DIAGNOSIS BASED SPECIALIST IDENTIFICATION IN THE HOSPITAL

BY

XUN LU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Carl A. Gunter

Abstract

Medical specialties provide essential information about which providers have the skills needed to carry out key procedures or make critical judgments. They are useful for training and staffing and provide confidence to patients that their providers have the experience needed to address their problems.

This work evaluates how machine learning classifiers can be trained on treatment histories to recognize medical specialties. Such classifiers can be used to evaluate staffing and workflows and have applications to safety and security. We focus on treatment histories that consist of the patient diagnoses. We find that some specialties, such as a urologist, can be learned with good precision and recall, while other specialties, such as anesthesiology, are less easily recognized. We call the former *diagnosis specialties* and explore four machine learning techniques for them, which we compare to a naive baseline based on the diagnoses most commonly treated by specialists in a training set. We find that these techniques can improve substantially on the baseline and that the best technique, which uses Latent Dirichlet Allocation (LDA), provides precision and recall above 80% for many diagnosis specialties based on a study with one year of chart accesses and discharge diagnoses from a major hospital. Furthermore, we explored several data mining techniques to discover valid but unlisted diagnosis specialties. We present the diagnosis specialty discoveries and their associated attributes that corroborate the discoveries.

To my father and mother, for their unconditional support and love.

Table of Contents

List of Tables vi

List of Figures vii

Chapter 1 Introduction 1
- 1.1 Problem Statement 1
- 1.2 Objectives of Study 2
- 1.3 Contributions 4
- 1.4 Thesis Structure 5

Chapter 2 Background 6
- 2.1 Clinical Terminology 6
- 2.2 Diagnosis Specialty 8
- 2.3 Related Works 10

Chapter 3 Dataset 12
- 3.1 Dataset Description 12
- 3.2 Taxonomy Code Statistics 13

Chapter 4 Diagnosis Specialty Prediction 15
- 4.1 Naive Approach 15
- 4.2 Supervised Latent Dirichlet Allocation 16
- 4.3 Baseline Classifiers 17
- 4.4 Cross-Validation 18

Chapter 5 Diagnosis Specialty Discovery 20
- 5.1 Topic Model Based Discovery 20
- 5.2 User Guided Discovery 24

Chapter 6 Experiment Results and Evaluation 27
- 6.1 Result of Diagnosis Specialty Prediction 27
- 6.2 Evaluation of Diagnosis Specialty Discovery 35

Chapter 7 Applications 39
- 7.1 Security and Privacy 39
- 7.2 Staffing Support 40
- 7.3 Patient Experience Enhancement 40

Chapter 8 Conclusion and Future Works 42

References 43

List of Tables

3.1 **Statistics for Audit Logs** 12
3.2 **Statistics for Patient EMR Records** 12
3.3 **Selected NPI Taxonomy Code** 14

5.1 **Topic Quality Comparison** 22
5.2 **Example topic for k set to 30** 23

6.1 **Experimental results of classifiers (in percentage). The best results are boldfaced. Superscripts n, d, s and k denote that the performance difference is statistically significant ($p < 0.05$) compared to Naive, Decision Tree, SVM and PCA-KNN for the respective metrics.** 29
6.2 **Diagnosis Features for *Pulmonary Disease*** 31
6.3 **Examples of Diagnosis Topics** 32
6.4 **User Cases** 34
6.5 **Potential New Diagnosis Specialties** 36
6.6 **Evaluation Statistics of Potential Diagnosis Specialtis** . 36
6.7 **Cluster Summaries for $\delta = 3$** 38

List of Figures

3.1 Frequency Distribution of Taxonomy Codes 13

4.1 Graphical Model for supervised Latent Dirichlet Allocation (sLDA) [20] 17

4.2 KNN with varying feature numbers 19

5.1 Graphical Model for (unsupervised) Latent Dirichlet Allocation (LDA) [24] 21

5.2 Network Schema of Dataset 25

6.1 Performance of the classifiers for each of the 12 core NPI taxonomy codes in the setting of *12-core*. 30

Chapter 1

Introduction

After completing medical school, physicians usually further their medical education in a specific specialty of medicine by completing a multiple year residency to become a medical specialist. Specialties are an important part of the medical profession. They provide information about which providers have the skills needed to carry out key procedures or make critical judgments. They are useful for training and staffing and provide confidence to patients that their providers have the experience needed to address their problems. There are many ways in which provider institutions express and take advantage of the specialties of their staff, including organizing them into groupings like departments or assigning them attributes like specialty codes. However, these expressions have limitations. For instance, at a given institution, some specialties may not be adequately expressed in these ways, or a specialty may be better assigned to one provider than another. In general, there could be a gap between the treatment history of a provider and the specialty expressed for the provider. Techniques for addressing these limitations can benefit staffing, quality control, building patient confidence, and other areas.

1.1 Problem Statement

Currently the Healthcare Provider Taxonomy Code [1] is used to describe providers' specialties. The code set is a Health Insurance Portability and Accountability (HIPAA) standard code set and it is the only code set that may be used in HIPAA standard transactions to report the type/classification/ specialization of a health care provider when such reporting is required [2]. Providers obtain their taxonomy codes by self-reporting when they apply for their National Provider Identifiers (NPIs) [3]. Providers usually report one or a few taxonomy codes according to their specialized training or the certifications they have.

Ideally, this mechanism would identify every health care provider with one or more taxonomy codes that most accurately describe their specialization. However, this is not always the case due to the following reasons:

1. The National Plan & Provider Enumeration System (NPPES) does not verify with the health care providers or with trusted sources that the taxonomy code selections made by health care providers when applying for the NPIs are accurate [1].

2. Some taxonomy codes do not correspond with any nationwide certifications that are approved by a professional board. For example, the specialty for "Men and Masculinity" is a well recognized area of interest, study and activity in the field of psychology, but there is no certification or credential available to identify psychologists who might work in this area [4].

3. Some national certifications are not reflected on the specialty code list. Since the specialty codes do not correspond to certifications within the field, health care providers will interpret these codes in different ways.

4. Health insurers may have access to the taxonomy code and they may use it in ways that affect health care provider's reimbursement. Therefore, health care providers are encouraged to choose taxonomy code with broader scope, like Internal Medicine as opposed to Cardiovascular Disease.

These observations argue that it is desirable to have additional systematic methods for the identification of provider's specialty.

1.2 Objectives of Study

In this work, we propose to identify providers' specialties by their treatment histories, particularly the diagnoses of patients whose Electronic Medical Record (EMR) the providers have accessed. Depending on whether a provider's specialty is listed in the Healthcare Provider Taxonomy Code Set there are two different problems to solve. If a provider have a listed specialty (*i.e.,* there is a taxonomy code corresponding to the specialty), we will need to correctly predict the provider's specialty. Otherwise, before we declare a provider as having no specialty we need to discover all the hidden specialties in hospital and check if the user has any of them. We denote the former problem as *Specialty Prediction* and the latter problem as *Specialty Discovery*.

For specialty prediction, we explore the question of whether a provider specialty can be *learned* by machine learning classifiers from the diagnoses of patients that were treated by a provider. Our specific focus is on a subset of Healthcare Provider Taxonomy Codes [1]. Such codes are often reflected in the history of the diagnoses of patients treated by the provider. For instance, a urologist (physician with a urology specialty code) might be expected to treat conditions like "retention of urine" and "calculus of kidney" more frequently than a provider that does not have a specialty in this area. By contrast, certain specialties, such as an anesthesiologist, seem likely to treat a wide range of patients where there may be less prevalence of any specific collection of diagnoses such as a provider's treatment record. The goal is to investigate the extent to which supervised learning techniques can be leveraged to learn, and subsequently classify, specialists based on their treatment histories with respect to diagnoses.

For specialty discovery, we explore the possibility of using data mining to discover some clusters of providers who don't have listed specialties but whose treatment histories nevertheless have pervasive themes within each cluster. For example, there is no taxonomy code for breast cancer specialist, but they are expected to work a lot with patients who have "neoplasm of breast" and "plastic surgery". Therefore, if there is a cluster of providers whose treatment histories have such common characteristics then we can argue that *breast cancer* deserves to be a specialty. Hidden specialties are very likely to exist since observation of the Healthcare Provider Taxonomy Code Set and medical expert opinion both suggest the incompleteness of the current code set [4]. Since the new specialties will be discovered base on treatment history, they should be diagnosis specialties as well.

To perform these investigations, we work with one year of treatment history for the healthcare providers in a major academic medical center.

1.3 Contributions

The main contribution of this project includes:

- Our findings reveal that provider specialties can be usefully classified into two general groups: (1) those which are strongly determined by diagnosis treatment history (*e.g.,* urology) and (2) those that are only weakly determined (*e.g.,* anesthesiology).

- For the specialties strongly determined by treatment history, which we call *diagnosis specialties*, we find that a naive classifier that provides a ranked list of diagnoses commonly treated by the specialty performs fairly well. Such a classifier generally models how a human might review the treatment history of a provider to determine the provider's specialty.

- We find that more statistically-informed machine learning techniques can provide significantly better precision and recall than the naive classifier for specialty prediction. The best results are obtained by supervise Latent Dirichlet Allocation (sLDA), which is a dimensionality-reduction technique that learn a basis of topics that represent diagnoses commonly grouped together in patient diagnosis lists. These techniques are able to achieve predictive capability on most diagnosis specialties with precision and recall in the range of 80% to 90%.

- We find that diagnoses frequently grouped together in patients' diagnoses histories provide a good basis for specialty discovery. As a result, we proposed a method for diagnosis specialty discovery using Latent Dirichlet Allocation (LDA) as well as two easy-to-follow criteria for evaluating potential new specialties. From the diagnosis topics generated by LDA, we discovered a valid *breast cancer* specialty in our dataset.

- Although user-guided clustering technique designed for Heterogeneous Information Network (HIN) such as PathSelClus [5] generally performs well on HIN, our experiment result shows that it is not effective at clustering users based on their treatment histories.

1.4 Thesis Structure

Chapter 2 describes the key concepts required to understand our methods and provides thorough background information. Section 2.1 describes various standard clinical terminology used in EMR system to denote patients' diseases and providers' specialties. Section 2.2 introduces the concept and a formal definition of *diagnosis specialty*. Section 2.3 provide an overview of related works.

Chapter 3 offers an insight into the Northwestern Memorial Hospital EMR and Audit log dataset. Section 3.1 describes the information contained in the dataset and Section 3.2 summarize important statistics of the providers' taxonomy codes in the dataset.

Chapter 4 highlights the methods we explored for predicting diagnosis specialty. Section 4.1 introduces the *naive* classifier we devised. Section 4.2 describes the *sLDA* classifier. And section 4.3 and Section 4.4 shows our baseline classifiers and the cross-validation method we adopted.

Chapter 5 aims at describing the data mining techniques we explored to discover new diagnosis specialties. Section 5.1 and Section 5.2 explains how to find new diagnosis specialties with LDA and PathSelClus respectively.

Chapter 6 presents experiment measurements and evaluation results. Section 6.1 displays the precision and recall of the classifiers we explored for specialty prediction. It also features a case study of the negative results. Section 6.2 shows the discovery result for each of the data mining techniques we explored.

Chapter 7 features some potentially useful applications of our specialty identification technique. Section 7.1 talks about the implications for EMR privacy , Section 7.2 explain how our work can be applied to hospital staffing support and Section 7.3 describes how our technique enhances patient experience.

Chapter 8 summarizes the results and discusses limitation and future scope.

The following components of this thesis were reported in [6]: Section 2.3, Chapter 3, Chapter 4, Section 6.1 and Section 7.3.

Chapter 2

Background

2.1 Clinical Terminology

Clinical terminology translates the complex medicine language to a compact and machine-readable format. Most data in the Electronic Medical Record (EMR) are stored in the form of some clinical terminology. In this section, we will introduce three terminologies that are relevant to our analysis: ICD-9 Diagnosis [7], CCS Diagnosis [8] and Healthcare Provider Taxonomy Code Set [1].

2.1.1 ICD-9 Diagnosis Code

The International Classification of Disease, Ninth Revision, (ICD-9) [7] is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. ICD-9 is the current official system of diagnosis code assignment in the United State, but it will be replaced by ICD-10 on October 1st, 2014 [9]; ICD-9 is a tabular form containing a numerical list of disease code number. It also includes an alphabetical index to the disease entries as well as a classification system for surgical, diagnostic and therapeutic procedures. ICD-9 diagnosis code is hierarchical with three levels. The highest level contains 19 diagnosis chapters bond together as a high level concept of different diagnosis. The middle level has more than 1000 diagnosis chapters where each diagnosis chapter can contain a group of correlated diagnosis codes. The lowest level consists of thousands of diagnosis chapters and each chapter is a leaf diagnosis code.

This hierarchy can be expressed in the number range of the first 3 digits of an ICD-9 code. For example, the range [390-459] represent the highest-level diagnosis chapter which contains all codes related to “Diseases Of The Circulatory System". Its sub-range [430-438] is a middle level diagnosis chapter containing all codes related to "Cerebrovascular Disease". And finally the code *432* denote the diagnosis "Other and unspecified intracranial hemorrhage". Further division of diagnosis is possible by appending the code by a dot and additional digits. For instance, code *432.1* represents "Subdural hemorrhage".

2.1.2 CCS Diagnosis Code

The Clinical Classifications Software (CCS) is based on ICD-9. CCS maps over 14,000 ICD-9 diagnosis codes into a much smaller number of clinically meaningful categories that are less redundant and more useful for presenting descriptive statistics than ICD-9 codes. CCS consists of two related classification systems, single-level and multi-level, which come with different levels of details. The single-level CCS aggregates diagnoses into 285 mutually exclusive categories. Meanwhile, multi-level CCS expands and splits the single-level CCS categories into 585 disjoint classes to provide more detail.

For instance, the single-level CCS code *98* (Essential hypertension) encompasses the ICD-9 codes *401.1* (Benign essential hypertension) and *401.9* (Unspecified essential hypertension). The corresponding multi-level CCS code is 7.1.1 (Essential hypertension) and its related codes are: 7 (Diseases of the circulatory system), 7.1 (Hypertension), 7.1.2 (Hypertension with complications and secondary hypertension), 7.1.2.1 (Hypertensive heart and/or renal disease), 7.1.2.2 (Other hypertensive complications).

For this study, we translate diagnosis data from ICD-9 to multi-level CCS to better analyze the pattern of diagnoses.

2.1.3 Healthcare Provider Taxonomy Code Set

The Healthcare Provider Taxonomy Code Set is designed to categorize the type, classification and specialization of healthcare providers. The code set is a Health insurance Portability and Accountability (HIPAA) standard code set. As such, it is the only code set that may be used in HIPAA standard transactions to report the type/classification/specialization of a health care provider when such reporting is required. Taxonomy codes are unique 10 character alphanumeric codes. The code set is structured into three distinct "levels" including Provider Type, Classification and Area of Specialization. As an example, the code *207RG0100X* describes providers whose type, classification and specialization are "Allopathic & Osteopathic Physician", "Internal Medicine" and "Gastroenterology" respectively.

2.2 Diagnosis Specialty

In this section we will define the notion of *Diagnosis Specialty* formally as it will be used throughout the rest of the paper. Let us begin with some examples before we attempt formal definitions. Consider the first few sentences of the descriptions of the specialties of urology and anesthesiology from the AAMC web site on Careers in Medicine (`https://www.aamc.org/cim/specialty/list`):

> *Urology focuses on the medical and surgical treatment of the male genitourinary system, female urinary tract, and the adrenal gland. Urologists treat patients with kidney, ureter, bladder, prostate, urethra, and male genital structure disorders and injuries.*
>
> *An anesthesiologist is trained to provide pain relief and maintenance, or restoration, of a stable condition during and immediately following an operation, obstetric, or diagnostic procedure. It is the anesthesiologist's foremost purpose and concern to protect the patient's well-being and safety just prior to, during, and after surgery.*

These short descriptions hint that it will be easier to characterize a urologist in terms of medical diagnoses for conditions, for example, of the kidney, ureter, and bladder, as opposed to an anesthesiologist, whose duties are more cross-cutting with respect to diagnoses, concerning essentially all conditions related to surgeries. Let us test this hypothesis with a naive classifier, which we will base on diagnosis codes. To see the general idea consider the following steps. First, take a dataset that indicates which patients have been treated by urologists and anesthesiologists and view each patient they treated as a document whose words are their diagnoses. Create a weighting for how many diagnoses of each kind were addressed by each provider, with some adjustment for how common the diagnosis is. This technique is typified by Term Frequency - Inverse Document Frequency (TF-IDF) and it gives a basic way to query for documents using terms. We can use it to query for specialists using the diagnoses they most often treat. For instance, using a dataset we will describe later, we find that urologists most often treat diagnoses like "retention of urine" and "urinary tract infection" whereas radiologist most often treat diagnoses like "other aftercare" and "other screening". When we take the 20 conditions most often treated by each of the two specialties and use them as a classifier, the results are fairly good for urology, yielding a precision of 83% and a recall of 71% in finding the urologists. However, the results for anesthesiologists are not so good, yielding a precision of 10% and a recall of 13%. If we try to get smarter and use a serious machine learning technique like a Support Vector Machine (SVM) we can do better at finding anesthesiologists with a 57% precision and 46% recall, but this is still weak compared to what we get for urologists, namely 100% precision and recall.

To think about this more generally one may ask: to what extent is it possible to train machine learning classifiers to recognize specialists based on the diagnoses of patients they treated? Which specialties are most amenable to training classifiers and which classifiers work the best? To address these questions formally we need some notation. Let $\mathsf{Bags}(X)$ be the collection of sets over X that respect multiplicity. For instance, if X is the set of numbers, then $\{1, 1, 2\}$ is the bag with two ones and a two. Let us suppose we are given a collection of *providers* U, a collection of *diagnoses* D, and a bag of sets of diagnoses $A \in \mathsf{Bags}(\mathsf{Subsets}(D))$ called the *diagnosis lists* that represent the lists of diagnoses of a collection of patients. The collection of diagnosis lists need to be a bag because it is possible that two patients have exactly the same diagnoses (hence giving a multiplicity of at least two for that set of diagnoses). Let us further assume that we are given a function T called the *treatment history* such that for each provider u, the value $T(u)$ is an element of $\mathsf{Bags}(\mathsf{Subsets}(D))$ that represents the diagnosis lists of patients treated by u.

A *specialty* is a subset S of U that describes a set of providers that have a given skill or certification. Based on this definition, the *specialty learning problem* is to use T, the treatment history function, together with the specialty, S, to produce a classifier $C_S : u \mapsto \mathbf{Bool}$ (where $\mathbf{Bool} = \{\mathsf{true}, \mathsf{false}\}$) that approximates S. Now, let us say that a *diagnosis specialty* is a speciality for which we are able to find a classifier that has precision and recall above a specified threshold.

It may be helpful to see these definitions by analogy to the classification of documents, an area that has inspired many of the techniques we will apply. The providers u can be likened to readers of documents, where A represents an archive of documents in which the words in each document correspond to diagnoses. The function T indicates for each user u the collection of documents that u has read. A specialty is a group of readers who (presumably) have something in common. The specialty learning problem is to develop a classifier that characterizes this common interest in terms of the documents they have read, if possible. So, for instance, if we were given a group of readers that are ophthalmologists and we find that they are inordinately interested in documents on disorders of the eyes, then we can use this proclivity to serve as a classifier for ophthalmologists, as we attempted to do for urologists and anesthesiologists using the naive classifier mentioned above. Such determinations are noisy and there may be no useful classifier. For instance, if our "specialty" consists of providers with blond hair, then we may not be able to determine any reasonable classifier for them from their reading history alone.

2.3 Related Works

A key driver behind our interest in inferring medical specialties is the analysis of audit logs for security and privacy purposes. Users have roles in the healthcare organization and if these roles are not respected by the online behavior of the users then there may be evidence of a security or privacy violation.

An early study on this theme examined the idea of looking at accesses to patient records to determine the position of an employee (doctor, nurse, dietician, and so on) [10]. This work used a Naive Bayes classifier and had generally poor performance on many positions, often because such positions could not easily be characterized in terms of the attributes used by the study. Experience Based Access Management (EBAM) envisioned such studies as part of a general effort to understand roles and least privilege by exploiting information about organizational behavior through the study of audit logs [11]. One other study in this direction sought to infer *new* roles from ways in which employees acted in their positions [12] by iteratively revising existing positions based on experience. Studies on Explanation Based Access (EBAS) [13, 14] addressed the problem of determining which *departments* are responsible for treating a given diagnosis, which is very similar to our problem of identifying an employee's specialty. In these studies the auditing system utilizes the access patterns of departments to determine diagnosis responsibility information in two ways: by analyzing (i) how frequent a department accesses patients with the diagnosis, and (ii) how focused the department is at treating the given diagnosis relative to others. For instance, EBAS can use this approach to determine that the Oncology Department is responsible for chemotherapy patients, while the Central Staffing Nursing Department is not responsible for those patients (even though they frequently access their records). The Random Topic Access Model (RTAM) [15] went beyond approaches based on conditional probabilities to work with topic models based on LDA that characterize the common behavior of employees in certain positions in the hospital.

The work in this paper can be seen as merging ideas from EBAS and RTAM to explore when a specialty can be described with a classifier based on LDA and whether new specialties can be discovered by data mining.

Chapter 3

Dataset

3.1 Dataset Description

The data for this study comes from the Cerner Powerchart EMR system in use at Northwestern Memorial Hospital (NMH). It contains all user accesses (audit logs) made over a one year period, as well as Electronic Medical Record (EMR) data for patient admitted during this period, specifically diagnosis lists, for patient encounters during this period. All data was de-identified for this study in accordance with the Safe Harbor standard of the HIPAA Privacy Rule and carried out under IRB approval. Since specialties are mainly concerned with respect to physicians, we filtered out users with other positions (*e.g.*, nurses, dieticians, and so on) from the dataset. The final dataset contains 4.8 million accesses made by more than three thousand physicians for almost 300,000 patients[1]. The data fields in the audit log are summarized in Table 3.1.

Attribute	**Value**
# of Timeframe	One year
# of Accesses	4,829,376
# of patients	291,562
# of Physicians	3,269
# of Patient Encounters	890,812
# of Taxonomy Codes	151

Table 3.1: Statistics for Audit Logs

Attribute	**Value**
# of User Roles	171
# of Locations	251
# of Services	104
# of Diagnoses	13,566
# of Procedures	2,165
# of Medications	1,822

Table 3.2: Statistics for Patient EMR Records

The EMR data consists of patient-encounter records, with each record corresponding to various diagnoses, procedures and medication. Approximately 13,000 diagnoses and 2,000 procedures are recorded in ICD-9 and ICD-9-CM

[1] This includes both hospitalization and hospital based outpatient encounters

format, respectively. Approximately 1,800 medications are recorded using RxNorm, a normalized naming system for generic and branded drugs [16]. Table 3.2 summarizes the fields of the EMR data. It is important to note that the dataset does *not* include information about which provider performed which procedure, prescribed which medication, or decided which diagnosis. We believe that our results would be similar if we did have this information, but such an analysis will need to await the opportunity to work with a more fine-grained dataset.

3.2 Taxonomy Code Statistics

A key feature of the dataset crucial to this study is that it also contains NPI taxonomy codes [1] for 60% of the physicians. About 150 NPI taxonomy codes are listed in the dataset, but most have fewer than 10 user instances, *i.e.*, providers who have the respective taxonomy code. Figure 3.1 shows frequency distribution of the taxonomy code in the dataset. Note the distribution is quite similar to the power-law distribution [17]. To ensure there were a sufficient amount of data for training the classifiers [18], we filtered out NPI taxonomy codes with fewer than 20 user instances. The resulting 23 NPI taxonomy codes[2] are shown in Table 3.3.

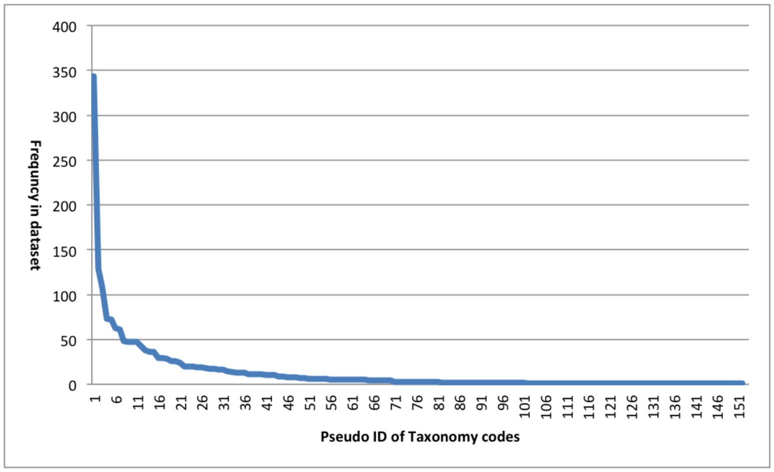

Figure 3.1: Frequency Distribution of Taxonomy Codes

[2] In this work, the terms *NPI taxonomy code name* and *NPI taxonomy code* are used interchangeably.

NPI Taxonomy Description	**NPI Taxonomy Code**
Internal Medicine	207R00000X
Pediatrics	208000000X
Obstetrics & Gynecology	207V00000X
Psychiatry	2084P0800X
Anesthesiology	207L00000X
Cardiovascular Disease	207RC0000X
Diagnostic Radiology	2085R0202X
Emergency Medicine	207P00000X
Neurology	2084N0400X
Nurse Anesthetist, Certified Registered	367500000X
Ophthalmology	207W00000X
Gastroenterology	207RG0100X
Physical Medicine & Rehabilitation	208100000X
Dermatology	207N00000X
Physician Assistant	363A00000X
Orthopaedic Surgery	207X00000X
Neonatal-Perinatal Medicine	2080N0001X
Infectious Disease	207RI0200X
Endocrinology, Diabetes & Metabolism	207RE0101X
Pulmonary Disease	207RP1001X
Neurological Surgery	207T00000X
Urology	208800000X
Acute Care	363LA2100X

Table 3.3: Selected NPI Taxonomy Code

Chapter 4

Diagnosis Specialty Prediction

Learning providers' diagnosis specialties from the diagnoses of patients they treated requires the training of some machine learning classifiers. This chapter introduces a range of classifiers we explored for diagnosis specialty prediction. We start off with an intuitive *naive* classifier in Section 4.1 that models how a human might review the treatment history of a provider to determine the provider's specialty. Then we move on to a more sophisticated topic modeling based classifier in Section 4.2. We also used three other state of the art classifiers to provide a baseline for classifier performance evaluation, and they are introduced in Section 4.3. In Section 4.4 we describe the cross-validation method we used for performance evaluation.

4.1 Naive Approach

Before delving into more sophisticated classifiers, let us first review an intuitive *naive classifier* of a kind that a human might imagine and apply to infer a specialty from a treatment history. The technique proceeds by finding the most relevant diagnoses of each diagnosis specialty (taxonomy code) and the most relevant diagnoses of each user. Users are classified according to the specialties with which they share the most common relevant diagnoses.

To describe this in detail, let U_n be the set of users with taxonomy code $n \in N$, where N is the set of NPI taxonomy codes. And let d_j be the j-th diagnosis and each user $u_i \in U_n$ has a vector $v_i = \{c_1, ..., c_k\}$ where c_j denotes the number of times that the user has accessed patients with d_j. We define

$$\mathsf{tfidf}_{d_j} = \log(\frac{v_i[d_j]}{a_i} + 1) \times \log(\frac{|U| + 1}{r_{d_j}})$$

where a_i is the total count of all diagnoses in v_i, and r_{d_j} is the number of users who are associated with the j-th diagnosis. The TF-IDF vector measures the relevance of diagnoses to users. In doing so, we can represent users by $v'_i = \{\mathsf{tfidf}_{\mathsf{d}_1}, ..., \mathsf{tfidf}_{\mathsf{d}_\mathsf{k}}\}$. For every user we collect their top m most relevant diagnoses R_i by sorting v'_i by value.

Similarly, we calculate the m most relevant diagnoses S_n for each taxonomy code using TF-IDF weighting. The TF-IDF values for taxonomy code n is $\sum_{u_i \in U_n} v'_i$. The similarity between specialty (taxonomy code) n and user u_i can be measured using S_n and R_i by the Jaccard Coefficient [19]:

$$J(R, S) = \frac{|R \cap S|}{|R \cup S|}$$

In this way, the naive classifier will predict user u_i's specialty according to:

$$n_{predict} = argmax_{n \in N} J(R_i, S_n)$$

4.2 Supervised Latent Dirichlet Allocation

The supervised Latent Dirichlet Allocation (sLDA) [20] is a generative model. The graphical model of sLDA is displayed in Figure 4.1. In the figure, α denotes the Dirichlet parameter, θ_d denotes per-document topic proportion, $Z_{d,n}$ denotes per-word topic assignment, $W_{d,n}$ denotes observed word, β_k denotes topics, Y_d denotes per-document response and (η, σ) denote the regression parameters. sLDA characterizes documents in a corpus as multinomials of a set of latent topics z_n where the topic proportions θ follows $\theta|\alpha \sim Dir(\alpha)$, and α is the Dirichlet parameter. These latent topics are modeled as multinomials over the words in a corpus as $w_n|z_n, \beta_{1:k} \sim Mult(\beta_{z_n})$. In this way, topics act as summaries of the different themes pervasive in the corpus and documents are characterized with respect to these summaries. The frequency of topics for each document will deterministically yield a response value y from a normal linear model as $y|Z_{1:k}, \eta, \sigma^2 \sim N(\eta^T \overline{z}, \sigma^2)$.

To apply sLDA to our problem, we model each physician as a document where the content is a bag of the diagnoses of patients he/she accessed. His/her primary NPI taxonomy is then treated as the response that corresponds to the document. The sLDA algorithm is adept at clustering words that frequently appear together in documents into topics. And in our case, this aspect translates into the ability to group correlated diagnoses together into diagnosis topics. After the sLDA model is trained, when it is provided with a physician (document) it will summarize the physician's associated diagnoses into a distribution over K topics. The number of topics K is determined from perplexity measure which can assess the effectiveness of different topic numbers. More details about perplexity measure can be found in Section 5.1.2. The allocation of the topics in a document then determine the response, which is the predicted specialty.

The effective generation of diagnosis topics makes sLDA model an effective tool for predicting diagnosis specialty, which is validated by the experiment results in the Evaluation chapter.

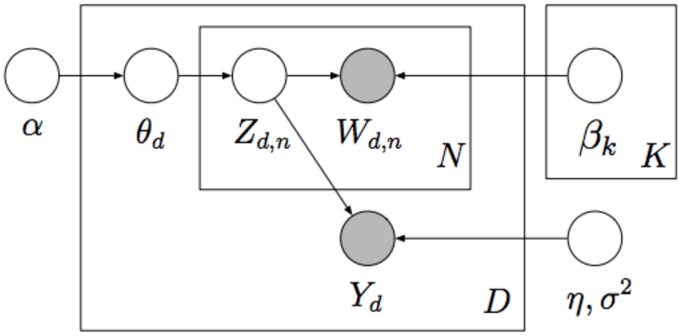

Figure 4.1: Graphical Model for supervised Latent Dirichlet Allocation (sLDA) [20]

4.3 Baseline Classifiers

We also evaluated three additional machine learning methods: decision tree (J48), Support Vector Machine (SVM), and K-Nearest-Neighbor with Principal Component Analysis (PCA-KNN) [21].

Decision trees are constructed in a top-down recursive divide-and-conquer manner. At start, all the training examples are at the root. Examples are partitioned recursively based on selected attributes. Test attributes are selected on the basis of a heuristic or statistical measure. Decision tree is a popular nonlinear classifier because it is convertible to classification rules than can be reviewed and interpreted by experts.

SVM is a classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane using support vectors that lie closest to the decision boundary. Particularly, SVM is effective on high dimensional data because the complexity of trained classifier is characterized by the number of support vectors rather than the dimensionality of the data.

KNN is an instance-based learning method—it stores training examples and delay the processing until a new instance must be classified. All instances correspond to points in the n-D space. The nearest neighbors are defined in terms of Euclidean distance. KNN returns the most common label among the K training examples nearest to the new testing instance. KNN suffers from the "curse of dimensionality"—distance between neighbors could be dominated by irrelevant attributes when the dimensionality of space goes higher.

To apply these methods, we map each user u_i to TF-IDF weighted diagnosis vectors v'_i, similar to the naive classifier. This vector along with u_i's primary taxonomy code serves as the input to these classifiers, with a length of 599. To address "curse of dimensionality" for KNN, we conduct dimensionality reduction for vectors before applying KNN. Here we use Principal Component Analysis (PCA) [22] to perform the dimensionality reduction and 50 most principal features are selected based on the parameter-tuning results in Figure 4.2. This experiment is conducted in the experiment setting of 12 core NPI taxonomy codes, which will be covered in the Evaluation chapter.

4.4 Cross-Validation

We will use 5×2 cross-validation for performance evaluation of the classifiers. In each of the 5 rounds, observations are split into two equal-sized sets A and B. Then a classier is trained with A and tested with B and *vise versa*. After 5 rounds, the average of the 10 results is returned. We choose 5×2 cross-validation over the n-fold cross-validation because 5×2 cross-validation is considered better at comparing the performance of algorithms [23].

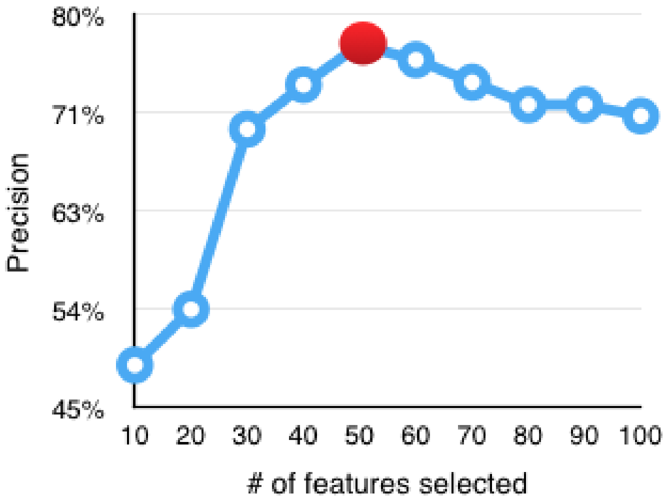

Figure 4.2: KNN with varying feature numbers

Chapter 5

Diagnosis Specialty Discovery

Discovering new diagnosis specialties from the treatment histories of providers requires effective clustering techniques that can divide a pool of providers into groups that have high inter-group distances (distinctiveness) but low intra-group distances (coherence). And new diagnosis specialties may emerge from these clusters. In this chapter, we describe two state of the art data mining techniques we use for diagnosis specialty discovery.

5.1 Topic Model Based Discovery

For this approach, we employee Latent Dirichlet Allocation (LDA) [24]. LDA is the base model for supervised Latent Dirichlet Allocation (sLDA) [20], which we introduced in Section 4.2. The graphical model of LDA is shown in Figure 5.1. In the figure, α denotes the Dirichlet parameter, θ denotes topic proportions, z denotes topic assignment, w denotes observed word and β denotes topics. As sLDA, LDA provides a set of topics, each represented as a bag of words that frequently appears in the same documents. And each document can be described as a distribution of topics.

The intuition behind using LDA to discover new diagnosis specialty is to find some LDA diagnosis topics that have coherent themes which correspond to some unlisted specialties in the hospital. And by representing providers as documents (*i.e.,* distribution of topics), we can cluster the providers based on which topic simplex[1] they are closest to. Details are provided in the following subsections.

[1] This can be visualized by plotting the providers by their topic distributions

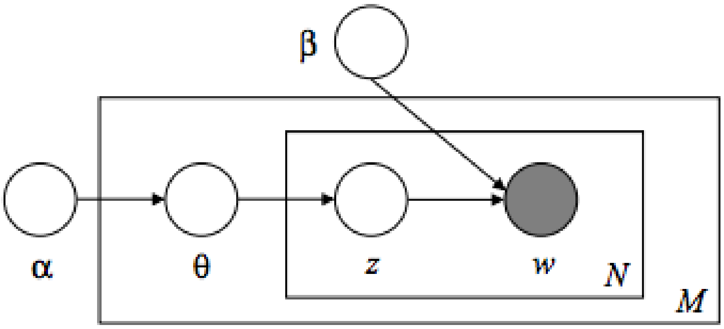

Figure 5.1: Graphical Model for (unsupervised) Latent Dirichlet Allocation (LDA) [24]

5.1.1 Representation of Providers

Since we are generating topics of diagnosis, the content of each document (provider) has to be diagnosis. However, diagnoses in our dataset are not provided with respect to providers, but patients. Therefore, we need to connect providers to diagnoses via patients. And there are two ways we can do it:

1. For any user[2] u_i, cross-referencing to find the set of patients P_i whose EMRs he accessed. Then for each patient $p_j \in P_i$ we can get the set of Diagnosis D_j that p_j has. From here, we can add diagnoses in D_j that occurred during the encounter of u_i and p_j to the "document" Doc_i representing u_i. In this way, the topics and their allocations for users can be found directly by training a LDA model with these documents.

2. We can also run the LDA on the patient dimension and get the topic distribution of patients first. Let T_{p_j} denote the topic distribution of patient p_j, T_{u_i} denote the topic distribution of user u_i and P_i denotes the set of patients whose EMRs are accessed by u_i. Then the topic distribution of users can be calculated as:

$$T_{u_i} = \frac{1}{|P_i|} \times \sum_{p_j \in P_i} T_{p_j}$$

[2] We will use *provider* and *user* interchangeably since providers are users of the EMR system.

We tested both approaches, Table 5.1a and 5.1b show one topic summary for each approach respectively.

Topic 10 diagnoses
Other hypertensive complications
Hypotension
Cancer of ovary
Coma, stupor, and brain damage
Hyposmolality
Ascites
Hematuria
Acute myocardial infarction
Backache, unspecified
Other connective tissue disease

(a) Example Topic From First Approach

Topic 10 diagnoses
Calculus of kidney
Elevated prostate specific antigen
Hematuria
Impotence of organic origin
Incomplete bladder emptying
Bladder neck obstruction
Urinary frequency
Hydronephrosis
Unspecified retention of urine
Other testicular hypofunction

(b) Example Topic From Second Approach

Table 5.1: Topic Quality Comparison

It is clear from the table that the topic obtained from the first approach has no clear theme and is even a little random, whereas the topic obtained from the second approach has a clear concentration on Urology. This is due to the fact that in the first approach each document contains the union of the diagnoses of all patients a user accessed, and in the second approach only the diagnoses of a single patient is in the document. The hodgepodge of many patients' diagnoses is likely to have multiple themes, thus rendering the topics generated by the first approach uninterpretable.

Therefore, we choose the second approach for our analysis.

5.1.2 Choice of Topic Number

An important parameter for LDA is the number of topics and we shall denote it as k. Unfortunately, there is no consensus on how to determine the value of k. A sign of good topic number is that the resulting topic summaries are semantically meaningful. And a rule of thumb for picking k is by utilizing the perplexity measure as we did in Section 4.2. The perplexity measure is an estimation of the expected number of equally likely words in the population, and minimizing perplexity corresponds to maximizing the

Topic 10 diagnoses
Other acne
Benign neoplasm of skin of trunk, except scrotum
Neoplasm of uncertain behavior of skin
Actinic keratosis
Viral warts, unspecified
Other seborrheic keratosis
Rosacea
Sebaceous cyst
Benign neoplasm of skin, site unspecified
Scar condition and fibrosis of skin

Table 5.2: Example topic for k set to 30

topic variance captured by the system [24]. The perplexity is calculated as:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d}\right\}$$

where D_{test} is a held-out collection of users to evaluate the models, w_d is diagnosis of patients.

We learn from the perplexity measure that 25,30 and 35 are good candidates for k. We then run LDA with all three topic numbers and compared the resulting topics. Overall when $k = 30$ the topics are most semantically meaningful. For example, Table 5.2 show a dermatology related topic generated when $k = 30$.

5.1.3 Topic-based Clustering

After LDA training, each user get an allocation in the topic simplex. A higher proportion in a topic indicates the user is more likely to access patient with diagnosis popular in that topic. Therefore, if we are trying to cluster users by their treatment history, it is reasonable to cluster them by their closest topic simplex which corresponds to the topic with the highest proportion among all topics. And the clustering follows:

$$C_i = argmax_{t \in T} P(u_i, t)$$

where C_i denotes the cluster assignment for user u_i, T denotes the set of topics and $P(u_i, t)$ denotes the proportion of topic t for user u_i.

In order to rule out the cases where the user is not leaning toward any topic, we set a threshold τ to cluster users whose $max_{t \in T} P(u_i, t) \leq \tau$ into a single cluster for users with no specialty. By default, we set $\tau = 0.5$.

5.1.4 Strategy for New Topic Evaluation

To find new diagnosis specialties from the topic generated by LDA model, we rely on our collaborator with medical expertise. Our expert will go through the diagnosis summaries of the topics and label each with one or a few medical themes that are pervasive in the topic.

After labeling, we compare the labeled topics with the Healthcare Provider Taxonomy Code Set [1] to see if there are topics that have pervasive themes but not listed in the code set. If such topics exist, they are potential new diagnosis specialties. For us to recognize a topic as new diagnosis specialty, there are two additional criteria:

1. The topic's cluster has to have at least 20 users in it.
2. The topic needs to be able to be learned by classifiers in Chapter 4 with performance comparable to existing diagnosis specialties.

The first criterion is required because it is not meaningful to discover specialties that have smaller crowd. Besides, this criterion is necessary for the second criterion. For the second criterion, if we mark the users in the topic's cluster with a new unique taxonomy code and include them in a training set with the incumbent diagnosis specialties, the classifiers from Chapter 4 have to perform well on both the incumbent and new taxonomy codes. This criterion is required by the definition of *diagnosis specialty*. And it requires the first criterion to be met to ensure there were a sufficient amount of data for training the classifiers [18].

5.2 User Guided Discovery

The structure of our data forms a typical heterogeneous information network (HIN) [25, 5]. Therefore, we use PathSelClus [5], a very effective technique in HIN, for user-guided clustering.

5.2.1 Preliminaries

A HIN is an information network with multiple types of objects and/or multiple types of links. For example, if a relation exists from type A to type B, denoted as ARB, the inverse relation R^{-1} holds naturally for $BR^{-1}A$. R and its inverse R^{-1} are usually not equal, unless the two types are the same and R is symmetric. Different from traditional network, HIN explicitly distinguishes object types and relationship types in the network. When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is called a *heterogeneous information network*; otherwise, it is a *homogeneous information network*.

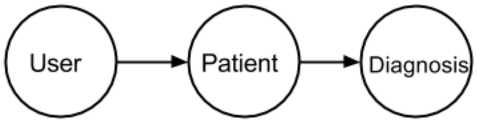

Figure 5.2: Network Schema of Dataset

Figure 5.2 shows the network schema of our dataset. It contains 3 types of objects, namely user (U), patient (P) and diagnosis (D). Links exist between user and patient by the relation of ”access” and ”accessed by”, between patient and diagnosis by the relation of ”diagnosed with” and ”given to”.

Link-based clustering in HIN clusters objects based on their connections to other objects in the network. The possible relations derived from a HIN between two types of objects in a meta level is called *meta-path* [26]. In our case, the *target object type* to cluster is U (users). And there are two meta-paths: $U - P - U$ and $U - P - D - P - U$.

5.2.2 Clustering and Specialty Discovery

During clustering, a decision has to be made about the weighted combination of different meta-paths to use. This is where user guidance comes in. We use PathSelClus [5] for user-guided clustering. In PathSelClus, user guidance is given in the form of object seeds in each cluster. For example, to cluster users based on the pattern of the diagnoses of patients treated by them, one can first provide several representative users as seeds for each pattern of diagnoses. On one hand, these seeds provide guidance for clustering the target object in the HIN. On the other hand, the seed information helps selecting the most relevant meta-paths for the clustering task.

The PathSelClus algorithm is designed to handle unseeded initial clusters since in practice there may not be enough information to seed all the clusters. This is the exact feature that makes it possible for us to use PathSelClus to discover new diagnosis specialties.

Let the number of listed diagnosis specialties be N and the number of new diagnosis specialties we are trying to explore be δ. Then we will create $N+\delta$ empty clusters at the initiation of PathSelClus and seed N of them with corresponding specialists. And the input to PathSelClus will include all the physicians in the hospital regardless of whether they have taxonomy code or not.

After PathSelClus has converged, all the users should have been assigned to an exclusive cluster. The δ unseeded clusters should also be filled with some users. We can analyze the semantics of the unseeded clusters by the users in them using the technique we described in Section 4.1. Namely, we treat a cluster as a taxonomy code and the users in t as having the taxonomy code. We can calculate the most relevant diagnoses for each cluster using the TF-IDF weighting technique. And then we can have our medical expert collaborator to label the clusters. As Topic model based discovery, the new clusters have to satisfy the two criteria in Section 5.1.4 to be recognized as a new diagnosis specialty.

Chapter 6

Experiment Results and Evaluation

6.1 Result of Diagnosis Specialty Prediction

6.1.1 Evaluation Setup

Since this part deals with diagnosis specialties, we have identified 12 NPI taxonomy codes from the 23 NPI taxonomy codes as likely diagnosis specialties. We chose these at first by expert opinion from the specialty description and the diagnosis rankings obtained from the naive classifier, but we also found that these are the specialties for which we were able to develop diagnosis-based classifiers that had recall and precision higher than 60%. These 12 NPI taxonomy codes are: *Obstetrics & Gynecology, Cardiovascular Disease, Neurology, Ophthalmology, Gastroenterology, Dermatology, Orthopaedic Surgery, Neonatal-Perinatal Medicine, Infectious Disease, Pulmonary Disease, Neurological Surgery, Urology.* We will denote them as the *12 core NPI taxonomy codes.*

Based on this distinction of NPI taxonomy codes, we designed four settings of experiments that explore the performance of the classifiers:

- **12 core NPI taxonomy codes only (*12-core*).** This experiment setting evaluates how the classifiers perform only on users with diagnosis specialties. In this setting the class number is 12.

- **12 core NPI taxonomy codes and 1 non-core class comprising of all the remaining 11 codes (*12-core and 1-non-core*).** This experiment setting does not distinguish NPI taxonomy codes that do not belong to the 12 core codes—it treats them as a single class denoted as *non-core.* In this setting the class number is 13. This experiment setting evaluates the classifiers' performance when there are noisy non-diagnosis specialties.

- **All 23 NPI taxonomy codes (*23-all*).** This experiment setting evaluates the classifier performance on users with all the 23 specialty codes. In this setting the class number is 23.

- **12 core NPI taxonomy codes in the setting of *23-all* (*12-core in 23-all*).** This experiment setting is the same as that of *23-all* except for average precision and recall of the 12 core NPI taxonomy codes are reported instead. In this setting the class number is 23. We plan to use results from this experiment and those of *23-all* to contrast the classifiers' performance on diagnosis specialties and non-diagnosis specialties.

As for classifiers' implementations, we used Weka [27] for SVM and Decision Tree (J48) with the default parameter values. For PCA-KNN, we used MATLAB for PCA and JAVA-ML [28] for KNN where the parameter K is set to 9[1]. For sLDA, we used an open source implementation [29].

6.1.2 Evaluation Results

The experimental results of the four sets of experiments are presented in Table 6.1. Table 6.1 shows the precision and recall for each of the experimental settings. The reported metrics are the average of the 5×2 cross-validation. In general, when the number of classes grows, the overall performance degrades because the multi-class classification problem gets more challenging.

Figure 6.1 shows the performance of the classifiers for each of the 12 core NPI taxonomy codes in the setting of *12-core*. The relative performance of the classifiers varies for different NPI taxonomy codes. Overall the classifiers perform best on the code *Obstetrics & Gynecology* and worst on code *Infectious Disease*. This is not surprising since *Obstetrics & Gynecology* can be described by a very distinctive set of diagnosis (*e.g.,* Ovarian Cyst, Outcome of delivery) whereas the diagnoses related to *Infectious Disease* are more general. The sLDA classifier show superior performance among all classifiers in Figure 6.1 as it get the highest precision for 8 out of 12 taxonomy codes and

[1] Parameter K is tuned based on the best performance and the value of 9 is reasonable since each class has at least 20 instances.

Experiment set	Classifier	Precision	Recall
12-core	Naive	70.2%	69.6%
	Decision Tree	72.6%	69.2%
	SVM	82.1%	81.8%
	PCA-KNN	77.0%	66.2%
	sLDA	$\mathbf{83.9\%^{ndk}}$	$\mathbf{83.6\%^{nsdk}}$
12-core & 1-non-core	Naive	67.4%	39.8%
	Decision Tree	75.4%	75.2%
	SVM	74.0%	77.0%
	PCA-KNN	51.7%	50.0%
	sLDA	$\mathbf{76.2\%^{nk}}$	$\mathbf{78.3\%^{nsdk}}$
23-all	Naive	43.3%	39.3%
	Decision Tree	57.7%	55.6%
	SVM	57.9%	61.5%
	PCA-KNN	61.2%	48.4%
	sLDA	$\mathbf{62.8\%^{nsdk}}$	$\mathbf{62.3\%^{nsdk}}$
12-core in 23-all	Naive	42.3%	42.2%
	Decision Tree	55.0%	52.7%
	SVM	62.2%	56.5%
	PCA-KNN	61.9%	53.6%
	sLDA	$\mathbf{66.4\%^{nsdk}}$	$\mathbf{70.1\%^{nsdk}}$

Table 6.1: Experimental results of classifiers (in percentage). The best results are boldfaced. Superscripts n, d, s and k denote that the performance difference is statistically significant ($p < 0.05$) compared to Naive, Decision Tree, SVM and PCA-KNN for the respective metrics.

the highest recall for 6 out of 12 taxonomy codes. The classifiers' average performance (weighted by the number of users with each taxonomy code) for the *12-core* experiment setting reported in Table 6.1 corroborated this finding. The SVM classifier and the Decision Tree classifier also did well and their performance are acceptable throughout all 12 taxonomy codes. The PCA-KNN classifier faltered markedly on the codes *Urology* and *Neurological Surgery* as reporting 0% precision and recall for both. This is possibly due to the dimensionality reduction with PCA. In Section 4.3, we choose 50 as the number of output features from PCA by maximizing the PCA-KNN classifier's overall performance in the *12-core* experiment setting. However, 50 may not be the best choice for all 12 core NPI taxonomy codes. Therefore, although PCA-KNN has shown good average precision and recall in Table 6.1 it can fail on a small portion of taxonomy codes such as *Urol-*

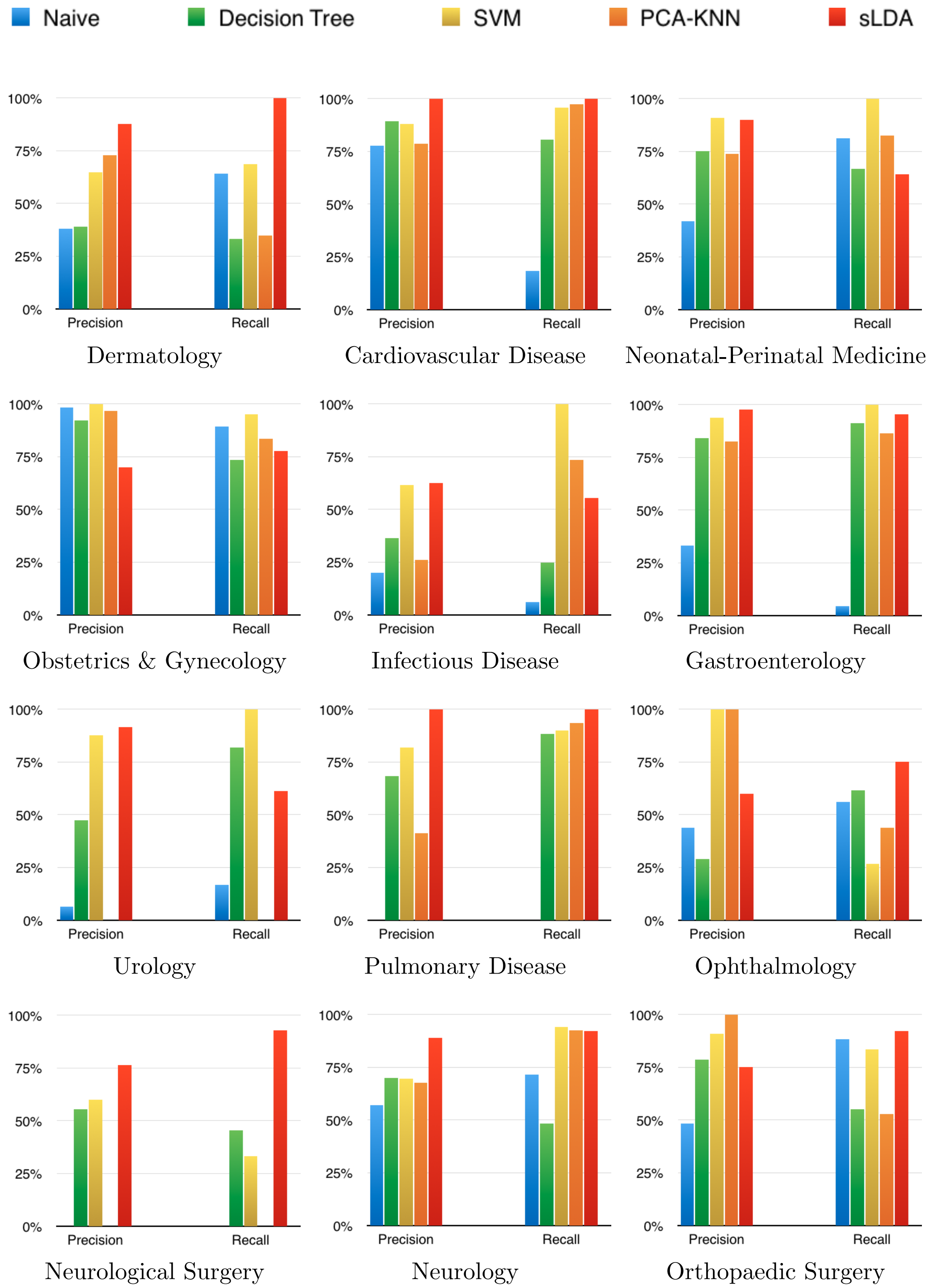

Figure 6.1: Performance of the classifiers for each of the 12 core NPI taxonomy codes in the setting of *12-core*.

ogy and *Neurological Surgery*. The Naive classifier performs the worst for most of the 12 taxonomy codes and it has reported 0% precision and recall for code *Pulmonary Disease* and *Neurological Surgery*. As PCA-KNN, the Naive classifier involves dimensionality reduction but it is done in a very crude way as described in Section 4.1—by selecting the m most relevant diagnoses as features. We picked $m = 20$ by maximizing the classifier's performance but the resulting features are not effective enough to describe the users and specialties. For instance, Table 6.2 shows the diagnosis features selected to represent the *Pulmonary Disease* code. It is not very specific to pulmonary-related conditions.

Diagnosis Features
Residual codes; unclassified; all E codes
Pleurisy; pleural effusion
Other and unspecified lower respiratory disease
Other and unspecified liver disorders
Chronic airway obstruction; not otherwise specified
Essential hypertension
Other injuries and conditions due to external causes
Congestive heart failure
Pneumonia; organism unspecified
Other connective tissue disease
Other fluid and electrolyte disorders
Anemia; unspecified
Emphysema
Respiratory failure
Other and unspecified circulatory disease
Pulmonary collapse; interstitial and compensatory emphysema
Atrial fibrillation
Other complications of surgical and medical procedures
Other back pain and disorders
Phlebitis and thrombophlebitis

Table 6.2: Diagnosis Features for *Pulmonary Disease*

To justify that sLDA's outperforming is statistically significant in general, we conducted the paired t-test and the results are also reported in Table 6.1. Although the precision of SVM classifier is close to sLDA in the experimental setting of *12-core*, the recall of sLDA is still significantly better than that of SVM. The better recall with statistical significance for sLDA than that of Decision Tree and SVM have also been observed in the setting of *12-core and 1-non-core*.

According to Table 6.1, the naive approach does not show promising results

Urology topic
Calculus of kidney
Elevated prostate specific antigen
Hematuria
Impotence of organic origin
Incomplete bladder emptying
Bladder neck obstruction
Urinary frequency
Hydronephrosis
Unspecified retention of urine
Other testicular hypofunction

(a) Urology Topic

Opthalmology Topic
Cortical senile cataract
Lens replaced by other means
Myopia
Unspecified tear film insufficiency
Cataract extraction status
Primary open-angle glaucoma
Unspecified cataract
Borderline glaucoma
Vitreous degeneration
Blepharitis, unspecified

(b) Opthalmology Topic

Table 6.3: Examples of Diagnosis Topics

especially when the number of classes grows and the multi-class problem gets more challenging. Smarter methods like Decision Tree demonstrates better performance than the naive approach due to its more sophisticated mechanisms in the attribute selection and partition. Moreover, in most cases both SVM and PCA-KNN perform better than Decision Tree because they are more effective on high dimensional data. Notably, SVM, PCA-KNN and sLDA reported better performance in the setting of *12-core in 23-all* than those in the setting of *23-all*, and this resonate well with our choice of the 12 core diagnosis specialties. Particularly, it is observed that PCA-KNN performs drastically worse in the setting of *12-core and 1-non-core* than in the other experimental settings. This may be explained by the fact that the "non-core" class in this experimental setting is very noisy since it is the mixture of 11 NPI taxonomy codes whereas KNN is extremely sensitive to noisy data [30].

The experimental results indicate that sLDA is the superior classifier on predicting physicians' specialties based on their associated diagnoses. The key reasons are as follows:

1. sLDA classifies users based on the weighting of diagnosis *topics* rather than simple feature weighting of diagnoses as in the case of other classifiers. The semantic topic summary filters out noises from individual diagnoses and generates more representative attributes.

2. The topics generated by sLDA are descriptive of diagnosis specialties.

Therefore, most of the diagnosis specialists are near the edge simplex of the diagnosis topics that reflects their specialties.

To highlight the second point, Table 6.3 shows two examples of the diagnosis topics generated by sLDA.

6.1.3 False Positives and Negatives

The previous section focused on the true predictions by the classifiers, but in some cases the false positives and negatives may be more interesting. A false negative represents an instance in which a user has a specialty but the classifier does not identify this specialty from the treatment history of the user. A false positive is an instance where a user does not have a specialty, but appears to have one based on the concentration of diagnoses in their treatment history.

False Negative Prediction

We found two primary reasons for false negatives. The first was relying on only a small value of $T(u)$, that is, making predictions for users who had treated only a small number of patients. The second was users who do not seem to be primarily treating patients in their specialty. This may be because: (1) they treat few patients in their declared area of specialty, or (2) they treat many patients outside their specialty, or (3) their specialty designation represents an error, or (4) some combination of these factors.

As an example of the first case, consider user u who is a physician and lists a taxonomy code in Ophthalmology (207W000000X). Table 6.4a lists the 10 diagnoses u mainly treats [2]. It is easy to observe that these are not specifically related to ophthalmology. However, upon closer examination, we find that u had only two patient-encounters during our one year of data collection. This contrasts with an average number of encounters in our cohort of 97. Hence the classifier misclassified user u because the user's data was not sufficient.

[2] For each user we compute the TF-IDF weights of their associated diagnoses as we did for the naive classifier, and the 10 diagnoses with the highest weights is reported in the table

Diagnoses
Acute pancreatitis
Bipolar disorders
Other thyroid disorders
Nausea and vomiting
Abdominal pain
Other and unspecified gastrointestinal disorders
Other screening for suspected conditions
Nonspecific chest pain
Benign neoplasm of ovary
Administrative/social admission

(a) User u most related diagnoses

Diagnoses
Cancer of bladder
Cystitis and urethritis
Other and unspecified diseases of bladder and urethra
Cancer of prostate
Cancer of testis
Cancer of kidney and renal pelvis
Other male genital disorders
Other and unspecified lower respiratory disease
Hematuria
Cancer of other urinary organs

(b) User w most related diagnoses

Diagnoses
Other nervous system symptoms and disorders
Other injuries and conditions due to external causes
Malfunction of device; implant; and graft
Codes related to mental health disorders
Cellulitis and abscess of leg
Paralytic ileus
Other and unspecified upper respiratory infections
Administrative/social admission
Other upper respiratory disease
Benign neoplasm of ovary

(c) User v most related diagnoses

Table 6.4: User Cases

To see another example illustrating the second case, we consider a user v who is a physician with a taxonomy code for Neonatal-Perinatal Medicine (2080N0001X). The top 10 diagnoses of user v are given in Table 6.4c. Like user u, they do not look closely related to the specialty, but unlike user u, user v is very active with 754 encounters. With this many encounters one is tempted to the view that user v may well practice neonatal-perinatal medicine but v also does a large number of other things that cause the classifier to fail to recognize the specialty.

False Positive Prediction

The false positive predictions are cases where users are predicted to have NPI taxonomy codes they do not have. These cases overlap with the false negative cases when the classifier is used to classify diagnosis specialists. But when we extend the classifier to roles other than physicians, we find many users with non-diagnosis specialties who nevertheless saw specialized diagnoses.

For example, user w is a Registered Nurse (RN) who has over 1,300 patient-encounters. However, the top 10 most related diagnoses as shown for w in Table 6.4b) are consistent with a Urology specialty, which is what the classifier for urology concluded about w. We did not have enough RNs in our dataset to tell if RN would have been a diagnosis specialty, but we conjecture that it would not have been one by itself. However, RNs can work in specialized fields, as w appears to do, even though it is not reflected by a taxonomy code.

The false positive specialty predictions may have significant derivative value. As in the case of the RN who focuses on urology patients, incidentally included in this analysis given an NPI entry, the latent expertise for other clinicians without NPI based specialties may be revealed through this analysis. With this new knowledge of staff members' clinical experience, staffing decisions and clinical coverage for short-staffed units becomes capable of automation yielding the possibility of more appropriately mapped staff allocations as the individual with the best experience match is chosen for coverage of a specific unit.

6.2 Evaluation of Diagnosis Specialty Discovery

6.2.1 LDA Outcomes

After running LDA on the physicians in our dataset, we found three promising candidate topics that can lead to new diagnosis specialties. The three candidates are: Breast Cancer, Obesity, Ear & Nose & Throat (ENT), and their respective topic summaries are provided in Table 6.5. It is clear from the topic summaries that these topics each have a distinct diagnosis pattern.

Then we cluster the users by the method described in Section 5.1.3, and

Topic 10 diagnoses
Obesity, unspecified
Morbid obesity
Obstructive sleep apnea
Unspecified sleep apnea
Hypersomnia with sleep apnea, unspecified
Paralysis agitans
Hip joint replacement by other means
Edema
Other dyspnea and respiratory abnormality
Body Mass Index 4

(a) Obesity

Topic 10 diagnoses
Allergic rhinitis due to other allergen
Allergic rhinitis due to pollen
Extrinsic asthma, unspecified
Unspecified asthma, with exacerbation
Intrinsic asthma, unspecified
Unspecified sinusitis (chronic)
Allergic rhinitis, cause unspecified
Chronic rhinitis
Polyp of nasal cavity
Cough variant asthma

(b) ENT

Topic 10 diagnoses
Personal history of malignant neoplasm of breast
Lump or mass in breast
Abnormal mammogram, unspecified
Other specified aftercare following surgery
Other sign and symptom in breast
Carcinoma in situ of breast
Family history of malignant neoplasm of breast
Other specified disorder of breast
Benign neoplasm of breast
Acquired absence of breast and nipple

(c) Breast Cancer

Table 6.5: Potential New Diagnosis Specialties

we put the three topics' clusters to test with the two criteria introduced in Section 5.1.4 to determine whether they are eligible to become diagnosis specialties. Table 6.6 show the evaluation statistics of the potential diagnosis specialties.

Topic Description		Breast Cancer	ENT	Obesity	Average *12-core*
Cluster Size		68	3	20	40 (20)
SVM	Precision	68.0%	N/A	13.9%	80.9% (60%)
	Recall	75.0%	N/A	10.3%	81.0% (50%)
PCA-KNN	Precision	62.1%	N/A	0%	76.0% (41.2%)
	Recall	55.9%	N/A	0%	66.2% (36.8%)

Table 6.6: Evaluation Statistics of Potential Diagnosis Specialtis

In Table 6.6, the column *Cluster Size* shows the number of users in each cluster and the column *Prediction* shows how well a potential diagnosis specialty can be learned by machine learning classifiers. The experiment setting is similar to *12-core* in Section 6.1 except we added these three potential specialties to make it *15-core*. Also we excluded the sLDA classifier from prediction. It is because sLDA classify users based on their topic allocations, and if the test cases are initially clustered by their topic allocations (as in our case of using LDA) then sLDA is almost guaranteed to have high performance. In this way, sLDA is not effective in evaluating the topics based on their semantics. We end up using the SVM and PCA-KNN classifiers because they have shown good performance at learning specialties in Section 6.1. Also note the reported values are result of the 5×2 cross-validation.

We also included the column *Average 12-core* to show the average statistics of the 12 existing core diagnosis specialties. And the values in the parenthesis represent the lowest respective values among the 12 existing core diagnosis specialties. Overall, this column provides a baseline that helps us decide whether a topic cluster can be considered as a new diagnosis specialty.

From Table 6.6 we can see that *ENT* don't satisfy the first criterion so it is out. Although *Obesity* have enough user instances, it is also eliminated because none of our classifiers was able to learn the topic well. The topic cluster of *Breast Cancer* not only has a good number of crowd but it is also able to be learned by the classifier with precision and recall comparable to the existing diagnosis specialties. Therefore, we shall tender *Breast Cancer* as the only discovered new diagnosis specialty.

6.2.2 PathSelClus Results

We run PathSelClus with the target number of new specialties δ from 1 to 10. Then we analyze the semantics of each unseeded cluster by method described in Section 5.2.2 for every experiment setting of different δ in hope of finding some potential diagnosis specialties. To our surprise, it turns out that none of the experiment settings could produce a semantically meaningful cluster of users that can form a new specialist group. For example, Table 6.7 show the summaries of the three unseeded clusters when δ is set to 3.

The top diagnoses within each clusters in Table 6.7 seem to be all over the

Topic 10 diagnoses
Other bacterial infections
Other non-traumatic joint disorders
Convulsions
Other upper respiratory disease
Phlebitis and thrombophlebitis
Malaise and fatigue
Other skin disorders
Fever of unknown origin
Cardiomyopathy
Substance-related disorders

(a) First Cluster

Topic 10 diagnoses
Chronic kidney disease
Essential hypertension
Other cardiac dysrhythmias
Abdominal pain
Phlebitis and thrombophlebitis
Other fluid and electrolyte disorders
Anemia; unspecified
Pleurisy; pleural effusion
Acute renal failure
Hyperpotassemia

(b) Second Cluster

Topic 10 diagnoses
Abdominal pain
Other and unspecified lower respiratory disease
Nonspecific chest pain
Urinary tract infection; site not specified
Diabetes mellitus without complication
Essential hypertension
Other nervous system symptoms and disorders
Pneumonia; organism unspecified
Phlebitis and thrombophlebitis
Other and unspecified circulatory disease

(c) Third Cluster

Table 6.7: Cluster Summaries for $\delta = 3$

map. We are not able to make any interpretation based on these clusters. Therefore, PathSelClus fail to discover any new diagnosis specialties. This is most likely because such link-based clustering techniques are generally susceptible to noises from objects (*e.g. Diagnosis*) that form inter-cluster bridges [31], which are abundant in our data.

Chapter 7

Applications

The experiment results have proven this work to be effective at identifying providers' genuine specialty as expressed in their treatment history. Therefore, this work has potential for a rich variety of applications in which accurate specialty information about providers are beneficial. This chapter shows some prospective applications in the healthcare domains.

7.1 Security and Privacy

As EMR systems expand in size, scope and distribution, it is critical to ensure the privacy of those whose information is stored. There have been many attempts to establish various access control policies in EMR system. In this respect, our work can be useful in assisting the detection of anomalous EMR accesses.

In general it is expected that EMR users, if having specialties, should mostly be accessing patients relevant to their diagnosis specialties. Therefore, if we compile a diagnosis summary of a EMR user either by TF-IDF weighting (described in Section 4.1) or topic modeling (described in Section 4.2), the diagnosis summary should match the user's putative diagnosis specialty. For instance, Cardiology specialists should have many heart and valve related diagnoses in their diagnosis summaries. Otherwise, there is an discrepancy between the user's putative specialty and his/her chart activity. In such case, there are two possible explanations: (i) The users have changed their interests and specialties after they got their NPIs but they failed to update the information; (ii) The users are accessing patients' records anomalously for illegitimate purpose. For the former cases, our work can help verifying users' Healthcare Provider Taxonomy Codes are accurate. And if a major discrepancy is detected, the users and his department can be automatically notified. For the latter case, our work helps filter out users whose activities are consistent with their putative diagnosis specialties, leaving only the suspicious ones for further investigation.

7.2 Staffing Support

Staff planning is crucial for the long-term development of any hospital. Uninformed staffing decisions can lead to not only under-staffing or over-staffing but also an evil combination of both — imagine a hospital with an overstaffed Neurology department and an understaffed Urology department.

The specialist identification technique in this work enables hospitals to have a clear and accurate accounting of different types of providers they have. Knowing "what we have" is first step of successful staff planning. Another crucial question to answer needs to be "what we need?". A easy way out is to learn from other well-staffed hospitals. To do so, hospitals need to understand each others' specialist structures, which may not be easy since physicians at different hospital may use the taxonomy differently [4]. Our work can be used to provide consistent specialist manifests across all hospitals.

7.3 Patient Experience Enhancement

There are many research projects and government initiatives aiming at providing better healthcare experience for patients, and our work can be applied to those projects to make the patient experiences even better.

To help patients finding suitable physicians there are research projects that enable patients to search the web for physicians with specific clinical expertise [32]. It is desirable that search results are ordered by the physicians' specialties so that the most suitable physician appears on top. Our work can help assure that the ordering is accurate.

Another great project is Health Information Exchanges (HIE) that allows healthcare professionals and patients across different institutions to appropriately access and securely share patients' EMRs. For HIE to work, the difference between institutions needs to be considered [33]. Our work can be useful in that it can help ensure consistent NPI taxonomy code usage across individual institutions participating in HIE, thus lowering the communication barrier.

Chapter 8

Conclusion and Future Works

This work shows the feasibility of identifying providers' specialties by their treatment histories. In the course of this study we defined a group of provider specialties as diagnosis specialties, which are strongly determined by diagnosis treatment history. We devised an intuitive *naive classifier* to illustrate how human may infer a specialty from treatment history. Then we demonstrated that diagnosis specialties can be learned with good precision and recall by machine learning classifiers. We measured four statistically informed machine learning techniques—Decision tree, SVM, PCA-KNN and sLDA—for predicting users' diagnosis specialties and found the dimensionality-reduction technique sLDA to have the best performance. sLDA is able to predict diagnosis specialty with precision and recall above 80%. Additionally, we explored the possibility of discovering hidden diagnosis specialties by clustering providers by their treatment histories. We evaluated a topic modeling approach with LDA and a user-guided approach for HIN with PathSelClus. LDA is able to discover a *breast cancer* specialty in our dataset and we are able to verify its validity.

A limitation of this work is that we only considered the providers' primary taxonomy code and we assumed they could only have one specialty. This is true for most providers, but some providers do have more than one taxonomy code. Therefore, our future work can explore method to identify providers' subordinate specialties based on treatment history and department information. Furthermore, it can lead to techniques to rank the specialties a provider have. Multi-label machine learning techniques will be used to achieve these goals.

References

[1] C. for Medicare & Medicaid Services, "Taxonomy code." http://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/MedicareProviderSupEnroll/Taxonomy.html.

[2] Centers for Medicare & Medicaid Services, "The Health Insurance Portability and Accountability Act of 1996 (HIPAA)." Online at http://www.cms.hhs.gov/hipaa/, 1996.

[3] N. Plan and P. E. S. (NPPES), "National provider identifier." http://nppes.cms.hhs.gov/NPPES/Welcome.do.

[4] "The npi taxonomy codes for psychology: Apa practice organization offers guidance, advocates for change." http://www.apapracticecentral.org/reimbursement/npi/select-code.aspx.

[5] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACM Trans. Knowl. Discov. Data*, vol. 7, pp. 11:1–11:23, Sept. 2013.

[6] X. Lu, A. Zhang, C. A. Gunter, D. Fabbri, D. Liebovitz, and B. Malin, "Learning a medical specialty from a provider treatment history," *Under review for publication.*

[7] C. for Disease Control and Prevention, "Icd-9." http://www.cdc.gov/nchs/icd/icd9.htm.

[8] H. Cost and U. Project, "Clinical classifications software (ccs) for icd-9-cm." http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[9] C. for Medicare & Medicaid Services, "Icd-10." http://www.cms.gov/Medicare/Coding/ICD10/index.html?redirect=/icd10.

[10] W. Zhang, C. A. Gunter, D. Liebovitz, J. Tian, and B. Malin, "Role prediction using electronic medical record system audits," in *AMIA Annual Symposium Proceedings*, vol. 2011, p. 858, American Medical Informatics Association, 2011.

[11] C. A. Gunter, D. M. Liebovitz, and B. Malin, "Experience-based access management: A life-cycle framework for identity and access management systems.," *IEEE Security & Privacy*, vol. 9, no. 5, pp. 48–55, 2011.

[12] W. Zhang, Y. Chen, C. Gunter, D. Liebovitz, and B. Malin, "Evolving role definitions through permission invocation patterns," in *Proceedings of the 18th ACM symposium on Access control models and technologies*, pp. 37–48, ACM, 2013.

[13] D. Fabbri and K. LeFevre, "Explaining accesses to electronic medical records using diagnosis information," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 52–60, 2013.

[14] D. Fabbri and K. LeFevre, "Explanation-based auditing," *Proceedings of the VLDB Endowment*, vol. 5, no. 1, pp. 1–12, 2011.

[15] S. Gupta, C. Hanson, C. A. Gunter, M. Frank, D. Liebovitz, and B. Malin, "Modeling and detecting anomalous topic access," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pp. 100–105, IEEE, 2013.

[16] "Rxnorm." `http://www.nlm.nih.gov/research/umls/rxnorm/`.

[17] "Power law." `http://en.wikipedia.org/wiki/Power_law`.

[18] R. Hogg and E. Tanis, *Probability and Statistical Inference.* Pearson Prentice Hall, 2006.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition).* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[20] D. M. Blei and J. D. McAuliffe, "Supervised topic models," *arXiv preprint arXiv:1003.0783*, 2010.

[21] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques.* Morgan kaufmann, 2006.

[22] I. Jolliffe, *Principal component analysis.* Wiley Online Library, 2005.

[23] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[25] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 565–576, ACM, 2009.

[26] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *VLDB11*, 2011.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[28] T. Abeel, Y. Van de Peer, and Y. Saeys, "Java-ml: A machine learning library," *J. Mach. Learn. Res.*, vol. 10, pp. 931–934, June 2009.

[29] C. Wang, "Supervised latent dirichlet allocation for classification." `http://www.cs.cmu.edu/~chongw/slda`.

[30] P. Cunningham and S. J. Delany, "k-nearest neighbour classifiers," 2007.

[31] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. The Kluwer International Series in Engineering and Computer Science, Springer, 2005.

[32] C. L. Cole, A. S. Kanter, M. Cummens, S. Vostinar, and F. Naeymi-Rad, "Using a terminology server and consumer search phrases to help patients find physicians with particular expertise," *Medinfo*, vol. 11, no. Pt 1, pp. 492–6, 2004.

[33] J. R. Vest and L. D. Gamm, "Health information exchange: persistent challenges and new strategies," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 288–294, 2010.