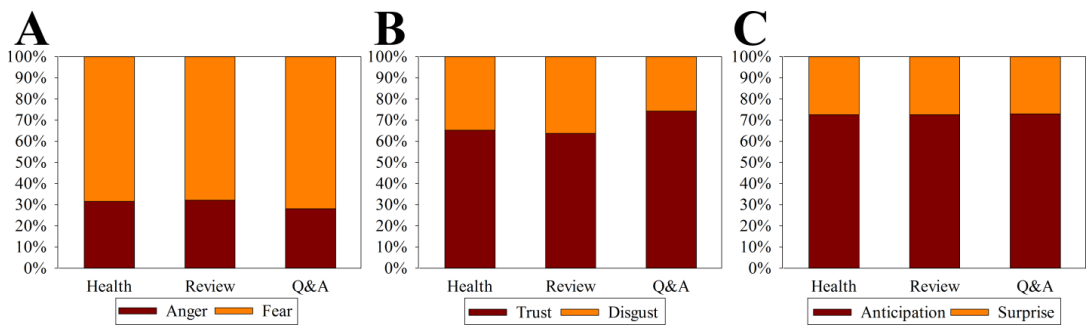


Figure F.1. An overview of the emotional analysis for health OSNs with a review format and a Q&A format. (A) The distribution of fear–anger,; (B) the distribution of surprise–anticipation.

Table F.2. Relatives changes of each item compared with the health OSN baseline shown in Figure F.1.

Emotion		
Disgust	Q&A	-26%
Trust	Q&A	+20%
Anger	Q&A	-10%

Sleep, anxiety, and depression are common and prevalent amongst both groupings of OSNs. Lisinopril, Lipitor, and NuvaRing are observed in health OSNs with a review format, but not those with a Q&A format. Lastly, there are frequent itemsets related to Xanax, Zoloft, hypothyroidism, and Synthroid in health OSNs with a Q&A format.

Table F.3. Frequent itemsets of size 1 for medical concepts.

Review Format Health OSNs		**Q&A Format Health OSNs**	
Sleep	10.47%	Female	9.02%
Depression	5.07%	Sleep	8.04%
Headache	4.32%	Eating	4.95%
Tired	4.17%	Disease	4.36%
Weight Gain	3.82%	Weight Gain	4.26%
Anxiety	3.67%	Anxiety	4.24%
Lisinopril	3.56%	Male gender	4.04%
Eating	3.45%	Depression	3.60%
Dizziness	3.35%	Mental Suffering	3.53%
Nausea	3.29%	Thyroid Gland	3.50%

Table F.4. Frequent itemsets of size 2 for medical concepts.

Review Format Health OSNs			Q&A Format Health OSNs		
Depression	Anxiety	1.04%	Sleep	Ambien	1.18%
Lisinopril	Coughing	1.01%	Anxiety	Depression	1.11%
Singulair	Asthma	0.85%	Thyroid Gland	Synthroid	1.10%
Sleep	Tired	0.76%	Female	Sleep	0.89%
Sleep	Sleeplessness	0.71%	Sleep	Anxiety	0.80%
Headache	Nausea	0.65%	Anxiety	Panic Attacks	0.72%
Sleep	Eating	0.63%	Anxiety	Xanax	0.71%
Sleep	Depression	0.62%	Sleep	Tired	0.70%
Singulair	Hypersensitivity	0.59%	Sleep	Xanax	0.70%
Lipitor	cholesterol	0.59%	Sleep	Sleeplessness	0.67%

Table F.5. Frequent itemsets of size 3 for medical concepts.

Review Format Health OSNs				Q&A Format Health OSNs			
Singulair	Hypersensitivity	Asthma	0.27%	Thyroid Gland	Synthroid	Hypothyroidism	0.30%
Sleep	Lisinopril	Coughing	0.18%	Anxiety	Depression	Zoloft	0.23%
Lisinopril	Coughing	Blood pressure	0.17%	Sleep	Anxiety	Depression	0.19%
NuvaRing	Libido	Sexual intercourse	0.16%	Sleep	Anxiety	Xanax	0.19%
Sleep	Depression	Anxiety	0.15%	Sleep	Sleeplessness	Ambien	0.19%
Depression	Anxiety	Lexapro	0.15%	Disease	Thyroid Gland	Synthroid	0.18%
Headache	Dizziness	Nausea	0.15%	Weight Gain	Thyroid Gland	Synthroid	0.18%
Depression	Singulair	Asthma	0.13%	Anxiety	Depression	Celexa	0.18%
NuvaRing	Libido	Contraceptives	0.12%	Thyroid Gland	Synthroid	Blood	0.17%
Lisinopril	Coughing	Dry cough	0.12%	Female	Thyroid Gland	Synthroid	0.16%

Appendix G Demographics of Providers

Table G.1 reports the distributions of single binary attributes for Castle Connolly Award=true and Referral Frequency=Very High for all providers; Wilcoxon signed-rank tests showed all differences to be significant with p less than 0.001. This table contains some interesting observations, in particular, providers who receive many referrals or a Castle Connolly award are more likely to accept new patients and Medicare patients; further, these providers also more likely to participate in PQRS, EHR, and eRx systems. The gender result is less surprising; according to a 2012 census of active physicians 70%

of doctors are male and 30% are female [143].

Table G.1. Distribution of single binary attributes.

	Percentage among providers with Castle Connolly Award=true (25,514 providers)	Percentage among providers with Referral Frequency=Very High (36,712 providers)	Percentage among All Providers
Gender=Male	79.6%	86.2%	69.2%
Accepting New Patients=true	84.0%	72.7%	55.5%
Accepts Medicare Insurance=true	71.1%	81.3%	56.7%
PQRS=true	34.7%	50.1%	24.6%
EHR=true	20.0%	23.9%	11.8%
eRx=true	32.8%	43.2%	21.5%

We visualized the ratio of providers with Referral Frequency=Very High and Castle Connolly Award=true over the total number of providers for each state using a heat map, shown in Figures G.1 and G.2. As shown in Figure G.1, Nevada and the mid and south Atlantic regions of the U.S. have the highest concentration of providers with Referral Frequency=Very High, which may imply that a majority of referral services are concentrated to a smaller number of providers in these areas due to a lack of specialists. As shown in Figure G.2, the north east region of the U.S. contains a higher concentration of providers with Castle Connolly awards than any other region in the U.S. Further, Florida, Washington, and Indiana also contain a considerably high ratio of Castle Connolly awards (greater than 5%). These results suggest that more providers in these states seek peer validation, which may result in a greater number of medical or clinical peer reviews. And peer review processes, such as accreditation programs, are tools to improve provider quality-of-care [140].

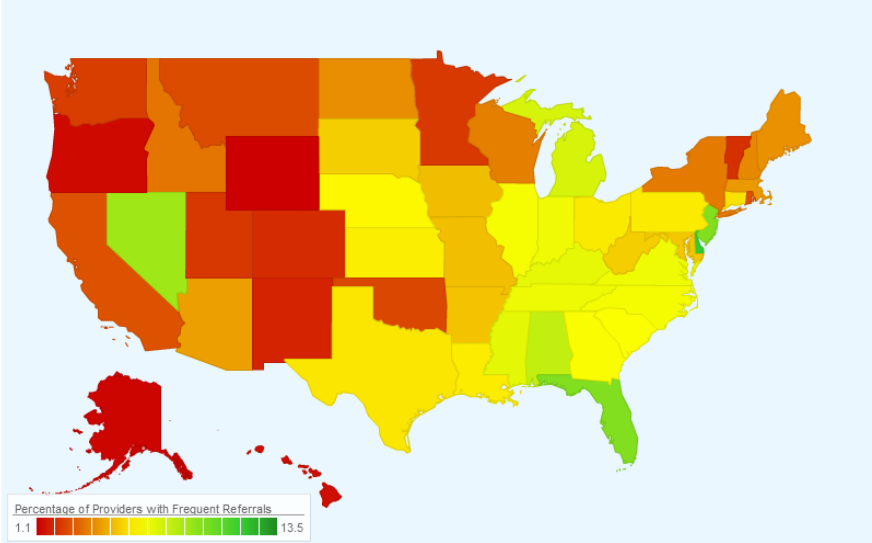


Figure G.1. Ratio of providers with Referral Frequency=Very High to the total number of providers by state.

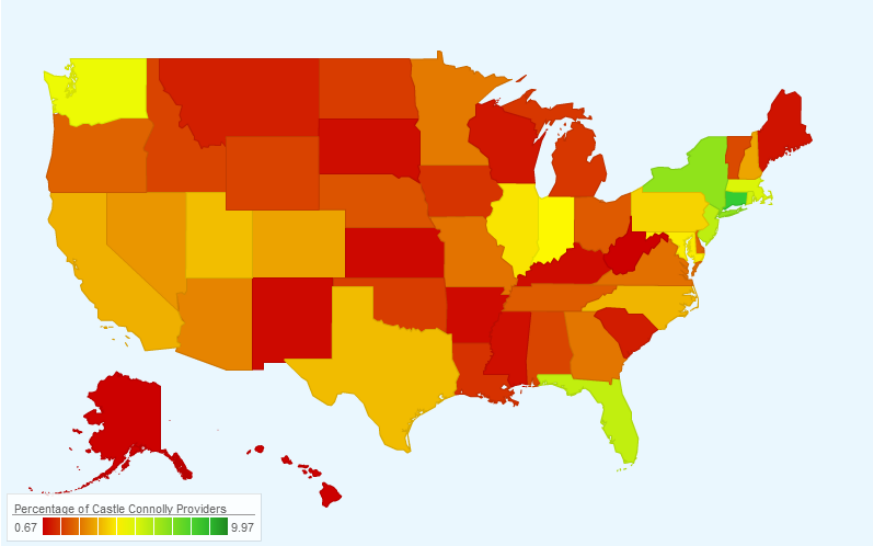


Figure G.2. Ratio of providers with Castle Connolly Award=true to the total number of providers by state.

Appendix H. State-Level Correlations

Here we present our analysis of state-level correlations with Referral Frequency=Very High in order to observe local trends in providers with frequent referrals. We found that 75 distinct attributes have a correlation greater than 0.05 when the data is stratified by each

state. A majority of these attributes had correlations greater than 0.05 in one or two states; Table H.1 lists the top 10 most frequently correlated attributes at the state level (note that the total number is 51 as Washington D.C. is included). Based on this table, there is indeed local influences on providers who are frequently referred, and these influencers are dominated by pediatric specialties.

Table H.1. The top 10 most frequently correlated attributes for Referral Frequency=Very High at the state level.

Attribute	Number of States
Pediatrics	50
Accepts Medicare Insurance	49
Emergency Medicine	49
Neonatal-Prenatal Medicine	48
Psychiatry	47
Child and Adolescent Psychiatry	45
Pediatric Critical Care Medicine	37
Pediatric Hematology-Oncology	37
Pediatric Cardiology	34
Obstetrics and Gynecology	33

We also examined correlations of Castle Connolly Award=true at the state level and found that 82 distinct attributes have a correlation greater than 0.05 when the data is stratified by each state. A majority of these attributes had correlations greater than 0.05 in one or two states; Table H.2 lists the top 10 most frequently correlated attributes at the state level. Based on this table, Castle Connolly awards indeed observe localized behavior and this behavior is influenced by the provider's specialty. This localized behavior could be explained by the peer-nomination process employed by Castle Connolly. Further, we also see local trends for certain types of drugs, such as Metformin for Type II diabetes and Cyclobenzaprine for muscle spasms. Lastly, despite the overrepresentation of males in Castle Connolly (79% versus 69% overall), we see that female has a correlation greater than 0.05 with Castle Connolly Award=true in nine states whereas male had zero states

with a correlation greater than 0.05.

Table H.2. The top 10 most frequently correlated attributes for Castle Connolly Award=true at the state level.

Attribute	Number of States
Family Medicine	32
Internal Medicine	21
Emergency Medicine	18
Anesthesiology	17
HCPCS: Emergency Department Visit	11
Accepts Medicare Insurance	10
Prescription: Metformin HCL	9
Gender=Female	9
Prescription: Cyclobenzaprine HCL	6
Prescription: Azithromycin	6

Appendix I. Most Discriminative Attributes for Referrals

To gain insight into attributes useful for classifying providers' referral frequency, we examined the top 10 most discriminative attributes for the discretized Referral Frequency attribute in Table I.1. This table shows that a provider's referral frequency may be discriminated by vascular-related prescriptions (e.g., Warfarin), if the provider offers electronic prescriptions, the provider's relative volume, if the provider is seeing new patients, and if the provider participates in PQRS. Note, the top three discriminative attributes from this table are also strongly correlated with Referral Frequency=Very High.

Table I.1. The top 10 most discriminative attributes for discretized Referral Frequency in terms of information gain.

Most Discriminative Attributes for *Referral Frequency=Very High*
Number of HCPCS Beneficiaries
Nonhospital
HCPCS: Initial Hospital Care
HCPCS: New Office/Outpatient Visit
PQRS
eRx
Relative Procedure Volume
Prescription: Furosemide
Prescription: Warfarin
Prescription: Plavix

Appendix J. Detailed Classification Results

Table J.1 reports the confusion matrix for the discretized Referral Frequency classifiers at the state level where each cell is tallied across all states. As with the national level, we see a majority of errors are relative to the ordering of categories. Further, we observe a significant improvement in sensitivity from 52% to 72% for Referral Frequency=Very High classifications, however there is no change in accuracy and some degradation in positive predictive value, from 78% to 70%. Other categories observed similar behavior except for Referral Frequency=Low, which observed a decrease in sensitivity. Thus, finding discriminative attributes to classify providers with high referral frequency is easier using attributes at the local level, and these local influencers should be modeled in each classifier separately. However, local influencers have less of an effect on classifying providers with very low referral frequency or no referrals.

Table J.1. Confusion matrix of Referrals at the state level. Each cell is tallied across all states.

Classified as →	Referral Frequency= None	Referral Frequency= Very Low	Referral Frequency= Low	Referral Frequency= High	Referral Frequency= Very High
Referral Frequency= None	232,331	13,972	6008	3453	858
Referral Frequency= Very Low	9134	17,417	14,293	720	3
Referral Frequency= Low	5781	11,172	64,310	19,896	137
Referral Frequency= High	1777	996	23,866	135,998	9923
Referral Frequency= Very High	112	12	210	9892	26,484

Table J.2 reports the confusion matrix for the Castle Connolly Award classifiers at the state level where each cell is tallied across all states. Compared to the national classifier,

we observed a degradation in sensitivity but an improvement in accuracy, specificity, and positive predictive value with 88%, 89%, and 18% respectively. Further, states with a high concentration of Castle Connolly awards had higher positive predictive values, namely New York, Florida and Connecticut all had positive predictive values over 30%. Thus, finding discriminative attributes to classify Castle Connolly providers is easier using attributes at the local level, and these local influencers should be modeled in each classifier separately.

Table J.2. Confusion matrix of Castle Connolly Award at the state level. Each cell is tallied across all states.

Classified as →	Castle Connolly Award=false	Castle Connolly Award=true
Castle Connolly Award=false	518,986	64,372
Castle Connolly Award=true	11,385	14,023

Appendix K. Rule Learning Results

In this section we report a summary of the rules found using the RIPPER algorithm on Castle Connolly Award and discretized Referral Frequency. For each dataset at the national and state level, we ran RIPPER with pruning, a maximum error rate of 50%, and the minimum number of items covered by a rule to 10; i.e., every rule evaluates to at least 10 positives and each rule has at most half the number of negatives. For every rule, at both the state and national levels, we computed its accuracy using the number of positives and negatives that the rule covers and present the rules that yield the highest accuracies; in the case of Referral Frequency, we only report rules that cover at least 100 providers as there are several rules that cover more than 100 providers with 90% or better accuracy. Essentially, each rule is identifying a cadre of providers with similar qualities who either have a high referral frequency or received a Castle Connolly award. This qualitative

analysis gives further insight into local influencers of highly referred providers and providers with a Castle Connolly award.

Table K.1 reports the top five most accurate rules that cover at least 100 providers for Referral Frequency=Very High. Based on the rules from this table, we indeed see that Number of Affiliated Hospitals and Number of HCPCS Beneficiaries are important factors in determining providers with Referral Frequency=Very High, but surprisingly, these rules do not consider specialties. Instead, every rule has an emphasis on Number of HCPCS Beneficiaries and four of the five rules contain Prescription: Hydrocodone-Acetaminophen=false. Thus—in addition to the number of hospital affiliations, and Medicare procedures and patients— providers who are highly referred perform specific laboratory procedures that differ based on locality and these same providers tend to avoid a specific medication unique to the locality.

Table K.1. The top five most accurate features for rules that imply Referral Frequency=Very High.

State	Rule	Positive	Negative	Accuracy
PA	Number of HCPCS Beneficiaries >= 2855 AND Number of Organization Members >= 13 AND Prescription: Hydrocodone-Acetaminophen=false AND Prescription: Avapro=false AND HCPCS: Electrocardiogram Report= true	154	1	99.3%
NC	Number of HCPCS Beneficiaries >= 1354 AND Prescription: Hydrocodone-Acetaminophen=false AND Number of Affiliated Hospitals >= 4 AND HCPCS: X-ray Exam of Abdomen=true	254	4	98.4%
MI	Number of HCPCS Beneficiaries >= 3038 AND Prescription: Hydrocodone-Acetaminophen=false AND Relative Cost of Procedures <= 0.16 AND HCPCS: CT Thorax with Dye	230	4	98.2%
NJ	NumHCPCSBeneficiaries >= 2706 AND NumHospitals >= 3 AND Prescriptions: Alendronate Sodium=false AND RelativeVolume <= 0.23 AND NumReviews <= 1	338	6	98.2%
TN	Number of HCPCS Beneficiaries >= 2590 AND Prescriptions: Hydrocodone-Acetaminophen=false AND Number of Affiliated Hospitals >= 5 AND Relative Cost of Procedures <= 0.19 AND Prescriptions: Klor-Con 10=false	125	3	97.6%

Table K.2 reports the top five most accurate rules that cover at least 10 providers for Castle Connolly Award=true. Based on the rules from this table, we see that Number of Fellowships and Years of Experience are important, the former appearing in four of the five rules and the latter appearing in all five rules. Further, we observed that three of the five rules contain attributes related to patient ratings. Thus, attributes that influence Castle Connolly awards differ from state to state, where attributes such as patient reviews or gender have differing influences in differing localities. Illinois presents an interesting rule that says doctors of Internal Medicine with a subspecialty in Pulmonary Disease who have at least one fellowship, participate in PQRS, use EHRs, and see less than 1380 Medicare beneficiaries each year are more likely to receive a Castle Connolly award. We also see an interesting rule in Washington, that says females with at least one fellowship, 20 to 35 years of experience, whose hospital affiliation score is in the top 53%, and who work at organizations with at least 189 employees are more likely to receive a Castle Connolly award.

Table K.2. The top five most accurate features for rules that imply *Castle Connolly Award=true*.

State	Rule	Positive	Negative	Accuracy
TX	Number of Fellowships > 0 AND Number of Organization Members >= 1290 AND Years of Experience > 30 AND Overall Rating >= 25 AND RelativeVolume >= 0.08	25	1	96.1%
IL	Years of Experience >= 25 AND Number of Fellowships > 0 AND EHR=true AND PQRS=true AND Internal Medicine, Pulmonary Disease=true AND Prescription: Levofloxacin=False AND Number of HCPCS Beneficiaries < 1380	15	1	93.7%
OK	Number of Patient Reviews >= 2 AND 86 < Number of Organization Members < 101 AND Years of Experience > 20 AND Medical School Rank >= 39	10	1	90.9%
FL	Years of Experience >= 24 AND Number of Fellowships > 0 AND Number of Patient Reviews >= 3 AND Knowledgeable >= 55 AND 310 < Number of Organization Members < 350 AND Number of Affiliated Hospitals < =1	28	4	87.5%
WA	Gender=Female AND 20 < YearsExp < 35 AND Number of Fellowships > 0 AND Hospital Affiliation Score > 46 AND Number of Organization Members >= 189	19	3	86.3%

Bibliography

1 Charles, D.: 'Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012'

2 http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html, accessed November 15 2014

3 http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalCompare.html, accessed November 15 2014

4 http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html?redirect=/PQRS/, accessed November 15 2014

5 http://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PartDData.html, accessed November 20 2014

6 Saeed, M., Lieu, C., Raber, G., and Mark, R.: 'MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring', in Editor (Ed.)^(Eds.): 'Book MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring' (IEEE, 2002, edn.), pp. 641-644

7 Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A.: 'Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation', Bioinformatics, 2003, 19, (10), pp. 1275-1283

8 Bodenreider, O.: 'The unified medical language system (UMLS): integrating biomedical terminology', Nucleic acids research, 2004, 32, (Database issue), pp. D267-D270

9 Stearns, M.Q., Price, C., Spackman, K.A., and Wang, A.Y.: 'SNOMED clinical terms: overview of the development process and project status', in Editor (Ed.)^(Eds.): 'Book SNOMED clinical terms: overview of the development process and project status' (American Medical Informatics Association, 2001, edn.), pp. 662

10 http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RXNORM/termtypes.html, accessed July 2013

11 Melton, G.B., Parsons, S., Morrison, F.P., Rothschild, A.S., Markatou, M., and Hripcsak, G.: 'Inter-patient distance metrics using SNOMED CT defining relationships', Journal of Biomedical Informatics, 2006, 39, (6), pp. 697-705

12 Cao, H., Melton, G.B., Markatou, M., and Hripcsak, G.: 'Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases', Journal of biomedical informatics, 2008, 41, (6), pp. 882-888

13 Ebadollahi, S., Sun, J., Gotz, D., Hu, J., Sow, D., and Neti, C.: 'Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for Near-Term prognostics', in Editor (Ed.)^(Eds.): 'Book Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for Near-Term prognostics' (American Medical Informatics Association, 2010, edn.), pp. 192

14 Plaza, L., and Díaz, A.: 'Retrieval of similar electronic health records using UMLS concept graphs', Natural Language Processing and Information Systems, 2010, pp. 296-303

15 Wang, F., Sun, J., and Ebadollahi, S.: 'Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment', Statistical Analysis and Data Mining, 2012, 5, (1), pp. 54-69

16 http://www.healthgrades.com/, accessed December 28 2014

17 http://www.vitals.com/, accessed December 28 2014

18 http://www.drugs.com/, accessed January 31 2013

19 http://www.webmd.com/, accessed January 31 2013

20 Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., and Gonzalez, G.: 'Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks'. Proc. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, Uppsala, Sweeden, 11-16 Jul 2010 2010 pp. Pages

21 Bian, J., Topaloglu, U., and Yu, F.: 'Towards large-scale twitter mining for drug-related adverse events'. Proc. Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, Maui, Hawaii, 29 Oct-2 Nov 2012 2012 pp. Pages

22 Zhang, Y., He, D., and Sang, Y.: 'Facebook as a Platform for Health Information and Communication: A Case Study of a Diabetes Group', Journal of medical systems, 2013, 37, (3), pp. 1-12

23 Frost, J.H., and Massagli, M.P.: 'Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data', Journal of Medical Internet Research, 2008, 10, (3)

24 Gao, G.G., McCullough, J.S., Agarwal, R., and Jha, A.K.: 'A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period', Journal of medical Internet research, 2012, 14, (1)

25 Mabotuwana, T., Lee, M.C., and Cohen-Solal, E.V.: 'An ontology-based similarity measure for biomedical data–Application to radiology reports', Journal of biomedical informatics, 2013, 46, (5), pp. 857-868

26 Bleeker, S.E., Derksen-Lubsen, G., van Ginneken, A.M., Van Der Lei, J., and Moll, H.A.: 'Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics', BMC medical informatics and decision making, 2006, 6, (1), pp. 29

27 Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., and Chute, C.G.: 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications', Journal of the American Medical Informatics Association, 2010, 17, (5), pp. 507-513

28 Aronson, A.R.: 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program'. Proc. AMIA Annual Symposium Proceedings, Washington D.C., 3-7 Nov 2001 2001 pp. Pages

29 Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., and Shahar, Y.: 'A comparative evaluation of full-text, concept-based, and context-sensitive search', Journal of the American Medical Informatics Association, 2007, 14, (2), pp. 164-174

30 Zhou, W., Yu, C., Smalheiser, N., Torvik, V., and Hong, J.: 'Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature', in Editor (Ed.)^(Eds.): 'Book Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature' (ACM, 2007, edn.), pp. 655-662

31 Ide, N.C., Loane, R.F., and Demner-Fushman, D.: 'Essie: a concept-based search engine for structured biomedical text', Journal of the American Medical Informatics Association, 2007, 14, (3), pp. 253-263

32 Lin, J., and Demner-Fushman, D.: 'The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine', in Editor (Ed.)^(Eds.): 'Book The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine' (ACM, 2006, edn.), pp. 99-106

33 Pedersen, T., Pakhomov, S.V., Patwardhan, S., and Chute, C.G.: 'Measures of semantic similarity and relatedness in the biomedical domain', Journal of biomedical informatics, 2007, 40, (3), pp. 288-299

34 Pesquita, C., Faria, D., Falcao, A.O., Lord, P., and Couto, F.M.: 'Semantic similarity in biomedical ontologies', PLoS Comput Biol, 2009, 5, (7), pp. e1000443

35 Zhang, X., Jing, L., Hu, X., Ng, M., and Zhou, X.: 'A comparative study of ontology based term similarity measures on PubMed document clustering': 'Advances in Databases: Concepts, Systems and Applications' (Springer, 2007), pp. 115-126

36 Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E.: 'Semantically enhanced information retrieval: An ontology-based approach', Web Semantics: Science, Services and Agents on the World Wide Web, 2011, 9, (4), pp. 434-452

37 Rada, R., Mili, H., Bicknell, E., and Blettner, M.: 'Development and application of a metric on semantic nets', Systems, Man and Cybernetics, IEEE Transactions on, 1989, 19, (1), pp. 17-30

38 Fagin, R., Lotem, A., and Naor, M.: 'Optimal aggregation algorithms for middleware', Journal of computer and system sciences, 2003, 66, (4), pp. 614-656

39 Pound, J., Ilyas, I.F., and Weddell, G.: 'Expressive and flexible access to web-extracted data: a keyword-based structured query language', in Editor (Ed.)^(Eds.): 'Book Expressive and flexible access to web-extracted data: a keyword-based structured query language' (ACM, 2010, edn.), pp. 423-434

40 Carpineto, C., and Romano, G.: 'A survey of automatic query expansion in information retrieval', ACM Computing Surveys (CSUR), 2012, 44, (1), pp. 1

41 Matos, S., Arrais, J.P., Maia-Rodrigues, J., and Oliveira, J.L.: 'Concept-based query expansion for retrieving gene related publications from MEDLINE', BMC bioinformatics, 2010, 11, (1), pp. 212

42 Lu, Z., Kim, W., and Wilbur, W.J.: 'Evaluation of query expansion using MeSH in PubMed', Information retrieval, 2009, 12, (1), pp. 69-80

43 Bast, H., and Weber, I.: 'Type less, find more: fast autocompletion search with a succinct index', in Editor (Ed.)^(Eds.): 'Book Type less, find more: fast autocompletion search with a succinct index' (ACM, 2006, edn.), pp. 364-371

44 Wang, H., Liang, Y., Fu, L., Xue, G.-R., and Yu, Y.: 'Efficient query expansion for advertisement search', in Editor (Ed.)^(Eds.): 'Book Efficient query expansion for advertisement search' (ACM, 2009, edn.), pp. 51-58

45 Ding, B., Wang, H., Jin, R., Han, J., and Wang, Z.: 'Optimizing index for taxonomy keyword search', in Editor (Ed.)^(Eds.): 'Book Optimizing index for taxonomy keyword search' (ACM, 2012, edn.), pp. 493-504

46 Farfan, F., Hristidis, V., Ranganathan, A., and Weiner, M.: 'XOntoRank: Ontology-aware search of electronic medical records', in Editor (Ed.)^(Eds.): 'Book XOntoRank: Ontology-aware search of electronic medical records' (IEEE, 2009, edn.), pp. 820-831

47 Tao, Y., Papadopoulos, S., Sheng, C., and Stefanidis, K.: 'Nearest keyword search in XML documents', in Editor (Ed.)^(Eds.): 'Book Nearest keyword search in XML documents' (ACM, 2011, edn.), pp. 589-600

48 Wu, Z., and Palmer, M.: 'Verbs semantics and lexical selection', in Editor (Ed.)^(Eds.): 'Book Verbs semantics and lexical selection' (Association for Computational Linguistics, 1994, edn.), pp. 133-138

49 Resnik, P.: 'Using information content to evaluate semantic similarity in a taxonomy', arXiv preprint cmp-lg/9511007, 1995

50 Lin, D.: 'An information-theoretic definition of similarity', in Editor (Ed.)^(Eds.): 'Book An information-theoretic definition of similarity' (1998, edn.), pp. 296-304

51 Leis, V., Kemper, A., and Neumann, T.: 'The adaptive radix tree: ARTful indexing for main-memory databases', in Editor (Ed.)^(Eds.): 'Book The adaptive radix tree: ARTful indexing for main-memory databases' (IEEE, 2013, edn.), pp. 38-49

52 Eysenbach, G.: 'Medicine 2.0: social networking, collaboration, participation, apomediation, and openness', Journal of Medical Internet Research, 2008, 10, (3), pp. e22

53 Van De Belt, T.H., Engelen, L.J., Berben, S.A., and Schoonhoven, L.: 'Definition of Health 2.0 and Medicine 2.0: a systematic review', Journal of medical Internet research, 2010, 12, (2), pp. e18

54 Swan, M.: 'Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking', International journal of environmental research and public health, 2009, 6, (2), pp. 492-525

55 Swan, M.: 'Scaling crowdsourced health studies: the emergence of a new form of contract research organization', Personalized Medicine, 2012, 9, (2), pp. 223-234

56 Wicks, P., Keininger, D.L., Massagli, M.P., la Loge, C.d., Brownstein, C., Isojärvi, J., and Heywood, J.: 'Perceived benefits of sharing health data between people with epilepsy on an online platform', Epilepsy & Behavior, 2011

57 Frost, J.H., Massagli, M.P., Wicks, P., and Heywood, J.: 'How the Social Web Supports patient experimentation with a new therapy: The demand for patient-controlled and patient-centered informatics', in Editor (Ed.)^(Eds.): 'Book How the Social Web

Supports patient experimentation with a new therapy: The demand for patient-controlled and patient-centered informatics' (American Medical Informatics Association, 2008, edn.), pp. 217

58 http://www.casesdatabase.com/, accessed Mar. 1 2013

59 Wongsuphasawat, K., and Gotz, D.H.: 'Outflow: Visualizing Patient Flow by Symptoms and Outcome', in Editor (Ed.)^(Eds.): 'Book Outflow: Visualizing Patient Flow by Symptoms and Outcome' (2011, edn.), pp.

60 Wongsuphasawat, K., and Gotz, D.: 'Exploring flow, factors, and outcomes of temporal event sequences with the Outflow visualization', Visualization and Computer Graphics, IEEE Transactions on, 2012, 18, (12), pp. 2659-2668

61 Zhang, Z., Gotz, D., and Perer, A.: 'Interactive Visual Patient Cohort Analysis'

62 Roitman, H., Yogev, S., Tsimerman, Y., Kim, D.W., and Mesika, Y.: 'Exploratory search over social-medical data', in Editor (Ed.)^(Eds.): 'Book Exploratory search over social-medical data' (ACM, 2011, edn.), pp. 2513-2516

63 Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., and Heywood, J.: 'Sharing health data for better outcomes on PatientsLikeMe', Journal of medical Internet research, 2010, 12, (2)

64 http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html, accessed March 12 2013

65 http://www.ncbi.nlm.nih.gov/books/NBK9685/ch03.I33_Descriptions_of_Each_File, accessed March 12 2013

66 Raghavan, P., Fosler-Lussier, E., and Lai, A.M.: 'Learning to Temporally Order Medical Events in Clinical Text'

67 Kapoor, B., and Kleinbart, M.: 'Building an Integrated Patient Information System for a Healthcare Network', Journal of Cases on Information Technology (JCIT), 2012, 14, (2), pp. 27-41

68 Ham, C., Dixon, J., and Chantler, C.: 'Clinically integrated systems: the future of NHS reform in England?', BMJ, 2011, 342

69 Newsham, A.C., Johnston, C., Hall, G., Leahy, M.G., Smith, A.B., Vikram, A., Donnelly, A.M., Velikova, G., Selby, P.J., and Fisher, S.E.: 'Development of an advanced database for clinical trials integrated with an electronic patient record system', Computers in Biology and Medicine, 2011, 41, (8), pp. 575-586

70 http://dailystrength.org/, accessed January 31 2013

71 http://www.druglib.com/, accessed January 31 2013

72 http://www.everydayhealth.com/, accessed January 31 2013

73 http://medications.com/, accessed January 31 2013

74 http://www.mediguard.org/, accessed January 31 2013

75 Denecke, K., and Nejdl, W.: 'How valuable is medical social media data? Content analysis of the medical web', Information Sciences, 2009, 179, (12), pp. 1870-1880

76 Swan, M.: 'Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem', Journal of Medical Internet Research, 2012, 14, (2), pp. e46

77 Swan, M.: 'Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen', Journal of Personalized Medicine, 2012, 2, (3), pp. 93-118

78 Fox, S.: 'The social life of health information, 2011', in Editor (Ed.)^(Eds.): 'Book The social life of health information, 2011' (Pew Internet & American Life Project, 2011, edn.), pp.

79 Lu, Y., Zhang, P., Liu, J., Li, J., and Deng, S.: 'Health-Related Hot Topic Detection in Online Communities Using Text Clustering', PloS one, 2013, 8, (2), pp. e56221

80 Weitzman, E.R., Cole, E., Kaci, L., and Mandl, K.D.: 'Social but safe? Quality and safety of diabetes-related online social networks', Journal of the American Medical Informatics Association, 2011, 18, (3), pp. 292-297

81 Shrank, W.H., Choudhry, N.K., Swanton, K., Jain, S., Greene, J.A., Harlam, B., and Patel, K.P.: 'Variations in structure and content of online social networks for patients with diabetes', Archives of internal medicine, 2011, 171, (17), pp. 1589

82 Greene, J.A., Choudhry, N.K., Kilabuk, E., and Shrank, W.H.: 'Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook', Journal of general internal medicine, 2011, 26, (3), pp. 287-292

83 Goeuriot, L., Na, J.-C., Min Kyaing, W.Y., Khoo, C., Chang, Y.-K., Theng, Y.-L., and Kim, J.-J.: 'Sentiment lexicons for health-related opinion mining'. Proc. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, Miami, FL, 28-30 Jan 2012 2012 pp. Pages

84 Chee, B.W., Berlin, R., and Schatz, B.: 'Predicting adverse drug events from personal health messages'. Proc. AMIA Annual Symposium Proceedings, Washington D.C., 3-7 Nov 2012 2011 pp. Pages

85 Yang, C.C., Jiang, L., Yang, H., and Tang, X.: 'Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media'. Proc. Proceedings of ACM SIGKDD Workshop on Health Informatics, Beijing, China, 12-16 Aug 2012 2012 pp. Pages

86 http://www.rxlist.com/script/main/hp.asp, accessed January 15 2013

87 http://hc.apache.org/, accessed January 4 2013

88 http://jsoup.org/, accessed January 4 2013

89 http://code.google.com/p/language-detection/, accessed February 25 2013

90 http://hunspell.sourceforge.net/, accessed 25 Feb 2013

91 http://www.drugs.com/drug-classes.html?tree=1, accessed March 4 2013

92 Fagin, R., Kumar, R., and Sivakumar, D.: 'Comparing top k lists', SIAM Journal on Discrete Mathematics, 2003, 17, (1), pp. 134-160

93 Denecke, K., and Soltani, N.: 'The Burgeoning of Medical Social-Media Postings and the Need for Improved Natural Language Mapping Tools': 'Where Humans Meet Machines' (Springer, 2013), pp. 27-43

94 http://semanticnetwork.nlm.nih.gov/SemGroups/, accessed April 2 2013

95 Baccianella, S., Esuli, A., and Sebastiani, F.: 'Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining'. Proc. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 17-23 May 2010 2010 pp. Pages

96 Toutanova, K., Klein, D., Manning, C.D., and Singer, Y.: 'Feature-rich part-of-speech tagging with a cyclic dependency network'. Proc. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Edmonton, Canada, 27 May-1 Jun 2003 2003 pp. Pages

97 Mohammad, S.M., and Turney, P.D.: 'Crowdsourcing a word–emotion association lexicon', Computational Intelligence, 2012

98 Agrawal, R., Imieliński, T., and Swami, A.: 'Mining association rules between sets of items in large databases', in Editor (Ed.)^(Eds.): 'Book Mining association rules between sets of items in large databases' (ACM, 1993, edn.), pp. 207-216

99 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H.: 'The WEKA data mining software: an update', ACM SIGKDD Explorations Newsletter, 2009, 11, (1), pp. 10-18

100 Han, J., Pei, J., and Yin, Y.: 'Mining frequent patterns without candidate generation', in Editor (Ed.)^(Eds.): 'Book Mining frequent patterns without candidate generation' (ACM, 2000, edn.), pp. 1-12

101 Davison, M.L.: 'Multidimensional scaling' (Wiley, 1983. 1983)

102 http://blog.twitter.com/2012/04/shutting-down-spammers.html, accessed April 2 2013

103 http://www.nbcnews.com/technology/technolog/google-wages-war-spam-comments-277331, accessed April 2 2013

104 http://blog.pinterest.com/post/37347668045/fighting-spam, accessed April 2 2013

105 Lagu, T., Hannon, N.S., Rothberg, M.B., and Lindenauer, P.K.: 'Patients' evaluations of health care providers in the era of social networking: an analysis of physician-rating websites', Journal of general internal medicine, 2010, 25, (9), pp. 942-946

106 Ellimoottil, C., Hart, A., Greco, K., Quek, M.L., and Farooq, A.: 'Online reviews of 500 urologists', The Journal of urology, 2013, 189, (6), pp. 2269-2273

107 Kadry, B., Chu, L.F., Kadry, B., Gammas, D., and Macario, A.: 'Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating', Journal of medical Internet research, 2011, 13, (4)

108 http://www.aaos.org/news/aaosnow/sep13/advocacy4.asp, accessed March 2 2014

109 Reimann, S., and Strech, D.: 'The representation of patient experience and satisfaction in physician rating sites. A criteria-based analysis of English-and German-language sites', BMC health services research, 2010, 10, (1), pp. 332

110 http://www.huffingtonpost.com/2012/02/21/find-the-best-doctors_n_1284564.html, accessed February 23 2014

111 Lee, Y.-M., and Ahn, D.-S.: 'A preliminary study for exploring the attributes of being a "good doctor"', Korean Journal of Medical Education, 2007, 19, (4), pp. 313-323

112 Lambe, P., and Bristow, D.: 'What are the most important non-academic attributes of good doctors? A Delphi survey of clinicians', Medical teacher, 2010, 32, (8), pp. e347-e354

113 Luthya, C., Cedraschia, C., Perrinb, E., and Allaza, A.-F.: 'How do patients define "good" and "bad" doctors?', 2005, 135, (5-6), pp. 82-86

114 Schattner, A., Rudin, D., and Jellin, N.: 'Good physicians from the perspective of their patients', BMC Health Services Research, 2004, 4, (1), pp. 26

115 Verhoef, L.M., Van de Belt, T.H., Engelen, L.J., Schoonhoven, L., and Kool, R.B.: 'Social Media and Rating Sites as Tools to Understanding Quality of Care: A Scoping Review', Journal of medical Internet research, 2014, 16, (2), pp. e56

116 Emmert, M., Sander, U., and Pisch, F.: 'Eight questions about physician-rating websites: a systematic review', Journal of medical Internet research, 2013, 15, (2)

117 Wan, X.: 'How Patients Rate Their Allergists Online: Analysis Of Physician-Review Websites', in Editor (Ed.)^(Eds.): 'Book How Patients Rate Their Allergists Online: Analysis Of Physician-Review Websites' (AAAAI, 2014, edn.), pp.

118 Segal, J., Sacopulos, M., Sheets, V., Thurston, I., Brooks, K., and Puccia, R.: 'Online doctor reviews: do they track surgeon volume, a proxy for quality of care?', Journal of medical Internet research, 2012, 14, (2)

119 http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-medical-schools/primary-care-rankings, accessed Februrary 28 2014

120 http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-medical-schools/research-rankings, accessed February 28 2014

121 http://health.usnews.com/best-hospitals/rankings, accessed February 28 2014

122 https://questions.cms.gov/faq.php?faqId=7977, accessed November 18 2014

123 https://www.medicare.gov/sign-up-change-plans/decide-how-to-get-medicare/original-medicare/how-original-medicare-works.html, accessed August 8 2015

124 https://www.medicare.gov/sign-up-change-plans/medicare-health-plans/medicare-advantage-plans/medicare-advantage-plans-referral-comparison.html, accessed August 8 2015

125 https://kaiserfamilyfoundation.files.wordpress.com/2013/01/8323.pdf, accessed January 28 2016

126 https://kaiserfamilyfoundation.files.wordpress.com/2013/06/8448.pdf, accessed January 28 2016

127 Groenewoud, S., Van Exel, N.J.A., Bobinac, A., Berg, M., Huijsman, R., and Stolk, E.A.: 'What Influences Patients' Decisions When Choosing a Health Care Provider? Measuring Prefernces of Patients with Knee Arthrosis, Chronic Depression, or Alzheimer's Disease, Using Dicrete Choice Experiments', Health Services Research, 2015

128 http://www.castleconnolly.com/, accessed February 28 2014

129 https://www.castleconnolly.com/about/pr.cfm?id=217019aa-6672-44b6-85e1-587176157eed, accessed January 28 2016

130 http://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/NationalProvIdentStand/DataDissemination.html, accessed November 15 2014

131 https://data.medicare.gov/data/physician-compare, accessed November 15 2014

132 http://www.census.gov/geo/maps-data/maps/pdfs/reference/us_regdiv.pdf, accessed Februrary 28 2014

133 http://www.wpc-edi.com/reference/, accessed December 2014

134 Cohen, W.W.: 'Fast Efective Rule Induction', in Editor (Ed.)^(Eds.): 'Book Fast Efective Rule Induction' (1995, edn.), pp.

135 Khoshgoftaar, T.M., Golawala, M., and Van Hulse, J.: 'An empirical study of learning from imbalanced data using random forest', in Editor (Ed.)^(Eds.): 'Book An empirical study of learning from imbalanced data using random forest' (IEEE, 2007, edn.), pp. 310-317

136 Chen, C., Liaw, A., and Breiman, L.: 'Using random forest to learn imbalanced data', in Editor (Ed.)^(Eds.): 'Book Using random forest to learn imbalanced data' (UC Berkeley, 2004, edn.), pp.

137 Murphy, S.L., Xu, J., and Kochanek, K.D.: 'Deaths: final data for 2010', National vital statistics reports, 2013, 61, (4), pp. 1-118

138 Frieden, T.R.: 'CDC Health Disparities and Inequalities Report-United States, 2013. Foreword.', Morbidity and mortality weekly report. Surveilance sumamries (Washington, DC: 2002), 2013, 62, pp. 1-2

139 http://www.beckershospitalreview.com/hospital-management-administration/executive-roundtable-a-high-level-look-at-hospital-affiliations.html, accessed October 22 2015

140 Alkhenizan, A., and Shaw, C.: 'Impact of accreditation on the quality of healthcare services: a systematic review of the literature', Annals of Saudi medicine, 2011, 31, (4), pp. 407

141 https://www.medicare.gov/hospitalcompare/Data/Overview.html, accessed August 8 2015

142 http://www.rxlist.com/script/main/hp.asp, accessed January 15 2013

143 Young, A., Chaudhry, H.J., Thomas, J.V., and Dugan, M.: 'A census of actively licensed physicians in the United States, 2012', Journal of Medical Regulation, 2012, 99, (2), pp. 11-24