Electronic Health Record Summarization over Heterogeneous and Irregularly Sampled Clinical Data

Rimma Pivovarov

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Electronic Health Record Summarization over Heterogeneous and Irregularly Sampled Clinical Data

Rimma Pivovarov

The increasing adoption of electronic health records (EHRs) has led to an unprecedented amount of patient health information stored in an electronic format. The ability to comb through this information is imperative, both for patient care and computational modeling. Creating a system to minimize unnecessary EHR data, automatically distill longitudinal patient information, and highlight salient parts of a patient's record is currently an unmet need. However, summarization of EHR data is not a trivial task, as there exist many challenges with reasoning over this data. EHR data elements are most often obtained at irregular intervals as patients are more likely to receive medical care when they are ill, than when they are healthy. The presence of narrative documentation adds another layer of complexity as the notes are riddled with over-sampled text, often caused by the frequent copy-and-pasting during the documentation process.

This dissertation synthesizes a set of challenges for automated EHR summarization identified in the literature and presents an array of methods for dealing with some of these challenges. We used hybrid data-driven and knowledge-based approaches to examine abundant redundancy in clinical narrative text, a data-driven approach to identify and mitigate biases in laboratory testing patterns with implications for using clinical data for research, and a probabilistic modeling approach to automatically summarize patient records and learn computational models of disease with heterogeneous data types. The dissertation also demonstrates two applications of the developed methods to important clinical questions: the questions of laboratory test overutilization and cohort selection from EHR data.

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are so very many people to thank, but I will limit myself to only two pages. First, Noémie. There would be no thesis without Noémie. I've always felt that the best decision I made at DBMI was choosing Noémie as my advisor. Noémie has spent countless hours helping me understand: (i) how to not break every server I log into, (ii) how to turn my convoluted ideas into a well thought out paper, (iii) how to be a kind and successful woman in science, and (iv) how nice roman numerals look in a paragraph of text. I'm certain that with time I will recognize more and more things that Noémie has taught me and I'm so thrilled that we can continue working together.

Thank you to my wonderful committee: George, Gil, Pete and Pete. I am so honored you agreed to be a part of my dissertation. Thank you for reading all of my work so thoroughly and genuinely helping me create a more cohesive, coherent, and impactful body of work.

Many many many people at DBMI have made this experience much brighter than it could have been. Hojjat, you have been the only person here for the entire time: from our first office together, to hours of co-TA office hours, to reunited cubicle buddies, to my brand new co-worker! I am in constant amazement of your kindness, which is so neatly wrapped in a very heavy layer of intelligence. Thanks for being my best DBMI buddy. Nicole, I'm certain that your presence helped keep my sanity intact. Watching you complete the PhD process with such poise (and such a curly cute baby) continues to motivate me to this day. Dave Albers, I don't know who else has been so encouraging and has believed in me from the very beginning to the very end. Thank you for your constant advice, support, encouragement, snacks, and humor. Adler, thank you for the innumerable hours you have spent showing me derivations, explaining Greek letters, teaching me how to look so cool when programming, and most importantly for your never-

ending patience. And finally, to Marina and Sharon who both make me extra cheerful whenever I see them in the hallway, and that always counts for so much.

Lizzy, my fellow (coffee-shop) dreamer, you always show me what I should strive for and with the PhD, it is no different. You are always able to hear me out, deeply empathize, and then make me laugh uncontrollably. Thank you for being my travel buddy and my favorite person to do many hours of nothing with. Mash, thank you to you as well. Your consistently honest, blunt, and absolutely hilarious way of interpreting any situation has helped keep me grounded and grateful.

And finally, the family. Garush, thank you for being the only one who can handle my constant barrage of crazy with so much ease, hilarity, patience, piggy back rides, and high fives. MaPa, thank you for always answering all of my phone calls…and texts and IMs and Facetimes and Skypes! Thank you for worrying about my safety and happiness and making it so wonderful to come home. Thank you for keeping me happy and sane.

Everyone, I really could not have done it without you.

Dedication

To my Babushkas and Dedushkas. I hope you would all be proud.

Chapter 1: Introduction

1.1 The Need for Summarization of EHR Data

The increased adoption of electronic health records (EHRs) has led to an unprecedented amount of patient health information stored in electronic format. Within the past decade, the number of healthcare practices that have some electronic capability to store patient data has grown to almost 80% and now hundreds of millions of patients across the United States have extensive medical histories in electronic form. As health information exchanges promise patient record integration across multiple care settings, the amount of available patient data will continue to explode (Adler-Milstein, Bates, and Jha 2011). In these health records, clinicians routinely document the care of their patients. Throughout the years, patient records accumulate medical history as a myriad of individual observations: results of laboratory tests and diagnostic procedures; interventions; medications; and detailed narratives about disease course, treatment options, and family and social history.

There is great potential for research in leveraging these large amounts of clinical data to learn about human health. The informatics community is posed to develop methods to mine the available information and ask questions such as: how can we further clinical knowledge, how can we assist clinicians in performing searches within and across patient records, how can we predict patient hospital course, and

how can we automatically condense records to provide succinct summaries of a patient's medical history? With this eruption of rich, complex, and essential health data for millions of patients, the informatics community has a new opportunity to tackle the challenges entailed in interpreting a wealth of health information.

However, this same availability of large records has raised concerns of information overload for the healthcare practioner (Farri 2012), with potential negative consequences on clinical work, such as errors of omission (McDonald 1976), delays (McDonald et al. 2014), and overall patient safety (Holden 2011). When caring for an individual patient, clinicians reason in the context of the patient's medical history. This is a cognitively difficult task. First, the search space for potential diseases that may account for the patients' symptoms is very large. Second, the individual clinical observations that form the patient's record are many, thus potentially overwhelming in aggregate, and at the same time each of them is potentially imperfect and uncertain.

Current EHR systems often do not present this tremendous amount of patient data in a way that supports clinical workflow or cognitive reasoning (Stead and Lin 2009). It is therefore imperative for patient care to automatically comb through the raw data points present in the records and detect timely and relevant information. Even more alarmingly, as the most chronically ill patients often have the largest datasets, their records are the most difficult to present coherently (Christensen and Grimsmo 2008). As an example, for a prevalent chronic condition in our institution, patients with chronic kidney disease have 338 notes on average in their record (from all clinical settings) gathered across an average of 14 years, with several patients' records containing over 4,000 notes. It is clear that during a regular medical visit, no practitioner can read hundreds of clinical notes. Fortunately, electronic storage of this health information provides an opportunity for EHR systems to "aid cognition through aggregation, trending, contextual relevance, minimizing superfluous data" (Schiff and Bates 2010). Currently available commercial EHR systems, however, inadequately address this need, sometimes providing organization of data but lacking in information synthesis (Laxmisan et al. 2012). Some vendor EHR dashboards display problem lists that aggregate billing codes but these are low in actionable knowledge (Van Vleck et al.

2008; Rosenbloom and Shultz 2012).

The focus of this thesis is to develop algorithms for automatically summarizing EHR data by creating methods to identify and address inherent challenges in the EHR data.

1.2 EHR Datasets

To develop robust methods for reasoning and learning over clinical datasets derived from EHRs, we must be cognizant of challenges that derive from the uniqueness of medical data. Here we list a set of issues confronted when developing methods for these clinical datasets:

Uncertainty: medicine inherently deals with uncertainty about both diagnoses and individual observations about patients. Clinicians observe patient symptoms and patient histories with different levels of certainty, and diagnoses themselves are probabilistic in nature.

Data Heterogeneity: clinical records are composed of different structured and unstructured data types, each with complementary information. Often, medications are listed as pharmacy orders and discussed in the clinical notes as well, and laboratory tests are reported in structured form and additionally recorded in clinical notes. To make matters more difficult, it is not uncommon for different data types to disagree; medication lists in clinical notes are often out of sync with structured medication lists from pharmacy orders.

Temporality: diseases evolve over different time scales. Some diseases are chronic in nature and their progression is documented in the record across many years, whereas other ailments are acute and appear in short spurts in the patient records.

Data sparsity: records are riddled with missing observations, and there is very often no regularity in the rate at which patients are observed. Many clinical observations are missing not at random (e.g., patients are observed more frequently when they are ill and come to see their physicians for treatment).

Healthcare features: in addition to the raw observations from clinical records (words in clinical notes, medications prescribed, etc.), features representative of the process of healthcare (rate of visitation, healthcare setting visited, etc.) can provide valuable information.

Inaccuracy: as clinicians themselves populate EHR information during the process of patient care, many mistakes are entered into the system: typos, selections from dropdowns, copy and paste errors that propagate old information, etc. This characteristic of EHR data differs greatly from clinical trial data, where there is a lot of manual effort involved in ensuring accuracy and validity of observations.

We specifically develop algorithms to take these challenges into account. For the experiments reported in this thesis, we use two different EHR datasets: the MIMIC II ICU dataset and the clinical data warehouse at NewYork Presbyterian Hospital (NYPH). All of the studies presented in this thesis are completed on varying subsets of these data, depending on the purpose of the study. Here we broadly describe both datasets and within each study detailed later, we specify the preprocessed subset used.

1.2.1 MIMIC II ICU Dataset

The MIMIC II Clinical Database (v2.26) (Saeed et al. 2011) is available at http://physionet.org/mimic2_clinical_overview.shtml. MIMIC II contains a de-identified set of over 23,000 adults from the Beth Israel Deaconess Medical Center's Intensive Care Units, including medical, surgical, and coronary care units. The dataset contains structured record data, and unstructured clinical note data. Patients have a broad set of existing conditions and reasons for being in the ICU. As this dataset is available to researchers who sign a data usage agreement, any work on this dataset can serve as a benchmark for future automated phenotyping algorithms.

1.2.2 NYPH Dataset

This dataset is derived from data collected at NewYork-Presbyterian Hospital (NYPH), a not-for-profit hospital in New York City. The clinical data warehouse at NYPH contains patient data for over 4 million patients from as early as 1985 to present day. The information in these records include pertinent information for medical care such as demographics, visits, medications ordered, laboratory tests performed, diagnoses that are billed for, problem lists, and clinical notes written by healthcare practitioners. For all of the experiments reported in this thesis, only adult (18 years old and older) patient records were used. Patient records include all of their visits, ranging from outpatient visits to hospital

admissions, to emergency department visits. Like the MIMIC patients, the NYPH patients have a broad set of conditions, but unlike them their records span years instead of days, and the documented conditions differ as well.

1.3 Thesis Approach

The dissertation describes mixed-methods approaches to dealing with some of the underlying issues in the summarization of both narrative and structured EHR data. Our goal was to use knowledge- and data-driven approaches to comb through clinical data and devise novel methods for dealing with redundancy, missing data, and automatic determination of saliency. Clinical records store patient information across a diverse set of data types. The work presented in this thesis examines a broad array of challenges in summarizing and making sense of clinical record data across many clinical data categories. The thesis explores the challenge of redundancy through the lens of clinical notes, examining how a patient's health history can influence which clinical concepts are redundant and which are not. The challenge of irregular sampling is explored through the lens of laboratory tests, examining whether the sampling rate of each laboratory test can provide important clinical insight. Finally, the thesis presents a method for combining diverse clinical data types to address the challenge of saliency, examining whether modeling laboratory tests, medications, clinical text, and diagnosis codes together can provide a clinically relevant summary of a patient's health states.

Aim 1 describes an approach for minimizing redundancy in the narrative part of a patient record; Aim 2 defines a method for identifying and mitigating the biases that exist in laboratory test measurements; and Aim 3 designs a summarizer that uses heterogeneous data to identify salient pieces of a patient's record.

Aim I: Develop a methodology to contextually reduce narrative redundancy in a patient's medical record by aggregating over-sampled concepts.

Aim II: Create a method for exploiting a patient's non-random patterns of missing laboratory test values for clinical modeling.

Aim III: Design a summarizer synthesizing heterogeneous data: clinical notes, laboratory tests, medications, and billing codes.

1.3.1 Aim I: Contextual Redundancy Removal in Clinical Notes

Objective: Develop a methodology to reduce over-sampling of concepts within patient records by identifying and aggregating similar concepts in clinical notes based upon the context of a patient's medical history.

Research Questions:

1. *Are there instances when clinical concepts should be aggregated?*

2. Can a patient's medical history alter which clinical concepts should be aggregated?

3. Does adding ontological knowledge to data-driven patient knowledge increase the accuracy of identifying concepts that should be aggregated?

Methods and Materials:

Aim 1 relies on a hybrid approach of data-driven and knowledge-driven complementary methods in order to identify contextually redundant concepts within a patient's longitudinal health record. The data-driven method of distributional semantics is able to capture the context surrounding each instance of a concept and has been successful in many settings including information retrieval (Manning and Schütze 2003), while ontological techniques are used to refine the results from the data. The methodology for assessing redundant concepts consists of three complementary similarity measures (Figure 1.1). One primary measure is data-based and relies on distributional semantics of patient notes authored by clinicians, while the other two are knowledge-based and rely on concept definitions and their relationships in the Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) terminology. Starting from a homogenous corpus of notes (i.e., notes about patients who share at least one clinical problem), notes are pre-processed to extract concepts mentioned in the corpus. The noted-based similarity measure ranks all pairs of concepts. The top-k pairs with the highest note-based similarity are then reordered using the two knowledge-based similarity measures.

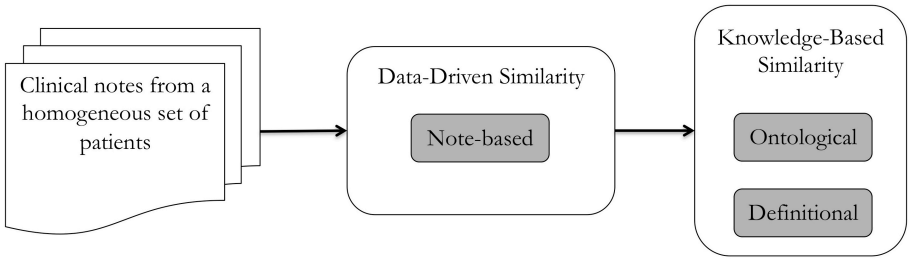


Figure 1.1 Our methodology for finding context-dependent similar concepts.

We choose to collect a homogenous and semantically coherent corpus of clinical notes; in order to ensure that concepts, which are clinically relevant to the patients, are likely to appear frequently enough and as the patient set is homogeneous we are able to find concepts that are similar specifically for the context of this particular disease. For this study, we collect a corpus of notes from patients with chronic kidney disease (CKD). The methods we employ are disease independent, but the fact that we select notes from patients all with at least one condition in common allows us to identify and aggregate concepts frequently mentioned when documenting a particular set of patients. Furthermore, CKD is a prevalent condition in our institution, thus allowing us to collect a large corpus of notes. Patients with CKD have many comorbidities and disorders, providing us with many different concepts to consider in our similarity computation.

Primary Findings:

Our approach scores the similarity of two input concepts by combining complementary information derived from usage patterns of clinical documentation, accepted definitions, and position of the concepts in an ontology. Our experiments show that given a coherent corpus of clinical notes, it is possible to determine automatically which concepts convey similar meaning in the context of the corpus. We demonstrate that by combining information from usage patterns in clinical notes and from ontological structure, the method can prune out concepts that are simply related from those which are semantically similar. Our method was evaluated against a gold-standard set of similar concepts: our method was able

to outperform the baseline methods and reached an AUC (area under the curve) of 92%. The results of the top 10 concept pairs identified by our method are demonstrated in Table 1.1.

UMLS Concept Preferred Term	**UMLS Concept Preferred Term**	**Similarity Score**	**Similar according to Expert Consensus**
C1998242 Traumatic injury of skeletal muscle	C0410256 Muscle Injury	1	Y
C1691215 Penile hypospadias	C0848558 Hypospadias	0.972	Y
C0240419 Muscle tenderness	C0575064 Skeletal muscle tender	0.966	Y
C2678517 Thrill (finding)	C0232269 Cardiac thrill (finding)	0.958	N
C0677659 Gastro-esophageal reflux disease with esophagitis	C0014869 Peptic Esophagitis	0.95	Y
C0149889 Anorectal fistula	C0205929 Anal fistula	0.945	Y
C0158458 Acquired hallux valgus	C0018536 Hallux Valgus	0.937	Y
C0520474 Aseptic Necrosis of Bone	C0029445 Bone necrosis	0.935	Y
C1261287 Stenosis	C0009814 Acquired stenosis	0.935	Y
C0243095 Finding	C0037088 Signs and Symptoms	0.933	Y

Table 1.1 Top-10 concept pairs found by the average of the data- and ontologically-based similarity measure.

1.3.2 Aim II: Exploiting Patterns of Missingness for Clinical Modeling of Laboratory Tests

Objective: Create a method for investigating and mitigating the effects of non-random sampling in laboratory tests on clinical modeling.

Research Questions:

1. Are there different missingness motifs across different laboratory tests?

2. Do the missingness motifs provide separate information to the laboratory test's numerical value?

3. Can adding the missingness motif to a laboratory value threshold increase phenotyping accuracy?

Methods and Materials:

The studies performed in Aim 2 categorize the different missing data patterns of irregularly sampled laboratory tests and evaluate the effect of these different patterns on EHR studies.

The first part of this study explores different patterns of irregularly sampled laboratory test data by considering the irregular sampling as a missing data problem. To understand the overall dynamics of laboratory tests recorded in the EHR, we categorize types of laboratory measurement motifs, identifying those more likely to cause biases in EHR-based research. We examine laboratory measurement patterns by looking at the distribution of days between consecutive measurements of each laboratory test, across the entire population. We examine and catalogue these resulting measurement gap histograms. If there is linearity in a measurement gap histogram when presented in log-log coordinates (i.e., a power-law) that implies scale-free measurement dynamics; in this situation all time scales are explainable by a single equation and likely represent a single context for ordering the laboratory test. If no approximately linear relationship between the frequency of measurement gaps exists, we looked for changes (e.g., peaks) that separate the different dynamics patterns; these different patterns may qualitatively imply different contexts of measurement based on either a change in health state or based on the healthcare documentation process. We catalogue the periodic measurement curves (measurement gap histograms)

based on observed approximate linearity and the presence of peaks in the histograms.

Missingness Motifs vs. Laboratory Test Values

In research with clinical data, there is an implicit assumption that a laboratory test's numerical value and the rate at which the test is ordered are highly correlated features. This assumption about value and measurement correlation likely stems from the existence of value-based guidelines and the widespread expectation that laboratory test values which fall outside of normal ranges prompt intervention and retesting. To investigate and quantify this assumption we ask: is there added information in looking at how a patient was measured, not only at the measurement value? Given a particular lab and all patients' time series for that lab, we construct a joint probability density function consisting of laboratory values and time between consecutive lab measurements (or gaps between measurement) in days. Using (i) linear correlation, (ii) mutual information, and (iii) differential entropy we explore the correlation of laboratory values to the laboratory test's time to repeat, examining whether the value and time between consecutive measurements (measurement gap) encode separate information or overlap in information content.

The Effect of Patterns of Missingness on Phenotyping

To demonstrate the utility of identifying and accounting for patterns of irregular sampling we conduct a typical EHR association test (Warner and Alterovitz 2012). In this task, we focus on a specific laboratory test as a use case for studying effect of measurement motifs on EHR-driven research. We examine lipase measurements and their impact on identifying acute pancreatitis. We asked the question: can the known association between an abnormal lipase value and acute pancreatitis be recovered from EHR data? To verify our hypothesis that patterns of missingness can impact the accuracy of identifying patients with acute pancreatitis, we considered three views of the data, based on the dynamics of lipase measurements within each patient's record: (i) only visits with short lipase measurement gaps, (ii) only visits with long lipase measurement gaps, and (iii) all visits independent of their lipase measurement gap. In each of these settings, we assessed the association between acute pancreatitis and lipase and studied the properties of visits that belong in the setting using ICD-9 codes and clinical notes. We hypothesize that as

acute pancreatitis is an acute disease, visits with short lipase measurement gaps will be more highly associated and relevant to acute pancreatitis.

Primary Findings:

We show that the context of a laboratory test measurement can often be captured by the way the test is measured through time. We perform three tasks to study the properties of these temporal measurement patterns. In the first task, we confirm that laboratory test measurement patterns provide additional information to the stand-alone numerical value. The second task identified three measurement pattern motifs across a set of 70 laboratory tests. Of these, two motifs exhibit properties that can lead to biased research results. In the third task, we demonstrate the potential for biased results on a specific example. We conduct an association study of lipase test values to acute pancreatitis. We observe a diluted signal when using only a lipase value threshold, whereas the full association is recovered when properly accounting for lipase measurements in different contexts (leveraging the lipase measurement patterns to separate the contexts) (Figure 1.2). We find that aggregating EHR data without separating distinct laboratory test measurement patterns can intermix patients with different diseases, leading to the confounding of signals in large-scale EHR analyses. This study results in a general methodology for leveraging measurement frequency to identify and reduce laboratory test biases.

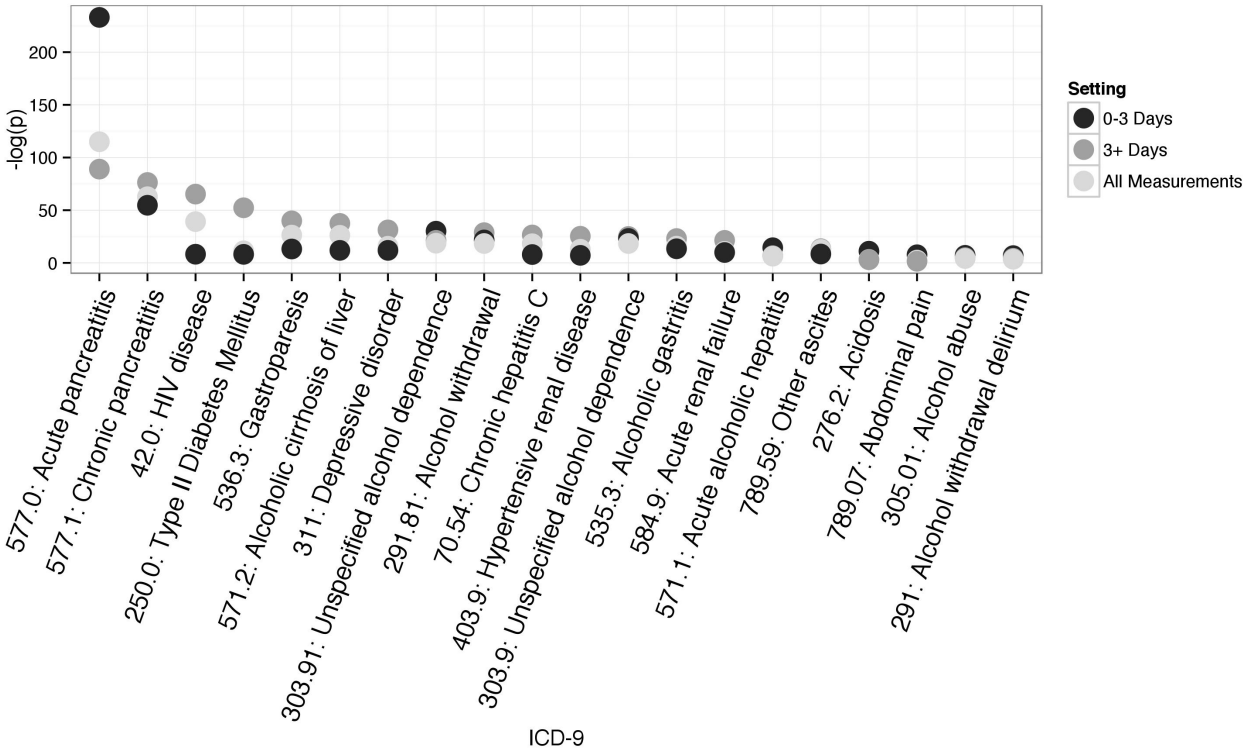


Figure 1.2 The results of a binomial association test between high lipase and ICD-9 codes. The binomial test was performed in all three settings (short gaps between measurements of 0–3 days, long gaps of more than 3 days, and all visits regardless of gaps between lipase measurements). The top 20 most significant associations are shown. For illustration purposes, the ICD-9 codes are sorted by association to high lipase in the 3 + days gap.

1.3.3 Aim III: Probabilistic Modeling of Patient Health States

Objective: Design a summarizer synthesizing heterogeneous data: clinical notes, laboratory test measurements, medications, and ICD-9 codes.

Research Questions:

1. *Is there a benefit to modeling different EHR data elements separately?*
2. *Are state-of-the-art evaluation techniques for probabilistic topic modeling applicable in the clinical domain?*

3. *Does the incorporation of ontological knowledge create inferred topics that better explain the data?*

Methods and Materials:

Building computational models of disease has been an active area of research, with approaches ranging from building ontologies and taxonomies of diseases based on clinical expertise, to creating highly precise model of specific diseases of interest through a mix of data-driven and clinical expertise, to discovering models directly from clinical observations. In this Aim, we create a model, the Phenome model, which works with the heterogeneous and irregularly sampled data types studied in Aims 1 and 2. The Phenome model is an unsupervised, generative model, which given a large set of EHR observations, learns probabilistic phenotypes. The model is a mixed membership probabilistic model of longitudinal patient records and phenotypes. Under this model, a patient record can be represented as a probabilistic mixture of phenotypes, and the phenotypes can be defined as a mixture of characteristics derived from a large, diverse population. Phenotype is defined here as a set of observations all related to a particular condition. The observations types used in this model are: medications, laboratory test instances, billing codes, and words from the narrative clinical documentation. The Phenome model is based on Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). The model enables the heterogeneous data elements encoded in a patient's record to be grouped through one phenotype definition and distributed across a predetermined number of P phenotypes. The phenotypes are learned jointly over heterogeneous EHR observations drawn from a large set of potential medications, diagnosis codes, laboratory tests, and free-text clinical notes. When applied to specific patient records, the Phenome model can provide actionable representation of the records, by describing them as a distribution over the patient's inferred phenotypes.

The goal of the Phenome model is to probabilistically identify sets of related heterogeneous clinical concepts that comprise phenotypes. Our goal is to learn phenotypes that are interpretable so that the model can be most useful in a clinical setting. To achieve this goal, we experiment with specifying informative prior distributions for the phenotypes such that each phenotype in the model is *a priori*

associated with concepts of clinical significance. We use a set of knowledge-bases to inject this clinical knowledge into the model.

We perform a set of different evaluations for this research. Automated evaluations include log-likelihood on a held-out set and automated topic coherence measures (Lau, Newman, and Baldwin 2014). A clinical expert performed evaluations about the coherence and granularity of learned phenotypes and did pairwise assessment of phenotypes learned by the Phenome model and the baseline methods. In addition, to examine how well the automated evaluations correlate with clinical expert judgment, we report on their correlation with the clinician. To evaluate the inferred phenotypes for patients given their EHR, we quantified the association between learned phenotypes and manually annotated disorders extracted from patient records. To evaluate the *a priori* grounding of phenotypes using clinical knowledge bases, we compared coherence and granularity scores between ungrounded and grounded phenotypes.

Primary Findings:

We demonstrate that the Phenome model can learn from different care settings and documentations of different healthcare institutions, without any adaptation needed. Our experiments show that the Phenome model yields phenotypes that (i) combine all these data types in a coherent fashion better than baseline models; (ii) are representative of single diseases, while baseline models tended to produce representations of either mixes of disease or high-level healthcare process; and (iii) when applied to unseen patient records, are highly correlated with the patients' ground-truth disorders. We also demonstrate that incorporating encoded clinical knowledge into the model leads to more coherent and cohesive phenotypes. Figure 1.3 presents an example of a phenotype that was learned after grounding the Phenome model. We demonstrate that large-scale probabilistic phenotyping is a promising approach to learning accurate and interpretable computational disease models.

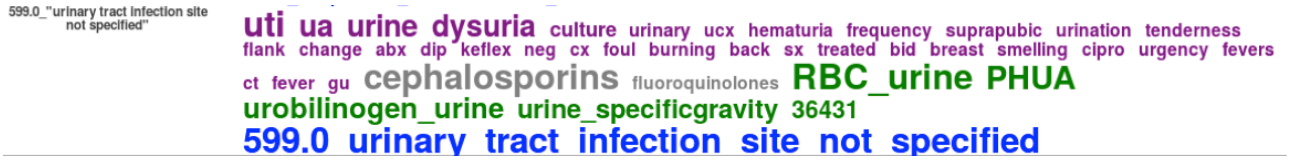


Figure 1.3 An example of a grounded phenotype learned by the Grounded Phenome model.

1.4 Contributions

The contributions of this dissertation are two-fold. The dissertation provides a literature review on work on clinical summarization as a whole and identifies important gaps that remain unsolved. The methodological contributions of this work include three novel methods for addressing some of the challenges in summarizing EHR data, namely: redundancy, irregular sampling, heterogeneity, and salience determination.

This dissertation also presents different ways in which raw data and ontological knowledge can be married, providing evidence for the benefit of combining these two data sources. The work presented here provides a platform for future work in applying latent variable models to EHR data as the models presented here are extendable in many different ways. Finally, the dissertation also presents two studies that demonstrate the applicability of the created methods for answering clinical questions.

1.5 Guide for the Reader

Chapter 2 is an in-depth review of the literature on EHR summarization. I identified published EHR summarization systems along with their inputs, outputs, method of summarization, methods of evaluation, and notable information. In addition, I highlight some of the remaining challenges in creating next generation EHR summarization systems.

Chapter 3 describes an approach that uses both knowledge-bases and clinical note structure to identify similar concepts in clinical text, as a way of finding which concepts can be aggregated for which groups of patients in an effort to reduce signal dilution and redundancy in studies with EHR data.

Chapter 4 reports on a study of the biases in laboratory test measurement and how these biases in measurement can be used to separate distinct patient populations.

Chapter 5 presents a probabilistic graphical model for summarizing patient data across heterogeneous data types along with an evaluation that demonstrates the utility of modeling different data types separately but jointly.

Chapter 6 details two separate experiments that showcase the applicability of the methods outlined in this thesis to clinical questions. One experiment applies laboratory test measurement pattern analysis to identify hemoglobin A1c testing trends over time, and the other uses the probabilistic graphical model to identify cohorts of diabetic patients.

Chapter 7 summarizes the contributions of the research described in this dissertation, as well as what I see as the next steps in EHR automated summarization.

Chapter 2: Background[1]

2.1 Approaches to EHR summarization

Given the unmet and well-recognized need for comprehensive EHR summarization (Powsner and Tufte 1994; Payne 2000), many research groups have designed and evaluated clinical data summarizers. In this chapter, we sample summarization applications to highlight different features including seminal work, different evaluation strategies, and various input/output data. We also examine the current work and future directions for six challenges of EHR summarization: information redundancy, temporality, missing data, salience detection, rules and heuristics, and deployment of summarization tools.

There are multiple theoretical frameworks for summarization in the clinical domain (Feblowitz et al. 2011) as well as for textual summarization in the general domain (Alterman 1991; Radev, Hovy, and McKeown 2002). In the broader field of summarization, there has been a lot of work in automated text

[1] A large part of this chapter was originally published in JAMIA. The full citation for this publication is: Pivovarov R, Elhadad N. Automated Methods for the Summarization of Electronic Health Records. JAMIA 2015.

summarization, specifically within the genres of news stories and scientific articles (see (Nenkova and McKeown 2012) for an in-depth review). Clinical summarization, "the act of collecting, distilling, and synthesizing patient information for the purpose of facilitating any of a wide range of clinical tasks" (Feblowitz et al. 2011), presents a different set of challenges from summarization in other domains and genres of texts.

While there exist other discussions on biomedical literature summarization methods (Afantenos 2006; Mishra et al. 2014) and EHR visualizations (Roque, Slaughter, and Tkatsenko 2010; Rind et al. 2010; West, Borland, and Hammond 2014), in this review we focus on characterizing existing clinical summarization systems by outlining the system outputs and evaluations as well as highlighting the remaining challenges that exist in automated summarization.

To categorize the summarizers highlighted in this review, we focus on two common dimensions used in the text summarization literature: extractive/abstractive summarization, and indicative/informative summarization. We define the four categories that describe summary types.

(1) *Extractive summaries* are created by borrowing phrases or sentences from the original input text. In the domain of clinical summarization, an extractive approach can identify pieces of the patient's record and display them without providing additional layers of abstraction.

(2) *Abstractive summaries* generate new text that synthesizes the original text. In the domain of clinical summarization, abstractive summaries may provide additional higher-level context to explain the data, such as computed quantities (e.g., trends) or automatically generated text.

Extractive and Abstractive summaries are further categorized as either indicative or informative.

(3) *Indicative summaries* point to important pieces of the original text, highlighting significant parts for the reader. In the domain of clinical summarization, indicative summaries may convey, for instance, when key tests were performed or diagnoses were made. Indicative summaries are meant to be used in conjunction with the full patient record.

(4) *Informative summaries* replace the original text. In the domain of clinical summarization, informative summaries are designed to be used independently of the full patient record, meaning they are used as a replacement for the original full set of raw data.

How to evaluate a summarizer, both its accuracy and its added value in supporting users carry out information-related tasks has also been the subject of investigation in general domain and clinical summarization. Intrinsic evaluations focus on the internal validity of a summarization tool. Typically, experts evaluate the quality of the automatically produced summaries; or themselves create gold-standard summaries, against which automatic ones are compared. In an extrinsic evaluation framework, the usefulness of the summarization tool is assessed through its effectiveness in helping individuals carry out a task. For instance, a clinical summary could be evaluated in an extrinsic fashion by comparing how quickly and accurately trial coordinators can identify patients eligible for a trial with access to patients' full records or with access to a summary instead.

Almost since the inception of EHRs, there has been an interest in creating meaningful succinct summaries for clinicians. The research on automated summary creation has spanned over 30 years and initiated with extracting recent structured events in a patient's history (Rogers and Haring 1979) evolving into performing natural language processing (NLP) (Liu and Friedman 2004) and automatically linking different data types (Cao et al. 2005; Klann et al. 2013) to create a more holistic view of the patient record. Table 1 lists clinical summarization systems proposed in the research literature in chronological order. We describe each system according to the following axes: the summarization approaches it implements, the type of input data it handles, the type of output summary, the way in which it was evaluated, and whether it was deployed in a clinical environment. Overall, summarization approaches investigated in clinical summarization have primarily been for indicative and extractive summarization. We also note a lack of evaluation, especially in the most recent years. We discuss in further detail the methods used for summarizing clinical data, along with the open research questions present in each of the summarization steps.

	Summarization approach	**Input**	**Output**	**Evaluation**	**Deployed (when is it generated)**	**General Notes**
NUCRSS (Rogers and Haring 1979; Rogers, Haring, and Watson 1979)	Extraction of clinical variables, indicative	Real structured EHR data	An 8 page summary of: Problem list, Vital signs, Cardiac-pulmonary-renal diagnoses, Treatments, Routine specialized laboratory examination, Suggestions to physicians regarding patient care	Laboratory study with medical students and physicians showed significant time savings and increased accuracy Randomized controlled trial showed that the NUCRSS improved process level (patient's length of stay and increased the amount of laboratory tests ordered) outcomes and may have improved care.	Yes (each patient visit)	Early example of a summarizer One of the few summary evaluations that demonstrate an impact on quality of care and process outcomes. The study found that the use of the summarizer was able to reduce patient's length of stay and increase the amount of laboratory tests that were ordered (perhaps because summarizer was able to remind clinicians of all the patients conditions that need monitoring).

STOR (O'Keefe and Simborg 1980)	Extraction of clinical variables, indicative	Real structured and unstructured EHR data	Loosely customizable, summary which included both time- and problem- oriented views	Clinical study found that clinicians were better able to predict their patient's future symptoms and laboratory test results when the using medical record in addition to STOR as opposed to just the medical record.	Yes (each patient visit)	Early example of a summarizer One of few examples of task-based evaluation The summary is context-dependent on the patient, but the context is manually determined by the clinician (what problems are active, what observations are relevant, etc.)
Powsner and Tufte (Powsner and Tufte 1994; Powsner and Tufte 1997)	Extraction of psychiatric variables and recent notes, indicative	Simulated structured, unstructured and genealogy data	A one page summary that visualizes the most salient content (as defined by recency) of the patient record.	None	No	A widely referenced prototype that continues to serve as a model for current EHR visualization and summarization applications.
Lifelines (Plaisant et al.	Extraction of	Simulated	Holistic	The original Lifelines	No	Lifelines is probably the most

1996; Plaisant et al. 1998)	clinical variables, indicative	structured data	interactive patient summaries using a temporal data view on top of the raw EHR data. Displays facts as lines on graphic time axis according to their temporal location and categories/significance are represented by color and thickness.	application was evaluated for work with juvenile youth records(Plaisant et al. 1996) by a small group of users who reported enthusiasm but mentioned potential biasing by the system's graphics.		well-known summarizer tool. The display has served as a model for future timeline-view clinical summarizers Lifelines2 was created for research and examining many patients together.
CliniViewer (Liu and Friedman 2004)	Extraction of concepts from text, indicative	Real unstructured EHR data	Combined NLP techniques and presented a tree view of a patient's	The system was able evaluated on accuracy and speed using real discharge summaries but no	No	One of the first examples of summaries created using NLP Allows for customizable user

			problems extracted from the narrative text to the clinician. Displays concepts in context when clicked.	evaluation with clinicians was conducted.		views Works on top of the MedLEE (Friedman et al. 1994) NLP engine which handles modifiers
IHC Patient Worksheet (Wilcox et al. 2005)	Extraction of clinical variables, indicative	Real structured EHR data	1-2 page outpatient summary of: demographics, problems, medications, laboratory tests, actionable advisories	A retrospective cohort study found that compliance with HbA1c testing was higher for patients who had a worksheet printed than for those who didn't.	Yes (each patient visit)	One of the few example of a clinical outcome tested in the evaluation
CLEF (Hallett and Scott 2005; Rogers, Puleston, and Rector 2006; Hallett 2008)	Abstraction from text and extraction of clinical	Simulated structured and unstructured	An interactive display of both navigational capabilities for the	None	No	One of the few natural language generation systems created for medical histories. Represents histories as a

	variables, indicative	cancer patient data.	EHR (indicative) and generates textual summaries (abstractive) to enhance comprehension. It uses information extraction techniques to identify classes of data and relationships between them.			semantic network of events organized temporally and semantically. Lists requirements that are very relevant to general designers of clinical summaries – the list was generated via initial requirements elicitation process. Uses a logical model of cancer history
KNAVE-II (Shahar et al. 2006)	Abstraction and extraction of clinical variables, informative	Real structured data on bone marrow transplant patients	Interactive data display of abstracted and raw protocol-based care data containing a tree-	A crossover study compared KNAVE-II with paper charts and Excel spreadsheet. Users produced quicker answers, had somewhat	No	Performs semantic, temporal, and context abstraction. Requires Domain-specific ontologies. Consists of a knowledge base, abstraction generator,

			browser and time chart.	better accuracy and preferred KNAVE-II however it did not achieve a very high system usability score.		navigation engine, and visualization. Lists 12 desiderata for interactive, time-oriented clinical data that should be used to guide future summarization work as well.
BabyTalk (BT-45) (Hunter et al. 2008; van der Meulen et al. 2010)	Abstraction of ICU data streams, informative	Real raw neonatal ICU data streams	Automatically generated natural language to describe ICU data streams for easier comprehension by the nursing staff.	A laboratory study found that human-generated text summaries of ICU streams helped nurses predict their patient's trajectories' better. The team is working to create automatically generated text summaries that perform as well as human-generated summaries.	No	A novel example of summarizing graphical ICU information by generating text.
Were et al. (Were et al. 2010)	Extraction of clinical	Real structured	Patient summary for use in an HIV	A pre-post study design using time-motion study	Yes (each	A largely successful process outcome.

	variables, indicative	EHR data from OpenMRS	clinic in Uganda	techniques and surveys. The authors found that providers who used the summary sheet were both able to spend more time directly with their patients and the average length of visit was reduced by 11.5 minutes.	patient visit)	Explores the utility of summaries in a low resource setting.
Timeline/AdaptEHR (Bui, Aberle, and Kangarloo 2007; Bashyam et al. 2009)	Abstraction from text and extraction of clinical variables, informative	Real structured, unstructured and image data on brain tumor patients	An interactive data display that summarizes and integrates various pieces of the EHR including images and free text.	A pilot study on Timeline found that although the initial learning curve was high, with time, the clinicians were able to perform image review quicker and were more confident in their clinical conclusions than when they used the EHR display	No	Timeline had manually coded rules while AdaptEHR aims to automatically infer rules and relationships from ontologies and graphical models, the publication states that the conditional probability tables are not yet defined. Has four dimensions of representing data: time space

						(where is the physical location of the tumor), existence (certainty), and causality (response to treatment)
HARVEST (Hirsch et al. 2014)	Extraction of concepts from text and clinical variables, indicative	Real structured and unstructured EHR data	A problem-based, interactive, temporal visualization of a longitudinal patient record.	A task-based, time evaluation found no difference in ability to extract, compare, synthesize and recall clinical information when using HARVEST in addition to the EHR, when carried out with subjects who had no prior experience with the summarization tool	Yes (real time)	Aggregates information from multiple care settings. Operates on top of a commercial EHR system using HL7 messages. Distributed computing infrastructure to enable real-time summarization.

Table 2.1 A sampling of clinical summarization applications. The table is organized by publication date. The inputs, outputs, methods, and evaluation strategies are listed along with notable additional information for each summarizer.

2.2 Methodological challenges to EHR summarization

The following sections present some unsolved challenges in clinical summarization. A conceptual framework proposed by Feblowitz et al. (Feblowitz et al. 2011) defines a set of actions that successful summarizers should accomplish with raw information: Aggregate, Organize, Reduce/Transform, Interpret, Synthesize. We discuss methodological challenges with automated summarization within the context of this framework.

Specifically,

- To successfully *aggregate* disparate clinical data sources, the ability to recognize and account for **similarity** is imperative. Such similarity occurs at different levels within narratives: from word-level similarity to concept to statement-level; as well as in other data types and across. We focus our discussion on textual similarity.

- The *organization* and *interpretation* of the aggregated data requires extraction and reasoning over clinical events and their **temporality**. We examine extraction of temporal information from text along with representation and reasoning over clinical events.

- The *organization* and *interpretation* of the aggregated data also requires that **missing data** points be accounted for. Patients are sometimes seen with predictable regularity but are most often seen at erratic intervals. Missing data points are often filled in by imputation, adding missing data indicators, deleting information with missing data, or other strategies.

- In the *reduction and transformation* of data and its *synthesis*, it is critical to decide which pieces of information are **important** and must be contained in the summary. Some methods for automatically detecting importance have relied on linguistic structure while others use probabilistic modeling techniques.

- To provide context for *interpretation* and *synthesis* of clinical data, it is useful to employ **existing knowledge** and create rules for the summarization. Knowledge-based heuristics often provide a way to specify time constraints, concept relationships, and abstractions.

- Finally, to successfully implement summarizers into clinical care, challenges of **deployment** need to be addressed. Because in vendor EHR systems there are limited opportunities to deploy innovative and experimental technology, there have been few attempts to translate patient record summarization systems into the clinic; however, to demonstrate utility, it is imperative to implement and study clinical summarization tools in the real world care setting.

2.2.1 Identifying and aggregating similar information

We review approaches to identifying and aggregating similar information on three different levels of language abstraction: words, concepts, and statements, as investigated within and outside the field of clinical summarization.

Word-level Similarity

In clinical NLP, much work has been devoted to identifying lexical variants that are similar in meaning (Friedman and Elhadad 2014). The Unified Medical Language System (UMLS), an agglomeration of different biomedical terminologies (Lindberg, Humphreys, and McCray 1993), for example, provides essential knowledge towards that goal by grouping words into concepts. The UMLS aggregates terms from different vocabularies and maps them to semantic concepts, each labeled with a Concept Unique Identifier (CUI). For instance, the terms MI, myocardial infarction, and heart attack all share lexical similarity, and map to the same underlying CUI. Within clinical summarization, normalization of words to concepts has only recently been investigated (Hirsch et al. 2014; Zhang et al. 2011).

An alternative, and most common approach in clinical summarization, is to identify word-level similarity by finding redundant strings of words. Patient records often contain redundant spans of text – this can be explained by the fact that documentation is often formulaic but also by the common habit of clinicians to copy and paste text from one note to another (Hirschtick 2006). Multiple different automated methods have been employed to identify copy and pasted words within clinical notes. A plagiarism detection tool called CopyFind has been used to identify overlapping phrases in input texts (Thornton et

al. 2013). More recently, global (Wrenn et al. 2010) and local (Zhang et al. 2011; Cohen, Elhadad, and Elhadad 2013) bioinformatics-inspired alignments have been proposed for identifying redundant sections along with language modeling techniques for assigning probabilistic similarity scores for phrase pairs (Zhang et al. 2011).

Concept-level Similarity

Concept-level similarity represents a more abstract level of similarity than similarity between words and strings. For instance, the concepts "epilepsy" and "seizure" – despite being two different UMLS concepts – share much semantic similarity when conveyed in a patient record.

In certain well-defined domains, clinical summarization approaches have relied on aggregating concepts, helping further the goal of synthesis (Shahar et al. 2006; Hsu et al. 2012) primarily through well-defined ontologies. For broader domains, how to identify that two semantic concepts are similar enough to be aggregated remains an open question. Furthermore, in text processing, mapping from words to concepts remains difficult because of the strong ambiguity of language (Friedman and Elhadad 2014).

Detection of semantic redundancy has been investigated through two approaches: knowledge-free and knowledge-based. Knowledge-free similarity metrics have been developed for textual input. They rely on Harris' 1968 hypothesis which stipulates that concepts that appear in similar contexts are similar (Harris 1968). In practice, concepts are compared in a vector space, where each concept is a vector representing the context in which the concept typically occurs. This method has been implemented multiple times in the clinical domain to identify similar UMLS concepts (Pedersen et al. 2007; Patwardhan and Pedersen 2006; Pivovarov and Elhadad 2012). Knowledge-free approaches are attractive when there is little ontological knowledge available. Alternatively, knowledge-based methods leverage existing resources to determine the similarity of two concepts. For instance, if the two concepts are present in an ontology, similarity can be assessed through the structure of the ontology. Other knowledge-based methods include examining similarity of the two concepts' definitions. We refer the reader to detailed reviews of concept-based similarity (Pedersen et al. 2007; Pesquita et al. 2009). We discuss our hybrid methodology for finding similar concepts using knowledge-free, knowledge-based, and

ontological definitions in Chapter 3. However, despite the active research on this topic, these concept-level similarity methods have not been yet translated to most clinical summarization systems.

Statement-Level Similarity

A pervasive aspect of a patient record is the high level of statement redundancy across notes. For instance, two pathology reports for a given patient share many similar statements. Beyond the formulaic nature of documentation, statement-level redundancy also occurs because of copying and pasting from previous notes with some minimal editing of the copied statements.

In clinical summarization, there has been little work on this important aspect of similarity identification. Recently, a topic modeling approach was proposed to identify and control for such redundancy across patient notes (Cohen et al. 2014). In the general NLP community, identifying statement level similarity has been studied through the tasks of paraphrasing identification and textual entailment (Androutsopoulos 2010). Many of the methods in text summarization for identifying both unidirectional (textual entailment) and bidirectional (paraphrasing) similarity employ a hybrid of methods for word-level and concept-level redundancy such as string similarity, logic-based methods, and context-vector (Dagan et al. 2010).

Along with the need for higher order language similarity work in the clinical domain, there is an ongoing push to personalize similarity detection. It is well established that semantic similarity is context-dependent (Janowicz 2008) and a recent study suggests that redundancy be examined as a function of the patient's previous history (Farri 2012). While identification of similar contexts based on the patient's health is an ongoing direction of research (see Chapter 3), there is further work to be done in identifying context-specific similarity on higher-order semantic levels. Identifying similar words, concepts, and removing redundancy by patient-tailored information aggregation is an important direction for future EHR summarization methodology.

2.2.2 Organizing and reasoning over temporal events

Patients' health evolves on many different time scales. Some health events such as pneumonia present themselves sporadically while chronic conditions like diabetes develop and worsen over a period of years. The importance of presenting clinical data in a time-dependent fashion has been recognized for a long time (Fries 1974; Cousins and Kahn 1991; Samal et al. 2011) however accurate temporal representation remains an open problem (Zhou and Hripcsak 2007; Sun, Rumshisky, and Uzuner 2013; Wu et al. 2014). Automatic creation of a clinical data timeline from textual and structured clinical records requires temporal event extraction, ordering, and reasoning.

Temporality is an active research area in the genre of news summarization given the quick news cycle and fast-paced evolution of news stories (Allan, Gupta, and Khandelwal 2001). However, news summarization research cannot always be readily translated into the health domain, as the challenges in health data are unique (Combi and Shahar 1997; Cios and Moore 2002). For example, different note types and specialties have different temporal relationships: pathology reports are often about one moment in time without reference to historical ailments whereas discharge summaries describe an entire inpatient hospital stay and instructions for future care. Styler et al. identified four complexities with extracting temporal information in clinical data: (i) diversity of time expressions, (ii) complexity of determining temporal relations among events, (iii) the difficulty of handling the temporal granularity of an event and (iv) general NLP issues (Styler, Bethard, and Finan 2014).

After the extraction of event time, there is a need for performing relative temporal ordering (Savova, Bethard, and Styler 2009). Event ordering is difficult in part due to inexact wording, but also because clinical knowledge is often needed to infer how long conditions may last (e.g., a diabetes diagnosis is often not discussed at every visit but a clinician is aware that diabetes is a chronic condition, not an intermittently reoccurring condition each time the "diabetes" term is mentioned or the diabetes ICD-9 code is recorded) (Hripcsak et al. 2009). Some recent work in event ordering includes the representation of temporal disease progression separately for each problem by Sonnenberg et al., an

approach they call “clinical threading” (Sonnenberg et al. 2012) and frame-like semantic representations with rule-based temporal extraction to arrange problems on a timeline (Jung et al. 2011). Raghavan et al. identify and temporally order cross-narrative medical events across documents in clinical text using weighted finite state transducers (Raghavan et al. 2014).

Reasoning and abstraction of extracted clinical events to highlight disease progressions and trends is critical for creating succinct clinical summaries. Abstractions of temporal data can include combining events within a certain time frame and performing interval-based abstractions such as combining multiple chemotherapy drug mentions into a chemotherapy regimen time span (Klimov, Shahar, and Taieb-Maimon 2010) or reasoning about the length of time that symptoms lasted and their relation to diagnosis (Zhou, Parsons, and Hripcsak 2008). The questions of which events should be combined and what an appropriate time frame is remain difficult and currently resolved by leveraging clinical knowledge and ontologies. Time-dependent clinical summarization is a continuingly evolving research area and there is opportunity for automatically identifying, accurately ordering, and performing reasoning over temporal clinical events.

2.2.3 Accounting for and interpreting missing data

Clinical records are sparse: documentation only occurs when a patient is seen by a clinician, thus clinical records miss the overwhelmingly large amount of observations about a patient across their lifetime. When summarizing sparse data, a critical complication is how to interpret and reason over the missing data (Wells et al. 2013). In some cases, missing data is not important and can safely be ignored by a summarization system (e.g., a patient has no change in health status in between visits). In other cases, the presence of missing data hints at a salient aspect about the patient that needs to be highlighted within the summary (e.g., patient is too sick to come to their visit). How to interpret and determine the salience of missing data is a challenge, and one not investigated thus far in clinical summarization.

In the field of general statistics, there are three types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Little and Rubin

2002). Most techniques for dealing with missing data assume that data are MCAR or MAR distributed, and include (i) variations of complete-case analysis, where only data with no missing values are used, (ii) single imputation, where missing data are imputed based on the values observed (using the mean, median, linear interpolation, etc.), and (iii) likelihood-based methods which compute maximum likelihood estimates for missing data (Enders 2006).

In the clinical domain, there is mounting evidence that most of the data are MNAR (Lin and Haug 2008; Pivovarov et al. 2014). For these data, the missingness is informative, meaning that there is an underlying reason that the data are missing but that this reason is simply unobserved. Some techniques that use informative missing data properties to infer properties about clinical data have been proposed. Chapter 4 of this dissertation presents our work on identifying patterns of missing data and using them to help infer healthcare status of patients. A common way of using missing data in the clinical domain has been to look at how long values should last based on recorded measurements or documentation frequency. For example, laboratory test measurements have been studied to gather appropriate imputation time (Hug 2006) and to infer health status features (Weber and Kohane 2013). Van Vleck studied duration and persistence of problems in notes (Van Vleck and Elhadad 2010) as a function of missing data, while Klann (Klann and Schadow 2010) and Perotte (Perotte and Hripcsak 2013) both studied the duration of ICD-9 codes. Klann estimated the durations for which each ICD-9 code remains valid and Perotte automatically classified ICD-9 codes into chronic and acute conditions. The modeling work that most explicitly demonstrates informativeness in missing data examined the accuracy of prediction models when: (i) ignoring missing data, (ii) interpolating missing data, or (iii) incorporating a missing data indicator, and reported that the missing data indicator method performed best (Lin and Haug 2008). To properly provide context and infer trend lines, as demonstrated by Poh and de Lusignan for kidney disease data (Poh and de Lusignan 2011b; Poh and de Lusignan 2011a), or to make predictions in clinical summaries it is critical to incorporate missing data literature and techniques into summarizer applications. The utility of modeling missing data explicitly is clear, however this conclusion is not being translated into clinical summarization research yet.

2.2.4 Reducing information to only the most salient

Salience identification has been heavily researched in the general domain text summarization literature. Early methods for identifying important topics relied on counts: frequency (Luhn 1958) and term frequency-inverse document frequency, which corrects for word specificity (Jones 1972). Other methods have focused on structure, such as document structure (Edmundson 1969) or syntax structure (Marcu 1997) to identify important phrases. Syntactic information gleaned from the input document can identify which parts of a sentence are salient and which may be safely removed from a summary (e.g., a relative clause). It is unclear, however, how these approaches translate to the clinical domain, where syntactic structure is unconventional. Using prior knowledge of the input document structure (e.g., biomedical papers have an introduction, followed by a methods section) to weigh the salience of information pieces based on where they are conveyed in the document is, however, promising in the clinical domain (yet not investigated thus far). Clinical notes follow a pre-specified structure; a diagnosis mention might be more relevant when conveyed in the past medical history than in the family history for instance. A different method for salience identification, still within the general domain summarization field, leverages discourse by considering sentences in input documents through a network, where lexical similarity between sentences is represented by the network edges. In this representation, salient sentences are the ones with the highest centralities (Radev, Jing, and Budzikowska 2000; Erkan and Radev 2004).

An alternative method for identifying relevant information relies on probabilistic modeling techniques such as Hidden Markov Models for identifying topics and topic changes in a set of documents (Barzilay and Lee 2004) or hierarchical Latent Dirichlet Allocation (LDA)-type models for identifying novel information with respect to older documents (Delort and Alfonseca 2012). Our work on employing Bayesian learning techniques to the construction of effective automated summaries is described in Chapter 5.

The one type of salience detection that has been explicitly studied in the clinical domain is based on cue phrases. Cue phrases are pieces of text that signify that what follows is likely to be important. For

example, “In conclusion” often precedes an important summarizing statement (Edmundson 1969). In clinical documentation, de Estrada et al. developed a system called Puya that found cue phrases indicating normality or abnormality in the physical exam sections of notes (de Estrada, Murphy, and Barnett 1997). Another way of detecting salience relies on n-gram language modeling to identify the most recent information in the record, under the assumption that the newest information is the most salient for the provider to see (Zhang, Pakhomov, and Melton 2012; Zhang 2014). A visualization prototype used this n-gram model to automatically highlight text that was found to be novel, drawing the provider’s attention to the new findings (Farri et al. 2012).

Defining salience in an operative fashion for automated summarization is an open question. In the general domain, there is evidence that humans sometimes disagree about what pieces of information are indeed salient, and that salience is often task-specific (Nenkova and Passonneau 2004). Similarly, in the clinical domain, determining what is important for a clinician is also probably quite task-specific. Nevertheless, it is safe to say that salience of elements in the patient record is related to capturing the health status of the patient and how it changes through time (Farri 2012; Suermondt et al. 1993). How to do so automatically, that is how to link textual and individual raw low-granularity observations to high-level clinical abstractions is one of the paramount challenge of informatics research. For instance, there has been little formal investigation of clinically specific markers of importance such as absolute change of a laboratory test value, the rate of change, the rate of mention of a particular concept, and other importance cues.

2.2.5 Using existing clinical knowledge

The informatics community has invested enormous effort into codifying clinical knowledge in a variety of terminologies and ontologies. This knowledge representation effort has been successful in helping efforts like phenotyping combine terminological knowledge, expert reasoning, and machine learning to create actionable disease definitions (Pathak, Kho, and Denny 2013). Similarly in summarization work, it is important to make use of these available clinical knowledge representations and

use them to generate rules and heuristics. Chapter 5 examines our efforts to incorporate coded clinical knowledge into an unsupervised, data-driven automated summarization engine.

Several holistic summarization efforts leveraged terminologies to identify concepts that are semantically related (e.g., medications that treat particular conditions) (Klann et al. 2013) or rules to determine salience (e.g., identify and highlight the salient results that are abnormal) (Plaisant et al. 1998). However, summarization engines built for particular diseases benefit most often from manually crafted rules and disease-specific knowledge bases as they enable tailored, task-dependent systems. The KNAVE-II application (Shahar et al. 2006), created for synthesis of bone marrow transplant patients, relies on an expert-maintained knowledge base for creating a semantic navigation system and concept abstraction. The Timeline system (Bui, Aberle, and Kangarloo 2007) is also built on a manually coded set of rules which identify salient concepts for different diseases, and perform temporal event reasoning. In addition, summaries that are setting and user specific often use expert-driven rules to ascertain which pieces of data should be shown at which time and to whom. Although the incorporation of clinical expertise into summarization is often a laborious process and sometimes only covers specific domains of expertise, it provides critical help in addressing some of the similarity, temporality and salience challenges. Of relevance to this review, we note that while existing summarizers rely on established knowledge resources, there is an active field of research to create these resources either by translating clinical expertise or acquiring the resources from data (Noy et al. 2009; Mortensen et al. 2012; Tao et al. 2013).

2.2.6 Deploying summarization tools into the clinic

The ultimate goal of any clinical summarization tool is implementation and usage by clinicians at the point of care. To date, however, there has been no widespread adoption of automated summarizers, especially for the large holistic temporal summarizers (Samal et al. 2011). Pervasive deployment is often hindered by the commercial EHRs systems that have been adopted across the country. Building real-time computational tools to work atop commercially built EHR systems is still a daunting task as these vendor

EHR systems are often not built to support interaction with outside applications. In addition, as the systems are closed off, dissemination of summaries across different hospitals and EHRs is a challenge as well. However, there is promising work with the i2b2-SMART platform that enables easier translation across institutions; researchers have developed a system to automatically link different data types across the EHR (mainly diseases and medications) and display a newly organized view of the patient record (Klann et al. 2013).

To create meaningful and practical summaries that assist clinicians during their point of care needs, summarizers need to provide real-time information with patient record updates immediately available in the summary. This is an especially difficult task when the summary tool works with natural language, as the processing must be completed quickly and accurately. Current work with distributed infrastructures, like Apache Hadoop, provides promising results for immediate summarization (Hirsch et al. 2014).

Another large barrier to translation of summarizer research into the clinical domain is rigorous evaluation. Hospitals often call for evidence of a useful summarizer before investing expensive resources into the implementation of the summarizer, but without adoption a summarizer is extremely difficult to evaluate. As is clear from Table 1, clinical summarization literature lacks standard evaluation metrics and there are very few extrinsic evaluations, a similar finding to a review of biomedical literature summarization by Mishra et al. (Mishra et al. 2014). Given the restriction of limited adoption, it is not clear on which dimensions clinical summarizers should be evaluated. Initially, in order to avoid costly development and implementations with marginal benefit, it is imperative to study the need for a summarizer tool, context of usage, and clinician workflow. However, without eventual implementation into clinical care, showing any process- or health-level outcomes is not possible and therefore how to perform useful evaluations remains unclear: should, for instance, summarization systems focus on accurate information extraction, facilitating information exploration (e.g., which concepts are most relevant to the clinician), or user-friendly designs? Although the rigorous user-interface and cognitive process evaluations that are necessary for creating new summarization systems often require deployment and study of actual use in practice, there exists guidance in the literature on cognitive aspects of clinical

reasoning that can inform summarization system creation. Prior work on general medical cognition (V. L. Patel, Arocha, and Kaufman 2001), clinical decision-making (Arocha, Wang, and Patel 2005; Kushniruk 2001), human-computer interaction for interface design (Patel and Kushniruk 1998; Jaspers et al. 2004; Thyvalikakath et al. 2014), handoff communication (Abraham et al. 2011; Abraham et al. 2014), clinical workflow analysis (Militello et al. 2014; Unertl et al. 2009) and some recent qualitative work specifically on clinical document synthesis which has identified common cognitive pathways for EHR document synthesis (Farri 2012) and patterns of EHR data access (Reichert et al. 2010) can guide the development of summarization systems. However, we emphasize that without actually studying the clinical context and manner in which clinicians use summarizers (either in the laboratory with prototype systems or in the clinic with deployed systems), it will be challenging to develop better evaluation strategies and better summarizers.

Chapter 3: Contextual Redundancy Removal in Clinical Notes[2]

3.1 Introduction to Similarity Detection

A standard way of approaching unstructured biomedical texts, such as patient notes written by clinicians, is to map mentions of biomedical terms, like symptoms and disease names, to semantic concepts in structured and standardized nomenclatures. The mapping helps group all lexical variants of the same biomedical concept under a unique semantic representation, thereby abstracting away from stylistic differences. For instance, the terms "heart attack," "myocardial infarction," and "MI" are all mapped to the same concept in UMLS. However, most biomedical ontologies and terminologies are designed based on a fine-grained organization of semantic concepts. As a result, when mapping term mentions in a text to semantic concepts, all too often semantically similar terms are mapped to different

[2] Most of this chapter was originally published in the Journal of Biomedical Informatics. The full citation for this publication is: Pivovarov R, Elhadad N. A Hybrid Knowledge-Based and Data-Driven Approach to Identifying Semantically Similar Concepts. J Biomed Inform. 2012;45(3)471-481.

concepts in the ontology. When the concepts are fed to data mining or pattern recognition analyses, this ontological granularity can result in problems of signal dilution (Popescu and Xu 2009). To enrich the sparse datasets and thus enable meaningful analysis, concepts that are semantically similar can be aggregated. The evaluation of whether two concepts are semantically similar enough for aggregation is often highly dependent on the context of the study itself (Dong, Hussain, and Chang 2010). For example, concepts such as "obese" and "morbidly obese" can be merged when studying Huntington's Disease, but should remain separate when investigating predictors for heart attack.

In this chapter, we examine the problem of concept aggregation in the context of a clinical data-mining task. We assess the value of corpus-driven and knowledge-driven methodologies to compute a similarity score for concept pairs. To evaluate similarity within a specific situation we rely heavily on context-specific data. Initial similarity calculations are compiled on a homogenous set of clinical notes, emphasizing the contextually dependent and corpus-driven methodology as a first step. The further refinement of the corpus-based measure is created on two types of ontological knowledge (path length and definitional word overlap), both aiming to differentiate related from semantically similar concept pairs. We evaluate the different methods, including a hybrid score that averages the three measures, on a large dataset of concepts. This work fits primarily within the field of clinical informatics with the goal of defining a comprehensive way to enrich the analysis of unstructured data located in EHRs.

3.2 Related Work on Similarity Detection

It has been shown that people generally agree upon the notion of similarity or relatedness between ideas (Pakhomov et al. 2010; Tversky 1977). As a result, there has been a large effort across various disciplines, including natural language processing (Rada et al. 1989; Androutsopoulos 2010), and biomedical informatics (Benabderrahmane et al. 2010; Verspoor, Dvorkin, and Cohen 2009; Elhadad and Sutaria 2007; Pedersen et al. 2007), to create automated methods that can find semantically similar concepts. Much of the research focuses on the identification of both similar and related concepts. Relatedness indicates a semantic association between concepts, such as "ear" and "nose," while similarity

specifies that two concepts can be used interchangeably (Budanitsky and Hirst 2005). The focus of this work is on similarity. Therefore, although many interesting methods have been published on relatedness identification, they are outside the scope of this chapter.

3.2.1 Methods for Semantic Similarity Calculation

Methods developed to identify semantic similarity among concepts fall loosely into two categories – knowledge-based (edge-based and syntactic) and corpora-based (distributional semantics), where information-content-based measures can span both. In this section, we review previous work with specific emphasis on the methods we later use for comparison (and are included in the publicly available UMLS-Similarity package) (McInnes, Pedersen, and Pakhomov 2009).

Edge-Based

Many methods have been developed for a hierarchical interpretation of similarity, based on the location of the concepts in an ontology and the paths among them. Some of the most common methods rely on edge counting, shortest path, and ontological depth (Rada et al. 1989; Caviedes and Cimino 2004; Leacock and Chodorow 1998), while others add the least common subsumer (LCS) to capture the granularity of a concept in the ontology (Wu and Palmer 1994; Al-Mubaid and Nguyen 2006). More recent advances have incorporated into similarity computation the distance to the LCS, assigning weights to the different path types (ontological depth, distance from concepts to LCS) (Matar and Egyed-Zsigmond 2008), as well as all of the superconcepts between two terms as a way to account for multiple inheritances (Batet, Sánchez, and Valls 2010). We list a few of them below.

Conceptual Distance (CDist) (Caviedes and Cimino 2004)

$$sim_{cdist}(C1,C2) = |\ shortest\_path(C1,C2)\ | \quad (1)$$

Leacock and Chodorow (lch) (Leacock and Chodorow 1998)

$$sim_{lch}(C1,C2) = -\log(|\ shortest\_path(C1,C2)\ |/(2*depth(ontology))) \quad (2)$$

Wu and Palmer (wup) (Wu and Palmer 1994)

$$sim_{wup}(C1,C2) = 2*depth(LCS)/(depth(C1)+depth(C2)) \quad (3)$$

Al-Mubaid and Nguyen (nam) (Al-Mubaid and Nguyen 2006)

$$sim_{nam}(C1,C2)=\log((|shortest\_path(C1,C2)|-1)*(depth(ontology)-depth(LCS))+2) \quad (4)$$

Information-Content (IC) Based

IC-based methods aim to create measures that incorporate the specificity of a concept within a similarity calculation. The IC calculation is based on the concept and all of its descendants' frequencies within a corpus of texts. The original measure proposed by Resnik evaluated the information shared by two concepts by measuring the IC of their LCS (Resnik 1995). As the Resnik measure can assign perfect similarity to any two concepts that share the same LCS, two other measures were proposed by Lin (Lin 1998) and Jiang and Conrath (Jiang and Conrath 1997). They also take into account the IC of the concepts themselves, Lin using ratios and Jiang and Conrath using subtraction. More recently, Pirro and Seco devised a similarity measure founded on the idea of "intrinsic IC" which quantifies IC values by relying on the structure of an ontology itself as opposed to a separate corpus (Pirro and Seco 2008).

Distributional Semantics

Distributional semantics follow the assumption that the meaning of a target word or concept can be acquired from the distribution of words surrounding it, as a whole over its many mentions in a collection of texts. Thus, similarity between two concepts can be quantified according to the amount of overlap between their overall contexts. Here, by context, we are referring to a weighted count of all the words in the sentences surrounding all the instances of a concept. Distributional semantics have been applied to several problems in biomedical informatics (Cohen and Widdows 2009). The distributional semantics methodology represents an abstraction of patterns over a larger corpus, where individual mentions of terms are agglomerated to derive an overall pattern of usage. As the abstraction occurs over many mentions and the words in the vocabulary are weighted (typically tf-idf weights), individual negations and other modifiers all contribute to the salient textual patterns present in the corpus. As distributional semantics allow us to compare two concepts in their usage and thus assess their semantic similarity,

conversely, such a representation can help perform word sense disambiguation as different senses of a word will appear with different words and phrases surrounding them.

The work of Pedersen et al. forms the basis of our context-based similarity measure (Pedersen et al. 2007). Pedersen et al. calculate similarity based on patterns of usage in text with the help of a context vector (which in their case, relied on the Mayo Corpus of Clinical Notes). Each concept of the corpus is represented as a sum of all word vectors that map to the concept, each of dimension the size of the vocabulary. The vector representing word *w* at index *t* is the number of times *w* and *t* co-occur in the same line of a note in the corpus. The similarity between two concepts is then computed as the cosine similarity between their corresponding context vectors. Pedersen found that "the ontology-independent Context Vector measure is at least as effective as other ontology-dependent measures" (Pedersen et al. 2007). Our note-based similarity approach differs mainly in the type of corpus we employ to derive the context vectors. Furthermore, we investigate to which extent this method and ontologically based methods, previously used independently of each other, can be used in complement.

Definitional

The idea of relying on the content of word definitions for assessing appropriate word senses was original proposed by Lesk (Lesk 1986). The Lesk algorithm selects the sense of a word in a text, which has the highest word overlap between its definition and its context in the text. Banerjee and Pedersen (Banerjee and Pederson 2002) adapted this method further using WordNet and essentially reversed the methodology for the assessment of semantic relatedness (they also added WordNet hyponyms into the computation). Given the Lesk measure, which identifies overlaps in the extended definitions of the two concepts, the relatedness score is defined as the sum of the squares of the consecutive word overlap lengths. A similar methodology was employed by Hamon and Grabar in the biomedical domain (Hamon and Grabar 2008).

Other Methods

Other published measures include similarity calculations between sets of concepts (Cordí et al. 2005), weights of different features in Gene Ontology (GO) (Benabderrahmane et al. 2010), and a

nonlinear model that is a function of various ontological features such as path length, depth, and local density (Li, Bandar, and McLean 2003). Alterovitz et al. implemented an information-theoretic technique to gauge similarity of GO terms and identify ontological inefficiencies (Alterovitz et al. 2010). Additionally, Rodriguez and Egenhofer focused on hybrid methods that compute both over term definitions and various hierarchical attributes such as features and neighborhoods. Petrakis et al. (Petrakis et al. 2006) refined the methodology further to compute neighborhood similarity.

3.2.2 Context-Aware Computing

The notion that the context surrounding information is important is not a novel one and many have thought about applying it in the medical context. Specifically, applications have been developed to facilitate context-aware data mining that would help provide background when evaluating the similarity of the data mining results (Singh and Vajirkar 2003).

Others have devised methods to convert traditional similarity measures into contextually dependent ones. Wu et al. propose a method in which given a similarity measure and a training set they are able to create a new distance function using the "kernel trick" (Wu, Chang, and Panda 2005). Dong et al. describe a method of ontological conversion designed to take context into account (Dong, Hussain, and Chang 2010). Other work has looked at the various types of context and how they affect similarity judgments specifically in the case of geospatial IR (Janowicz 2008; Kessler, Raubal, and Janowicz 2007). Most of these context measures are created to enhance personalization across information retrieval systems where the context consists of user, system, and background information. Our method is designed to incorporate an aggregate disease context over many patient records to create disease-specific similarity calculations.

3.3 Method for identifying similar concepts

Our composite methodology consists of three complementary similarity measures (Figure 3.1). One primary measure is context-based and relies on distributional semantics of patient notes authored by clinicians, while the other two are knowledge-based and rely on concept definitions and their

relationships in the SNOMED-CT ontology. Starting from a homogenous corpus of notes (i.e., notes about patients who share at least one clinical problem), notes are pre-processed to extract concepts mentioned in the corpus. A three-way filter is applied to prune out the extracted concepts and keep a homogeneous set of concepts to be aggregated. The context-based similarity ranks all pairs of concepts. The top-k pairs with the highest context-based similarity are then reordered using the two knowledge-based similarity measures. This section describes the dataset and its pre-processing, the filtering of concepts, the three measures, and the experimental setup for our experiments.

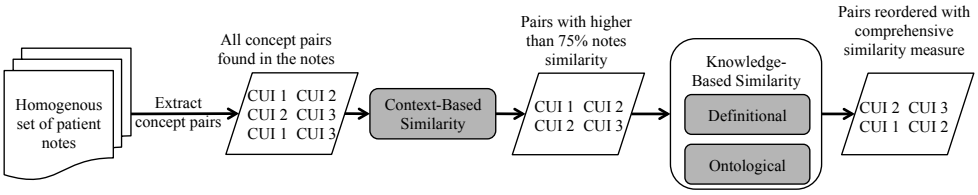


Figure 3.1 Our methodology for finding context-dependent similar concepts. An overview of the entire pipeline, beginning with a set of patient notes and ending with a ranked list of similar CUI pairs.

3.3.1 Data and Knowledge Sources

Distributional similarity techniques assume that the meaning of a word or concept can be represented by the context in which it appears, across a large number of mentions. As such, the more frequent a concept is in a corpus, the more accurate its context will be at representing the meaning of the concept. Corpus selection is important – a random sample of notes from a random sample of patients might provide a large set of concepts pairs for which to assess similarity, but the concepts might be too sparse and the resulting contexts might be misleading. Our corpus selection process follows.

We chose to collect a homogenous and semantically coherent corpus of clinical notes, in order to ensure that concepts, which are clinically relevant to the patients, are likely to appear frequently enough. For this study, we collected a corpus of notes from patients with chronic kidney disease (CKD). The methods we employ are disease-independent, but the fact that we select notes from patients all with at least one condition in common allows us to identify and aggregate concepts frequently mentioned when documenting a particular set of patients. Furthermore, CKD is a prevalent condition in our institution,

thus allowing us to collect a large corpus of notes. Patients with CKD have many comorbidities and disorders, providing us with many different concepts to consider in our similarity computation. Using ICD-9 codes as evidence of CKD, all notes for CKD patients recorded between 1990 and September 2010 were extracted from the NYPH dataset. Each note was processed by our in-house NLP pipeline (Lipsky-Gorman and Elhadad 2011), which identifies document structure (section boundaries and headers, list items, paragraph boundaries and sentence boundaries) (Li, Lipsky-Gorman, and Elhadad 2010), performs shallow syntactic analysis (part-of-speech tagging and phrase chunking), and named-entity recognition of UMLS concepts through dictionary matches. The named-entity recognition in our corpus was performed using the 2010AA UMLS version and restricted to the SNOMED-CT terminology. The full pipeline was tested on a manually curated gold standard of 31 notes and yielded an F-measure of 88.55. The pipeline processed a patient note in 0.26 seconds on average.

The knowledge-based part of our similarity computation relies on the SNOMED-CT. SNOMED-CT is a terminology of clinical terms and is a primary resource for concept standardization in the clinical domain. SNOMED-CT is particularly useful for our purposes because it provides term definitions and synonyms, as well as semantic relations among concepts. The relationship types have very specifically defined attributes and lend themselves well to our ontological similarity measure. We utilize the concept definitions and synonyms encoded in SNOMED-CT for the definitional similarity. The version of SNOMED-CT we use in this study is from the July 31st 2010 release and consists of over 292,000 active concepts, 760,000 concept descriptions and 824,000 inter-concept relationships.

3.3.2 Filtration

Given the pre-processed corpus of CKD notes, we can extract a list of all mentioned concepts. In an effort to create a concise and unambiguous list of similar concept pairs, however, we perform a three-tiered filtration step. The filtration relies on the concepts (semantic types), the structure of the notes (section types), and the note category (note types).

The concept filtration follows the hypothesis that two concepts are more likely to be similar if they belong to the same semantic group, whereas two concepts from different semantic groups can indicate semantic relatedness only. For example, it is unlikely that an anatomical concept is similar to a disorder, whereas it is possible that the disorder and the anatomical concept are semantically related (a stroke occurs in the brain, for instance). In practice, limiting the set of semantic types constrains the set of potential relationship types among CUIs, by reducing potential meronyms and focusing instead on hyponyms and metonyms (Elhadad and Sutaria 2007). In this study, we filter in all concept mentions that belong to the Disorder (DISO) semantic group, as defined by McCray et al. (McCray, Burgun, and Bodenreider 2001) The DISO group contains 12[3] out of the total 133 semantic types. We chose this group as it represents the types of concepts we are interested in examining (e.g., diseases, findings, and signs and symptoms), but the method is agnostic to the chosen semantic group(s).

The section filtration's aim is two-fold: to ensure the pool of input concepts is homogenous and to mitigate the potential CUI mapping errors that occurred during pre-processing. To keep the list of input concepts homogenous and specific to the patients under study, we filter out concepts mentioned in the Family History section of the notes. The concepts mentioned in the Medication section are filtered out in an attempt to reduce pre-processing mistakes, arising from the ubiquitous medication abbreviations that are prone to erroneous UMLS mapping. Because our pre-processing pipeline does not perform word-sense disambiguation and keeps all possible CUI mappings instead, the input concept list to our similarity computation can contain incorrect concepts. For instance, "mg," a common abbreviation in the Medication section of notes, which stands for "milligram", is mapped erroneously to "Madagascar," and "Magnesium."

[3] Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom.

The note filtration operates at a higher level and selects the notes that contain the "richest" content for our purposes. Our initial corpus of CKD notes has more than 1,700 different note types (e.g. Primary Provider note, Cardiology Consult note, and Miscellaneous Nursing note). We identified the note types that had more than 60,000 occurrences of SNOMED concepts in the DISO semantic group over the entire set of patients. The filter helps producing a homogenous, focused list of input concepts, on which to compute pairwise similarities.

3.3.3 Note-Based Similarity

The note-based similarity takes two UMLS CUIs as input and returns a similarity score defined as the cosine similarity between the two concepts' context vectors (Manning, Raghavan, and Schütze 2008). The context vectors are derived from the filtered, pre-processed notes for each CUI. They have V elements, where V is the size of the vocabulary. In our experiments, the vocabulary consists of all stemmed words present in the filtered notes.

Given a CUI c and a stemmed word w, the value of the context vector for c at index w is defined as the tf-idf value of w, when c and w occur in the same sentence over the input filtered notes. The tf-idf value is based upon the number of times c and w appear both individually and together. Note that our metric differs from previous work slightly, as we operate over sentences, as opposed to lines in the note (Pedersen et al. 2007).

The note-based similarity is computed for all concept pairs. However, because highly infrequent CUIs have very sparse context vectors, which do not represent their context accurately, we only considered CUIs that occur more than five times in the input corpus. Calculation of note-based similarity for all CUI pairs on our corpus took 5.28 hours on a linux machine with CentOS 5.4 16-core, 2.93GHz Xeon X5570 model, 24GB RAM, with a Hitachi 10,000 RPM drive.

Following our hypothesis that contextual similarity is the basis for semantic similarity, concept pairs with high note-based similarity are kept as candidates for similarity, while all the other pairs are discarded. In our experiments, we set the threshold for note-based similarity at 75% which resulted in 794

CUI pairs. For instance, in the context of CKD patients, the concepts pairs (Muscle Injury, Traumatic injury of skeletal muscle) and (Acne, Common Acne) both have a cosine similarity above 75%, and are considered for further processing. The concept pairs with lower similarity scores are filtered out, such as (Localized mass, Nodule) and (Chronic Low Back Pain, Pain, NOS) which each have similarity scores in the low 50%.

3.3.4 Ontological Similarity

We describe a novel method for semantic similarity using ontologically defined relationships. We look at the SNOMED-CT ontology as a flat terminology and concentrate on edge types rather than the hierarchy itself. With this view of SNOMED-CT we are able to look at all of the layers that consist of deleted, moved, and retired concepts. This was highlighted as important in the Dong et. al paper that stated: "…in an ontology environment, the types of relations are various, and relations can be defined by multiple restrictions. Obviously, the two factors cannot be ignored when computing similarity for ontology concepts." (Dong, Hussain, and Chang 2010) The method is based upon the types of relationships between concepts where the different types are broken down into three tiers. The tiers are defined by the characteristic and refinability features of SNOMED-CT relationships (Table 3.1) and used to group the relationships into ones resembling the most to least closely related. The tiers were used to define weights for various relationships types and the weights were chosen to reflect the strength of each relationship type. There seems to exist a natural hierarchy of relationship strengths that we chose to exploit in the method, such as the observation that non-refinable relationships are of a stronger nature than ones that are optionally or mandatorily refinable. In addition, the weights chosen reflect a system to ensure a score between 0-1 for each relationship type and to delineate between tiers, a twice-larger difference between tiers was introduced (.2) in comparison to the weight difference within each tier (.1). To reduce complications every term listed as both qualifier and defining (Associated Finding, Access, Priority, Clinical Course, Laterality, Associated Procedure, Using Device, Surgical Approach) was grouped in the Defining, Optionally Refinable tier and always given a .5 weight. The weights described

here can be tweaked to assess similarity alternatively or even a different semantic relationship; our main contribution lies in proposing a novel way to consider ontological path length calculations.

Characteristic	Refinability	Example	Tier	Weight
Defining	Not Refinable	Is A	1	1
Historical	Not Refinable	Replaced By	1	0.9
Additional	Not Refinable	Part Of	2	0.7
Defining	Optionally Refinable	Associated With	3	0.5
Qualifier	Optionally Refinable	Measurement Method	3	0.4
Qualifier	Mandatory Refinable	Associated Finding	3	0.3

Table 3.1 Relationship weights for our algorithm. Relationship tears are defined by concept Refinability and Characteristic, which are both assigned by SNOMED-CT.

To find the path between two UMLS CUIs, each CUI was mapped to all of its SNOMED-CT concepts and all possible combinations of pairwise shortest paths were calculated. The average of these paths was taken and assigned as the official path length between the CUIs. The ontological similarity was calculated in 2.18 seconds for all 794 CUI pairs (15,187 SNOMED concept pairs) on a Linux machine with Ubuntu 10.04.03 12-core, 2.4GHz Opteron 4176 model, 32GB RAM, with a Dell 7,200 RPM drive. The following algorithm (Eq 5) was used to assign ontological weights for each individual pairwise path:

$$Sim_0 = \sum_{e=1}^{E} weight_e \ / |E| - \alpha(|E|-1) \quad (5)$$

$E = \{e_1, e_2, \ldots e_n\}$ where e_i = edge in path

$weight_e$ = assigned weight for edge e

$\alpha = .2$

For example, one path between C0002622 (Amnesia) and C0751295 (Memory Loss) is between the SNOMED-CT concepts 247606008 and 162199006, with a path similarity of 0.5333 as illustrated in Figure 2 and Eq 6.

$$Sim_O(247606008,162199006) = (.9+1+.9)/3 - .2(2) = .5333 \quad (6)$$

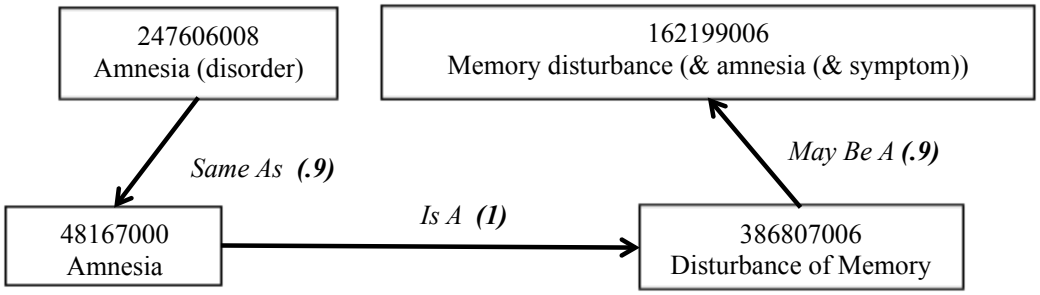


Figure 3.2 An example of the relationship-weighted path calculation. This is the shortest SNOMED-CT path between 247606008 and 162199006. An example of a path between two UMLS concepts that were mapped back to SNOMED-CT, the "may be a" edge is not found by hierarchical path methods.

3.3.5 Definitional Similarity

The third similarity metric we used was definitional similarity. Definitional similarity is a measure of lexical commonality between two concepts – a metric widely used in word sense disambiguation, which can be seen as a reverse goal of our task (Banerjee 2002). We focused on lexical inclusion as the metric and we used the following formula (Eq 7):

$$Sim_D = |(C_1+C_2)| - |\{C_1+C_2\}| / Min(|C_1|,|C_2|) \quad (7)$$

Where C_i={words in definition and synonyms of all mappings of CUI "C_i" in SNOMED-CT}

We chose to use this metric as a way to capture complete subsets as being perfectly similar while adequately capturing the amount of discordance between the two sets of words. While the Lesk methods look at consecutive words, we treat the definitions as a bag of words. For example, the similarity between C0240419: Tender Muscles and C0575064: Muscle Soreness would be between SNOMED-CT concepts 22166009 and 278018006 with a definitional similarity of 1.

22166009 Skeletal muscle tender (finding), Skeletal muscle tender, Muscle tenderness, Muscle soreness, Tender muscles

278018006 Tender muscles (finding), Tender muscles, Tender muscles

C_1= [skeletal, muscle, tender, tenderness, soreness, muscles]

C_2= [tender, muscles]

$Sim_D(22166009,278018006)=((6+2)-6)/2=1$ (8)

The definitional similarity calculation for all 794 pairs took 0.31 seconds on the same Ubuntu machine used for ontological calculations.

3.3.6 Experimental Setup

Our method finds similar pairs from a very large number of pairs (in our experiments, approximately 14 million); therefore, it would be impossible to create an annotated gold standard list for full evaluation of true negatives and positives. In addition, because the set of input pairs is extracted from a corpus of notes, any gold standard is bound to be corpus-dependent. Therefore, for an evaluation method, we instead assess the accuracy of our methods, its variants and a baseline on a subset of all 14 million pairs, namely the ones with high note-based similarity (i.e., above 75%). In our experiments, there were 794 such pairs.

The evaluation of all three methods was calculated on the 794 pairs (those already filtered by the note-based similarity). The definitional and ontological similarity measures were used and evaluated as secondary metrics. The first evaluation was performed on the note-based method alone to assess its individual contribution. Next, the average of the note-based and ontological methods as well as the average of the note-based and definitional methods were calculated to see the added benefit of each. Finally, the average of all three methods was computed.

To further assess whether a threshold on note-based similarity at the 75% level is appropriate, we calculated similarities and collected gold-standard annotations for 100 random CUI pairs from the 25%-50% note-based similarity bracket and 100 random pairs from the 50-75% bracket as well.

Annotations

Two physician annotators evaluated the results of the similarity calculations. The annotators were presented with the 794 CUI pairs in random order along with all of their SNOMED-CT definitions and synonyms. The annotators were specifically asked about the similarity of the concepts within the context

of a general population of CKD patients. The annotators were asked to answer the following question with yes, no, or maybe: "Considering a patient with CKD, from a clinical standpoint, would you say that these two concepts could be used interchangeably?" The annotators were not shown any actual medical notes but only a pair of terms. Such an evaluation was chosen as the purpose of our method was to summarize the term usages from the corpus as a whole and use the shared kidney disease framework to find similarity specific to the kidney context overall. The inter-annotator agreement between the physicians was calculated using Cohen's Kappa (Cohen 1960) and after converting all Maybe's to Yes, this resulted in a kappa of .68 without any further adjudication. A kappa of .68 is accepted as representing a substantial amount of agreement between annotators (Landis and Koch 1977). A conversion of Maybe's to No's resulted in a slightly lower Kappa of .67. The final conversion from Maybe to Yes was appropriate in this instance not only because of the higher Kappa, but also as the "Maybe" was used to annotate terms that are similar in some cases and would require more specific knowledge on the particular patient to determine definitive similarity. A few examples of terms that were marked as "Maybe" are (Swollen Foot, Swollen Ankle), (Acute and Chronic Inflammation, Focal Chronic Inflammation), (Radiation Burn, Effects of Radiation). The accepting of a "Maybe" annotation as "Yes" gives us more opportunity to create a comprehensive list of similar concepts for large-scale concept aggregation. Given the two annotators' answers for a particular pair, a consensus gold-standard answer was defined as "Yes" if both annotators answered yes, and "No" otherwise. Under this setup, the original list of 794 concept pairs contains 145 pairs annotated as similar.

Baseline

UMLS-Similarity (McInnes, Pedersen, and Pakhomov 2009) is a Perl package which encodes ten different similarity methods based on ontologies and corpora. UMLS-Similarity encodes path-based measures, information-content measures, definitional measures, and note-based measures. We ran five of the measures on the UMLS-2010AA as baseline, consisting of the path-based and definitional measures, as described in Table 3.2.

Category	Measure	Used for Baseline?	Why not?	Parameters
Path	path	No	Same as the inverse of cdist	-
	cdist	Yes	-	SNOMED-CT vocabulary Parent/Child and Broader/Narrower
	lch	Yes	-	
	nam	Yes	-	
	wup	Yes	-	
Information content	jcn	No	IC was calculated on different corpus	-
	lin	No		-
	res	No		-
Definitional	lesk	Yes	-	Definitions from SNOMED-CT vocabulary All relationship types
Note-based	vector	No	Very similar to our vector-based method	-

Table 3.2 All of the UMLS-similarity measures and their inclusion or exclusion in our baseline.

Evaluation Metrics

As we are interested in evaluating each similarity measure independently as well as their combined effect, we created receiver operating characteristic (ROC) curves for each combination of similarities at every similarity threshold (Hanley and McNeil 1983) for both our method and the baseline comparisons. We used the areas under the curves as the single measure to compare how well each similarity measure did.

3.4 Results

The total dataset collected for this experiment consisted of 2,569 patients and their notes, which covered a span of over 20 years (1990-2010). In total, there were 403,819 notes from which we extracted 8,869 unique UMLS concepts that are in SNOMED-CT. The minimum unit of computation used in this study was a sentence containing a CUI and the corpus can be described as the set of these sentences. Figure 3.3 shows a histogram of the notes with respect to the number of sentences they contain.

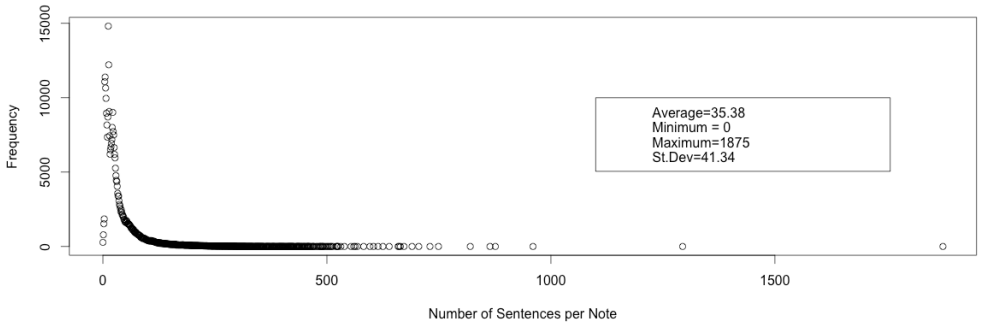


Figure 3.3 Descriptive Statistics for sentences in our CKD corpus. This graph illustrates the variety of sentence lengths found and chunked by the ClinNote pipeline.

The note filtration was used to narrow down the total notes used for the experiment, while keeping rich content. Initially, the corpus consisted of 403,819 total notes with 1,739 unique note types. Selecting the note types with more than 60,000 DISO concepts occurrences overall resulted in keeping 17 concept-rich note types (Table 3.3). This filtration led to a total of 170,775 notes used for the analysis (that is, less than 1% of the note types captures over 40% of the notes). As many institutions have a similar problem of unrestricted note type changes within their medical departments, the rapidly growing and changing EHR note structures make it difficult to keep track of which notes are most important. As the importance of note types is dependent on the question being asked, it is possible to vary the semantic groups included in the analysis, thereby altering the concepts found and the note types considered to be most important. This simple way of ranking note importance by concept density is a dynamic and institution-independent way to identify the most contextually specific salient notes in their EHRs.

Note Type	Number of CUIs that map to SNOMED-CT from the DISO group
Clinical_Note	743268
Discharge_Summary_Note	432138
Admission_Note	341975
Signout	305453
Physical_Therapy	258589
Progress_Note	253971
Nursing_Adult_Admission_History	239946
Surgical_Pathology_Event	223064
Follow-up	204593
Consult_Note	154823
12-Lead_Electrocardiogram	135571
Transthoracic_Echocardiography	112197
Ambulatory_Internal_Medicine_Structured_Note	110324
AMB_Internal_Medicine_Follow-Up_Note	107129
Operative_Note	93521
Primary_Provider_Clinic_Note	91803
X-Ray_of_Chest,_Portable	63968

Table 3.3 Note types selected through the note filter

3.4.1 Concept Similarity Results

To determine the best way to create the context vectors, we performed all experiments with stemming and without stemming the words in the corpus. The stemming approach showed a minor improvement over the unstemmed version and therefore we chose to report the stemmed similarity results.

Experiments for Concept Pairs with High Note-based similarity

We report in this subsection results on a full list of pairs with a note-based similarity above 75% similarity, corresponding to 794 pairs.

The lexicographical comparison of definitions and synonyms of the concepts created a second layer of similarity which we used in addition to the ontological method to move pairs which are simply related to further down on the ranked list than those which are semantically similar. Given the algorithm used for definitional similarity, we often found high similarity between parent-child concepts or concepts with a short definition or list of synonyms.

Figure 3.4 shows the ROC curves on the 794 pairs annotated for similarity with different combinations of the similarity measures (combinations represent an average of the individual measures). Although not reported in this chapter, we also experimented with the effect of simply applying the ontological and definitional scores to rank and re-order the 794 pairs (without averaging in the note-based score itself). These resulted in slightly smaller AUCs than their note-based averaged counterparts.

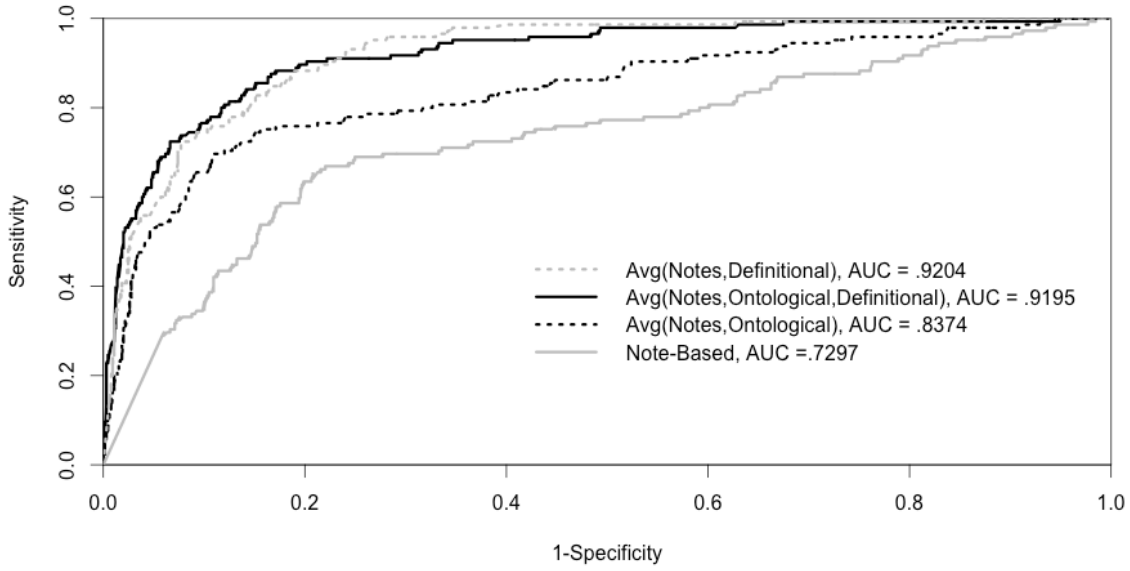


Figure 3.4 ROC curves comparing different parts of our methodology. We compared the curves of all similarity combinations between note-based, ontological-based, and definition-based measures.

Table 3.4 shows the top-10 concept pairs ranked by the composite method, which averages the three similarity scores for each pair.

UMLS Concept Preferred Term	UMLS Concept Preferred Term	Similarity Score	Similar according to Expert Consensus
C1998242 Traumatic injury of skeletal muscle	C0410256 Muscle Injury	1	Y
C1691215 Penile hypospadias	C0848558 Hypospadias	0.972	Y
C0240419 Muscle tenderness	C0575064 Skeletal muscle tender	0.966	Y
C2678517 Thrill (finding)	C0232269 Cardiac thrill (finding)	0.958	N
C0677659 Gastro-esophageal reflux disease with esophagitis	C0014869 Peptic Esophagitis	0.95	Y
C0149889 Anorectal fistula	C0205929 Anal fistula	0.945	Y
C0158458 Acquired hallux valgus	C0018536 Hallux Valgus	0.937	Y
C0520474 Aseptic Necrosis of Bone	C0029445 Bone necrosis	0.935	Y
C1261287 Stenosis	C0009814 Acquired stenosis	0.935	Y
C0243095 Finding	C0037088 Signs and Symptoms	0.933	Y

Table 3.4 Top-10 concept pairs found by the composite (average of the note-based, ontological-based, and definitional-based) method.

It is difficult to assess the coverage of our approach to identify similar concept pairs over the original set of 14 million pairs. Instead, we investigate to which extent the automatically discovered candidate pairs are relevant for the input corpus. Since the goal of this method is to aggregate concept

pairs that are semantically similar, it is important to know whether the discovered pairs are frequent enough in the input corpus. We assessed the coverage of the concepts that made it into our annotated list of 794 pairs. The concepts made up for 6% of the total number of concepts in the corpus. They are common concepts however, as they cumulatively make up for 30% of the concept frequencies in the corpus. This confirms that our approach is appropriate for discovering pairs of similar concepts, which are frequent in the corpus, and thus important to discover.

We compared our comprehensive method with 5 methods (lch, wup, cdist, nam, lesk) packaged in the UMLS-Similarity perl program. As all of the methods we used except Lesk rely upon the hierarchical relationships present in the UMLS only (Parent/Child or Broader/Narrower), they frequently missed paths between concepts. This happens because many concepts are linked with non-hierarchical relationships such as "moved to" or "deleted from". Table 3.5 shows the number of missing paths (from the 794 total pairs, 145 of which are similar) when using PAR/CHD, RB/RN, or both. RB/RN misses the largest amount of paths and while PAR/CHD does quite a bit better, even combining the two, leaves over 10% of the total and 20% of the truly similar concept paths as null. We present the ROC curves for the combined PAR/CHD and RB/RN hierarchical methods as well the Lesk method, although the Lesk and CDist scores are coarse whole number measures and lead to fairly few data points for the ROC curves (Figure 3.5). The baselines all underperform compared to our three similarity metrics and their combinations.

We compared our final list of similarities with the list of 566 concept pairs collected by Pakhomov et al. (S. Pakhomov et al. 2010) in their semantic similarity study. There were only four pairs in common (Rhonchi, Rales), (Polydipsia, Polyuria), (Vertigo, Dizziness), and (Constipation, Diarrhea). Only Rhonchi/Rales were found as similar by the annotators.

	Broader/Narrower (RB/RN) Relationships	Parent/Child (PAR/CHD) Relationships	PAR/CHD or RB/RN
Pairs with no path	77.8%	15.6%	11.3%
Pairs with no path that are similar	75.9%	24.13%	20.0%

Table 3.5 Missing paths in hierarchical methods. For the "PAR/CHD or RB/RN" method, we used a PAR/CHD path if it existed and RB/RN path otherwise.

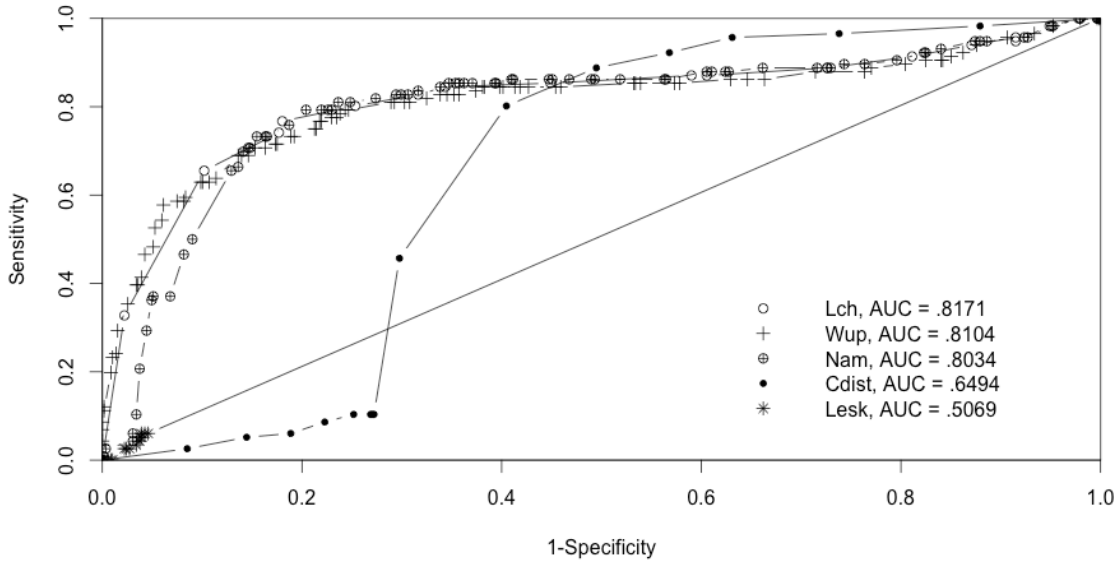


Figure 3.5 ROC curves of the baseline methods. We used five methods encoded in UMLS-Similarity and calculated the ROC curves including only the pairs between which paths were found, leaving only 116 similar pairs and 704 pairs in total.

Experiments for Concept Pairs with Low Note-Based Similarity

In the experiments described above, we focus on concept pairs with a high note-based similarity (above 75%). This assumes that the context from notes is the most salient cue towards semantic similarity

compared to definitional and ontological similarities. As a sanity check that note-based similarity does not miss pairs that are semantically similar, we expanded our gold standard with 200 more concept pairs and collected similarity assessment from our judges, following the same methodology as in the above data set: 100 random pairs from the set of pairs with a note-based similarity between 50% and 75% similarity and 100 random pairs from the set of pairs with a note-based similarity between 25% and 50% similarity. We calculated the path-based similarity and the ontological similarity for the 200 concept pairs. Out of the 200 pairs, only three were scored as interchangeable, and thus similar: (Respiratory alkalosis, Alkalosis), (Disturbance in sleep behavior, Sleep disorders) and (Liver palpable, Liver edge). Furthermore, none of the three were unanimously assigned a "Yes" by our experts. While, this is only for a random sample, it provides face validity to the claim that note-based similarity is the primary factor to assess context-based semantic similarity.

3.5 Discussion

3.5.1 Impact of Context

The experiments confirm that context plays a crucial role in assessing similarity of medical concepts: the writing patterns of clinicians provide valuable information to determine which concepts are mentioned in similar contexts and thus are good candidate pairs for aggregation. However, these patterns are all the more visible because the information used for the note-based similarity is derived from a large corpus, with a coherent set of concepts, all related to a particular topic (chronic kidney disease in our experiments). For example, consider the two concepts "Difficulty Hearing" and "Complete Deafness." Generally, the two might be similar enough for aggregation but not given a history of kidney disease. One of our physician annotators pointed out that difficulty hearing might serve as a clue of an adverse drug event caused by an overly high dosage of medication. Complete deafness does not offer the same reaction, as total deafness is rarely an adverse drug event. Such examples illustrate the need for context-dependent similarity measures.

In our experiment, we found the inter-annotator agreement between the physicians was quite high given the subjective nature of the question. The fair amount of agreement indicates that physicians generally concur on medical concept similarity within a particular context. It is a testament to the fact studied generally by Tversky (Tversky 1977) and in the medical arena by Pakhomov et al. (Pakhomov et al. 2010; Pakhomov et al. 2011) that there is a universal concept of similarity that most people agree upon.

3.5.2 Impact of the Ontological-based Similarity

The relationship-based ontological measure we proposed was able to locate many more paths than other popular methods encoded in the UMLS-Similarity package. Because the baseline methods rely upon hierarchical relationships only, they are often unable to find the complicated trajectories among concepts and focus primarily on more straightforward pairings. This limits the types of paths that can be found between two concepts. In contrast, our ontological similarity incorporates all types of ontological relationships into its path computation, and due to the disease/disorder and SNOMED-CT only filtrations, ensures that a path will always be found between two concepts. Our method takes care to assign greater weight to more significant relationships (Is-A), but does incorporate edges between concepts that others do not, such as “may be a” or “moved to”.

3.5.3 Impact of the Definitional-based Similarity

The definitional similarity measure also provided important information about similarity and did surprisingly well, serving as the best individual similarity measure when averaged with the context-based primary. Nevertheless, it must be emphasized that this measure performs well as a discriminatory measure after the initial contextual similarity threshold, and would probably not perform well on its own to discover candidate similar pairs. In fact, when applied to the 14 million pairs, 85% of them share no words in their definitions, and thus have a 0.0 definitional-based similarity measure and conversely, 2,099 pairs have a perfect 1.0 definitional-based similarity measure. On its own, this measure does not have enough granularity to rank pairs. It performs well combined with more nuanced similarity metrics.

Leveraging the concept definitions for assessing similarity can be viewed as a word sense disambiguation method (the original proposition by Lesk was to use definitions in exactly this way (Lesk 1986)). Given the potential tagging ambiguity that may arise during the entity-recognition phase and would result in a perfect note-based similarity score, the definitional similarity helps to move apart those incorrectly perfectly similar concepts.

3.5.4 Impact of Combining Data-driven and Knowledge-driven Similarity Measures

The note-based similarity measure, which relies on patterns of clinicians writing their notes, and the knowledge-based similarity measures, which rely on ontological knowledge, provide complementary cues to the assessment of concept similarity. Concepts that appear far away from each other in the ontology, but are used comparably in the clinical notes are generally irrelevant in the CKD context and can be aggregated for the purpose of data mining. For instance, the concepts C0025874 (Metrorrhagia) and C0312414 (Menstrual spotting) have a low ontological similarity score of .126, but a note-based score of .904, and thus can be correctly identified as similar. Alternatively, concepts may appear very close in a medical ontology but be used vastly differently in context. Such use indicates general similarity but a notable difference in the context of CKD. For example, the concepts C1444079 (Focal chronic inflammation) and C0021376 (Chronic inflammation) had a full ontology-based similarity of 1, but the note-based score was 0.814. In general, we found that it was essential to incorporate ontological and definitional similarity to separate pairs misguided by abbreviations used in medical text. Although the note-based similarity helps anchor variants of the same concept it has no mechanism for word-sense disambiguation specifically when an abbreviation maps to multiple SNOMED concepts. Out of the 794 pairs that were evaluated, 80 of them had a note-similarity of 1, indicating that they were used in exactly the same way throughout the entire corpus. Most of this is due to a shared "trigger," as our named-entity recognition tool maps each word to all possible UMLS CUI matches. The letters RA could trigger both Rheumatoid Arthritis and Refractory Anemia, giving them a note-based similarity of 1, but clearly they

are semantically different. Of the 80 pairs with perfect note similarity, 42 of them were actually similar such as (Incomplete Spontaneous Abortion, Incomplete Abortion Unspecified) both triggered by "Incomplete Abortion" and (Tonsillar Carcinoma, Malignant neoplasm tonsil) both triggered by "Tonsil Cancer"; the rest were disambiguated thanks to the addition of ontological and definitional similarity.

Chapter 4: Exploiting Patterns of Missingness for Clinical Modeling of Laboratory Tests[4]

4.1 Introduction

Automating feature selection from EHR variables is a difficult task, as the EHR is an inherently biased data source: EHR data are collected with the primary goal of delivering and documenting patient care, not with the primary goal of creating a curated research dataset (Hripcsak and Albers 2013; Hersh et al. 2013). Identifying and then mitigating such biases will result in not only the development of more accurate methods for deriving computational models of disease but also in learning better prediction

[4] This chapter was originally published in the Journal of Biomedical Informatics. The full citation for this publication is: Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform. 2014;51:24-34.

models from EHR data. Currently, laboratory tests are one of the most widely used features in EHR disease-modeling research and are therefore the focus of this chapter.

In this work, we hypothesize that (i) the specific context of a laboratory test order can be derived from EHR-observed measurement patterns, and (ii) that this context can be leveraged for better disease modeling. While a laboratory test's numerical values can help distinguish healthy from sick patients, test values themselves cannot separate sick patients by their ailment when the test is associated with multiple diseases. For instance, while the numerical results may be comparable, the rate of measurement for a gestational diabetes screening glucose test and a chronic diabetes monitoring glucose test will differ greatly. We predict that how often a laboratory test is ordered within a particular time window can help correctly separate one disease state from another. We further hypothesize that laboratory test measurement patterns provide complementary and independent information from the numerical values indicated by the laboratory tests. We formally explore the relationship between both laboratory test measurement gaps and laboratory test values to determine whether the context in which a laboratory test is ordered alters the way its value should be interpreted, and is therefore a critical feature for disease modeling. While analyses of laboratory measurement patterns have been conducted (Lyon et al. 2009; Saxena et al. 1993; Weber and Kohane 2013; van Walraven 2003), the analyses and interpretations have focused on resource overutilization and informing clinical practice rather than on EHR-driven research.

Before describing our methods and findings, we provide background on laboratory testing from an informatics standpoint and report on previous work in the emerging research area of EHR bias identification and mitigation.

4.1.1 Capturing the Context of Laboratory Testing

At the point of patient care, different laboratory tests are ordered at different rates, often dictated by what physiologic process the test is measuring, and very often there exist multiple reasons for ordering a particular laboratory test. The three most common reasons for ordering a test are (i) diagnosing a condition, (ii) screening for a condition, or (iii) monitoring a pre-existing condition. Some laboratory tests

are ordered for one specific, clinical reason: for instance, the prostate-specific antigen (PSA) test is ordered exclusively to screen patients for prostate cancer. Others serve multiple clinical purposes: TSH, for instance, is used both to diagnose and monitor patients with disorders associated with the thyroid hormone. Finally, some tests, such as creatinine, are ordered both for clinical purposes like monitoring chronic disease progression and diagnosing acute conditions, and for healthcare process purposes like following guidelines as part of a routine panel for preventive testing (McPherson and Pincus 2011). When hospital protocol dictates measurement times, as is the case with routine preventative panels, creatinine's measurement patterns arguably reflect healthcare processes more than they reflect the health status of a patient. Thus, the context in which a laboratory test is ordered depends both on its clinical purpose and the surrounding healthcare processes.

Deriving the context of a laboratory measurement is a challenge, however. EHR data lack an explicit indication for why each laboratory test was ordered, and using other dimensions of EHR data for derivation of such information (such as ICD-9 codes and clinical notes) is equally problematic. ICD-9 codes are notoriously non-specific to patient disease state and are often not recorded for all patient ailments (Birman-Deych et al. 2005; Farzandipour, Sheikhtaheri, and Sadoughi 2010). Clinical notes rarely explicitly state the exact reason a test has been ordered.

The specific description of the context in which a laboratory test is measured, therefore, is not included in most computational models of disease. In fact, most often models include only a laboratory test's numerical value, a range of values (Chen, Dudley, and Butte 2010; Lasko, Denny, and Levy 2013), or the presence or absence of a laboratory test (Lin and Haug 2008) as features, but no contextual information about the situation surrounding the order. In this chapter, we investigate whether aggregating numerical values of laboratory tests taken in multiple separate contexts without explicitly separating the contexts can lead to the confounding of research conclusions.

In research with clinical data, there is an implicit assumption that a laboratory test's numerical value and the rate at which the test is ordered are highly correlated features. This assumption about value and measurement rate correlation likely stems from the existence of value-based guidelines and the

widespread expectation that laboratory test values which fall outside of normal ranges prompt intervention and retesting. Value-based guidelines for laboratory test ordering dictate measurement frequency based on a test's numerical value. For instance, the guideline for performing a diagnostic PSA test states that if a patient's PSA is slightly over 4.0 ng/mL in the initial measurement, the PSA test should be remeasured within 48 hours to confirm the need for a biopsy. Our work formally investigates the linear and nonlinear relationship between numerical value and measurement patterns in EHR-recorded data.

4.1.2 EHR Biases

EHR data are biased because they are gathered in an uncontrolled environment and are not carefully curated for research purposes. EHR data are noisy, sometimes erroneous, and often sparse (Lasko, Denny, and Levy 2013). At the same time EHR data contain sometimes conflicting (e.g., notes and coded data provide differing medication lists) and redundant information (e.g., clinicians often copy-and-paste from previous notes). From a temporal standpoint, the EHR contains data about elements that evolve at different time scales and often evolve over time, as treatment affects patient state. Because of these complexities, assessing the impact of EHR biases and correcting for their impact on data-driven methods is an emerging research topic.

Recent research has shown that naive EHR statistical analyses can lead to the reversals of cause and effect (Hripcsak, Albers, and Perotte 2011), induction of spurious signals (Albers and Hripcsak 2010), large errors when predicting optimal drug dosage (Sagreiya and Altman 2010), value cancellation of temporal signals when aggregating different cohorts (Albers, Hripcsak, and Schmidt 2012; Albers and Hripcsak 2012; Albers and Hripcsak 2011), and model distortion when not accounting for redundancy in the narrative part of the EHR (Cohen, Elhadad, and Elhadad 2013).

One particularly problematic bias inherent to the EHR is the prevalence of data points that are missing not at random (Rubin 1976) (e.g., patients are seen and measured more often when they are sick, and measured less often when they are healthy). Inferring missing information, such as values when the

patient is not seen, is a challenging research area. While there have been different approaches to mitigating this type of missingness (Schafer and Graham 2002; Lasko, Denny, and Levy 2013; Sammon et al. 2015), mostly researchers ignore missing values or interpolate them (Farhangfar, Kurgan, and Pedrycz, 2007.; Abdala and Saeed 2004; Hug 2006), with some recent work on classifying which variables should be interpolated and which should be ignored. Lin and Haug demonstrated that some missing values are themselves informative by creating Bayesian networks that explicitly model the absence of clinical variables; these models were able to predict medical problems better than those that ignored or interpolated missing values.(Lin and Haug 2008) Our work builds upon Lin and Haug's findings and focuses on leveraging the temporal missingness within laboratory measurement data. We see the patterns of laboratory test measurement as patterns of missing data. As a way to mitigate the EHR biases, we explore the use of different missing-not-at-random patterns to classify different patient health states and stratify heterogeneous populations into homogenous patient groups.

4.2 Materials and Methods

Our study is carried out in three consecutive tasks, as described below:

Task 1 We explored the correlation of laboratory values to the laboratory test's time to repeat, examining whether the value and time between consecutive measurements (measurement gap) encode separate information or overlap in information content.

Task 2 To understand the overall dynamics of laboratory tests recorded in the EHR, we categorized types of laboratory measurement patterns, identifying those more likely to cause biases in EHR-based research.

Task 3 We used lipase as a case study for how rates of measurement can be used to account for biases in laboratory test measurement data.

We narrowed our population to patients that have visited the NYPH Ambulatory Internal Medicine clinic at least 3 times. The full longitudinal records (i.e., all inpatient and outpatient data points) for these patients were gathered.

Three physicians reviewed and edited a list of frequently measured laboratory tests. They

constructed a set of 70 laboratory tests of interest to primary care and internal medicine. We extracted the time series for these tests between September 1990 and September 2010 for all of the patients in the population.

4.2.1 Task 1: Correlation between Measurement Gap and Numerical Value

We quantified the relationship between value and measurement gap, asking: in the patient population, is there added information in looking at how a patient was measured, not only at the measurement value? Given a particular laboratory test and all patients' time series for that test, we constructed a joint probability density function (PDF) using a kernel density estimate in Matlab. The PDF consisted of laboratory values and time between consecutive lab measurements (or gaps between measurements) in days.

To assess the degree of correlation between a laboratory test's numerical values and its measurement gaps, we experimented with (i) linear correlation (estimated at the 95% confidence interval) and an associated p-value and (ii) a non-linear measure of correlation, mutual information (MI) between laboratory test values and gaps between measurements. Mutual information attains a value of zero when the random variables underlying the distributions (values and measurement patterns) are completely independent. Mutual information attains a maximum when the two distributions are deterministic.

4.2.2 Task 2: Finding Laboratory Test Measurement Motifs

We explored the different types of laboratory test measurement dynamics that exist in the EHR data by creating measurement gap histograms for 70 different laboratory tests. We computed the following quantity: for each laboratory test taken on each patient, we calculated the days between two consecutive measurements of that laboratory test on that patient. A histogram was created for each test, mapping the day gap between consecutive measurements and number of such gaps, when aggregated across the entire patient dataset. For example, if a patient had a creatinine test taken on February 3rd and another creatinine test taken on February 5th, the count of creatinine tests with a measurement gap of 2 days would be

incremented by one.

We visualized the measurement gap histograms in different coordinate systems to explore the measurement dynamics across laboratory tests. We uncovered differences when examining the histograms in the logarithmic coordinate system. Using log-log coordinates, we visually looked for modes present in the histograms. If there is linearity in a measurement gap histogram when presented in log-log coordinates (i.e., a power-law) that implies scale-free measurement dynamics and that all time scales represent a single context or reason for ordering the laboratory test. If no approximately linear relationship between the frequency of measurement gaps exists, we visually looked for changes (e.g., peaks) that separate the different dynamics patterns; these different patterns may qualitatively imply different contexts of measurement based on either a change in health state or based on the healthcare documentation process. We catalogued the measurement gap histograms based on observed approximate linearity and the presence of peaks in the histograms, as determined by a manual review of the curves.

4.2.3 Task 3: Studying the Potential Effect of Measurement Motifs on Research

In this task, we focus on a specific laboratory test as a use case for studying the effect of measurement motifs on EHR-driven research.

For the use case, we chose to study lipase and acute pancreatitis. Acute pancreatitis is a well-understood condition and because its diagnosis is largely laboratory-based it is a good test case to validate our hypotheses. Both amylase and lipase tests have been used for acute pancreatitis diagnosis but they are not specific for this condition: both tests are also used for monitoring of chronic pancreatitis and diagnosing pancreatic cancer. We conducted all experiments on both laboratory tests; in this chapter, we focus on lipase as recent literature has shown it to have higher diagnostic sensitivity and specificity (Banks, Freeman, and Gastroenterology 2006). Our results for amylase were similar to those for lipase.

We asked the question: can the known association between an abnormal lipase value and acute pancreatitis be recovered from EHR data? To verify our hypothesis that laboratory measurement

dynamics can impact the accuracy of identifying patients with acute pancreatitis, we considered three views of the data, based on the dynamics of lipase measurements within each patient's record: (i) only visits with short lipase measurement gaps, (ii) only visits with long lipase measurement gaps, and (iii) all visits independent of the length between lipase measurements. In each of these settings, we assessed the association between acute pancreatitis and lipase and studied the properties of visits that belong in the setting using ICD-9 codes and clinical notes. We hypothesize that as acute pancreatitis is an acute disease, visits with short lipase measurement gaps will be more highly associated and relevant to acute pancreatitis.

Settings

We divided each patient record into individual visits (defining a full inpatient admission as one visit). Each record (represented as a set of visits) was divided into bins of visits with short lipase measurement gaps and visits with long lipase measurement gaps. Figure 4.1 shows a schematic diagram of an individual's longitudinal record: visits 1 and 5 belong to the short-gap bin because the lipase measurements were taken in rapid succession, the other visits belong in the long-gap bin because they show a long time between consecutive lipase measurements.

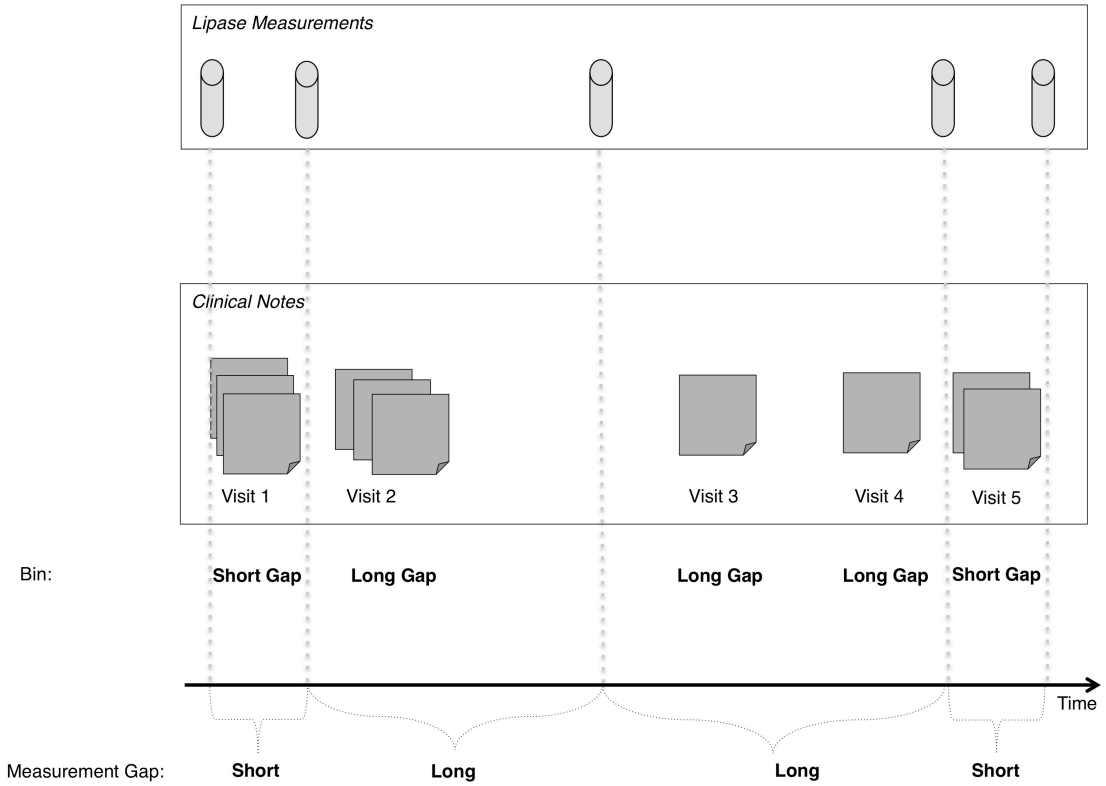


Figure 4.1 A schematic of a longitudinal record. A single patient's longitudinal record is divided up into visits (represented here as a set of notes written during the visit) and visits are binned into short or long gaps with respect to lipase measurements. For instance, the first and fifth visit are binned as short-gap visits, because they contain at least two consecutive laboratory test measurements that occur within a short time period. Visits 2, 3, and 4 are long-gap visits because they occur between two consecutive lipase measurements that are taken over a longer period of time.

To determine the threshold for how many days define a short-gap and long-gap, we used the laboratory test's measurement pattern histogram. Under the hypothesis that peaks in the histogram are indicative of the context within which a laboratory test is ordered we use the location of peaks and trends around them as thresholds for separating visits into short- and long- gap bins. This visit-separation method is generally insensitive to the exact threshold and is heuristically defined for each laboratory test as a function of the location and number of peaks in the measurement histogram. The measurement histogram of lipase (Figure 4.2) had one peak at one day and showed a change in measurement pattern at 3 days (the lipase histogram is only nearly linear after the 3-day gap).

As there is only a measurement gap when a patient has two or more tests, only patients with at least two lipase measurements were included in the analysis. Similarly, no visits after the last recorded lipase test were added to either bin. Our visit-binning method is not limited to two (short- and long- gap) bins and there can be more granular bins such as bins of regular weekly visits. In the case of lipase, the measurement pattern histogram illustrated two distinct measurement patterns.

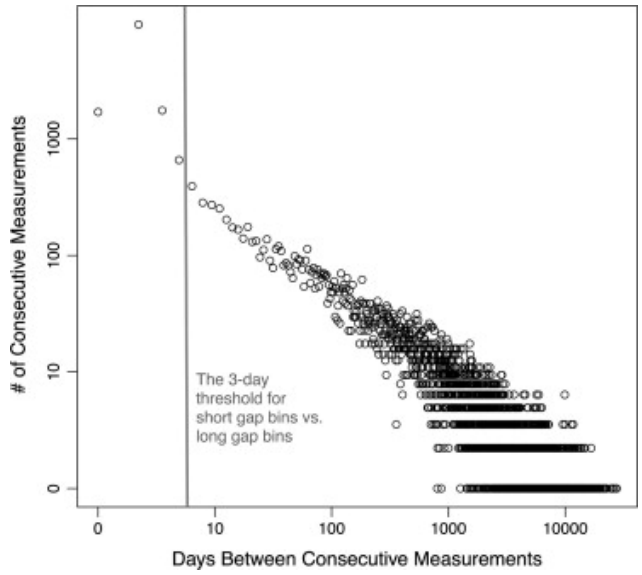


Figure 4.2 The measurement gap histogram curve for the lipase laboratory test. The measurement curve is presented on log–log scale as a histogram of the days between consecutive lipase test measurements for each patient, aggregated across the full population. We examined the figure visually and found that the pattern in measurement gap frequency changes at approximately 3 days; after 3 days the histogram curve is nearly linear. We used 3 days as a threshold for separating measurements on a short time scale (0–3 days) from measurements on a long time scale (over 3 days).

To analyze the differences between bins of short-gap measurements and long-gap measurements, we separated all ICD-9 codes and clinical notes created during short-gap lipase measurements from those collected during long- gap bins. We conducted association studies for lipase and acute pancreatitis in all three settings: (i) only short-gap visits, (ii) only long-gap visits, (iii) all visits.

Analyses

To assess in what ways the visits with short lipase measurement gaps differ from visits with long

lipase measurement gaps, we ran three analyses using ICD-9 codes and clinical notes. For all of the following analyses, we used patients who exist in both the short-gap and the long-gap bins. This filtration reduced the confounders and ensured that differences we uncovered were from genuine separate health states. All p-values were Bonferroni-corrected.

Note Types. We looked at note types across the short and long measurement gap bins. Note types can inform the status of the patient. For example, a high frequency of admission and discharge notes indicate many inpatient visits, while primary provider notes are indicative of outpatient doctor visits. We performed a chi-squared test to assess the strength of association between the frequency of each note type and the gap bin; this test was chosen because we are comparing counts across different bins.

Note Content. We analyzed the frequency and coverage of all words across the notes in each bin. Differences in note content indicate differences in topics and hint at different contexts of measurement across gap bins. To correct for the redundancy within notes, we calculated word coverage. Redundancy across notes within a patient was implicitly handled, as individual patient records were divided into both bins. We look at the note content, both frequency and coverage, to check the separation created by long and short lipase gaps. The presence of certain words that relate to specific health states can hint at the level of separation across the gap bins and provide clues as to whether relevant parts of the patient's record are indeed being separated out.

For each word we calculated:

$$\text{Frequency} = \frac{\#\text{ times the word appears across all notes in the gap bin}}{\text{total \# of words across all notes in the gap bin}}$$

$$\text{Coverage} = \frac{\#\text{ notes in which the word appears}}{\text{total \# of notes in the gap bin}}$$

We performed a chi-squared test to assess the strength of association between the coverage of each word and the gap bin.

Association Study. To test whether laboratory measurement dynamics affect a typical EHR association study (Warner and Alterovitz 2012), we conducted a phenome-wide search using the binomial

test to find ICD-9 codes associated with an elevated lipase. For every ICD-9 code, we compared its frequency of occurrence with high lipase in all three settings. The binomial test was used to assess the statistical significance of deviations due to high lipase from the expected distribution of ICD-9 codes. The variables were defined as follows, per ICD-9:

$$H_0: P(\text{ICD} - 9) = \frac{\#\text{ visits where the ICD} - 9 \text{ is recorded}}{\#\text{ visits where a lipase lab tests is done}}$$

$$\#\text{ trials} = \#\text{ visits where median lipase} > 43$$

$$\#\text{ successes} = \#\text{ visits where (median lipase} > 43 \text{ AND the ICD} - 9 \text{ is recorded)}$$

4.3 Results

Our final dataset consisted of 14,141 patients, their notes, ICD-9 codes, and laboratory test times and values for 70 tests, spanning 20 years. On average each patient had 150.4 ICD9 codes [95% CI: 147.7-153.1], 825.8 laboratory tests [95% CI: 807.7-847.0], and 133.5 clinical notes [95% CI: 130.8-136.3] over the entire study period.

4.3.1 Task 1: Correlation between Measurement Gap and Numerical Value

Both correlation metrics, linear and nonlinear (through mutual information), between measurement patterns and test values were carried out for the 70 laboratory tests (see Appendix A Tables 1 and 2 for full results). When we did not separate by time scale, there was little correlation: with the linear measure, all laboratory tests had a correlation very close to zero. With the mutual information measure, although also very low, a few laboratory tests demonstrated some level of correlation. The highest mutual information was 0.15, detected for the albumin laboratory test and only nine other tests had a mutual information higher than 0.1 (see Appendix A Tables 1 and 2). Overall, these very low correlations indicate that there is separate information encoded in the laboratory test measurement pattern and the laboratory test's numerical value.

Exploring the correlation statistics separately for different time scales (short and long gaps), some

laboratory tests such as LDL displayed no correlation (Figure 4.3(a)), using either metric, on any time scale, while other laboratory tests such as HbA_{1c} and creatinine, showed some degree of correlation. For LDL, the numerical value does not affect its testing rate. We interpret the absence of correlation in measurement patterns and value as a result of healthcare process, such as adherence to guidelines for testing (Grundy et al. 1993). HbA_{1c} displays a clear negative linear correlation of -0.193 only on the slow time scale, a higher HbA1c value is correlated to a shorter time until next measurement (Figure 4.3(b)). Creatinine also displays a clear negative linear correlation of -0.208, but on the short time scale.

The results from the linear correlation calculations were consistent with earlier work on the relationship between laboratory value and measurement frequency (Weber and Kohane 2013). Weber and Kohane assigned categories to laboratory tests based on how numerical values were perceived (e.g.: "Bad-Good" represented a laboratory test where a low value was bad, and a high value was good). In our work, a positive linear correlation indicates a "Bad-Good" test where a low value prompts rapid retesting and a high value has a longer measurement gap, similarly a negative linear correlation represents a "Good-Bad" test.

The interplay between correlations on the full time scale and separately on the short and long time scales revealed interesting findings about measurement dynamics. For example, bicarbonate showed a positive linear correlation on both long and short time scales (the higher the value, the longer the gap between measurements), but when aggregating the time scales together and computing the total linear correlation, the correlation disappeared (Figure 4.3(c)). By contrast, the PSA screening test had very similar mutual information on the full and long time scales. The differences between tests such as bicarbonate and the PSA screening test hint at differences in the contexts in which laboratory tests are ordered (Figure 4.3(d)). As PSA is measured in a single context, it is not subject to signal dilution due to timescale aggregation. Alternatively, the correlation results for laboratory tests such as bicarbonate, which are measured for multiple reasons and in various contexts, indicate that it is sometimes necessary to separate laboratory measurements by underlying context of measurement.

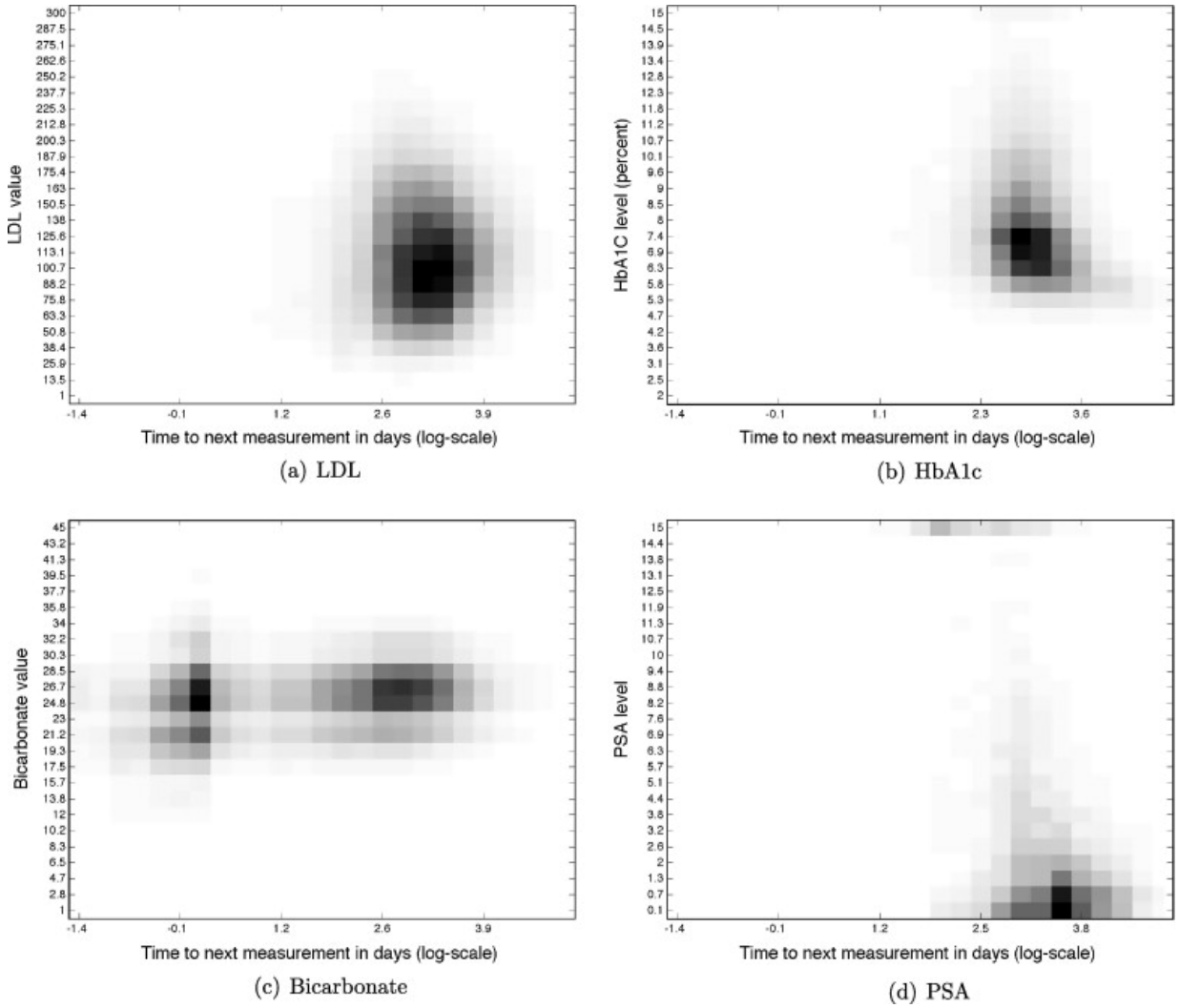


Figure 4.3 Density plots of the PDFs (consisting of laboratory values and time between consecutive measurements) for four laboratory tests shown on the full time scales. The *x* axis represents the log time to next measurement in days: log (1 h) = −1.38, log (1 day) = 0, log (1 week) = .85, log (1 month) = 1.47, log (1 year) = 2.56. Each graph shows different levels of correlation: LDL has no correlation on any time scale as shown by the mostly round ball, HbA1c has a negative correlation as shown by the L-shape of the curve, Bicarbonate separates along two time scales while PSA is almost exclusively measured on the long time scale of over a year between consecutive measurements.

4.3.2 Task 2: Laboratory Test Measurement Motifs

Manual cataloguing of the measurement gap histograms for 70 laboratory tests uncovered a set of three motifs that were most common. The three motifs of laboratory test ordering are influenced by two factors: patient health state and the healthcare process (Figure 4.4). These two factors contribute to create the shape of the laboratory test's histogram. Certain histogram motifs highlight the presence of multiple contexts in which the test is being ordered. The histogram shape can determine whether further population stratification is necessary for conducting analyses or whether the laboratory measurements already represent a mostly homogenous patient set. The contributions of the patient health state and the healthcare process is dependent on the laboratory test itself and define the three motifs of test ordering: (i) primarily inpatient, (ii) primarily outpatient, (iii) a mixture of in- and outpatient. Figure 5 shows typical graphs from each of these three categories.

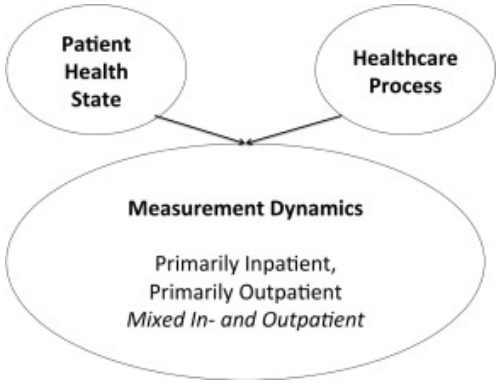


Figure 4.2 A Bayesian network describing the two factors that influence a laboratory tests measurement pattern. The extent to which each factor contributes, changes which motif the test belongs to, thereby changing how it can be used in research settings. The "mixed in- and outpatient" motif represents laboratory tests taken across patients with multiple health states. Laboratory tests with mixed motifs may contribute to biased results when computing over the multiple health states as one population of patients.

In general, laboratory tests that show peaks at very short time gaps in their measurement histograms are representative of tests taken during inpatient stays; laboratory tests with measurement graphs that peak

at longer gaps of a few months are representative of measurements obtained during an outpatient visit. For some tests, the documentation (outpatient vs. inpatient) reason is aligned with the clinical reason; related to the fact that the numerical value obtained from a particular test is valid for a specific time period only. For example, troponin levels are representative of a patient state at the hour level, and thus their measurements are on the timescale of days. Because troponin is measured for patients suspected of suffering a myocardial infarction, and such a diagnosis has a high rate of inpatient admission, the troponin measurement dynamics are representative of an inpatient stay.

Troponin's measurement dynamics represent a primarily inpatient laboratory test, motif (i). Other laboratory tests such as microalbumin and HbA_{1c} change at a slower time scale and are ordered primarily in outpatient settings; as the values change slowly, there is no need to repeat their measurement during a short-term hospital admission. Therefore, HbA_{1c} and microalbumin measurement dynamics represent primarily outpatient visits, motif (ii).

Laboratory tests that follow motif (iii) represent a set of tests whose measurement dynamics result from a mixture of both clinical and documentation reasons. For instance, glucose changes rapidly and is widely used in inpatient settings to monitor short-time scale changes but is also a regular test performed during outpatient visits to monitor chronic diabetics. The glucose dynamics are evident by the histogram diagram in Figure 5 where there is a fast time scale peak at 1 day, and a smaller slow time scale peak at 91 days. The peak at 91 days shows quarterly patient monitoring. Many other laboratory tests have motif (iii) measurement dynamics: for example, creatinine displays an almost identical histogram as glucose because they belong on the same basic metabolic panel and lipase along with amylase also have a mixture motif because of their use in both inpatient settings for acute events and outpatient settings for long-term monitoring.

Triglycerides is also a laboratory test with a mixture motif but with very different mixture weight (iii.b). This type of mixture laboratory test represents a dynamic that is also a result of mixed documentation and clinical reasons but with much heavier weight on the outpatient component mixing.

Most of the population receives triglycerides at 3-month time scales to assess heart health but a small subset of patients have their triglycerides monitored on a much shorter time scale, these are ICU patients with feeding tubes. The large portion of outpatient testing is seen in the triglycerides measurement gap histogram because the peak at 91 days is at a similar height as the peak at 1 day.

These three laboratory measurement motifs determine how to use different laboratory tests in EHR research. The tests for which clinical and documentation reasons align (laboratory tests used almost exclusively for in- patients or outpatients) represent a homogenous set of contexts, or patient states. Thus, with laboratory tests in motif (i) or (ii), aggregating across patient values is a safe approach. In contrast, the laboratory tests with the mixed measurement motifs, might represent several separate patient states (such as patients receiving triglyceride measurements as outpatient patients and patients receiving triglyceride measurements as ICU patients). Aggregating patient's values without separating the different patient state contexts in a large-scale study may introduce biases. The next section shows results of selecting patient cohorts by relying on a laboratory test with mixed dynamics (lipase) and how it impacts disease modeling (acute pancreatitis).

4.3.3 Task 3: Measurement Patterns Highlight Clinical State

We considered the histograms of measurement dynamics as plots of missing measurements. We presumed that for laboratory tests in the mixed motif category (iii), the data are missing not at random and may be informative of the patient's health state. Following intuition from Little's pattern mixture models (Little 1993) we hypothesized that different missing data patterns (or varying gaps between measurements) define different health states.

We explored this idea using lipase and hypothesized that (i) lipase measurement dynamics indicate two distinct missing values patterns, each representative of a clinical condition, rather than a documentation state; and (ii) separating the dataset by missingness patterns (visits with short gaps between measurements vs. visits with long gaps between measurements) helps recover the association between elevated lipase and the health state of acute pancreatitis.

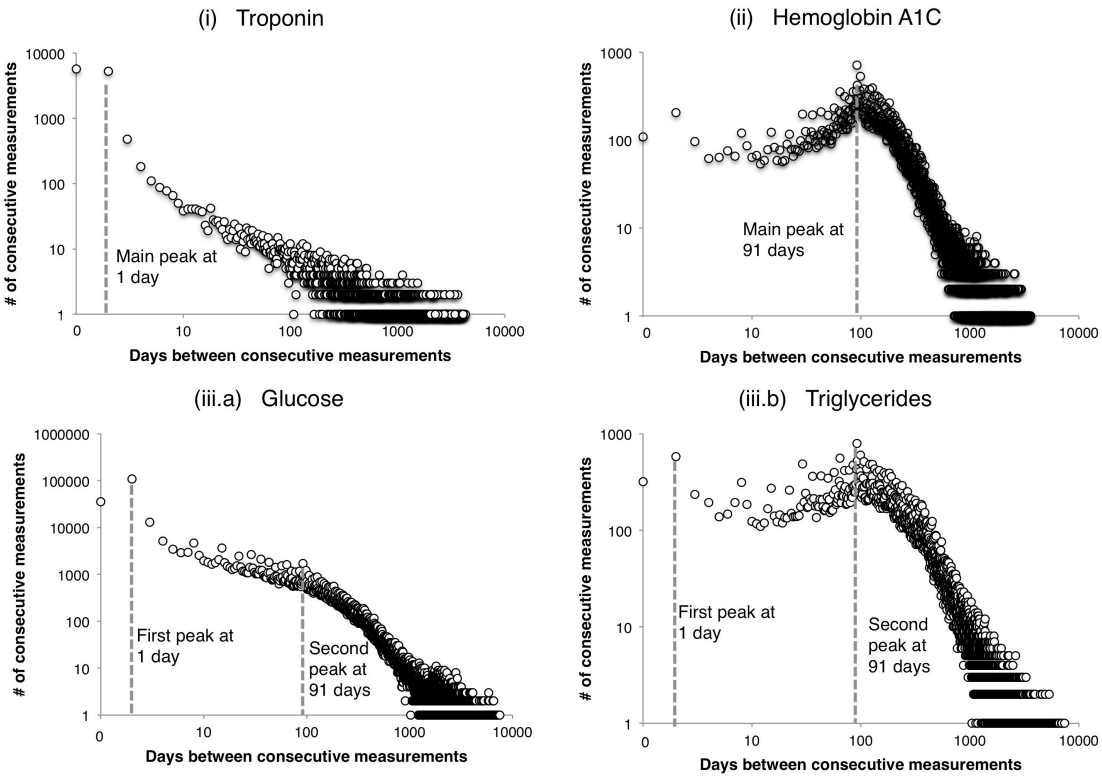


Figure 4.3 Representative examples of the three measurement gap motifs identified. Each laboratory test motif is presented on a log–log scale as a histogram of the days between consecutive test measurements for each patient, aggregated across the full population. (i) Troponin represents a primarily inpatient laboratory test, with a peak at 0 days and displays an approximately linear relationship in the coordinate system; (ii) HbA_{1c} is an example of a primarily outpatient laboratory test, showing a highly peaked distribution around 91 days; (iii.a) Glucose represents a mixture of in- and outpatient measurements, evidenced by the complex histogram: a high peak at on a short time scale (less than 10 days) and another peak at long time scales (multiple months); (iii.b) Triglycerides is another example of mixed laboratory test dynamics but with a slightly different mixture type: triglycerides has a high outpatient component and shows two different time scale peaks with a large quantity of measurements on the long time scale.

Note Types

The note-type analysis indicated a significant difference between the common note types used in the short-gap and long-gap bins. The note types in Table 4.1 showed that lipase measurement dynamics

can highlight true clinical differences, rather than documentation differences between inpatient and outpatient visits.

Note type	**0-3 Day measurement gap**		**3+ Day measurement gap**		**Difference in %**
	Raw frequency	% Total note types	Raw frequency	%Total note types	
Signout	6,269	0.144	21,345	0.039	0.105
Miscellaneous Nursing Note	4,443	0.102	14,944	0.027	0.075
12-Lead Electrocardiogram	2,126	0.049	9,217	0.017	0.032
X-ray of Chest, Portable	1,460	0.034	3,901		0.027
Discharge Summary	783	0.018	3,030	0.006	0.012
Progress Note	837	0.019	3,937	0.007	0.012
Adult Social Work Progress Note	662	0.015	2,383	0.004	0.011
Physical Therapy	605	0.014	2,325	0.004	0.01
Admission Note	579	0.013	2,154	0.004	0.009
Respiratory Care Patient Assessment	512	0.012	1,434	0.003	0.009

Table 4.1 Note types indicative of healthcare setting (in vs. outpatient) are spread across short and long gap bins, hinting that the measurement gap-based separation is not representative of healthcare, but rather of health states. This table shows the top 10 normalized differences in % of total note types in each bin, each has a highly statistically significant difference in % total note types.

There are note types that are only written during inpatient stays: an admission note, a signout note during a hospital shift change, and a discharge note. If separating visits based on lipase measurement dynamics were separating on a purely documentation basis with inpatient visits in the short-gap bin and outpatients in the long-gap bin, we would expect these inpatient-specific notes to be exclusively present in

the short-gap bin. Instead, there are more inpatient notes in the longer gap bin but the coverage of inpatient notes is larger in the 0-3 day bin; the signout, admission, discharge account for 14%, 2%, and 1% of the total note-types, respectively. The long-gap bin contains a large amount of inpatient data demonstrating that the laboratory test measurement dynamics are able to isolate visits based on health not hospital status in the short-gap bin. The note types present in the short-gap bin are relevant for diseases associated with lipase measurement as well. Elevated lipase often leads to testing for inflamed pancreas by ordering ultrasound scans, CT scans, ERCP or Chest X-Rays. Common note-types for all 4 procedures were significantly more frequent in the 0-3 gap bin and ranked in the top 10% of Table 4.1. The results from the note type analysis suggest that lipase measurement dynamics are able to separate by patient health status and find specific visits that more likely pertained to acute pancreatitis events.

Note Content

The words with the largest difference in coverage between the short-gap and long-gap bins are very relevant to pancreatitis (Table 4.2): both lipase and amylase can be used to diagnose pancreatitis, Librium is an anti-anxiety drug often given to alcoholic patients with withdrawal symptoms, and alcoholic patients often have pancreatitis. There are more references to “pancreatitis” in the long-gap bin but the normalized frequency and coverage of the word is much higher in the short-gap bin. This indicates that the notes written during shorter gaps in measurement are more focused on the pancreatitis diagnosis.

The word “pancreatic” modified “cancer” with a high prevalence in both gaps. This results from many patients with long-term pancreatic disorders experiencing acute episodes during their illness. We found that the word “cancer” has a coverage and frequency about 3 times higher in the long-gap bin than in the short-gap bin.

	0-3 Day measurement gap			3+ Day measurement gap			
Words	Raw frequency	%Total words	% Notes containing that word	Raw frequency	%Total words	% Notes containing that Word	Difference in coverage
pancreatitis	10,732	0.14	6.94	13,303	0.01	1.41	5.52
lipase	4,908	0.06	4.28	9,728	0.01	1.5	2.79
amylase	3,855	0.05	3.48	98,246	0.01	1.3	2.19
withdrawal	4,139	0.05	2.8	12,562	0.01	1.28	1.52
librium	3,303	0.04	2	6,064	0	0.59	1.42
pancreatic	4,393	0.06	2.76	15,992	0.01	1.38	1.38
epigastric	3,668	0.05	2.92	15,767	0.01	1.89	1.04

Table 4.2 Words associated with a pancreatitis health state are more frequently found in the short gap bins, suggesting that the separation is grouping notes written during visits that are more concentrated on the pancreatitis diagnosis into the 0–3 day measurement gap. The raw, normalized, and coverage frequencies of words in each gap sorted by the difference in coverage between the two bins is shown in this table. The words with a larger than 1% difference are shown, each of the words shown is highly associated with the 0–3 gap.

4.3.4 Recommendation for EHR Research with Laboratory Measurements

Knowing that acute pancreatitis is associated with high lipase levels, we used the binomial test to investigate whether separating visits by lipase measurement gaps highlight this association more prominently. In each setting, we performed a phenome-wide analysis to see the association between high lipase and ICD-9 577.0 (Acute Pancreatitis).

In the 0-3 day gap setting, the binomial test found the top association to be acute pancreatitis with an extremely significant Bonferroni corrected p-value of $< 1 \times 10^{-234}$. In the other two settings (greater than 3 day gap, and no separation by gaps), acute pancreatitis was also found to be the top association but

with a much smaller p-value (Figure 4.3). The long-gap setting had the smallest association and therefore the no separation setting also showed a much lower p-value. These p-value differences demonstrate that signal dilution is a consequence of ignoring measurement frequency bias during phenome-wide analyses. The process of binning laboratory values by their gaps between measurements (Table 4.3) can reduce confounding by not mixing different patient health states. Our results also demonstrate the generalizability of this measurement gap separation method. The measurement bins, created based on lipase measurement dynamics, were able to differentiate levels of association in other diseases as well. Type II diabetes (which may reflect clinicians screening type II diabetics with high triglycerides for pancreatitis) and HIV were differently associated with each setting. Without the lipase measurement-based separation of visits these disease associations are confounded by the bias of short-term lipase measurements. Interestingly, the ICD-9 for chronic pancreatitis is similarly associated with all three settings. The consistency of chronic pancreatitis is from patients having acute episodes during their chronic illness, conversely, not all acute pancreatitis patients have chronic pancreatitis. This asymmetry is demonstrated in the association patterns for acute and chronic pancreatitis.

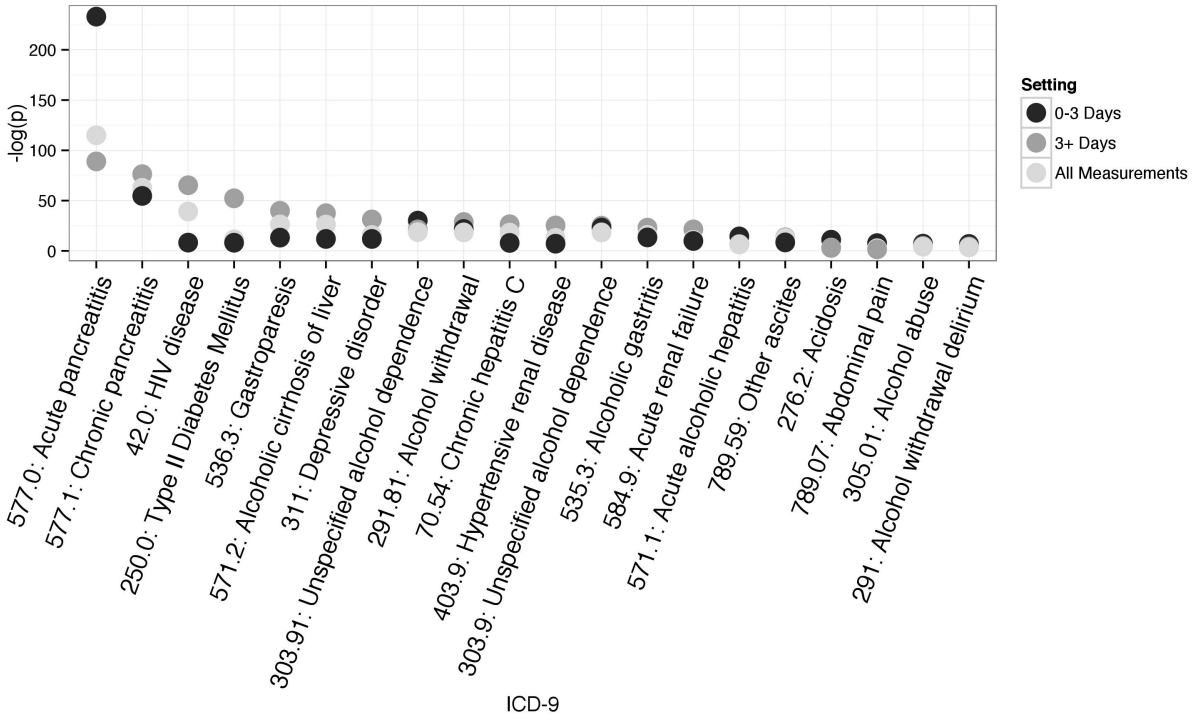


Figure 4.4 The results of a binomial association test between high lipase and ICD-9 codes. The binomial test was performed in all three settings (short gaps between measurements of 0–3 days, long gaps of more than 3 days, and all visits regardless of gaps between lipase measurements). The top 20 most significant associations are shown. For illustration purposes, the ICD-9 codes are sorted by association to high lipase in the 3 + days gap.

Step	Action	Motivation
1	Plot a histogram of the frequency and measurement gap in log–log coordinates	The histogram provides a method to visually examine the laboratory tests measurement dynamics
2	Examine the modality of the plot; looking for multi-modality	If the histogram is multi-modal, it may imply a difference in patient health states or a healthcare process bias
3	If there are multiple peaks, define a measurement gap threshold to separate the peaks	This separation defines multiple settings for the EHR experiment, creating sets of homogenous data points with respect to their measurement gaps
4	Perform the EHR experiment separately for each setting	Separately performing experiments for different settings may remove confounding bias

Table 4.3 **Measurement gap separation method.** The actionable measurement gap separation method for finding and removing a confounding bias in laboratory test EHR data.

4.4 Discussion

EHR research studies rely heavily on laboratory tests and their numerical values. We studied how laboratory test's pattern of measurements may provide additional information to a laboratory test's values. We discovered there is very limited correlation between how often a test is ordered and the value of the test. This lack of correlation implies that the value and measurement gap are informationally orthogonal to each other and are both important features to include when looking at laboratory tests, and specifically when using laboratory tests as features to represent patient disease state. In addition, there is evidence of different correlation results when examining laboratory test values on different time scales, showing that temporality plays a crucial role in the use of clinical laboratory test data.

We found evidence that measurement patterns of laboratory tests are dictated by clinical and physiological knowledge, but often confounded by the healthcare process, such as hospital documentation practices whether from workflows or guidelines.

One clear artifact of hospital document patterns was revealed by examining the measurement gap

histograms of 70 laboratory tests. Many laboratory test histograms have peaks at exactly 91 days, although their histogram height is varied. Our hospital serves a highly captive and sick population of patients in the surrounding neighborhoods. As the population is sick, 3-month checkups are a common practice and as the population lives nearby, the general adherence to a strict 3-month (91 day) schedule is high. It is clear that this 91-day peak is caused by the operations of the hospital and makeup of the population – not the clinical state of each individual patient. Although this particular healthcare process bias is specific to our institution, we postulate similar types of biases exist across the country and should be mitigated before using the EHR-recorded data for research.

Upon cataloguing all 70 laboratory tests we uncovered three types of measurement dynamics motifs, one which represents "mixed" laboratory tests where clinical factors and documentation standards are misaligned.

4.4.1 Separation by Measurement Pattern Mitigates EHR Bias

Laboratory tests with multiple ordering reasons, such as those used for both diagnosis and monitoring, present challenges to EHR-based research. When using laboratory values without accounting for the laboratory test's frequency of measurement, confounders can dilute the results of a study. The dynamics of laboratory measurement gaps across a population can reveal different measurement patterns present in the data; examining the measurement patterns of a particular test and then performing patient record decomposition in a strategic manner reduces signal dilution. For example, filtering the full patient cohort by visits within a particular measurement pattern of missingness will provide a more focused dataset of patient states.

We demonstrated our method of visit separation through measurement gap analysis on lipase as a specific use case. Using note types, note content, and ICD-9 codes we showed that our method could group inpatient visits pertinent to a particular clinical condition (acute pancreatitis in our example) away from inpatient visits pertinent to other clinical reasons. We also performed a secondary analysis to test that separating on hospital status does not yield the same results as separating by measurement gap. When

looking exclusively at inpatient visits (without accounting for measurement gap), the association between high lipase and acute pancreatitis was only as high as the ”3+ Days” association found using the measurement gap separation method (Figure 4.6). Therefore, we infer that the measurement gap separation method can indeed separate on health status, not simply healthcare process.

The work presented in this thesis is highly relevant to researchers working on cohort identification algorithms, especially with the recent push for more automated ways to perform high-throughput phenotyping (Chen et al. 2013; Wei et al. 2010; Lussier and Liu 2007). We present the stratification of an individual’s medical record by laboratory test measurement frequency as a new conceptual paradigm for studying EHR data with EHR-recorded laboratory tests.

Chapter 5: Learning Probabilistic Phenotypes from Heterogeneous EHR Data[5]

5.1 Introduction to the Phenome Model

Computational tools and techniques that reduce the dimensionality of many individual patient characteristics documented in the EHR, that discover underlying clinically meaningful latent states of the patient, and that allow reasoning in a probabilistic fashion over them would be powerful allies to clinicians. In fact, these tools would also facilitate many analytics tasks when applied to entire patient populations, including predicting disease progression, comparing effectiveness of treatments, and studying disease interactions (Hripcsak and Albers 2013; Wei and Denny 2015; Liao et al. 2015). For such tools to operate in a robust fashion across varied patients and enable high-throughput search over many diseases, modeling from large amounts of patient records is critical. How to build these tools is an open research question.

Here, we tackle this challenge through jointly modeling a very large set of diseases and an overwhelmingly large set of clinical observations. The observations are drawn directly from the heterogeneous EHR data, and the diseases are modeled in an unsupervised fashion. We refer to this task as large-scale probabilistic phenotyping, in essence building computational models of diseases from patient records.

We introduce the Phenome model in its first iteration, a graphical model for large-scale probabilistic phenotyping. The key contributions of the model are:

---

[5] This chapter was originally published in the Journal of Biomedical Informatics. The full citation for this publication is: Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning Probabilistic Phenotypes from Heterogeneous EHR data. 2015: 58: 156-165.

- It models diseases and patient characteristics as a mixture model, thus scaling easily to large sets of diseases and clinical observations;
- It derives phenotypes—individual disease distributions over patient observations—from raw data and diverse data types common to most EHRs: text, laboratory tests, medications, and diagnosis codes;
- It is unsupervised, and as such, can learn disease models across datasets from different institutions and different care settings, such as intensive care, emergency care, and primary care;
- It can incorporate clinical knowledge through informative priors; thus enabling the phenotypes learned in an unsupervised manner to be guided towards more established models of disease.
- It leverages a topic modeling approach, handling issues inherent to EHR data such as sparsity and noise, and capturing relations amongst observations that are implicit in the records.

5.2 Related Work

We discuss related work according to two areas of research: computational models of disease and probabilistic graphical models in the clinical domain.

5.2.1 Related work in Computational models of disease

One of the promises of the EHR is to enable reasoning and decision support over patient record data. Therefore, deriving a computationally actionable representation of patients based on their clinical records has been a grand research challenge, with proposed solutions from several disciplines and research fields. Since healthcare is driven almost entirely by the presence and/or severity of disease, representing diseases in an actionable fashion has been much investigated over the years.

The eMERGE phenotyping effort aims to model individual diseases one at a time. It relies heavily on expert consensus to build disease definitions that can be applied over a large set of EHRs. While time consuming, this effort yields precise phenotypes of single diseases (Newton, Peissig, and Kho 2013). More recently, single disease modeling efforts have experimented with automated feature extraction from

knowledge sources to reduce the manual effort involved in creating precise phenotypes (Yu et al. 2015). In addition to single disease phenotyping, researchers have also explored the use of clustering techniques to identify subtypes of a given individual disease (Doshi-Velez, Ge, and Kohane 2014; Lasko, Denny, and Levy 2013; Marlin et al. 2012; Schulam, Wigley, and Saria 2015).

When it comes to modeling a very large set of diseases at once, most of the work to date has been heavily reliant on manual knowledge curation. Ontologies like SNOMED- CT encode information about diseases such as potential treatments, symptoms, and the relationships amongst them. Bayesian networks which encode relationships amongst diseases and symptoms have also been developed (Miller, Pople, and Myers 1982). The Internist 1/QMR-DT resource was created manually and allows for computational reasoning about diseases and symptoms (Shwe et al. 2005; Jaakkola and Jordan 1999). One drawback of these resources is that, while their content is curated, they do not necessarily link to observation types documented in patient records.

Approaches that leverage healthcare data—whether claims data or EHR data—to represent diseases and their interactions have also been proposed in the literature (Hanauer, Rhodes, and Chinnaiyan 2009; Hidalgo et al. 2009; Roque et al. 2011). However, these approaches focus on interactions amongst diseases rather than modeling the diseases themselves.

Recently, a novel method to learn representations of multiple diseases across a large set of patients was proposed based on matrix factorization (Ho, Ghosh, and Sun 2014). In this framework, unseen patients can be assigned phenotypes, as defined by a collection of diagnosis codes.

Our work aims for a similar goal to that of Ho and colleagues (Ho, Ghosh, and Sun 2014): the Phenome model learns phenotypes for a wide range of diseases, and derives the model based on EHR data. However, the Phenome model departs from previous work in disease modeling in the following ways: (i) it learns a representation jointly over heterogeneous EHR data types (as it has been demonstrated that unstructured data adds valuable information to purely structured data in the context of phenotyping (Liao et al. 2015)), (ii) it operates in a probabilistic framework, thus enabling modeling of the uncertainty inherent in noisy EHR observations, and (iii) the model is not inherently limited to a set

number of data types, in this current iteration the Phenome model incorporates four different data types but provides a framework to expand and include many other clinically relevant data types as well.

5.2.2 Probabilistic graphical models in the clinical domain

With the growing amount of EHR data in electronic format, modeling the EHR with latent variable models has been an increasingly active area of research. The well-established Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003) has been applied to raw clinical text for several tasks, such as identifying sets of similar patients (Arnold et al. 2010), correlating disease topics with genetic mutations (Chan et al. 2013), analyzing themes in patient safety event data repositories (Fong and Ratwani 2015), and predicting ICU mortality (Ghassemi et al. 2014), with success, suggesting that topic modeling can act as a powerful and reliable dimensionality reduction technique. Such unsupervised modeling of topics is attractive as the language of clinical notes is particularly noisy with much paraphrasing power, and styles (e.g., abbreviations) that vary widely from one institution to another and from one care setting to the next. More recently, researchers applied LDA to billing codes from disparate EHRs and found sets of phenotypes that remain consistent across multiple institutions, further demonstrating the power and portability of unsupervised learning techniques (Chen et al. 2015). Researchers have also investigated novel probabilistic graphical models, also used in a variety of tasks, including ICU illness severity scoring (Saria, Koller, and Penn 2010), diagnosis code prediction (Perotte et al. 2011), redundancy-aware topic modeling (Cohen et al. 2014), disease progression modeling (Wang, Sontag, and Wang 2014), and disease subtypes identification (Schulam, Wigley, and Saria 2015).

While some clinical latent variable models have been evaluated in task-based settings (Ghassemi et al. 2014), evaluating the intrinsic value of the learned latent variables beyond their face validity, as well as their ability to infer meaningful latent states on unseen data can yield much information about the models. For general domain texts, evaluation methods of topic modeling have been much investigated. Experiments to obtain human judgments have been proposed (Chang et al. 2009), and automatic metrics that aim to correlate with such judgments (i.e., held-out likelihood and automatic topic coherence) have

been explored (Wallach et al. 2009; Newman, Lau, and Grieser 2010; Lau, Newman, and Baldwin 2014). Reliable and valid human judgments of topics are difficult to obtain, and as such automatic metrics are attractive. In the clinical domain, these metrics have not been validated fully, and quality judgments from clinicians are critical. When it comes to our phenotyping task, we want to evaluate how well the latent variables represent individual diseases. Because clinicians are trained to think about diseases as probabilistic mixture of symptoms, treatments, and comorbidities, we can leverage this training towards collecting qualitative judgments of phenotypes.

Our goal in developing the Phenome model is to build interpretable disease models that are clinically valid and actionable. As such, we developed the model with knowledge of the EHR characteristics in mind and designed experiments to test for clinical relevance of the learned phenotypes.

5.3 The Phenome Model

The Phenome model proposed in this chapter is a mixed membership model, inspired by the topic model literature (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). This model learns computational representations of disease based on observations from patient records, as encoded in the EHR. In this chapter we present two versions of the Phenome model, the fully unsupervised (UPhenome) model, and the grounded (GPhenome) model which incorporates clinical knowledge via known relationships between diseases and medications, laboratory tests, words, and diagnosis codes.

5.3.1 Inputs and Outputs of the Phenome Model

The input to the Phenome model consists of a large set of patient records, where each record is composed of free-text notes, medication orders, diagnosis codes, and laboratory tests. Each data type is treated as a bag of elements. The words in the notes are tokenized and simple filtering of vocabulary is based on token frequency as well as stop words removal. The medication orders are mapped whenever possible to bag of medication classes. The diagnosis codes are also encoded as a bag of codes.

There are two outputs to the Phenome model: learned phenotypes and an inference mechanism to identify a specific phenotype distribution for an unseen patient record. The learned phenotypes can act as

computational models of disease, and can be evaluated according to their interpretability and clinical relevance. In addition, the top-ranked diagnosis for a given phenotype can be used as a label proxy, thus supporting the interpretability of the disease model.

The inference mechanism acts as a dimension reduction technique, where each new patient record can now be represented as a distribution over a concise set of clinically meaningful variables (the learned phenotypes). Such a representation can be leveraged in many health analytics tasks, including patient record summarization, risk prediction, and patient cohort selection. The variables of the model are listed in Table 5.1.

P	Number of phenotypes
R	Number of patient records
β_r	Phenotype distribution for patient record r
η_p	Diagnosis code distribution for phenotype p
I_r	Number of diagnosis codes in record r
$v_{i,r}$	Diagnosis code instance i in record r
$\gamma_{i,r}$	Phenotype assignment for diagnosis code i in record r
θp	Words distribution for phenotype p
Nr	Number of words in record r
$w_{n,r}$	Word instance n in record r
$\delta_{n,r}$	Phenotype assignment for word instance n in record r
ιp	Medications distribution for phenotype p
Or	Number of medication orders in record r
$x_{o,r}$	Medication instance o in record r
$\varepsilon_{o,r}$	Phenotype assignment for medication instance o in record r

κ_p	Laboratory test distribution for phenotype p
M_r	Number of laboratory tests in record r
$y_{m,r}$	Laboratory test instance m in record r
$\zeta_{m,r}$	Phenotype assignment for laboratory test instance m in record r

Table 5.1 Variables in the Phenome model.

5.3.2 Baseline Models to compare against the Phenome Model

Before describing the Phenome model, we first describe two baseline models. The baseline models use LDA and are built on (1) clinical text and (2) all data types. Although billing codes are often used for identifying patient cohorts, it has been demonstrated that individual billing codes are not a reliable proxy for disease modeling (Wei et al. 2015). In addition, in our preliminary experiments (not shown) we were unable to produce clinically-reasonable phenotypes when running LDA with only billing codes.

The first baseline, LDA-text, considers only the notes in the records, following the hypothesis that clinical information about the diseases of a patient will be documented in the notes and thus can be captured through standard topic modeling. This is a state-of-the-art approach in several models for clinical data. Our first baseline is thus a vanilla LDA applied to the bag of words in the notes. We additionally experimented with modeling the other data types separately but as they yielded much less interpretable phenotypes due to sparsity (each patient record has many fewer diagnosis codes than words), we only used the LDA-text as a baseline.

The second baseline, LDA-all, learns topic models based on all observations in the record (words, medications, diagnosis codes, and laboratory tests). In this baseline we apply a vanilla LDA on all observation types in a single bag. The working hypothesis for this baseline is that diseases can be represented across the different data types in the record. This baseline model has the exact same input as the Phenome model.

5.3.3 Graphical Model representation of the Phenome model[6]

In the Phenome model, a patient record is represented as a probabilistic mixture of phenotypes, and the phenotypes are defined as a mixture of characteristics derived from a large, diverse population of patients. More specifically, a phenotype is defined as a set of distributions over the observation vocabularies, one for each of the four heterogeneous data types. As in the LDA baselines, we model the observations and phenotype assignments as multinomial distributions, and the phenotype distributions as sets of Dirichlet distributions. The details of the probabilistic latent variable model are shown in Figures 5.1 and 5.2.

The Phenome model departs from the LDA-text baseline by considering all data types in the record. It departs from the LDA-all baseline by treating each data type on its own, and learning the type-specific phenotype distribution separately (η_p , θ_p , ι_p , κ_p variables). There are multiple advantages to treating the data types this way: (i) this formulation adheres to the genre and characteristic of EHRs; (ii) it allows for future specification of different levels of sparsity and distributions for each data types; (iii) it enables a platform for incorporating domain knowledge specific to each data type which we explore further in this chapter; and (iv) it enforces the principle that conditioned on phenotype assignments, the per-data type phenotype distributions are independent of each other. Since the four phenotype distributions must separately sum to one, this mitigates potential imbalance in data type prevalence, and thus hinders one frequent observation type (e.g., words) from overwhelming the less frequent ones (e.g., diagnosis codes).

[6] The notation used in Figure 5.2 and in Section 5.3.4 differs slightly from the notation used in the journal publication version of this work. The thesis notation is more precise, however, for readability purposes it was simplified for the journal article.

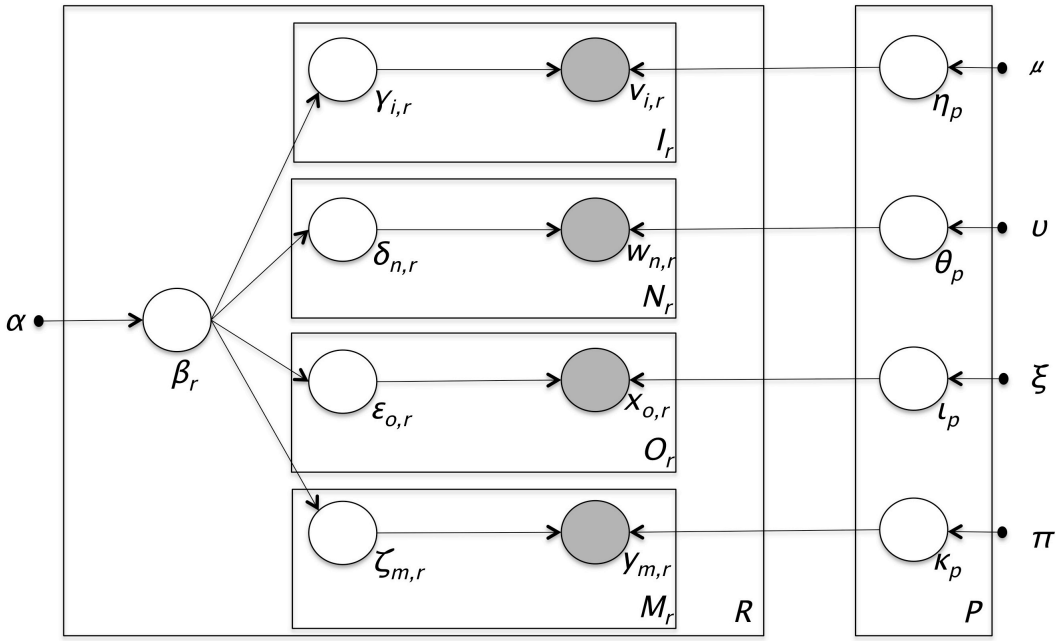


Figure 5.1 Graphical representation of the Phenome model.

foreach *phenotype p in* $\{1...P\}$ **do**
  Sample phenotype distributions for each data type
  $\eta_p \sim Dir(\mu)$; $\theta_p \sim Dir(\nu)$; $\iota_p \sim Dir(\xi)$;
  $\kappa_p \sim Dir(\pi)$
end
foreach *record r in* $\{1...R\}$ **do**
  Sample patient phenotype composition
  $\beta_r \sim Dir(\alpha)$
  Sample instance assignments
  foreach *diagnosis instance* $i \in I_r$ **do**
    $\boldsymbol{\gamma_{i,r}} \sim Mult(\beta_r)$; $\boldsymbol{v_{i,r}} \sim Mult(\eta_{\gamma_{i,r}})$
  end
  foreach *word instance* $n \in N_r$ **do**
    $\boldsymbol{\delta_{n,r}} \sim Mult(\beta_r)$; $\boldsymbol{w_{n,r}} \sim Mult(\theta_{\delta_{n,r}})$
  end
  foreach *medication instance* $o \in O_r$ **do**
    $\boldsymbol{\epsilon_{o,r}} \sim Mult(\beta_r)$; $\boldsymbol{x_{o,r}} \sim Mult(\iota_{\epsilon_{o,r}})$
  end
  foreach *lab test instance* $m \in M_r$ **do**
    $\boldsymbol{\zeta_{m,r}} \sim Mult(\beta_r)$; $\boldsymbol{y_{m,r}} \sim Mult(\kappa_{\zeta_{m,r}})$
  end
end

Figure 5.2 Generative story for the Phenome model

5.3.4 Inference in the Phenome Model

To perform inference on the Phenome model, we derived a collapsed Gibbs sampler which collapses the parameters β, η, θ, ι, κ. Due to the similarity of the models, inference in the Phenome model follows methods previously outlined for inference in LDA (Griffiths and Steyvers 2004). The derivations for the sampler are available in Appendix B. For illustrative purposes, in equation (1), we show the conditional distribution necessary for Gibbs sampling for one of the phenotype assignment variables, the laboratory test phenotype assignment, and note that the conditionals for the other phenotype assignment variables follow closely.

$$p(\zeta_{m,r} = p \mid \zeta_{-(m,r)}, \ldots) = \frac{\left(\pi + C^{p,-(m,r)}_{(\cdot):,l^*,1:Mr}\right)}{\sum_{l=1}^{|\kappa|}\left(\pi_l + C^{p,-(m,r)}_{(\cdot),l,1:M_r}\right)} \times$$
$$\left(\alpha + C^{p}_{r,(\cdot),1:I_r} + C^{p}_{r,(\cdot),1:N_r} + C^{p}_{r,(\cdot),1:O_r} + C^{p,-(m,r)}_{r,(\cdot),1:M_r}\right). \quad (1)$$

Here, the notation $C^{p}_{r,(.),1:I_r}$ represents the total count of all diagnosis code tokens I_r in patient record r, that are assigned to phenotype p. Similarly for $C^{p}_{r,(.),1:N_r}$ (word tokens) and $C^{p}_{r,(.),1:O_r}$ (medications). The counts for laboratory tests are performed without counting the particular laboratory test instance, m, which belongs to the type l^* and for which the conditional distribution is being evaluated. Therefore, $C^{-p,-(m,r)}_{(.),l^*,1:M_r}$ represents the total count of all laboratory tests, M_r, in patient record r, that are assigned to phenotype p, except (represented by a minus) for the current laboratory test instance m in record r. Finally, we note that the denominator of this conditional only contains counts for the laboratory test, demonstrating the differences across the phenotype distributions for each data type.

5.3.5 Grounding the Phenome Model

A primary goal of our work is to generate phenotypes compatible with clinician's mental models of diseases. Incorporating a human in the loop, like in interactive topic modeling (Hu, Boyd-Graber, and Satinoff 2011) and clinical anchor learning (Halpern et al. 2014) is a promising approach. Another

approach to support this goal is to incorporate knowledge from existing clinical knowledge resources, inspired by the advances in constrained topic modeling (Andrzejewski, Zhu, and Craven 2009; Andrzejewski et al. 2011; Hu, Boyd-Graber, and Satinoff 2011), and incorporating known semantic relations in disease modeling (Doshi-Velez, Wallace, and Adams 2015).

In this work, we experiment with automatically grounding the unsupervised Phenome model algorithm using existing clinical knowledge resources. Links between different data types and diseases are derived from a variety of curated terminologies. We used the following sources: RxNorm and NDFRT for medications linked to diseases by the "May Treat" or "May Prevent" relationship, Medical Entities Dictionary (MED - the terminology that underlies the NYPH data warehouse) for laboratory tests linked to diseases by the "Procedure indicates Patient Problem" relationship, UMLS for billing codes to diseases links through the ICD-9 to SNOMED-CT link, and the Observational Medical Outcomes Partnership Standard Vocabulary for words to diseases by collecting at all first-order relationships from that disease to all other entities and tokenizing the results. Although the Logical Observation Identifiers Names and Codes (LOINC) terminology is more widely used for encoding laboratory tests, we chose to implement the GPhenome model using the MED because the link between laboratory tests and diseases was readily available from the MED. The link information was encoded into the model priors by creating informative, asymmetric priors with pseudo counts when the links were identified. To ensure that this prior information would not be completely overwhelmed by the data, every observation-disease link found in the clinical ontologies had the pseudo count for the prior set to 100.

5.4 Experimental Setup for the Phenome Model

We now describe our datasets, parameter settings and model selection, as well as the different evaluation experiments we carried out**.**

5.4.1 Datasets

To investigate the generalizability of the Phenome model, we experimented with two qualitatively different mixtures of patients and patient diseases: (1) records of extremely sick patients who

are in the intensive care unit (ICU) with constant monitoring, which usually spans a few days; and (2) records for outpatients regularly followed by care providers over multiple years. These datasets are also from different institutions using different EHR systems. In each dataset, 80% of the records were used for training, and 20% for testing. Descriptive statistics about the training sets for each dataset are given in Table 5.2.

	MIMIC ICU Dataset	**NYPH Outpatient Dataset**
	Total / Unique	**Total / Unique**
Patients	18,697 / 18,697	9,828 / 9,828
Words	13,086,278 /12,919	12,840,334 /12,295
Medications	1,044,541 / 855	22,146/ 353
Lab Tests	7,499,446 / 309	549,699 / 318
Diagnoses	159,740 / 985	233,214 / 931

Table 5.2 Descriptive statistics for the MIMIC ICU and NYPH Outpatient training datasets. In this work, the number of patients and the number of input records is equivalent. Each MIMIC patient has one input record consisting of all data gathered during one ICU stay and each NYPH patient has one input record consisting of all data gathered during four consecutive outpatient visits.

MIMIC II ICU Dataset

We included all adult patients in the dataset, independently of their present or absent conditions. For each record, we selected one ICU admission and all of its corresponding observations: discharge summary, all medications, all diagnosis codes, and all laboratory tests. Medications and laboratory tests were mapped to the standard vocabulary definitions provided by MIMIC. For all data types, we limited the vocabulary to observations that appeared at least 20 times in the training dataset.

NYPH Outpatient Dataset

For this study, the NYPH data was subset to select only outpatient visits, providing a contrast to the MIMIC ICU dataset. We included all patients, independently of their conditions. Since their records span decades and often hundreds of visits, we considered slices of records for each patient that capture a somewhat stable health status. We selected the most recent time slice of each record that contained four different primary provider notes with no intervening inpatient stays. We defined record lengths by number of notes and not absolute time to account for different rates of visit (Hripcsak, Albers, and Perotte 2015). Any patient whose record slice lasted less than 1 month or greater than 4 years was removed and this resulted in patient records with mean length of 10 months (7 months standard deviation).

As in MIMIC, we collected all observations related to primary provider notes, medications, diagnosis codes, and laboratory tests. The range of medications is much more diverse than MIMIC medications, and thus we mapped all medications to their therapeutic class when possible (e.g., "Tylenol" was mapped to "Analgesic"). Similarly, the laboratory tests were mapped to groups of tests when possible (e.g.,"Glucose finger stick" was mapped to "Glucose"). As this dataset (in contrast to MIMIC), is not deidentified, we applied a simplistic method for trying to automatically remove as many patient and physician names as possible by removing any first and last names that appeared in the "Top 1000 Names" as listed by the US Census. Thus, the vocabulary sizes for these observations were dramatically reduced. We applied frequency thresholds to the data to closely match MIMIC vocabulary sizes whenever possible.

5.4.2 Model Parameters and Model Selection

MIMIC II ICU Dataset

With the MIMIC dataset, we focused our model selection on identifying the best number of latent variables (P). We ran the UPhenome model with the following P settings on the MIMIC ICU dataset: 50, 75, 100, 250, 500, 750. All of the hyperparameters (α, μ, ν, ξ, π) were set to 0.1 for all models. To ensure appropriate burn-in time for each model, the Gibbs sampler was run 7,000 iterations and the log-

likelihood curves on training data were examined to verify burn-in. The learned phenotype assignment settings for each model were selected as the ones that produced the maximum log-likelihood over all 7,000 iterations on the training set.

For model selection, we optimized for interpretability of phenotypes, and thus relied on both held-out likelihood and coherence of the learned phenotypes. The automated phenotype coherence calculation was performed using normalized pointwise mutual information (NPMI) over all observation types, as described by Lau et al. (Lau, Newman, and Baldwin 2014) and using the provided open source code. For each model, we used the average NPMI across all phenotypes to represent the overall coherence of the learned phenotypes from the model.

For this experiment, NPMI is calculated for each observation (z_i) in each phenotype, limited to the top N=40 most probable observations. The per-phenotype NPMI is the average of the NPMI of each word, and the model NPMI is the average of each per-phenotype NPMI value, defined by:

$$NPMI(z_i) = -\sum_{j=1}^{N-1} \frac{log\frac{P(z_i,z_j)}{P(z_i)P(z_j)}}{\log P(z_i,z_j)} \quad , \qquad (2)$$

where $P(z_i)$ = probability of seeing observation z_i and

$P(z_i,z_j)$ = probability of seeing both observation z_i and observation z_j in the same patient record.

To calculate the likelihood on the held-out set, we implemented a Chib-style estimator as described by(Murray and Salakhutdinov 2009). The likelihood was calculated using 1,000 iterations of the estimator for every setting of P (likelihood curves and the NPMI coherence are provided in the online supplement material). The combination of the two metrics suggested that P=250 was likely to produce the best phenotypes given the current parameter settings on the MIMIC dataset.

For both baseline models LDA-text and LDA-all, we used the MALLET software package (McCallum 2002) with similar number of topics as the selected UPhenome model (K=250) and similarly with 7,000 iterations to ensure burn-in. The hyperparameter settings for the baseline models were 0.01 to increase sparsity due to the larger size of the combined vocabulary.

NYPH Outpatient Dataset

With the NYPH dataset, we experimented with grounding the Phenome model and with hyperparameter optimization.

Grounding

To identify the correct phenotypes on which to ground the Phenome model, we explored the disorders frequently documented in the NYPH dataset. Using counts of SNOMED-CT Core Problem List disorders most commonly found across the entire NYPH dataset, we found that 695 disorders were common enough in the NYPH dataset and contained enough information upon which to ground (ie. enough medications, laboratory tests, words, or diagnosis codes that are linked to this disorder across clinical ontologies). We ran GPhenome on 750 phenotypes: 695 of them had augmented counts for the identified grounding and the remainder were left ungrounded to learn other phenotypes present in the data.

Hyperparameter Setting

We ran several iterations of the UPhenome model with different hyperparameter settings to determine the best setting for α and the rest of the hyperparameters (μ, ν, ξ, π). The models were run until the burn-in and the log- likelihood curves on training data were examined to verify burn-in. The learned phenotype assignment settings for each model were selected as the ones that produced the maximum log-likelihood over all iterations on the training set.

To find which hyperparameter setting produced the best model, we used the same Chib-style estimator as described by (Murray and Salakhutdinov 2009). The likelihood was calculated using 1,000 iterations of the estimator for every setting of the hyperparameters.

For the baseline models LDA-all baseline model, we used the MALLET software package (McCallum 2002) with similar number of topics as the selected GPhenome model (K=750) and with a similar number of iterations to ensure burn-in. The hyperparameter settings were set to be the same as the optimal settings found for the NYPH dataset (α =0.1, [μ, ν, ξ, π]=0.001).

5.4.3 Evaluation Experiments

In addition to the automated evaluation metrics, such as held-out likelihood and automatic coherence of the learned phenotypes, we carried out the following set of experiments: qualitative assessment of learned phenotypes (manual coherence, manual granularity, pairwise phenotype comparison, label quality) and ability of the phenotypes to characterize ground-truth disorders present in a set of unseen patients (disorders to phenotypes associations).

All qualitative judgments were obtained from a clinical expert. The learned phenotypes were displayed using a modified version of the interactive topic modeling interface (Hu, Boyd-Graber, and Satinoff 2011). The interface is particularly useful to us because it enables users to edit learned phenotypes by adding/removing observations and marking them as important or ignorable. For each phenotype, the top 40 most probable observations were displayed and weighted by their phenotype probability (normalized by their data-type specific phenotype probability). The observations were also color-coded to signify their data types (purple for words, grey for medications, green for laboratory tests, and blue for diagnosis codes).

Manual Coherence

This experiment allows us to capture the intrinsic quality of the learned phenotypes across its probable observations, very much like the automatic coherence metric aims to. In our case, a coherent phenotype is one that describes a single condition and that does not contain observations that are not typically seen in patients with this condition.

In the MIMIC ICU dataset we compared the UPhenome model with the baseline LDA-all model each. In the NYPH dataset, we compared the UPhenome model, the GPhenome model, and the baseline LDA model. For each experiment, 50 phenotypes were randomly selected from each model and presented one at a time to the clinical expert. The expert was asked to score the coherence of a given phenotype (without being told which model produced the phenotype) according to a 1-5 Likert scale, where 1 designated no coherence (i.e., uninterpretable) and 5, perfect coherence. To help the expert in his

assessment, we instructed him to use the interface to edit the phenotypes, and use his number of edits as a cue for lack of coherence.

Manual Granularity

Because both the baseline and the Phenome models are unsupervised, there is no guarantee that the learned phenotypes are good representations of clinically meaningful diseases. In particular, it is possible that the modeling of patient observations generates clusters that are reflective of the documentation processes of healthcare rather than of the documentation of clinical status of a patient. For instance, patient discharge to a nursing home in the MIMIC dataset contains very specific documentation patterns, which in an unsupervised setting could be easily grouped.

We asked the clinical expert to categorize the random phenotypes into one of the following three granularities: (i) a non-disease phenotype (e.g., a healthcare process or uninterpretable phenotype phenotype); (ii) a mix of diseases; (iii) a single disease.

Pairwise phenotype comparison

In this experiment, we assess the compared overall quality between phenotypes learned from the different models. In the MIMIC dataset, we compared the learned UPhenome phenotypes and learned LDA-baselines phenotypes. To ensure that the comparison is fair, we selected the 50 random Phenome phenotypes, and identified the most similar corresponding LDA-all phenotype for each. For the NYPH dataset we compared the UPhenome and GPhenome phenotypes and selected 50 random UPhenome phenotypes and identified the most similar corresponding GPhenome phenotype for each. To compute pairwise phenotype similarity, we used Jensen-Shannon divergence over the posterior probabilities of all observations.(Manning and Schütze 2003) The clinical expert was presented pairs of phenotypes without him knowing which model generated which phenotype. The expert was asked to choose which phenotype was more clinically coherent or to code the comparison as impossible when both phenotypes were unintepretable, equivalent, or paired improperly.

Label Quality

One of the data types used in the Phenome model is diagnosis codes. The diagnosis codes are ICD-

9 codes, which often describe specific diseases (e.g., “breast cancer”), but can also describe classes of diseases (e.g., “malignant neoplasms”), as well as generic statements about a patient (e.g., “personal history of other diseases”). Since the diagnosis codes are often used in the clinical world as proxies for conditions present in a patient, this experiment assesses to which extent the most probable diagnosis code for a learned phenotype is a clinically appropriate label for the phenotype as a whole.

Since the LDA-text baseline does not include diagnosis codes, this experiment was skipped for this model. Similarly, since in the baseline LDA-all model it is possible that the top-40 most probable observations do not include any diagnosis code (in our experiment, this happened actually very often), we assessed label quality for the 50 Phenome phenotypes only.

We asked the clinical expert to categorize the top diagnosis code with respect to the phenotype as a whole as (i) related; (ii) unrelated; or (iii) actionable. An actionable label is one that accurately represents the phenotype at the right granularity and it can be relied upon when making a decision about a patient with the phenotype assigned.

Disorders to Phenotypes Associations

While the previous experiments assessed the quality of the learned phenotypes, this experiment assesses the ability of the Phenome phenotypes to characterize clinically relevant and ground-truth disorders present in a set of unseen patient records. If there is a strong association between phenotypes and the disorders in these records, then the phenotypes can be considered clinically relevant for a given patient.

A potential application of the Phenome model to identify present disorders for a given patient by inferring the most probable phenotypes for the record. To validate this point, we relied on a gold-standard set of records, which contain manual annotations of the disorders present and mentioned in the records’ notes. The ShARe gold standard is based on MIMIC notes and contains such annotations (Semeval-2015 task 14: Analysis of clinical text, 2015.; Pradhan et al. 2014). We included the 350 discharge summaries from the ShARe corpus in our test set. For each record, gold-standard annotation provided a list of SNOMED-CT Disorder concepts, along with modifiers such as negation and uncertainty. Ground-truth

disorders for a record were defined as concepts with no negation or uncertainty.

In total, for each of the 350 records, we have ground-truth disorder concepts (from a set of 2,000+ unique concepts in the corpus) and inferred phenotype assignments. We created an association matrix, similar to work by Griffiths and Steyvers (Griffiths and Steyvers 2004), visualizing the degree of association between present concepts and phenotypes. We selected concepts, which occurred in at least 50 records. We selected the concepts that occurred in at least 50 records. This experiment examines associations between common disorders and learned phenotypes; with 350 patients, there are not enough annotations to associate phenotypes with rare disorders. The association was computed using normalized pointwise mutual information, and for each concept the top phenotype was selected.

5.5 Results

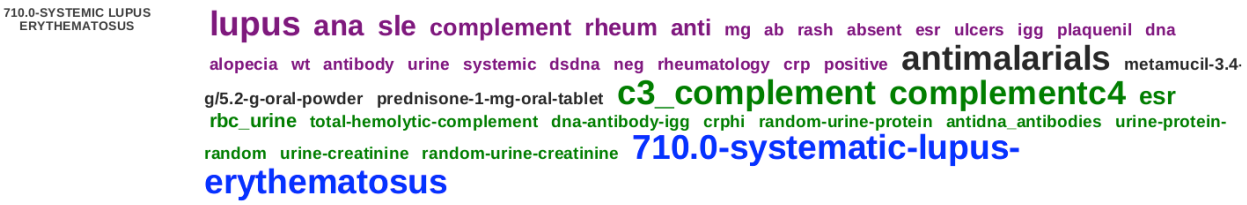


Figure 5.3 An example of a learned phenotype. The top 40 most probable observations for the phenotype are listed.

The Phenome model is able to produce interpretable results: Figure 5.3 shows an example of learned phenotype on the NYPH dataset. The label on the left is the most probable diagnosis code, in this case SLE (Systematic Lupus Erythematosus). The words (in purple) are indeed related to this disease and refer to abbreviations in clinical notes ("rheum") or mentions of important laboratory tests for SLE ("ana", "esr"), as well as mentions of specific drugs indicated for SLE ("plaquenil"). The medications are also related (plaquenil, one of the most common drug for SLE is an antimalarial medication), as well as prednisone. Both C3 and C4 levels are used as diagnosis tests for SLE, while ESR and others test for level of inflammation in a patient. SLE is a phenotype learned from the NYPH dataset, which represents outpatient records over long periods of time. Since SLE is a chronic disease it makes sense that it was discovered in this dataset. For comparison, there is no phenotype learned on the MIMIC dataset that

captures characteristics of SLE. Because the Phenome model is unsupervised, it models the diseases that are of interest to a given input clinical setting/patient cohort.

5.5.1 Model Selection

MIMIC II ICU Dataset

Using the NPMI (Table 5.3) and held-out likelihood (Figure 5.4) calculations together identified P=250 as the best number of latent variables to explain the MIMIC dataset. The likelihood was calculated on a held-out dataset of 4,685 (20%) MIMIC ICU patients.

Model	Number of Latent Variables	Average Normalized Pointwise Mutual Information
Phenome Model	50	-0.0009
	75	0.0029
	100	0.0057
	250	0.0140
	500	0.0128
	750	0.0
LDA-all	250	0.0750
LDA-text	250	0.0896

Table 5.3 Average NPMI for different numbers of latent variables

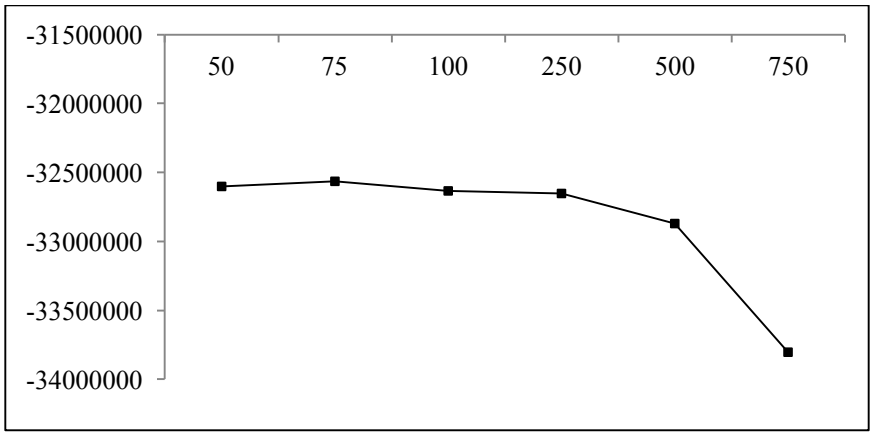


Figure 5.4 Held-out likelihood calculation for the MIMIC dataset for different numbers of latent variables.

NYPH Outpatient Dataset

The number of phenotypes (750) for the NYPH dataset was determined by disease concept frequency in order to perform grounding. Nine different hyperparameter settings were tested (Table 5.4) on a validation set to determine the optimal setting for the NYPH dataset. To ensure that (0.1, 0.001) is the optimal setting, the held-out likelihood for (0.1, 0.0001) was also calculated; the result was -2.522×10^7.

	α parameter setting			
μ, ν, ξ, π parameter settings		**0.1**	**0.01**	**0.001**
	0.1	-2.610×10^7	-2.569×10^7	-2.57×10^7
	0.01	-2.610×10^7	-2.519×10^7	-2.531×10^7
	0.001	-2.511×10^7*	-2.512×10^7	-2.530×10^7

Table 5.4 Held-out likelihood on a test set for different parameter settings. * represents the optimal setting.

5.5.2 Evaluation 1: Coherence

MIMIC ICU Dataset

For the comparison of the Phenome model with baselines, we report its comparison to the LDA-all baseline alone, but note that the results were similar as for the comparison to the LDA-text baseline.

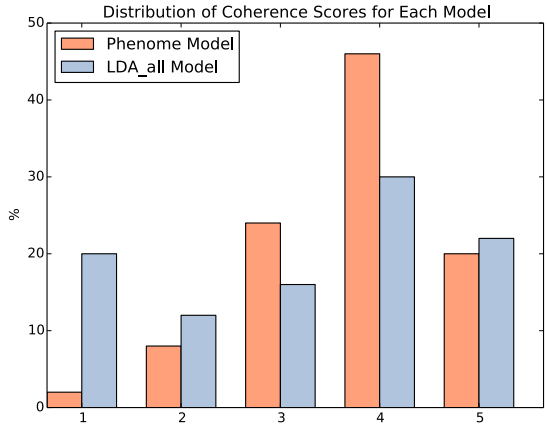


Figure 5.5 Distribution of manual coherence scores for the UPhenome and LDA-all phenotypes on MIMIC data. A score of 1 represents an unintepretable phenotype and 5, a perfect phenotype.

Figure 5.5 shows the distribution of coherence scores for the 50 phenotypes from the Phenome model and the 50 phenotypes from the LDA-all model. The LDA-all phenotypes contained many more unintepretable phenotypes than the Phenome ones, and at the same time contained slightly more perfectly coherent phenotypes than the Phenome ones. Overall, about 66% of the Phenome phenotypes were scored as good (coherence score above 4), while only about 52% of the LDA-all phenotypes were scored as good.

NYPH Dataset

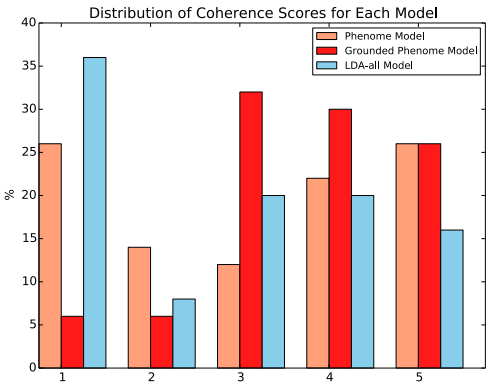


Figure 5.6 Distribution of manual coherence scores for the UPhenome, GPhenome, and LDA-all phenotypes on NYPH data.

Figure 5.6 shows the distribution of coherence scores for the 50 phenotypes from the Phenome model, 50 phenotypes from the grounded Phenome model, and 50 phenotypes from the LDA-all model. The LDA-all phenotypes contained many more uninterpretable phenotypes than either of the Phenome ones. The GPhenome model was able to create more coherent phenotypes than the fully unsupervised Phenome model. Overall, about 48% of the Phenome phenotypes were scored as good (coherence score above 4), 56% of the Grounded Phenome model phenotypes were good, but only about 36% of the LDA-all phenotypes were scored as good. We note, however, that by performing ad-hoc post-processing on the LDA-all results such as data-type mapping and data-type normalization (a step that the Phenome model does inherently during learning), we were able to achieve manual coherence results comparable to the Phenome model results.

5.5.3 Evaluation 2: Granularity

MIMIC ICU Dataset

The UPhenome model yielded more phenotypes that are clinically well-defined as representing a single disease than the LDA-all baseline. Furthermore, the LDA-all baseline suffered from a high number of unintepretable phenotypes (as confirmed by the Manual coherence experiment).

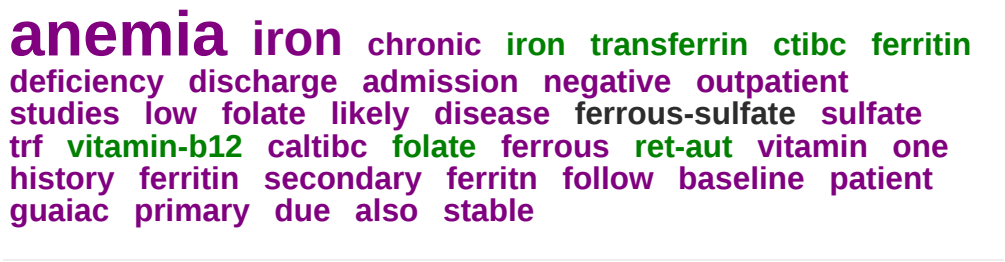
For the UPhenome phenotypes, the clinical expert categorized 10% as non-disease phenotypes, 10% as a mix of diseases, and 80% as representing a single disease. In contrast, for the LDA-all phenotypes, the clinical expert categorized 42% as non-disease phenotypes, 6% as a mix of diseases, and 52% as representing a single disease.

NYPH Dataset

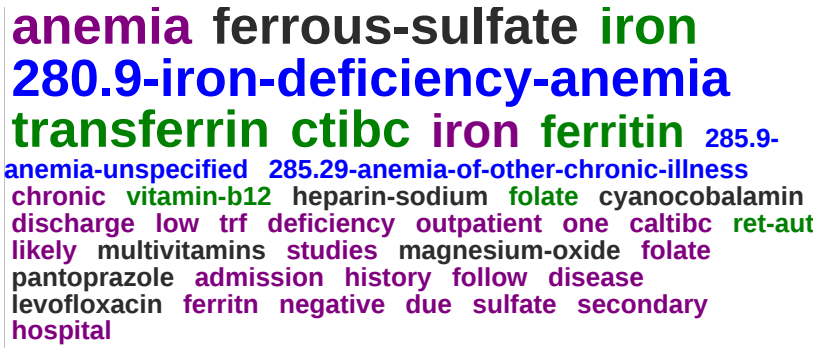
For the UPhenome phenotypes, the clinical expert categorized 46% as non-disease phenotypes, and 54% as representing a single disease. In contrast when the model was grounded, the clinical expert categorized 14% as non-disease phenotypes and 86% as representing a single disease. For the LDA-all model, the expert found that 62% of the phenotypes were non-disease, 4% represented a mixture of diseases, and only 34% of the phenotypes represented a single disease.

5.5.4 Evaluation 3: Pairwise Phenotype Comparison

MIMIC ICU Dataset



(a) LDA-all phenotype



(b) Phenome phenotype

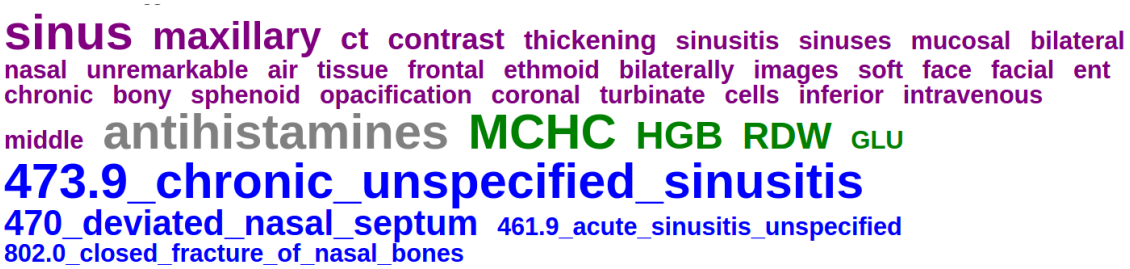
Figure 5.7 An example of LDA-all and Phenome phenotypes, both about Iron Deficiency Anemia, as paired automatically by Jensen-Shannon divergence.

Figure 5.7 displays an example of paired phenotypes shown to the clinical expert. We can see that the most probable observations in the LDA-all phenotype are words and laboratory tests (except for one medication), with only a few highly probable and relevant words/tests and most irrelevant to iron deficiency anemia. In comparison, the most probable observations in the Phenome phenotype are spread across data types, with a majority of observations relevant to iron deficiency anemia.

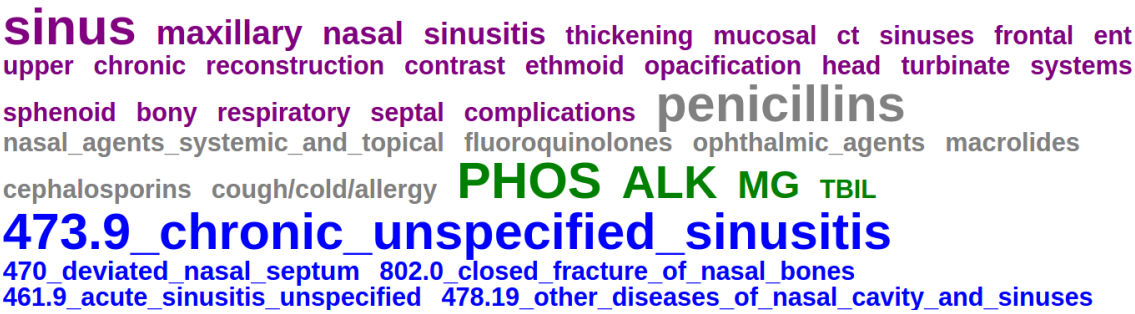
Overall, out of the 50 pairs of phenotypes assessed, the clinical expert considered comparison impossible for 9 pairs. The Phenome phenotype was superior to the LDA-all comparison in 80.4% of the remaining pairs and superior to LDA-text in 68.3% of the remaining pairs.

NYPH Dataset

For this outpatient dataset, we performed a pairwise comparison between grounded and ungrounded phenotypes, to assess the effect of grounding on the quality of the learned phenotypes. Overall, the JS divergence was not able to match the phenotypes as well for this task and the clinical expert found that 12 pairs were not matched well or represented uninterpretable phenotypes. Of the remaining pairs, the clinician found that 55% of the time, the ungrounded Phenome model was superior to the grounded Phenome model.



(a) Ungrounded Phenotype for Sinusitis.



(b) Grounded Phenotype for Sinusitis

Figure 5.8 Grounded and ungrounded Phenome model phenotypes.

Figure 5.8 displays two phenotypes paired for the clinical expert evaluation. Since the NYPH data is not de-identified, names sometimes appear in phenotypes (especially clinicians that specialize in particular diseases as they often sign the clinical notes). A name has been blurred in the Figure 5.8. The expert chose the grounded phenotype (b) over the ungrounded phenotype (a) as a better representation of sinusitis. In both of the phenotypes, the learned laboratory tests are not ideal, however in the grounded

phenotype displays a more relevant set of terms and medications. The grounding of the sinusitis phenotype (Figure 5.9), demonstrates how the augmented priors that were inserted into the GPhenome model, the figure illustrated how the grounding is able to guide the phenotypes to better disease models using clinical knowledge encoded in available knowledge resources. We note, however, that the informative priors specified by the grounding, apply equal weight to all of the words and medications displayed in Figure 5.9, but the phenotype (Figure 5.8b) has differing weights for the different medications and words; this confirms our hypothesis that the grounding can guide the modeling procedure but statistics of the input data are still well represented.

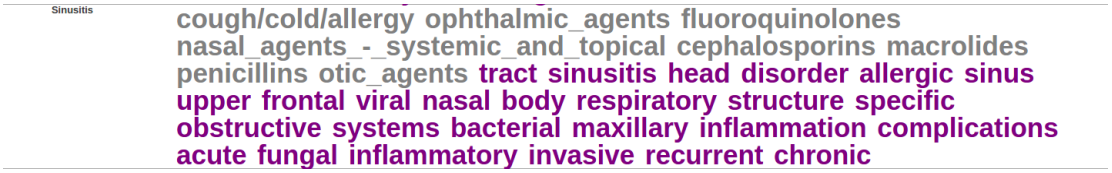


Figure 5.9 The observations that had augmented counts for the grounded Sinusitis phenotype.

5.5.5 Evaluation 4: Label Quality

MIMIC ICU Dataset

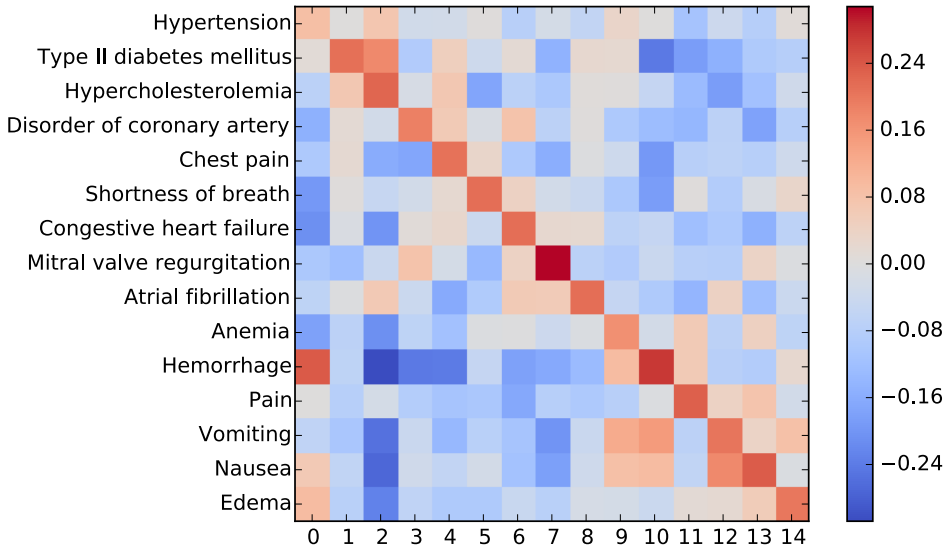
Overall, the most probable diagnosis code for a given phenotype was evaluated as good proxy for a phenotype label. In the MIMIC ICU dataset, the clinical expert assigned a label as actionable for 48% of the UPhenome phenotypes, 44% of them were considered related, while only 8% of them were considered unrelated.

NYPH Dataset

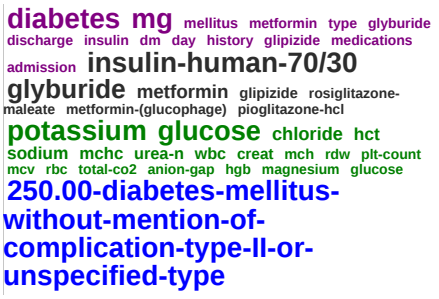
The clinical expert compared the UPhenome labels to the GPhenome labels. The expert found that the UPhenome model, 42% of the labels were actionable, 26% were related and 32% were unrelated. The GPhenome labels were found to be superior: 56% actionable, 28% related, and 14% unrelated.

5.5.6 Evaluation 5: Disorders to Phenotypes Comparison

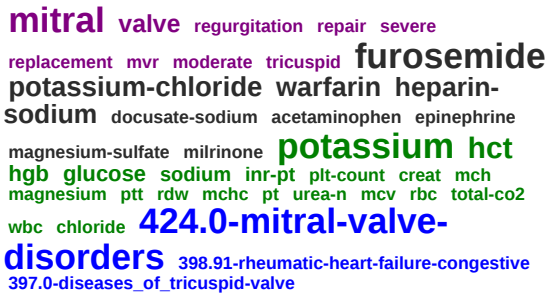
As the ground truth disorders were available for the notes from the MIMIC ICU dataset, this evaluation was performed only on the MIMIC dataset, not the NYPH outpatient data.



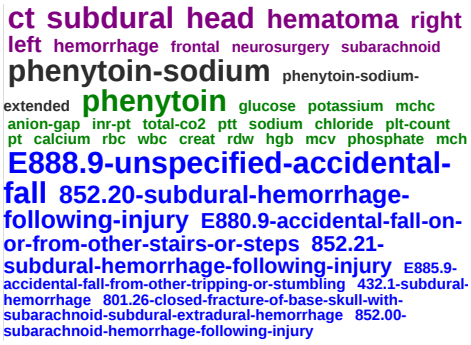
(a) Associations amongst ShARe annotations (y-axis) and phenotypes (x-axis). The x-axis is sorted to demonstrate the highest associations on the diagonal.



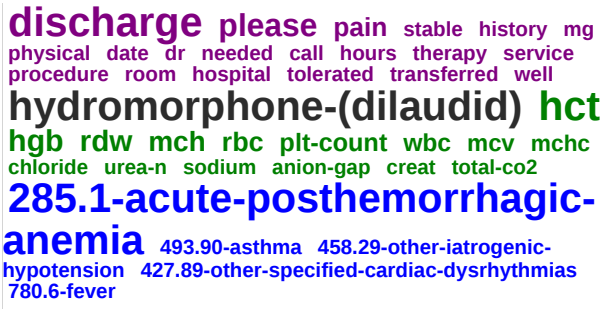
(b) Phenotype 1



(c) Phenotype 7



(d) Phenotype 10



(e) Phenotype 11

Figure 5.10 Association of manually identified ground-truth concepts and automatically inferred phenotypes over a set of patients, along with four example phenotypes.

Concept Unique Identifier (CUI)	Concept Description
C0020538	Hypertension
C0011860	Type II diabetes mellitus
C0020443	Hypercholesterolemia
C1956346	Disorder of coronary artery
C0008031	Chest pain
C0013404	Shortness of breath
C0018802	Congestive heart failure
C0026266	Mitral valve regurgitation
C0004238	Atrial fibrillation
C0002871	Anemia
C0019080	Hemorrhage
C0030193	Pain
C0042963	Vomiting
C0027497	Nausea
C0013604	Edema

Table 5.5 UMLS concept unique identifiers for the ShARe annotations in Figure 5.10.

Figure 5.10 shows the association matrix between the present disorders in the gold standard ShARe corpus and the inferred phenotypes on the gold-standard records. There were 15 disorder concepts that were present for at least 50 patients in the dataset. As such the matrix is 15x15. Borrowing from the visuals displayed by (Griffiths and Steyvers 2004), we sorted the matrix to clearly display the ShARe annotated disorders and their mostly highly associated inferred phenotypes on the diagonal. For clarity purposes, we grouped the 15 disorder concepts based on each other's clinical similarities. For instance,

hypertension, hypercholesterolemia, and type II diabetes are often seen together in patients, and similarly for the symptoms nausea and vomiting.

The figure indicates there is an association between the disorders present in the gold-standard records and their inferred phenotypes. Upon inspection of the inferred phenotypes, they are good representations of the present disorders. For instance, Phenotype 7 has the highest association with the disorder Mitral Valve Regurgitation, and its most probable observations are perfectly coherent with respect to this disease.

When concepts shared clinical characteristics (e.g., nausea and vomiting or the cluster of hypertension, diabetes, and hypercholesterolemia), the associated phenotypes are also shared amongst them (e.g., phenotypes 12,13 and phenotypes 1,2).

We display two more examples of phenotypes. Phenotype 10, which is associated with hemorrhage shows several diagnosis codes, all potentially leading or describing hemorrhages. Phenotype 11 is an unintepretable topic, containing highly prevalent observations throughout the MIMIC dataset.

5.5.7 Evaluation 6: Quantitative Metrics

The quantitative metrics for the MIMIC ICU dataset were used to inform model selection and as such we report only the quantitative calculations for the NYPH dataset in this section.

	Held-out likelihood	**NPMI**
UPhenome Model	-2.512×10^7	0.021
GPhenome Model	-2.543×10^7	0.031
LDA-All	-3.501×10^7	-0.062

Table 5.6 Quantitative evaluation for the unsupervised Phenome model, the grounded Phenome model, and the LDA-all model. All of the models were run on the same input set of Columbia patients, with 750 latent variables, 0.1 for the per-patient phenotype distribution hyperparameter and 0.001 for the per-phenotype data distribution hyperparameter.

The UPhenome model is able to outperform both the GPhenome and LDA-all model when predicting on a held-out test set (Table 5.6). As the GPhenome model is seeded with ontological knowledge that in some cases may not be perfectly aligned with the existing data, the model's ability to predict may be hindered

and can account for the lower held-out likelihood score. The very large benefit that the Phenome models have over the LDA-all models is due to the separation of data types. The LDA-all model is not inherently separating laboratory tests from medications from diagnosis codes from words, therefore the LDA-all model has a much more difficult task of predicting infrequent data types as they are competing with the overwhelming number of words that exist in the clinical notes.

In contrast to the MIMIC ICU dataset, the Phenome models outperform the LDA-all baseline with the NPMI metric.

5.6 Discussion

The results of the various automated and manual evaluations suggest that the Phenome model is a promising approach to discovering models of disease. We discuss next two characteristics of our model— the joint modeling of heterogeneous data types and their unsupervised modeling— as well as the differences between automated and manual coherence assessments for the task of phenotyping and the effects of grounding.

5.6.1 Joint modeling of heterogeneous EHR data

Previous work has shown that single data types are not sensitive enough to perform phenotyping at a high accuracy. Wei et al. demonstrated that when phenotyping ten chronic conditions, algorithms that only used single instances of billing codes as a phenotyping algorithm achieved an average F-score of 0.17, algorithms that used multiple billing code instances achieved an average F-score of 0.60. However, when multiple heterogeneous data types were leveraged (billing codes, clinical text, and medications), the algorithm was able to achieve an F-score of 0.70 on average (Wei et al. 2015). The Phenome model builds upon this knowledge as it is designed to jointly model multiple, different EHR data types.

The Phenome model leverages the innate heterogeneity and incompleteness of EHR data. By modeling each data type separately as opposed to a bag of observations like in the LDA-all baseline, the model can accommodate for imbalance of observations from each data type. For instance, there are many more words than diagnosis codes, even after stop words removal and vocabulary filtering. By design, the

Phenome model ensures that each data type is represented in the learned phenotypes, thus truly modeling across data types. This explains for instance, why all Phenome phenotypes given as examples in the chapter have a mix of data types, while most LDA-all phenotypes are overwhelmed by words and laboratory tests (i.e., the most common data types in our observations).

The separation of each data type in the Phenome model allows for specification of informative priors that are different for each data type. The ability to enhance the Phenome model with outside clinical knowledge, and even more specifically knowledge that is different for each data type is a powerful mechanism that will be explored further in the GPhenome model work. Although the model separates the different data categories, this separation does not render the data types independent. On the contrary, because the patient-specific phenotype distribution (β) is unobserved, the data types become dependent from the explaining-away effect. Therefore, the information learned about the billing code distribution across phenotypes is dependent on the medications distribution, the laboratory test distribution is dependent on the distribution of words, and so on.

5.6.2 Generative unsupervised modeling of EHR data

Like all unsupervised models, the Phenome model is exciting in its ability to discover patterns in input datasets, such as disease models. When applied to the MIMIC corpus, the learned phenotypes are representative of diseases that are documented in an intensive care unit, like acute kidney failure, while when applied to the NYPH dataset, the learned phenotypes are more representative of chronic and acute conditions that do not require intensive care, such as SLE.

Without careful modeling however, unsupervised models can yield unwanted results. The LDA-all phenotypes often highlighted information about hospital course in aspects nonspecific to any underlying condition, such as coagulation status, palliative care status, type of drug exposure, and plan for discharge. All of these topics make sense and represent distinct patterns in the input datasets, but they do not represent diseases. In contrast, the Phenome phenotypes represented a greater number of distinct disease

than the LDA-all phenotypes. Furthermore, they were more clinically relevant, as multiple aspects of a given disease were included such as secondary complications or potential treatments.

5.6.3 Automated coherence metrics vs. human judgments

When comparing the average coherence of the LDA-all phenotypes and the Phenome phenotypes for MIMIC ICU data, LDA-all yielded a significantly higher average Normalized Pointwise Mutual Information (NPMI) (.07 vs. .014). By contrast, in the NYPH dataset, the Phenome models had much higher NPMI scores than the LDA-all baseline (0.02 vs. -0.06). NPMI was established as a valuable automated evaluation metric of learned topics and was shown to correlate with human judgments of topic coherence (Lau, Newman, and Baldwin 2014). In our experiments however, we found little correlation between the clinician's judgments and the NPMI of the learned phenotypes (Pearson R=0.31 and Spearman=0.33 over the MIMIC ICU UPhenome phenotypes and Pearson R=0.34 and Spearman =0.31 in the NYPH UPhenome phenotypes). In our settings, the Phenome model is a mixture model over text but also coded data (e.g., diagnosis codes, medications, and laboratory results). It is possible that the computationally coherent (often co-occurring terms) are not actually clinically relevant. For instance, the LDA-all phenotype with the highest NPMI contained the following most probable observations: "pm total co pt potassium gap sodium urea chloride anion glucose creat hct hgb rbc mcv mchc mch wbc rdw", a mix of routine, nondiscriminatory laboratory tests. By contrast, the NYPH data deals with much more varied note types and patient diseases and may explain the low NPMI scores yielded by the LDA-all model.

It is also possible that different observations within coded data types may not occur frequently together. For instance, there are several diagnosis codes which are highly clinically relevant with each other, and yet do not get coded together in patient records: different stages of pancreatic cancer for example, would make sense in a single phenotype for the disease, but will not be seen jointly over many patients at a time.

5.6.4 Effects of Grounding the Phenome model

Augmenting the Phenome model with known associations that exist in clinical ontologies is able to provide more coherent and cohesive phenotypes. A phenotype may be grounded by any combination of words, diagnosis codes, laboratory tests, or medications. We found that often when a phenotype was grounded exclusively by words, the final learned phenotype often did not reflect the data it was grounded with. However, when a phenotype was grounded with diagnosis codes or medications or laboratory tests, the phenotype would resemble the prior grounding information much more strongly. We attribute these results to (i) the difference in observations counts and (ii) vague grounding of words. The differences in the observation counts means that prior pseudo count of 100 does not apply very strong grounding to words as there are many word instances but a prior pseudo count of 100 does make a large difference to a much smaller set of observations such as medications or laboratory tests. The vague grounding of the words is due to the chosen method of word grounding: taking all first-order relationships that are encoded in the knowledge base as linked with the disease and tokenizing the results. This method often results in comorbidities or vague terms such as “complications” and “chronic” being grounded to a phenotype.

Although this work presents a first pass at how to incorporate clinical ontologies into the Phenome model, future work in pseudo count refinement (perhaps having a different pseudo count augmentation for each data type) and newer methods for identifying links for the chosen ontologies (perhaps only chosing words that have particular ontological relationships) may result in phenotypes that are even closer to clinicians mental models of disease.

Chapter 6: Applications to Clinical Questions

6.1 Introduction

In an effort to apply some of the techniques outlined in this thesis to open questions in clinical informatics, we used two techniques on two different questions. High-throughput screening of laboratory test usage using measurement motifs was a helpful technique for identifying and mitigating biases that exist in irregularly sampled data (Chapter 3), and in this chapter, we apply the same idea and similar method to identifying laboratory test overutilization by examining measurement motifs.

The other clinical question tackled in this section is one of high-throughput phenotyping (Hripcsak and Albers 2013), a question of identifying cohorts of patients using EHR data. We employ the grounded version of Phenome model, which was developed to automatically learn computational models of disease (Chapter 5), to the task of automatically finding a set of type II diabetic patients from raw and heterogeneous EHR data.

6.2 Leveraging measurement motifs to study inappropriate use of laboratory tests[7]

6.2.1 Introduction

A recent report from the Institute of Medicine estimates that as much as 30% of healthcare costs in the United States are a result of unnecessary care. Finding ways to reduce unnecessary care can ease some of the healthcare cost burden without affecting the quality of patient care (Smith et al. 2013). One major contributor to excessive healthcare costs is the over-ordering of laboratory tests.

Laboratory test orders recorded in an institution's EHR can be analyzed to identify patterns of ordering across a large patient population, study adherence to existing ordering guidelines, and quantify potentially unnecessary care. This approach is especially attractive for high-volume tests, for which robust pattern analysis can be conducted and for which guidelines have been specifically constructed through detailed analysis of the latest research and expert panel discussions to maximize the test's utility.

One frequently ordered laboratory test with specific ordering guidelines is Glycated Hemoglobin A1c (HbA1c). HbA1c is the measure of average blood sugar control over 6-12 weeks. The healthy range of HbA1c is between 4-6% and diabetic patients have higher HbA1c values. Although diabetic classification as controlled and uncontrolled is usually measured with blood glucose measurements, it is commonly reported that the desired HbA1c level for a controlled diabetic is less than 7%. For uncontrolled diabetic patients, HbA1c levels often rise much higher.

Historically, HbA1c has been a standard test for the monitoring of diabetes: in 2002 the American Diabetes Association (ADA) established that uncontrolled diabetic patients should have their HbA1c measured every 3 months and controlled diabetic patients should have it measured every 6 months (Sacks

[7] This chapter was originally published in JAMIA. The full citation for this publication is:
Pivovarov R, Albers DJ, Hripcsak G, Sepulveda JL, Elhadad N. Temporal Trends of Hemoglobin A1C Testing. JAMIA. 2014;21:1038-44.

et al. 2002). Recently, new evidence suggests HbA1c can be used for the diagnosis of diabetes as well (Selvin et al. 2010; Handelsman et al. 2011). The 2009 ADA guidelines incorporated this finding and began recommending the use of HbA1c for the diagnosis of diabetes (International Expert Committee 2009; American Diabetes Association 2010). These guidelines state that if a patient has a HbA1c value of 6.5% or more for the first time, the patient should be retested (on a different day) to confirm the diabetes diagnosis; unless the patient exhibits clinical symptoms or has a blood glucose >= 200 mg/dl (American Diabetes Association 2013), then no retesting is necessary. The presence of guidelines (both in 2002 and 2009), along with the sharp distinction of how HbA1c should be ordered for monitoring and for diagnosis, both provide a point of comparison when analyzing patterns of HbA1c ordering.

Despite these widely publicized guidelines for diabetes care, there are numerous reports of over-ordering of HbA1c labs. In a study focusing on newly diagnosed diabetes patients, HbA1c orders were analyzed over a period of two years (Laxmisan, Vaughan-Sarrazin, and Cram 2011). It was found that 8.4% of patients (N = 11,003) received at least one repeat HbA1c within 30 days of their initial test and 30.8% (N = 40,162) within 90 days. A more recent 10-year retrospective analysis at a UK university hospital found that 21% of 519,664 HbA1c orders were ordered too soon (as defined by sooner than 6 months for patients with < 7% HbA1c and less than 2 months for patients with 7% or over) (Driskell et al. 2012).

Striking differences have been shown in the frequency of HbA1c orders across different healthcare settings. In a study at a Turkish university hospital, 10.3% of all 10,496 HbA1c orders over a two year study period were performed within less than a month of one another and when looking only at inpatient orders, 33.8% were found to be ordered within less than a month (Akan et al. 2007). Other studies have also found inappropriate repeat testing more frequent in hospitalized patients (van Walraven 2003; Salvagno et al. 2007).

In this study, we focus on the overall temporal trends of HbA1c ordering across a 15-year span. Thus, we explore the ordering patterns across both inpatient and outpatient data points, as there might be an impact on ordering patterns for patients transitioning between outpatient and inpatient settings.

In this work, we analyze HbA1c laboratory test order data over a 15-year longitudinal time scale that covers the release of two separate ADA guidelines; thereby giving us an opportunity to retrospectively study the influence of both guidelines over time. Guidelines for HbA1c provide instruction on when to measure HbA1c for different types of patients. Specifically, we study how well both the diagnosis guidelines and the monitoring guidelines are being followed at our institution.

Because of the large patient population with diabetes, HbA1c is a high-volume test and is one of the most frequently ordered tests in our institution. We comprehensively examine all of the HbA1c measurements in our clinical data warehouse; regardless of who is coded for having diabetes to generate a more complete view of the HbA1c measurement trends. We are also in position to link clinical notes to ordering patterns in order to qualitatively assess the reasons behind the ordering patterns we observe.

We ask the following research questions: (i) What are the patterns of ordering HbA1c in a large patient population? (ii) Do HbA1c orders follow guidelines with respect to frequency of measurement? and (iii) If patterns of ordering do not follow guidelines, in which ways do they depart from the guidelines and what are potential explanations for the departure?

6.2.2 Methods for Identifying HbA1c Temporal Trends

We collected all HbA1c measurements in the NYPH data between January 1996 and December 2010. All measurements were included in the dataset, i.e., there were no selection criteria for the patients. Each data point consisted of a tuple (patient identifier, timestamp of individual measurement, and corresponding value).

Patterns of HbA1c Ordering through Time

To capture total yearly ordering patterns, we looked at all HbA1c orders performed in our institution. Across all 15 years, we report counts of HbA1c orders and stratify the counts by the numerical HbA1c values.

HbA1c Ordering Patterns Pre vs. Post 2002 Guidelines

The 2002 ADA guidelines established that patients with controlled diabetes should be monitored every six months and patients with uncontrolled diabetes should be monitored every three months. To verify the extent to which these guidelines had an impact on practice, we aggregated our institution's HbA1c measurement data into two subsets: pre-guideline measurements (1996-2001) and post-guideline measurements (2003-2010).

For each year, we include only tests that are conducted within the same calendar year; therefore, tests that are conducted in December and repeated in January of the next year are not included in our results.

To visualize repeat ordering patterns, we aggregated data in the following way. For each patient's HbA1c time series we calculated the days between two consecutive measurements within a year and aggregated across all patients. A histogram was created for each year mapping the gap between consecutive measurements and number of such measurement pairs across the dataset that year. For instance, if a patient had two consecutive measurements 132 days apart and both measurements occurred in 2007, this would contribute twice to the 132-day gap for 2007.

To test whether there was a change between the pre- and post-guideline ordering patterns, we performed a two-sample Kolmogorov-Smirnov (K-S) test (Massey 1951). The K-S test measures whether the pre- and post-guideline samples come from the same distribution. To quantify the specific differences in pre- and post-guideline periods, we performed an L_1 distance calculation separately on the 0-90 day and the 91-365 day sections of the distribution. For the L_1 distance calculation, we transformed the discrete measurement counts for the pre- and post-guideline periods to probability density functions (PDFs) representing the probability of every measurement gap across the time period. We quantified the difference between the pre- and post guideline PDFs using the absolute difference between the distributions (i.e., the L_1 distance) defined by:

$$d(a,b) = \int_a^b |f_1(x) - f_2(x)|dx$$

Here, f_1 and f_2 are the probability density functions and a, b represent two time points.

Additionally, to visualize measurement frequency differences between controlled and uncontrolled diabetic patients we created a density plot of the joint probability of HbA1c value and time to next measurement. This comparison is important as ADA guidelines are defined based on controlled and uncontrolled diabetic patients, that correspond to glucose and HbA1c levels. We hypothesize that high HbA1c values alert the physicians to an uncontrolled diabetic thereby influencing the patient's time to next measurement (likely closer to 3 months). Alternatively, a well-controlled diabetic patient would likely not be tested for another 6 months.

Finally, to enable comparison to prior work that focuses on over-testing of well-controlled diabetics, we used HbA1c <7.0% as a proxy measure for well-controlled diabetes and calculated how many of the total HbA1c orders each year, are unnecessary (repeated within 180 days for a patient with HbA1c < 7.0%).

Diagnostic Use of HbA1c Ordering

Considering the 2009 ADA's recommendation for diagnostic use of HbA1c (when a patient has an HbA1c of >=6.5% for the first time, it is then encouraged to have the patient's HbA1c retested on another day to confirm a diabetes diagnosis), we tracked the values of the measurements across gaps. Our hypothesis is that rapid re-measurement of HbA1c (within 10 days) is due to the occurrence of a first high value. To test this hypothesis we looked at the proportion of rapidly retested HbA1c's that follow the diagnostic guidelines. Finally, we qualitatively assessed reasons that guideline deviations may occur by reading a sampling of clinical notes of patients who had HbA1c tests repeated within ten days.

6.2.3 Results of Temporal Analysis of HbA1c Measurement

Overall, our dataset consisted of 397,926 HbA1c orders, measured for 119,691 unique patients across 15 years. The maximum number of orders per patient was 150, and the average was 3.32, with a large standard deviation of 5.83 orders. The high variance we observed is due to the various characteristics of our EHR data. As we are not filtering our population for only diabetic patients we have

large differences between regularly monitored patients and those who were tested once for screening purposes. In addition, we have a sparse dataset of patients some of whom may not be regularly followed up at the outpatient clinics and others who receive their care in other institutions.

Patterns of HbA1c Ordering

Over the 15-year period from 1996-2010, there was an increase in the raw number of HbA1c tests at NYPH (Figure 6.1). This increase follows a general increase in measured patients (6,232 patients who had their HbA1c measured in 1996 to 31,765 patients in 2010). Over the 15-year period, the rates of HbA1c testing have remained fairly steady, between 2.09-2.7 tests per patient per year, with a very slight upward trend. The increase of tests with <6.5% values is consistent with the use of HbA1c for screening purposes, as most screened patients will be normal.

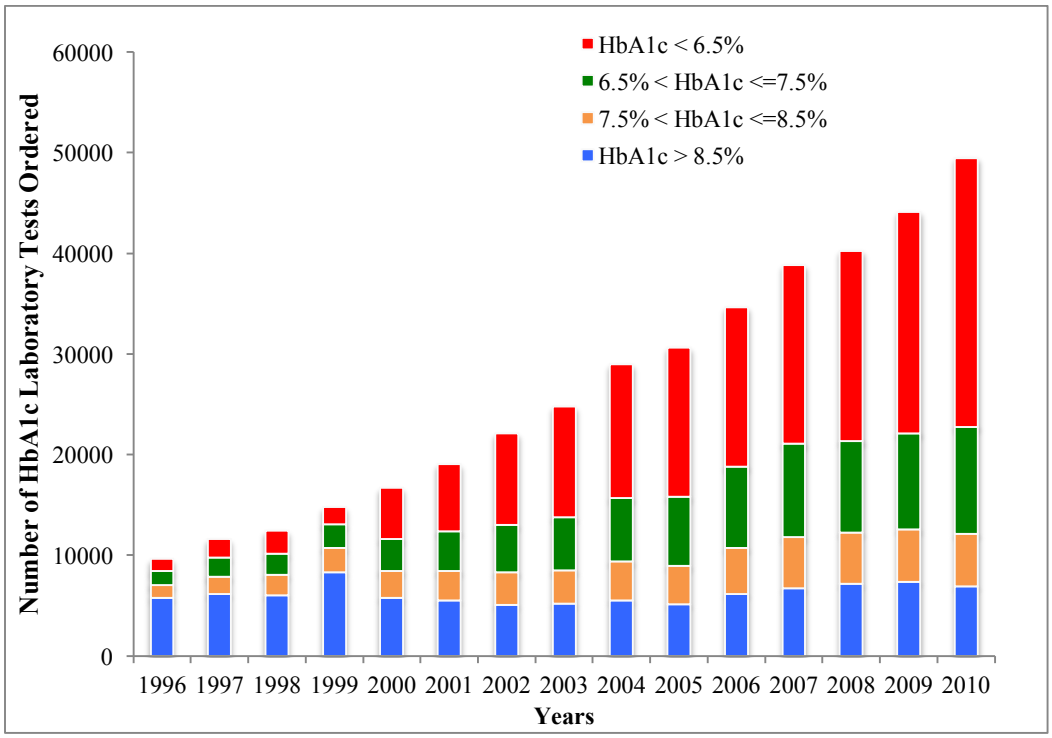


Figure 6.1 Counts of all HbA1c orders over the years 1996-2010, stratified by HbA1c numerical value. This figure includes all patients with at least one HbA1c measurement.

To adjust for diabetes screening, we report statistics for patients who have had at least two HbA1c measurements within their record. There is still a steady increase of measured patients (4,434 in 1996 to 19,302 in 2010) and a steady increase from 2.3 tests per patient per year in 1996 to 3.09 in 2010.

Pre and Post 2002 Guideline Ordering Patterns

The two-sample K-S test showed a statistically significant difference between the measurement gap counts in the pre and post guideline time periods ($p<0.001$). By estimating the probability density functions of the measurement gap counts, we found the measurement gap distribution changed from a fairly unimodal to a bimodal distribution (Figure 6.2). The modality shift shows that a mostly homogenous dataset of measurement gaps transformed into a heterogeneous dataset. We observed that that the probability density functions of the pre and post guidelines measurement gaps were highly similar after 3 months, both sharply peaking at the 3-month time frame and the post-guideline curve having a slight peak at 180 days.

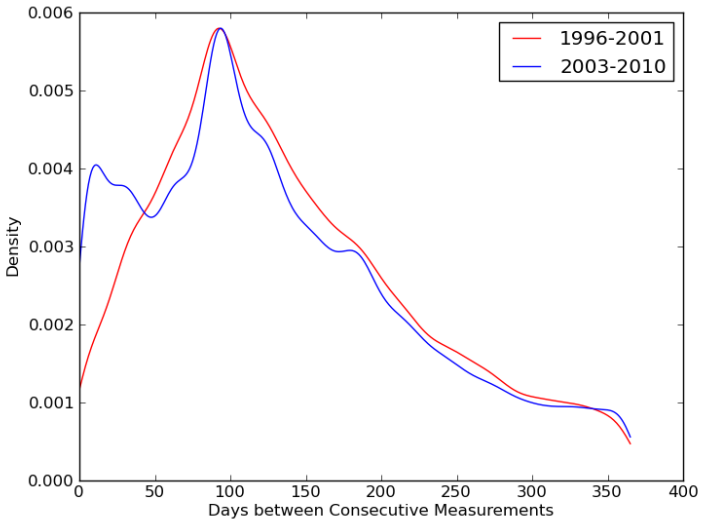


Figure 6.2 Probability density function estimated using a kernel density estimate on the aggregated gaps between HbA1c measurements for both the pre-guideline period (1996-2001) and the post-guideline period (2003-2010).

Noting that the two distributions look very different in the 0-90 days time frame, we performed a separate L_1 distance calculation for two parts of the distributions (0-90 day gap, 91-365 day gap) to

quantify the differences in the distributions. We were able to detect an order of magnitude difference in the L_1 distances: L_1(0-90 days) = .0012 and L_1(91-365 days) = 0.00029.

We observed a sharp transition in 2002 in how patients are measured on short time scales of less than 90 days. This implies that how patients' HbA1c values are measured on time scales of longer than three months has not changed in 1.5 decades whereas there was a dramatic change in how patients' HbA1c values are measured on time scales of less than three months starting in approximately 2002.

In addition to measurement gaps, we evaluated how values correlated to measurement gap in both pre and post 2002 guideline periods. Figure 6.3 is a density plot representation of the joint probability of each HbA1c level and the time to next measurement for both time periods.

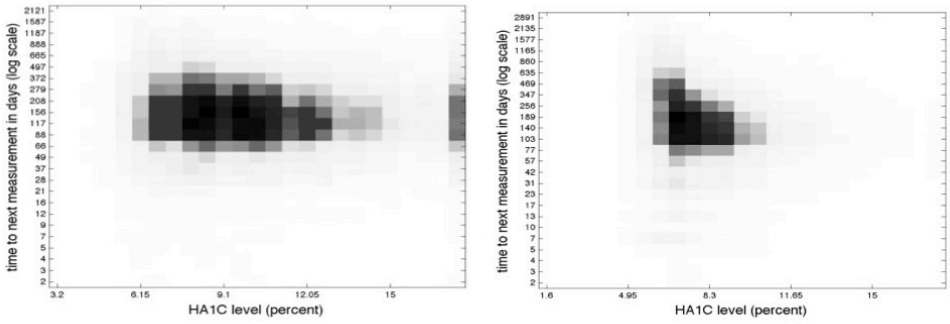


Figure 6.3 Joint probability between each HbA1c percentage and time to next measurement before the 2002 guidelines (left) and after the 2002 guidelines (right).

The density plots indicate a change in ordering after the 2002 guidelines as well. Before the guidelines were released, there is almost no correlation between HbA1c value and time to next measurement; most of the population is measured between 60 and 500 days irrespective of the patient's value. After the 2002 guidelines, we see a much more prominent correlation between HbA1c value and time to next measurement as shown by the L-shaped curve; having a higher HbA1c value prompts quicker retesting (around 100 days) while lower HbA1c values have higher probabilities of being measured at longer time scales. If one's diabetes is more controlled (towards the left of the x-axis), there is a longer time to next measurement.

However, when looking specifically at unnecessary repeat tests for well-controlled diabetics (repeats within 180 days for patients with HbA1c < 7%), we still find that rates of inappropriate use increase over time. Between 1996-2000, 3.8-6% of the HbA1c tests were repeated inappropriately. After 2000, inappropriate testing rose to 11.8-19.5%, growing to over 20% in 2004 and remaining stable between 19-20% until 2010.

Diagnostic Use of HbA1c

After 2002, we found growing numbers of HbA1c tests repeated within 10 days. Orders repeated within 10 days accounted for 1-2% of all HbA1c orders in 1996-2001, and grew to 2.8-3.8% after 2001. Therefore, we examined the dataset in multiple ways to investigate whether the repeats were for justified diagnostic purposes or whether overutilization may be occurring. The 2009 guidelines for diabetes diagnosis allow for HbA1c rapid retesting if the patient exhibits a HbA1c value of >=6.5% for the first time.

Looking across all 15 years, we tested whether the patients with rapid retests meet the criteria for a diagnostic measure. Surprisingly, in 1996, 37% of retests were justified, whereas only 13% in 2010 were justified according to the 2009 guidelines (Figure 6.4).

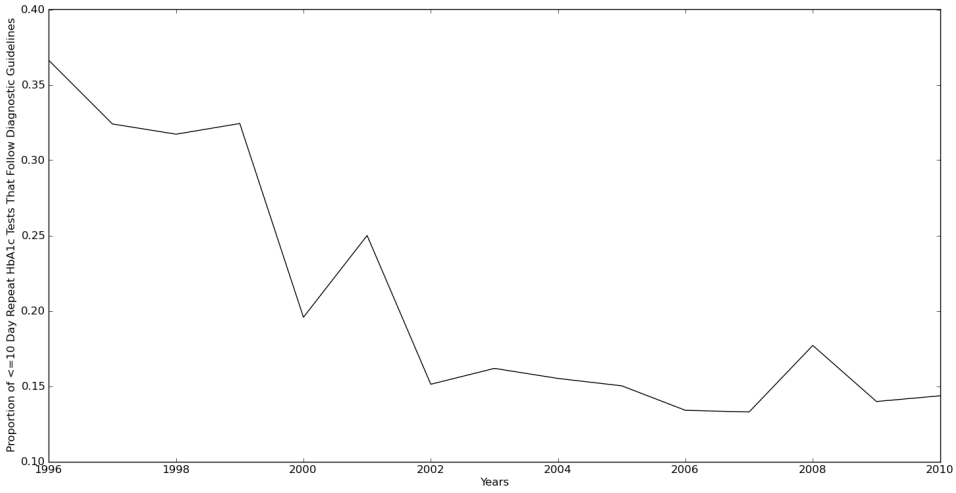


Figure 6.4 Proportion of HbA1c measurements taken within 10 days that follow the appropriate guidelines for diagnostic use. This graph only examines HbA1c repeated orders that are within 10 days. Each point was calculated per year as: (# of ≤10-day repeats where the initial test result was a patient's first HbA1c of ≥6.5%)/(all ≤10-day repeats).

We find that the rapidly repeated tests cannot be explained by adherence to guidelines. In fact, over time, a larger portion of rapid repeats is conducted on HbA1c tests with lower values even as there is no justification for any <=10 day repeats when a value is <6.5%. In particular, beginning around the year 2000, between 40-50% of the rapidly repeated tests were ordered for the lowest of HbA1c values (Figure 6.5).

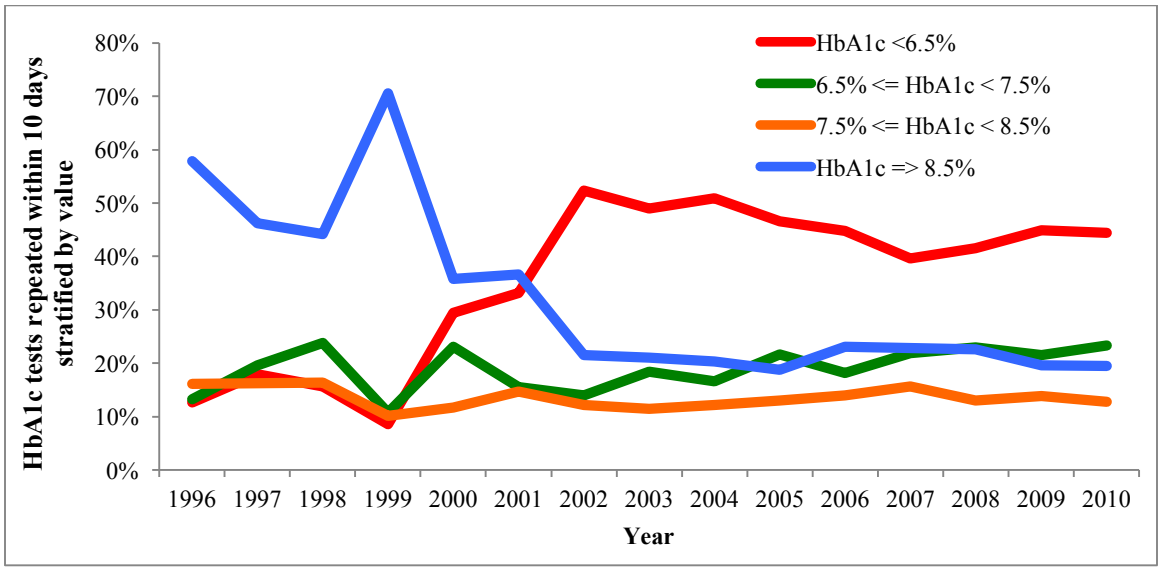


Figure 6.5 Numerical stratification of HbA1c tests reordered within 10 days over the years 1996-2010.

The overall rate of rapid retesting has grown similarly across all HbA1c values (Figure 6.6), but as the number of initial <6.5% orders has increased over time so has the number of inappropriately retested HbA1c tests. From 1996-2001, we find that 1-2% of HbA1c tests with an initial <6.5% value are rapidly retested and after 2001, 2-3.6% are rapidly retested.

As a means to better understand the rapidly retested HbA1c phenomenon, we conducted a manual chart review for a subset of 100 randomly chosen patients that had HbA1c rapid repeats. A common pattern we identified was outpatients receive a HbA1c test, are admitted to the hospital a few days later and have a repeat HbA1c upon admission to the floor. This could be indicative of care-coordination issues where inpatient physicians are unaware of their outpatient counterparts ordering the HbA1c test only a few days prior (even though the laboratory results are visible in the EHR). Multiple times, consecutive outpatient visits within 10 days resulted in HbA1c being tested during several of the visits. This could also point to care-coordination issue amongst clinicians with different specialties (even though, once again, the laboratory results are visible in the EHR). Such care-coordination issues could lead to providers ordering a second "initial" HbA1c test, without realizing there already exists an initial test result.

Another pattern we identified was a retest of HbA1c because the value was much lower than expected; specifically, a HbA1c result seemed too low given a patient's previous history and therefore the same physician re-ordered the test; sometimes the clinical notes even referred to the surprisingly low first value. Finally, we found that between 5-10% of the rapid repeat tests were a result of physicians conducting point-of-care HbA1c testing and ordering a confirmatory laboratory-run test HbA1c test.

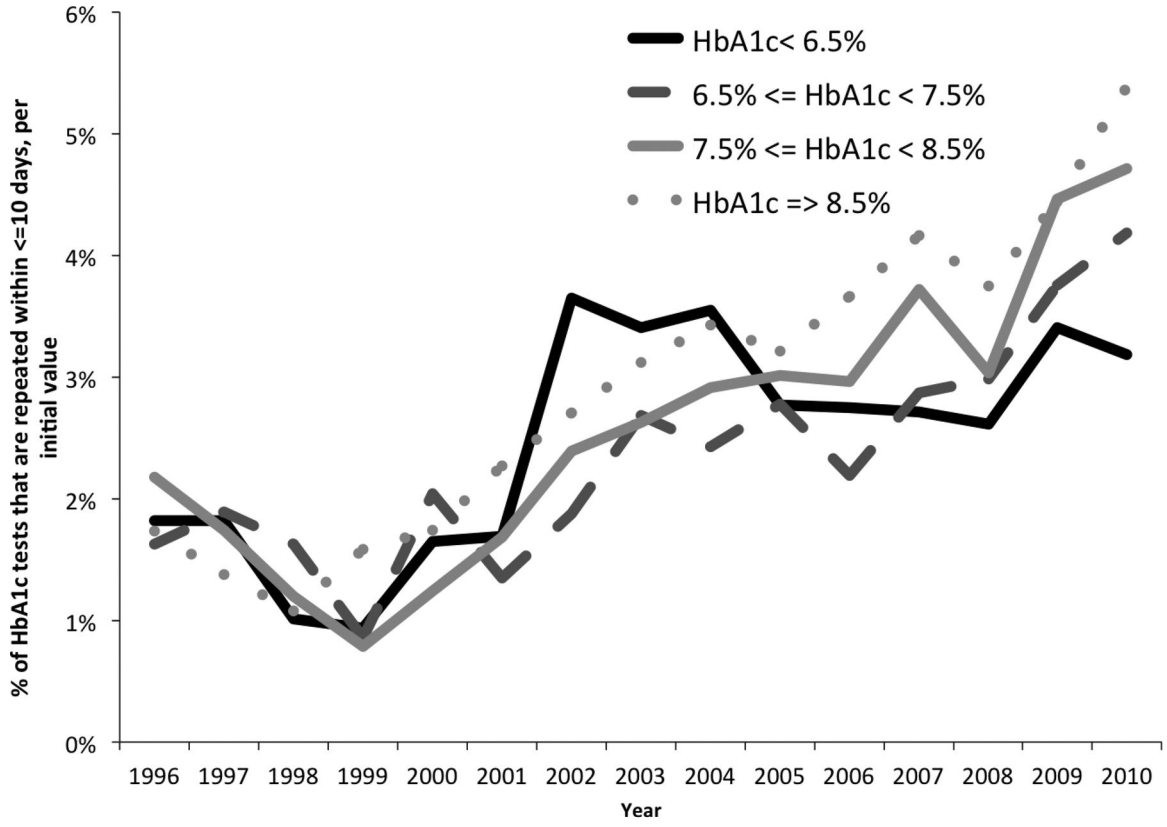


Figure 6.6 Percentage of HbA1c tests that are reordered within 10 days over the years 1996–2010, stratified by numerical value.

6.2.4 Discussion

Distributional analysis over time of consecutive HbA1c measurements frequencies has uncovered a number of phenomena. Overall, the raw number of HbA1c tests has increased over time. This can be explained in part by the natural increasing number of patients coming to our institution for care and in part by the increasing number of patients with diabetes. In addition to the general increase in testing, we

also see evidence for increase of inappropriate use from the growing number of very short gap measurements (between 0-10 days) over the 15-year period.

We note that some of the measurement patterns established by the 2002 and 2009 guidelines predate their releases, although there were publications that hinted at diagnostic and monitoring recommendations before the guidelines as well (Goldstein et al. 2004; Singer et al. 1989; Peters et al. 1996). It seems that physicians have been using HbA1c as a diagnostic measure for at least 6 years before the official ADA guidelines which recommended HbA1c's usage for diagnostic purposes were released; we can see this by the prominent peak of short gap measurements that begins to appear around 2003. Additionally, we are able to see that during the mid 1990s, before the 2002 guidelines specifying 3 and 6-month measurement intervals were released, there were already slight peaks at those two measurements gaps.

Overuse of HbA1c Measurements: Retesting Within 10 Days

While clinicians roughly follow the 3-month and the 6-month guidelines, we find that there is a strong signal for seemingly unnecessary repeated measurements within 10 days. The findings from this study are well aligned with recent literature about HbA1c measurement dynamics and the overutilization of HbA1c over short periods of time. For instance, Lyon et al (Lyon et al. 2009) also uncovered a highly prevalent short-gap peak, but their peak was at approximately 30 days, not as short as the time gap in our dataset.

By comparing ordering distributions in our dataset we find the trend towards repeated testing over very short time periods (<= 10 days) is increasing in volume and is the most frequent in 2010. In conjunction with a general increase in tests with values HbA1c <6.5%, we uncover a troubling trend: a growing proportion of rapid retests are conducted on tests with values <6.5%, despite ADA guidelines only recommending rapid retests on a subset of >=6.5% tests. Moreover, the signal stays salient, even after controlling for the usage of HbA1c as a diagnostic tool through two tests a few days part.

Based on the results illustrated in Figure 6.4, we postulate that between 65-85% of the HbA1c repeat measurements occurring within 0-10 days are unnecessarily ordered – a total of 9,491 tests from

1996-2010. The 9,491 tests represent HbA1c repeat test overutilization rate of 2.3% across the entire 15-year period. We use the Medicare reimbursement of $13.24 per test to estimate an unnecessary expenditure of approximately $125,600 at just one hospital in New York City. This figure does not account for personnel, laboratory time, ordering or interpretation time, any patient care costs, or over ordered laboratory tests as a result of transferred patients and a lack of health information exchange, which has also been shown to contribute to laboratory test overutilization (Stewart et al. 2010).

There are many potential explanations for the increasing rate of overuse of HbA1c tests. The addition of HbA1c to preset laboratory order panels may lead to the retesting, as often it is more efficient for the clinician to order an entire panel rather than remembering to exclude the HbA1c test. Another reason for this trend could be a consequence of guideline-induced over-vigilance; providers are intent on following the monitoring HbA1c guidelines but do not always remember to check whether an HbA1c result has been recently recorded.

In this chapter, we provided a methodology for studying utilization patterns that serves as a useful diagnostic approach for identifying trends of inappropriate laboratory test use over time. We analyzed the laboratory measurements of HbA1c for all the patients in our institution, over 15 years. Our study replicates prior work on HbA1c overutilization and offers new insight into the trends of laboratory ordering over time, in particular pre- and post-guidelines for diabetes monitoring, correlations between HbA1c value and testing frequency, and the use of HbA1c as a diabetes diagnostic tool. With the number of diabetes patients expected to continue to grow (Huang et al. 2009), it is essential to identify the ways in which HbA1c ordering is misused. Our study contributes to this effort.

Impact on Informatics Research

As the number of EHRs across the country increases, there is a growing potential for pertinent and effective IT interventions to help optimize healthcare resources, as well as to ensure that that clinicians adhere more closely to national guidelines for testing (Baron and Dighe 2014). For instance, identifying rapid retests of HbA1c as a strong pattern in an institution, the EHR could now implement a module that denies rapid retest of HbA1c without appropriate reasoning from the ordering physician, and perhaps

require a phone call to the laboratory to verify the need for a second test. Alternatively, there could be a systematic review and removal of HbA1c from laboratory test order sets.

Our work on trends in measurement gaps is relevant to large-scale data analysis work in informatics as well. To properly study large patient populations (in our case, across the entire EHR), it is helpful to stratify or decompose the population into homogenous data points (Pivovarov et al. 2014). In the pre-2002 guideline period, the distribution of HbA1c data points by gap is mostly unimodal, but the post-guideline period shows that the distribution changes towards a bi-modal distribution. Thus, the population of laboratory measurements is decomposable according to different reasons, some health-related (measurements 3 months apart for uncontrolled diabetic patients), but some for reasons that might not be health-related (measurements less than 10 days apart). The ability to quantify and recognize decomposable and not decomposable distributions of data points when performing large-scale research on the EHR is critical to ensure precise and robust inference.

6.3 Leveraging the Grounded Phenome model for cohort identification

6.3.1 Introduction

Performing cohort selection, or phenotyping, for clinical research studies is an important area of informatics research. The ability to accurately gather cohorts is necessary to perform studies on the exploding amount of electronic patient record data; studies on pharmacovigilance, comparative effectiveness, disease progression, and many others often rely on robust ways to measure cohorts. Recently, a consortium of informatics researchers (the eMERGE network) began creating repository of algorithms for selecting EHR patients with certain diseases. (https://phekb.org/). The algorithms use rule sets based on recorded laboratory tests, medications, billing codes, etc. to identify a set of case patients and control patients. The algorithms have been shown to have high specificity and positive predictive value and have been replicated across a handful of institutions. However, these algorithms have a few

drawbacks: (i) the phenotype definitions are very labor-intensive, include iterative design processes, and require the insight of clinical experts; (ii) as the process is so involved, there is very limited coverage of diseases that have an eMERGE algorithm created; (iii) the diseases are treated as binary events, a patient has the disease or not – there is no severity scale or ranking or probability of having the disease.

The Phenome model (presented in Chapter 5), presents an alternative to rule-based phenotyping. As the Phenome model learns phenotypes directly from the structure of the data (with the GPhenome model incorporating encoded clinical knowledge to provide prior information that can guide the phenotypes to be more clinically relevant), the model does not require heavy expert involvement. The model is able to learn the composition of many diseases at once, and because it was constructed in a probabilistic graphical modeling framework, it represents diseases states as probabilities: each patient has a certain probability of being ill with each disease. Because of the benefits outlined above, the Phenome model is an attractive alternative, leveraging unsupervised computational techniques for creating initial phenotypes, which can then be further refined by experts.

In this study, we assess the ability of the GPhenome model to gather patient cohorts and directly compare the GPhenome model method with the eMERGE algorithm. We focus on identifying a cohort of Type 2 Diabetes Mellitus (T2DM) patients as it is a highly prevalent disease in our population and there exists a validated eMERGE phenotyping algorithm for it.

6.3.2 Methods for Identifying Type II Diabetics from EHR data

The grounding of the Phenome model creates nonexchangeable phenotypes by augmenting counts for certain observations and for certain phenotypes. These observation-phenotype links that inform the augmented counts are automatically derived from clinical ontologies. In turn, these augmented counts form informative priors and encourage the formation of more cohesive and representative phenotypes. For this phenotyping study, we chose to use the grounded version of the Phenome model (GPhenome).

The specific observations that were automatically identified from the clinical ontologies as related to T2DM are displayed in Figure 6.7. These observations are weighted more heavily in the GPhenome model prior for the T2DM phenotype.

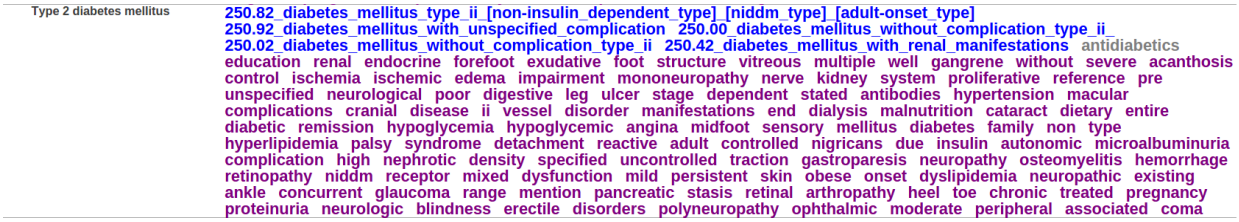
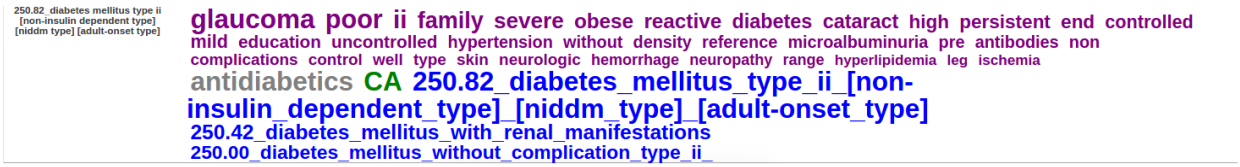
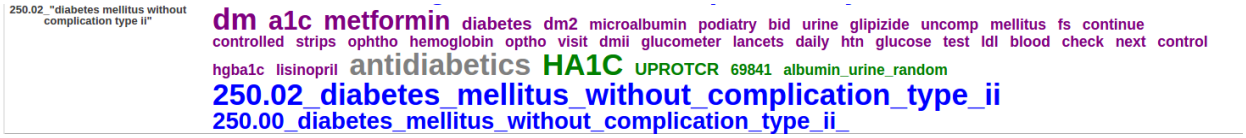


Figure 6.7 The diabetes prior that was input to ground the Phenome model for T2DM. Billing codes are in blue, words are in purple, and medications are in grey.

The GPhenome model was applied to the outpatient NYPH dataset described in section 5.4.1; each patient is represented by a recent outpatient time slice of their record. GPhenome was learned on the training set of 9,828 patients and tested on held-out test set of 2,457 patients. The GPhenome model was run with the same parameters as described in section 5.4.2 (P=750, α =0.1, [μ, ν, ξ, π]=0.001). After the 750 phenotypes were learned, a clinical expert reviewed the phenotypes and identified two that represent T2DM (Figure 6.8).



(a) The T2DM phenotype learned around the grounded priors for T2DM



(b) The additional T2DM phenotype learned by the GPhenome model.

Figure 6.8 The two T2DM phenotypes used for identifying T2DM case patients.

Using the GPhenome model inference mechanism, we inferred a distribution over phenotypes for each patient in the test set. As the GPhenome model assigns a weight to each phenotype for each patient, we

were able to rank the patients identified by the weight of their T2DM phenotype. The T2DM weight of each patient was calculated as the sum of their weights for both of the T2DM phenotypes (Figure 6.8). This ranking provides a way to list patients by how prevalently the T2DM diagnosis is discussed in their record with respect to other diseases.

Study 1: Precision-Recall of the GPhenome model

Evaluation of the GPhenome model's ability to identify patients who are Type II diabetics included the calculation of a precision-recall curve for a random 10% subset of the the test patients (246 patients). Gold-standard labels were assigned based on a chart review. A clinician was given the patient MRN, the patient record-slice date range, and asked to answer the question: "If all you have is the information contained in the EHR within this date range, do you think this patient has T2DM?"

Study 2: Direct comparison of GPhenome and eMERGE

In order to directly compare the GPhenome model with a well-accepted algorithm for phenotyping, the eMERGE algorithm for T2DM was implemented. The eMERGE T2DM case algorithm presents five ways in which a patient can be classified as a T2DM patient; these five ways use four data types: medications, ICD-9 codes, laboratory tests, and problems lists. For example, one way that a patient can be classified as a T2DM case is by having no Type I diabetes mellitus (T1DM) diagnosis codes, at least one T2DM medication and at least one abnormal glucose or HbA1c laboratory test result. The full algorithm flowchart can be found at: https://phenotype.mc.vanderbilt.edu/phenotype/type-2-diabetes-mellitus. The algorithm applies a set of strict criteria aiming to ensure that well-documented T2DM patients are captured and no T1DM patients are captured; the algorithm stipulates that if a patient has any T1DM diagnosis codes recorded, they cannot be considered at T2DM patient. Although the algorithm generally performs with a 98-100% positive predictive value, it has been demonstrated that the strict criteria for T2DM case definition may actually omit many true cases resulting in low sensitivity of 32% (Fort, Wilcox, and Weng 2014). To ensure proper comparison of the GPhenome model and eMERGE, both algorithms are applied on the held-out test set of 2,457 patients (the GPhenome model is learned on the training set of 9,828 patients).

To compare GPhenome to the eMERGE model (which outputs a cohort of patients), a probability threshold must be chosen. The GPhenome model assigns each patient a non-zero weight for every phenotype, therefore, a threshold was chosen to separate patients without T2DM from those with T2DM, we examined the distribution of T2DM weights visually (Figure 6.9) and chose a threshold of 0.01, excluding the patients at the head of the distribution.

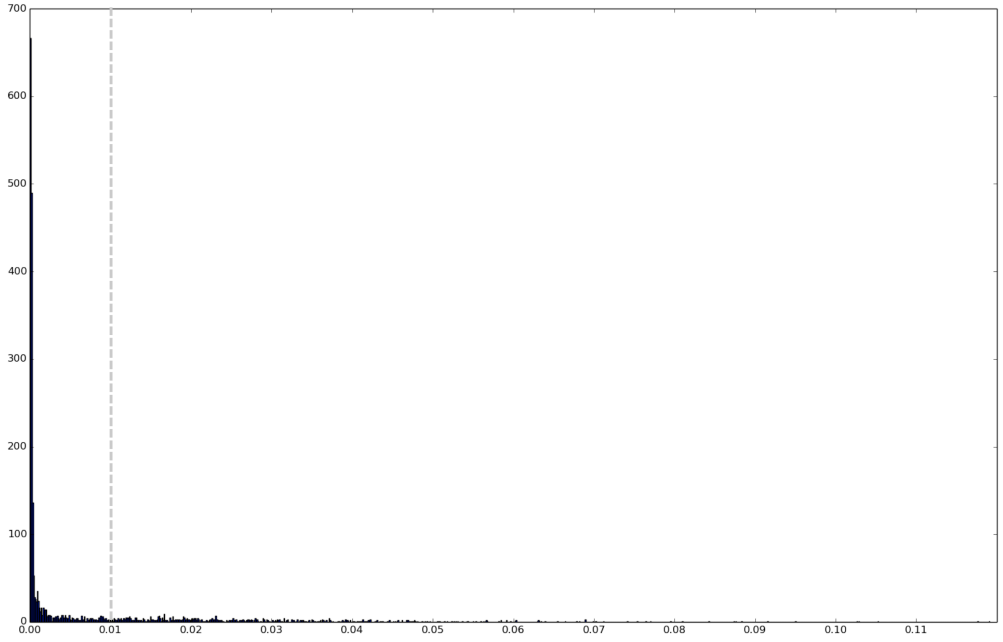
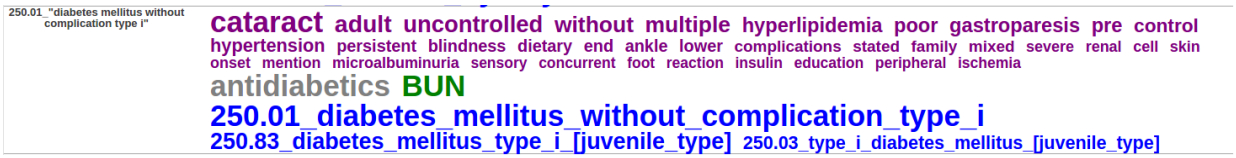


Figure 6.9 Distribution of T2DM phenotype weights. The dotted line represents the chosen threshold of 0.01, those patients with a weight at or over the threshold were chosen as T2DM case patients.

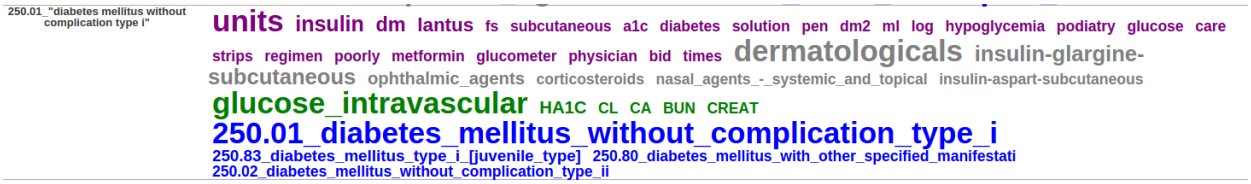
Additional constraints were applied to GPhenome in order to match the eMERGE algorithm's removal of T1DM patients (to ensure that the T2DM cohort was not diluted by type I diabetics). The final selection criteria for a T2DM case patient was:

Combined weight ≥ 0.01 for the T2DM phenotypes and

Combined weight for the T1DM phenotypes (Figure 6.10) < combined weight for the T2DM phenotypes.



(a) The T1DM phenotype learned around the grounded priors T1DM.



(a) The additional T1DM phenotype learned by the GPhenome model.

Figure 6.10 The two T1DM phenotypes used for ruling out case patients that may have T1DM instead of T2DM.

The evaluation of the patient cohorts selected by the two different models included calculation of sensitivity, specificity, positive predictive value, and F-measure. To estimate these metrics, a clinician conducted chart reviews of 50 random patients selected as cases by eMERGE, 50 random patients selected by GPhenome, and 50 patients that neither model selected. The clinician was given the patient MRN, the patient record-slice date range, and asked to answer the question: "If all you have is the information contained in the EHR within this date range, do you think this patient has T2DM?" The MRN list was randomly shuffled and the clinician was not told which algorithm selected which patient.

As the GPhenome model assigns a weight to each phenotype for each patient, we are able to rank the patients identified by the weight of their T2DM phenotype. This ranking provides a way to list patients by how prevalently the T2DM diagnosis is discussed in their record in contrast to other diseases. We evaluated the GPhenome model ranking by using the precision at K metric, where K=50. To calculate the precision at K, the clinician was asked to also to do a chart review on the ranked top 50 patients (again, with no knowledge that these were derived from GPhenome).

Finally, to assess the threshold of 0.01 as a phenotype weight for who a true T2DM patient is, a precision-recall curve was created on a subset of the 2457 test patients. A randomly selected 10% subset of all test patients was labeled as having T2DM or not having T2DM with the same question as before, "If

all you have is the information contained in the EHR within this date range, do you think this patient has T2DM?".

6.3.3 Results of the Type II Diabetes Cohort Identification

Study 1: Precision-Recall of the GPhenome model

The 10% subset of test patients were ranked by their combined probabilities for the two T2DM phenotypes. Using the chart review results, a precision-recall curve was created (Figure 6.11); the associated area under the curve was 0.87.

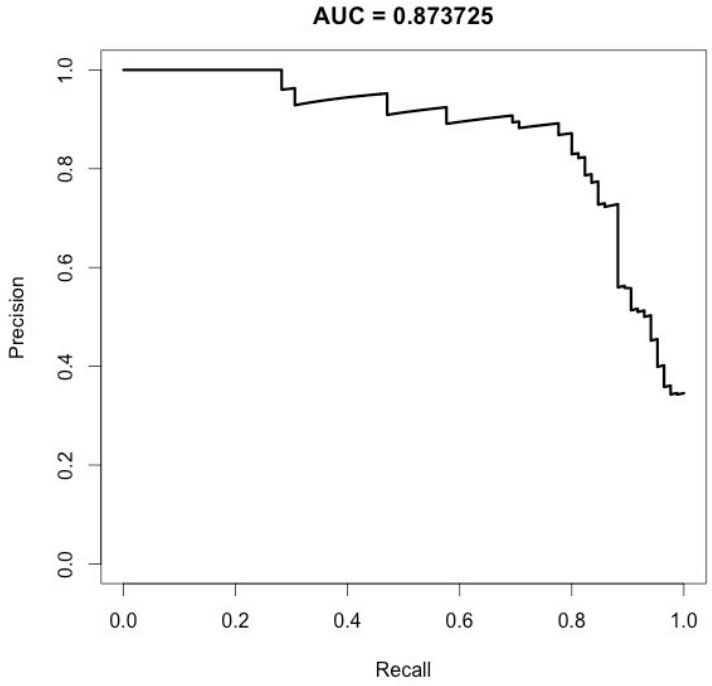


Figure 6.11 Precision-Recall Curve for T2DM Cohort Selection.

Study 2: Direct comparison of GPhenome and eMERGE

Out of the 2457 test patients, the eMERGE algorithm identified 402 diabetic patients (16.4% of the total test patients). The GPhenome model algorithm identified 497 diabetic patients (20.2%). Overall, the probabilities of T2DM inferred for the test patients ranged from 1-12%. Although there was some

overlap between the identified patients, almost 70% were only identified by one of the two algorithms (Figure 6.12).

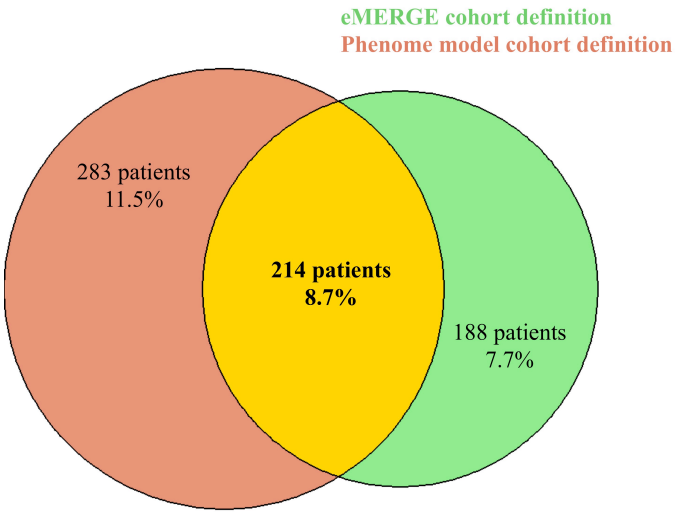


Figure 6.12 Overlap between cohorts identified by GPhenome and eMERGE cohorts.

The performance measures (Table 6.1) were calculated by extrapolating from the randomly sampled patients evaluated by the clinician. A bootstrap procedure was used to estimate the 95% confidence intervals.

	Sensitivity	**Specificity**	**PPV**	**F-measure**
Grounded Phenome Model	0.283 [95% CI: 0.25-0.31]	0.953 [95% CI: 0.91-0.99]	0.920 [95% CI: 0.84-0.98]	0.433 [95% CI: 0.39-0.47]
eMERGE algorithm	0.241 [95% CI: 0.22-0.27]	0.973 [95% CI: 0.94-1]	0.941 [95% CI: 0.86-1]	0.384 [95% CI: 0.35-0.42]

Table 6.1 The statistical measures of performance for eMERGE and GPhenome.

In addition to the measures of classification accuracy, the expert performed an evaluation of the GPhenome model ranking. The clinician found that 100% of the top 50 GPhenome-identified T2DM patients were true T2DM patients.

6.3.4 Discussion

The results of cohort selection by the two models suggest that cohorts selected by eMERGE rule-based algorithms could be greatly augmented by the use of the GPhenome model. We discuss the benefits of using GPhenome for cohort selection below.

The GPhenome model maintains the portability of other cohort selection tools. eMERGE algorithms are specifically constructed around widely-used terminologies for the purposes of portability; any institution can apply the eMERGE rules and gather a patient cohort. As the GPhenome model is an unsupervised algorithm that learns computational disease models specifically from input data and publically available clinical ontologies, it can also be implemented at any institution.

As an inherent benefit of being a probabilistic graphical model, the GPhenome model assigns a weight to each phenotype, for each patient. In contrast to the eMERGE algorithms which present a binary result of a patient being a case or not a case, the GPhenome model has the ability to rank patients by their phenotype weight. This ranking enables researchers to set thresholds based on their desired cohort size, sensitivity/specificity constraints, and even the presence/absence of other diseases and comorbidities as we demonstrated by the T1DM weight constraint.

To examine why some patients were captured by GPhenome and not by eMERGE, we conducted a small and an informal chart review. We found that in some cases, most of the diabetic evidence was in the clinical notes and therefore missed by the eMERGE algorithm. In many other cases, the reason for exclusion was the coding of T1DM diagnosis code. T1DM codes are often recorded by mistake for T2DM patients because of how similar the diagnosis code numbering is for the two diseases (250.0 vs. 250.1); the eMERGE algorithm filters these patients out by design. The GPhenome model, however, weighs the T1DM diagnosis code along with other T1DM evidence and assigns each patient a probability of being ill with T1DM. Unfortunately, it is more difficult to assess the reason that some patients were included by eMERGE but excluded by GPhenome given the probabilistic nature of the model and lack of

explicit rule sets. However, the results of the precision-recall curve demonstrate that reducing the current threshold of 0.01 could yield a larger set of true T2DM patients and have higher accuracy results.

In future work we will continue to refine the grounding, as ideally the grounding regimen would provide one phenotype representing the T2DM disease. However, the current grounding mechanism employed in GPhenome already enables a smaller set of relevant disease models. In the UPhenome model, there exist seven phenotypes that represent T2DM, three phenotypes that represent T1DM, and one phenotype that is a mixture of both diabetes types. In contrast, the GPhenome model only has two phenotypes for T2DM and two phenotypes for T1DM.

These experiments demonstrate that a probabilistic technique and model such as the GPhenome model provide a promising platform for high-throughput phenotyping. The results obtained by the GPhenome model are comparable to eMERGE results and carry the following additional benefits: (i) the unsupervised nature of the model learns directly from the data; (ii) the algorithm learns many computational models of disease at the same time, greatly reducing the amount of expert time needed; and (iii) the model enables ranking of patients by probability of having the phenotype. With further model refinements in areas such as data-type specific modeling, incorporation of new data, and grounding technique enhancements, we believe this model will provide a new platform for phenotyping across many different institutions.

Chapter 7: Conclusions and Future Work

7.1 Conclusion

The concluding chapter of this thesis focuses on summarizing the research presented, highlighting the contributions of this work, detailing the limitations of the studies described, and the offering potential directions for future work.

Chapter 2 outlined the research conducted in the field of clinical summarization and identified a set of gaps that remain in the published literature including methods for dealing with redundancy, methods for mitigating the biases caused by irregular sampling, and methods for automatically determining what pieces of a patient record should be highlighted and what pieces should be discarded from a summary. The first study (Chapter 3) dealt with the question of abundant textual redundancy in clinical text: the chapter presented a hybrid approach to identifying redundant concepts in patient records.

The second study (Chapter 4) focused on irregular sampling in laboratory tests: the study demonstrated the importance of patterns of missingness in laboratory tests and displayed the benefit of including these patterns in clinical modeling. The third study (Chapter 5) focused on the question of saliency and featured a probabilistic graphical model created for summarizing EHR data. Finally, Chapter 6 described two separate studies: one which explored the patterns of missingness and temporal trends of

laboratory tests in the context of overutilization and guideline adherence, and another which examined the practicality of the using the probabilistic graphical model for automated high-throughput cohort selection.

7.2 Contributions

The immense amount of EHR data collected as a byproduct of routine patient care provides a unique platform to perform large-scale research studies of human health. Through careful analysis of the variables in this vast dataset, researchers can conduct a variety of multifaceted studies such as prediction of future patient health state, evaluation of intervention effectiveness, computational disease modeling, and identification of dangerous drug-drug interactions (Prokosch and Ganslandt 2009; McCarty et al. 2011; Xiaoyan Wang et al. 2008). This dissertation contributes to the growing body of research on how to best use EHR data for a variety of tasks. The work described here specifically investigates the challenges in leveraging EHR data for automated patient record summarization. The specific contributions of this thesis are:

- **Context-specific concept aggregation**. Mapping clinical term mentions to semantic concepts in an ontology provides valuable abstraction from lexical variants present in text. But some concepts might need to be further aggregated in order to avoid problems of signal dilution. Our approach described in Aim I scores the similarity of two input concepts by combining complementary information derived from usage patterns of clinical documentation, accepted definitions, and position of the concepts in an ontology. Our experiments show that, given a homogeneous corpus of clinical notes, it is possible to determine automatically which concepts convey similar meaning in the context of the corpus with accuracy above that of previously proposed methods. Our work provides evidence to show that the similarity of clinical concepts depends on a patient's health history, an insight that can help guide future semantic similarity work for the clinical domain. Our method for identifying context-specific similar concept can be used as a pre-processing step in

many clinical NLP applications. In addition, we contribute a novel way of combining insights from text written by clinicians and from ontological concept definitions and locations.

- **Demonstrating EHR biases in laboratory tests using missingness patterns.** In Aim II, we show that surprisingly, there is often little shared information between laboratory test values and the laboratory test's rate of measurement in time. Furthermore, measurement patterns are useful features to use in disease modeling and they can result from a combination of hospital workflow practices and clinical states. When the clinical and documentation biases are not in concert (as is often the case with laboratory tests used for multiple purposes), EHR-driven association studies may produce biased results. We catalogued the measurement dynamics of laboratory tests into three motifs, one of which has mixed patterns of measurement and is prone to biases. For the re-use of clinical data to facilitate novel data-driven informatics research, understanding salient features and correcting for EHR biases is a necessary step. We demonstrate and present a method for how to control for the biases by disambiguating patient health states based on laboratory measurement frequency, using the laboratory test lipase and acute pancreatitis as an illustrative example. With this work we show that incorporating the temporality of laboratory tests, and specifically by capturing the phenomenon of irregular sampling and missing data, we can better understand the clinical state of a patient. Laboratory tests were used to demonstrate the usefulness of missingness patterns, however this intuition can be applied to other clinical data as well, such as diagnosis or procedure codes.
- **Identifying HbA1c overutilization using temporal distribution analytics.** With the growing interest in detecting redundant and inappropriate laboratory test utilization (Zhi et al. 2013; Bates et al. 1998), we apply the methodology and insight learned in Aim II to identify overutilization of HbA1c in a large academic medical center. The methods described in this study can easily be replicated to other high-volume tests. As more data-derived patterns of use are detected across labs, we can start to better understand key aspects of clinicians' workflow and what informatics solutions can be put in place to support clinicians for quality care all the while ensuring cost effectiveness.

- **Creating a probabilistic model to learn computational representations of disease across heterogeneous EHR data types.** We show that latent variable models can be leveraged to handle some issues with EHR data. For instance, the model's probabilistic framework enables modeling and reasoning over uncertain data; they can handle sparse data; they come equipped with mechanisms for tractable and scalable computation, and many of them are able to produce interpretable results. Aim III details the creation of a new latent variable model that uses probabilistic graphical modeling techniques for identifying salient parts of patient records. We show that learning phenotypes directly from raw and heterogeneous patient records by separately but jointly modeling different EHR data types, is able to automatically derive coherent and useful computational models of disease.
- **Display a methodology for combining data from different and disparate EHR systems.** The methods used to create the Phenome model are capable of combining data from different EHR systems to create a unified patient summary. The input patient data to the Phenome model was derived from two different systems used by Columbia University physicians: SCM CROWN and Allscripts. This work demonstrates the potential for probabilistic graphical models to easily integrate and aggregate data from separate EHR systems which is an important contribution to the Health Information Exchange initiatives across the nation.
- **Presenting a method to enhance computational models of disease with clinical knowledge.** The informatics community has invested a lot of effort in creating large and comprehensive ontological repositories of clinical knowledge. We describe a method for automatically leveraging the relationships in these ontologies when phenotyping directly from EHR data. We show how providing informative priors derived from clinical ontologies can help guide unsupervised mixture models to identify more coherent and cohesive phenotypes.
- **Identifying challenges in topic model evaluation for the clinical domain.** The machine learning community has developed some algorithms for automatic evaluation of topic models and

demonstrated that they correlate with human judgment of coherence. In our work, we show that this result may not be applicable to clinical data. We find that clinical expert knowledge does not correlate with the automated measures of coherence and that healthcare process detritus that exists in patient records (formulaic note structure, copy and paste redundancy, etc) can create phenotypes that automated methods mark as highly coherent but are clinically irrelevant.

- **Designing a new method for high-throughput cohort identification from EHR data.** The Grounded Phenome model built and discussed in this thesis can provide a new avenue for phenotyping from EHR data. We show that the GPhenome model can identify cohorts as well as well-established rule-based algorithms but with additional benefits such as limited expert involvement, the ability to identify many cohorts at once, and the ranking of patients by disease probabilities. This new method can significantly augment results from rule-based methods for cohort identification.

7.3 Limitations

We acknowledge that there exist many limitations to the studies explored in this thesis. The specific limitations are listed and described below:

- **Ontology Usage.** For many studies reported in this work, we rely on ontological knowledge to enhance our data-driven techniques. For the work on reducing textual redundancy, we depend on accurate relationships between concepts and up to date definitions for each SNOMED-CT concept. In the grounding of the Phenome model, we assume the links between diseases and observations (medications, laboratory tests, diagnosis codes), are current and comprehensive. Errors that may exist in the ontologies would propagate into our methods and potentially hurt the modeling results.
- **Procedure Code Absence.** The work in this thesis explores characteristics of four different EHR data types: clinical notes, ICD-9 codes, prescribed medications, and ordered laboratory tests. The thesis also presents a method for unsupervised grouping of these data. However, it is clear that

many important clinical events are encoded in procedure codes, most often using the Current Procedural Terminology (CPT). In our work, we did not include procedural data and this exclusion potentially limited our ability to learn cohesive and comprehensive phenotypes. The addition of procedure codes could be especially powerful for phenotypes that have well-defined procedures associated with them.

- **Text Processing.** The preprocessing of textual data and extraction of variables is crucial for much of the work in this thesis. We used HealthTermFinder as the clinical NLP software to perform processing such as chunking, parsing, and named-entity recognition. As clinical text is highly unstructured and has varied grammar, the software is imperfect. We recognize that the imperfection at the text processing phase is carried through the work and can account for errors and signal dilution in many of studies described. We also made two other modeling choices that affect the study results: a bag-of-words assumption and removal of modifers. We use the processed text in a bag-of-words fashion which does not preserve the word order and additionally, we chose not to capture concept modifiers that appear in the text. These design choices give us the ability to represent the frequency and co-occurrence of various medical concepts, however negation and subtle meaning may be lost due to these simplifying assumptions.
- **Generalizability.** Much of the work conducted in this thesis is carried out in a single institution. We applied some of our methods to another dataset (MIMIC ICU) and have presented institution-agnostic methods and workflows. However, we recognize that while the populations we study are large, our findings may be limited by the data within our data warehouse. Additionally, in the laboratory tests biases study, we acknowledge the limitation of using only one test to demonstrate the measurement gap biases. We only present the lipase and acute pancreatitis association study as a single proof-of-concept example to exhibit the potential importance of separating laboratory test values by measurement pattern. Finally, we recognize that using the Medical Entities Dictionary (MED) as the terminology for laboratory tests limits the generalizability of the GPhenome model. Known laboratory test to disease associations were easily available through the MED and although

there is an older version of the MED incorporated into UMLS, we recognize that using the more standard Logical Observation Identifiers Names and Codes (LOINC) vocabulary would increase the reproducibility of this work.

- **Heuristics.** Within some of the work in this thesis, we chose heuristic-based thresholds. In the concept aggregation research, we implemented a cutoff for note-based similarity at 75%. This threshold was set by looking at the curve of the similarity values and picking enough to demonstrate the methodology and provide the annotators with a manageable set of pairs to evaluate. Similarly, we chose the 3-day threshold for the laboratory test biases study by examining the histogram of days between consecutive measurements. Finally, in the cohort-identification study, the 1% threshold for choosing patients as diabetic cases was determined by inspecting the distribution of phenotype weights across the entire population. In future work, finding data-driven ways to substantiate these different thresholds will be important.
- **Temporality.** Although we demonstrate in Chapter 4 that temporal features are very important for modeling EHR data, the Phenome model is currently designed to work on slices of patient records when the patient is at a relatively stable health status. It is possible to simulate a changing patient phenotype by running the Phenome model on discrete time windows within a patient record, but ideally, the Phenome model would be augmented to fully understand the evolution of health states in a patient. We fully acknowledge that developing a model, which overlooks the temporal component present in EHR data, is a large limitation in our work.

7.4 Future Work

The findings reported in this thesis along with the current methodological limitations point to many different directions for future research. We present a set of interesting problems for future examination below.

- **Combining insights and methods from all three Aims.** The goal of this thesis is to create different

methods for dealing with quirks and pathologies of EHR data, specifically in the context of EHR summarization. Combining the work from these different studies is an exciting arena for future research. The concept aggregation techniques can be applied as pre-processing for the Phenome model. Applying the redundancy removal technique to clinical notes and thereby reducing the dimensionality of the input text can help guide the Phenome model to produce more coherent and interpretable results. Additionally, as we have shown, laboratory tests contain significant information both in their numerical value and in their sampling frequency. Incorporating this insight in the Phenome model can be done in multiple ways. We can pre-process laboratory tests to separate them into different sampling frequencies or identify a way to model laboratory measurement patterns along with laboratory test values. The joint modeling proposal is both an interesting technical challenge that may reveal more latent structure and associations between measurement patterns and patient diseases. Finally, including other data (such as flowsheets, vital signs, demographics, procedure codes, etc.) into the Phenome model structure may yield different and exciting insights into disease pathologies and phenotype structure that exists in EHR patient populations.

- **Beyond bi-directional, pairwise similarity.** When examining the pairs of similar concepts produced by our method in Aim I, we noticed potential for expanding our method to higher dimensions of similarity and clustering concepts. The pairwise similarity often produced triangulated results, which suggest clustering could be carried out as an extension of the pairwise similarity methodology to identify groups of concepts that are semantically similar enough to be aggregated. For instance, we located multiple triplets (three concepts vaguely describing the same concept with each pair achieving high similarities) and one five-pair cluster with the five different concepts describing sputum of different colors (yellow, green, brown, clean, and white). We found each of the 10 combination pairs scoring similar. Extending the method to find clusters of similar concepts is exciting future work. In addition, considering defining a direction of similarity as discussed by Kotlerman et al. to incorporate more pairs and pairs with a non-symmetric similarity (Kotlerman, Dagan, and Szpektor 2010) is an attractive area for further research.

- **Increased Generalizability.** Two large limitations to generalizability include (1) data-driven thresholding and (2) incorporation of more standard terminologies. Many of the studies described in this thesis relied on visual examination of different graphs to determine appropriate thresholding and filtering. Applying techniques for density estimation, change-point analysis, or appropriately fitting mixture models would result in more generalizable methodologies that can be applied in other institutions. Using more standard terminologies for data types will also result in the broader applicability of the work – future work will focus on mapping laboratory tests to LOINC codes instead of MED codes.
- **Temporality in the Phenome Model.** In its current version, the Phenome model does not explicitly encode any temporality about given patient records. Because longitudinal records and diseases themselves are often not time invariant (Pivovarov et al. 2014), and progress at different time resolutions, it is a non-trivial task to model temporality across all diseases at once. A simple approach to further the Phenome model work would include experimenting with incorporating temporality by inferring phenotypes over time, much like the approach of dynamic topic models(Blei and Lafferty 2006).
- **Data-type specific modeling.** Each of the considered data types in the Phenome model has specific characteristics that can be exploited further. The EHR text, especially when learning from several notes for each patient, has much redundancy that can be accounted for (Cohen et al. 2014). Medications and diagnosis codes are hierarchical in nature, with evidence that incorporating that structure helps in modeling clinical information (Perotte et al. 2011). Finally, each laboratory test has associated values, and it is clear that different distributions of the same test can describe different diseases; for instance, glucose with a distribution biased towards high values would belong to a phenotype describing diabetes, while glucose with a distribution with mean towards low values would be probable in a hypoglycemia phenotype.
- **Refinement of ontological grounding.** In this thesis we show that unsupervised disease modeling can

be influenced to learn more coherent phenotypes by grounding with informative priors. The question of how best to employ grounding for this task remains future work. For example, it may be important to augment the priors differently for different data types; as there are many more words than laboratory tests, grounding both with the same weight may give too much weight to laboratory tests. In addition, questions remain about how best to chose words specific to each phenotype when performing grounding.

References

Abdala OT, and Saeed M. 2004. "Estimation of Missing Values in Clinical Laboratory Measurements of ICU Patients Using a Weighted K-Nearest Neighbors Algorithm," *Computers in Cardiology* 31:693–96.

Abraham J, Thomas KG, Almoosa KF, Patel B, and Patel VL. 2014. "Comparative Evaluation of the Content and Structure of Communication Using Two Handoff Tools: Implications for Patient Safety." *Journal of Critical Care* 29 (2): 311.e1–7.

Abraham J, Nguyen V, Almoosa KF, Patel B, and Patel VL. 2011. "Falling through the Cracks: Information Breakdowns in Critical Care Handoff Communication." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium* 2011: 28–37.

Adler-Milstein J, Bates DW, and Jha AK. 2011. "A Survey of Health Information Exchange Organizations in the United States: Implications for Meaningful Use." *Annals of Internal Medicine* 154 (10):666-671.

Afantenos, SD. 2006. "Automatic Text Summarization from Multiple Sources for Time Evolving Events." National and Kapodistrian University of Athens.

Akan P, Cimrin D, Ormen M, Kume T, Ozkaya A, Ergor G, and Abacioglu H. 2007. "The Inappropriate Use of HbA1c Testing to Monitor Glycemia: Is There Evidence in Laboratory Data?" *Journal of Evaluation in Clinical Practice* 13 (1): 21–24.

Albers DJ, and Hripcsak G. 2010. "A Statistical Dynamics Approach to the Study of Human Health Data: Resolving Population Scale Diurnal Variation in Laboratory Data." *Physics Letters A* 374 (9): 1159–64.

Albers DJ, and Hripcsak G. 2011. "Estimation of Time-Delayed Mutual Information and Bias for Irregularly and Sparsely Sampled Time-Series." *arXiv.org*, October.

Albers DJ, and Hripcsak G. 2012. "Using Time-Delayed Mutual Information to Discover and Interpret Temporal Correlation Structure in Complex Populations." *CHAOS* 22 (1): 013111.

Albers DJ, Hripcsak G, and Schmidt M. 2012. "Population Physiology: Leveraging Electronic Health Record Data to Understand Human Endocrine Dynamics." *PLoS ONE* 7 (12): e48058.

Allan J, Gupta R, and Khandelwal V. 2001. “Temporal Summaries of News Topics,” In *SIGIR*. 10–18.

Al-Mubaid H, and Nguyen HA. 2006. “A Cluster Based Approach for Semantic Similarity in the Biomedical Domain.” *IEEE Int. Conf. on Granular Computing*, July, 2713–17.

Alterman R. 1991. “Understanding and Summarization.” *Artificial Intelligence Review* 5 (4): 239–54.

Alterovitz G, Xiang M, Hill DP, Lomax J, Liu J, Cherkassky M, Dreyfuss J, et al. 2010. “Ontology Engineering.” *Nature Biotechnology* 28 (2): 128–30.

American Diabetes Association. 2010. “Diagnosis and Classification of Diabetes Mellitus”. *Diabetes Care* 33(S62-S69*)*.

American Diabetes Association. 2013. “Standards of Medical Care in diabetes–2013.” *Diabetes Care* 36:S11–66.

Androutsopoulos I. 2010. “A Survey of Paraphrasing and Textual Entailment Methods.” *arXiv.org*, May, 1–53.

Andrzejewski D, Zhu X, and Craven M. 2009. “Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors.” In *ICML*.

Andrzejewski D, Zhu X, Craven M, and Recht B. 2011. “A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-Order Logic.” In *IJCAI*.

Arnold, CW, El-Saden SM, Bui AT, Taira R. 2010. “Clinical Case-Based Retrieval Using Latent Topic Analysis.” *AMIA 2010 Symposium Proceedings*, July, 1–5.

Arocha JF, Wang D, and Patel VL. 2005. “Identifying Reasoning Strategies in Medical Decision Making: A Methodological Guide.” *Journal of Biomedical Informatics* 38 (2): 154–71.

Banerjee, S. 2002. “Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet,” December, 1–98. University of Minnesota.

Banerjee S and Pederson T. 2002. “An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet.” *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, November, 136–45.

Banks PA, Freeman ML, and the Practice Parameters Committee of the American College of Gastroenterology. 2006. “Practice Guidelines in Acute Pancreatitis.” *The American Journal of Gastroenterology* 101 (10): 2379–2400.

Baron JM, and Dighe AS. 2014. “The Role of Informatics and Decision Support in Utilization Management.” *Clinica Chimica Acta; International Journal of Clinical Chemistry* 427 (January): 196–201.

Barzilay R, and Lee L. 2004. “Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization.,” in *ACL*. 113–20.

Bashyam V, Hsu W, Watt E, Bui AAT, Kangarloo H, and Taira RK. 2009. “Informatics in Radiology: Problem-Centric Organization and Visualization of Patient Imaging and Clinical Data1.” *Radiographics* 29 (2): 331–43.

Bates DW, Boyle DL, Rittenberg E, Kuperman GJ, Ma’luf N, Menkin V, Winkelman JW, and Tanasijevic MJ. 1998. “What Proportion of Common Diagnostic Tests Appear Redundant?” *The American Journal of Medicine* 104 (4): 361–68.

Batet M, Sánchez D, and Valls A. 2010. “An Ontology-Based Measure to Compute Semantic Similarity in Biomedicine.” *Journal of Biomedical Informatics* 44 (1): 118–25.

Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, and Devignes M. 2010. “IntelliGO: A New Vector-Based Semantic Similarity Measure Including Annotation Origin.” *BMC Bioinformatics* 11 (1): 588.

Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, and Gage BF. 2005. “Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors.” *Medical Care* 43 (5): 480–85.

Blei DM, and Lafferty JD. 2006. “Dynamic Topic Models.” In *ICML*.

Blei DM, Ng NY, and Jordan MI. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3: 993–1022.

Budanitsky A, and Hirst G. 2005. “Evaluating WordNet-Based Measures of Lexical Semantic Relatedness.” *Comput Linguist* 32 (1): 13–47.

Bui, AAT, Aberle DR, and Kangarloo H. 2007. “TimeLine: Visualizing Integrated Patient Records.” *IEEE Transactions on Information Technology in Biomedicine* 11 (4): 462–73.

Cao H, Markatou M, Melton GB, Chiang MF, and Hripcsak G. 2005. "Mining a Clinical Data Warehouse to Discover Disease-Finding Associations Using Co-Occurrence Statistics." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, 106–10.

Friedman C, and Elhadad N. 2014. "Natural Language Processing in Health Care and Biomedicine." In *Biomedical Informatics. Computer Applications in Healthcare*, 255–84. Springer Science & Business Media.

Caviedes, JE, and Cimino JJ. 2004. "Towards the Development of a Conceptual Distance Metric for the UMLS." *Journal of Biomedical Informatics* 37 (2): 77–85.

Chang J, Boyd-Graber J, Gerrish S, Wang C, and Blei D. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Neural Information Processing Systems*.

Chan KR, Lou X, Karaletsos T, Crosbie C, Gardos S, Artz D, and Ratsch G. 2013. "An Empirical Analysis of Topic Modeling for Mining Cancer Clinical Notes," In *ICDMW* 56–63.

Chen, DP, Dudley JT, and Butte AJ. 2010. "Latent Physiological Factors of Complex Human Diseases Revealed by Independent Component Analysis of Clinarrays." *BMC Bioinformatics* 11 (Suppl 9): S4.

Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, and Xu H. 2013. "Applying Active Learning to High-Throughput Phenotyping Algorithms for Electronic Health Records Data." *Journal of the American Medical Informatics Association*, 20(e2):e254-9.

Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, Liebovitz D, Sun J, Denny J, and Malin B. 2015. "Building Bridges across Electronic Health Record Systems through Inferred Phenotypic Topics." *Journal of Biomedical Informatics*, 55:82-93.

Christensen T, and Grimsmo A. 2008. "Instant Availability of Patient Records, but Diminished Availability of Patient Information: A Multi-Method Study of GP's Use of Electronic Patient Records." *BMC Medical Informatics and Decision Making* 8 (1): 12.

Cios KJ, and Moore GW. 2002. "Uniqueness of Medical Data Mining." *Artificial Intelligence in Medicine* 26 (1-2): 1–24.

Cohen J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*. 20(1):37-46.

Cohen R, Aviram I, Elhadad M, and Elhadad N. 2014. “Redundancy-Aware Topic Modeling for Patient Record Notes.” *PLoS ONE* 9 (2): e87555. doi:10.1371/journal.pone.0087555.

Cohen R, Elhadad M, and Elhadad N. 2013. “Redundancy in Electronic Health Record Corpora: Analysis, Impact on Text Mining Performance and Mitigation Strategies.” *BMC Bioinformatics* 14 (1): 10.

Cohen T, and Widdows D. 2009. “Empirical Distributional Semantics: Methods and Biomedical Applications.” *Journal of Biomedical Informatics* 42 (2): 390–405.

Combi C, and Shahar Y. 1997. “Temporal Reasoning and Temporal Data Maintenance in Medicine: Issues and Challenges.” *Computers in Biology and Medicine* 27 (5): 353–68.

International Expert Committee. 2009. “International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes.” *Diabetes Care* 32 (7): 1327–34.

Cordí V, Lombardi P, Martelli M, and Mascardi V. 2005. “An Ontology-Based Similarity between Sets of Concepts.” In *Proceedings of WOA*, 16–21.

Cousins SB, and Kahn MG. 1991. “The Visual Display of Temporal Information.” *Artificial Intelligence in Medicine* 3 (6): 341–57.

Dagan, I, Dolan B, Magnini B, and Roth D. 2010. “Recognizing Textual Entailment: Rational, Evaluation and Approaches.” *Natural Language Engineering* 15 (4): 1–16.

de Estrada WD, Murphy S, and Barnett GO. 1997. “Puya: A Method of Attracting Attention to Relevant Physical Findings.” *Proceedings : AMIA Fall Symposium*, 509–13.

Delort JY and Alfonseca E. 2012. “DualSum: A Topic-Model Based Approach for Update Summarization,” In *ACL* 214–23.

Dong H, Hussain FK, and Chang E. 2010. “A Context-Aware Semantic Similarity Model for Ontology Environments.” *Concurrency and Computation: Practice and Experience* 23 (5): 505–24.

Doshi-Velez F, Ge Y, and Kohane I. 2014. “Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis.” *Pediatrics* 133 (1): e54–63.

Doshi-Velez F, Wallace B, and Adams R. 2015. “Graph-Sparse LDA: A Topic Model with Structured Sparsity.” In *Arxiv*. October.

Driskell OJ, Holland D, Hanna FW, Jones PW, Pemberton RJ, Tran M, and Fryer AA. 2012. "Inappropriate Requesting of Glycated Hemoglobin (Hb A1c) Is Widespread: Assessment of Prevalence, Impact of National Guidance, and Practice-to-Practice Variability." *Clinical Chemistry* 58 (5): 906–15.

Edmundson, HP. 1969. "New Methods in Automatic Extracting." *Journal of the ACM (JACM)* 16 (2): 264–85.

Elhadad N, and Sutaria K. 2007. "Mining a Lexicon of Technical Terms and Lay Equivalents." In *Proceedings of the ACL BioNLP Workshop*, 49–56.

Enders CK. 2006. "A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research." *Psychosomatic Medicine* 68 (3): 427–36.

Erkan G, and Radev DR. 2004. "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization." *J. Artif. Intell. Res. (JAIR)* 22: 457–79.

Farhangfar A, Kurgan LA, and Pedrycz W. 2007. "A Novel Framework for Imputation of Missing Values in Databases." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37 (5): 692–709.

Farri OF. 2012. "Understanding Clinician Information Demands and Synthesis of Clinical Documents in Electronic Health Record Systems." University of Minnesota.

Farri O, Rahman A, Monsen KA, Zhang R, Pakhomov SV, Pieczkiewicz DS, Speedie SM, and Melton GB. 2012. "Impact of a Prototype Visualization Tool for New Information in EHR Clinical Documents." *Applied Clinical Informatics* 3 (4): 404–18.

Farzandipour M, Sheikhtaheri A, and Sadoughi F. 2010. "Effective Factors on Accuracy of Principal Diagnosis Coding Based on International Classification of Diseases, the 10th Revision (ICD-10)." *International Journal of Information Management.* 30(1):78-84.

Feblowitz, JC, Wright A, Singh H, Samal L, and Sittig DF. 2011. "Summarization of Clinical Information: A Conceptual Model." *Journal of Biomedical Informatics* 44 (4): 688–99.

Fong A, and Ratwani R. 2015. "An Evaluation of Patient Safety Event Report Categories Using Unsupervised Topic Modeling." *Methods of Information in Medicine* 54 (3).

Fort D, Wilcox AB, and Weng C. 2014. "Could Patient Self-Reported Health Data Complement EHR for Phenotyping?" *AMIA Annual Symposium Proceedings*.

Friedman C, Alderson PO, Austin JH, Cimino JJ, and Johnson SB. 1994. "A General Natural-Language Text Processor for Clinical Radiology." *AMIA* 1 (2): 161–74.

Fries JF. 1974. "Alternatives in Medical Record Formats." *Medical Care* 12 (10): 871–81.

Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, and Szolovits P. 2014. "Unfolding Physiological State." In *The 20th ACM SIGKDD International Conference*, 75–84. New York, New York, USA: ACM Press.

Goldstein DE, Little RR, Lorenz RA, Malone JI, Nathan D, Peterson CM, and Sacks DB. 2004. "Tests of Glycemia in Diabetes." *Diabetes Care* 27 (7): 1761–73.

Griffiths TL, and Steyvers M. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl 1 (April): 5228–35.

Grundy SM, Bilheimer D, Chait A, and Clark LT. 1993. "Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)." *JAMA*. 269 (23):3015-3023.

Savova S, Bethard S, and Styler W. 2009. "Towards Temporal Relation Discovery from the Clinical Narrative," *Proceedings: AMIA*, 568.

Hallett C. 2008. "Multi-Modal Presentation of Medical Histories" In *ACM*. 80–89.

Hallett C, and Scott D. 2005. "Structural Variation in Generated Health Reports," In *ACL* 1–8.

Halpern Y, Choi Y, Horng S, and Sontag D. 2014. "Using Anchors to Estimate Clinical State without Labeled Data." In *AMIA*. Washington, DC.

Hamon T, and Grabar N. 2008. "How Can the Term Compositionality Be Useful for Acquiring Elementary Semantic Relations?" *Proceedings of the 6th International Conference on Advances in Natural Language Processing*.

Hanauer DA, Rhodes DR, and Chinnaiyan AM. 2009. "Exploring Clinical Associations Using `-Omics' Based Enrichment Analyses." *PLoS ONE* 4 (4): e5203.

Handelsman Y, Mechanick JI, Blonde L, Grunberger G, Bloomgarden ZT, Bray GA, Dagogo-Jack S, Davidson JA, Einhorn D, and Ganda O. 2011. "American Association of Clinical

Endocrinologists Medical Guidelines for Clinical Practice for Developing a Diabetes Mellitus Comprehensive Care Plan." *Endocrine Practice* 17: 1–53.

Hanley JA, and McNeil BJ. 1983. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology* 148 (3): 839–43.

Harris ZS. 1968. *Mathematical Structures of Language*. New York: Interscience Publishers John Wiley & Sons.

Hersh WR, Weiner MG, Embi PJ, and Logan JR. 2013. "Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research." *Medical Care*.

Hidalgo CA, Blumm N, Barabási AL, and Christakis NA. 2009. "A Dynamic Network Approach for the Study of Human Phenotypes." *PLoS Computational Biology* 5 (4): e1000353.

Hirsch JS, Tanenbaum JS, Gorman SL, Liu C, Schmitz E, Hashorva D, Ervits A, Vawdrey D, Sturm M, and Elhadad N. 2014. "HARVEST, a Longitudinal Patient Record Summarizer." *Journal of the American Medical Informatics Association : JAMIA*, amiajnl – 2014–002945.

Hirschtick RE. 2006. "Copy-and-Paste." *JAMA: The Journal of the American Medical Association* 295 (20): 2335–36.

Ho JC, Ghosh J, and Sun J. 2014. "Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization," In *ACM* 115–24.

Holden RJ. 2011. "Cognitive Performance-Altering Effects of Electronic Medical Records: An Application of the Human Factors Paradigm for Patient Safety." *Cognition, Technology & Work (Online)* 13 (1): 11–29.

Hripcsak G and Albers DJ. 2013. "Next-Generation Phenotyping of Electronic Health Records." *Journal of the American Medical Informatics Association: JAMIA* 20 (1): 117–21.

Hripcsak G, Albers DJ, and Perotte A. 2011. "Exploiting Time in Electronic Health Record Correlations." *Journal of the American Medical Informatics Association : JAMIA* 18 Suppl 1 (December): i109–15.

Hripcsak G, Albers DJ, and Perotte A. 2015. "Parameterizing Time in Electronic Health Record Studies." *Journal of the American Medical Informatics Association : JAMIA*, 22(4):794-804.

Hripcsak, G, Elhadad N, Chen YH, Zhou L, and Morrison FP. 2009. "Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts." *AMIA* 16 (2): 220–27.

Hsu W, Taira RK, El-Saden S, Kangarloo H, and Bui AAT. 2012. "Context-Based Electronic Health Record: Toward Patient Specific Healthcare." *IEEE Transactions on Information Technology in Biomedicine* 16 (2): 228–34.

Huang ES, Basu A, O'Grady M, and Capretta JC. 2009. "Projecting the Future Diabetes Population Size and Related Costs for the U.S." *Diabetes Care* 32 (12): 2225–29.

Hug CW. 2006. "Predicting the Risk and Trajectory of Intensive Care Patients Using Survival Models." MIT.

Hunter J, Freer Y, Gatt A, Logie R, Mcintosh N, van der Meulen M, Portet F, Reiter E, Sripada S, and Sykes C. 2008. "Summarising Complex ICU Data in Natural Language." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, 323–27.

Hu Y, Boyd-Graber J, and Satinoff B. 2011. "Interactive Topic Modeling." In *Association for Computational Linguistics*. Portland, Oregon.

Jaakkola TS, and Jordan MI. 1999. "Variational Probabilistic Inference and the QMR-DT Network." *Journal of Artificial Intelligence Research*.

Janowicz K. 2008. "Kinds of Contexts and Their Impact on Semantic Similarity Measurement." *Sixth Annual IEEE International Conference on \ldots*.

Jaspers WM, Steen T, van den Bos C, and Geenen M. 2004. "The Think Aloud Method: A Guide to User Interface Design." *International Journal of Medical Informatics* 73 (11-12): 781–95.

Jiang JJ, and Conrath DW. 1997. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy." *Proc. of the International Conference on Research in Computational Linguistics*, 19–33.

Jones KS. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28 (1): 11–21.

Jung H, Allen J, Blaylock N, de Beaumont W, Galescu L, and Swift M. 2011. "Building Timelines from Narrative Clinical Records: Initial Results Based-on Deep Natural Language Understanding," In *ACL* 146–54.

Kessler C, Raubal M, and Janowicz K. 2007. "The Effect of Context on Semantic Similarity Measurement." *On the Move to Meaningful Internet Systems: OTM 2007 Workshops*.

Klann JG, McCoy AB, Wright A, Wattanasin N, Sittig DF, and Murphy SN. 2013. "Health Care Transformation Through Collaboration on Open-Source Informatics Projects: Integrating a Medical Applications Platform, Research Data Repository, and Patient Summarization." *Interactive Journal of Medical Research* 2 (1): e11.

Klann JG, and Schadow G. 2010. "Modeling the Information-Value Decay of Medical Problems for Problem List Maintenance," In *ACM* 371–75.

Klimov D, Shahar Y, and Taieb-Maimon M. 2010. "Intelligent Visualization and Exploration of Time-Oriented Data of Multiple Patients." *Artificial Intelligence in Medicine* 49 (1): 11–31.

Kotlerman L, Dagan I, and Szpektor I. 2010. "Directional Distributional Similarity for Lexical Inference." *Natural Language Engineering*. 16(4):359-389.

Kushniruk KW. 2001. "Analysis of Complex Decision-Making Processes in Health Care: Cognitive Approaches to Health Informatics." *Journal of Biomedical Informatics* 34 (5): 365–76.

Landis RJ., and Koch GG. 1977. "Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.

Lasko, TA, Denny JC, and Levy MA. 2013. "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data." *PLoS ONE* 8 (6): e66341.

Lau JH, Newman D, and Baldwin T. 2014. "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality." In *ACL*. 530-539.

Laxmisan A, Vaughan-Sarrazin M, and Cram P. 2011. "Repeated Hemoglobin A1C Ordering in the VA Health System." *The American Journal of Medicine* 124 (4): 342–49.

Laxmisan A, McCoy AB, Wright A, and Sittig DF. 2012. "Clinical Summarization Capabilities of Commercially-Available and Internally-Developed Electronic Health Records." *Applied Clinical Informatics* 3 (1): 80–93.

Leacock C, and Chodorow M. 1998. *Combining Local Context and WordNet Similarity for Word Sense Disambiguation*. An Electronic Lexical Database. WordNet: An electronic lexical database.

Lesk M. 1986. “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone.” In *Proceedings of the 5th Annual International Conference on Systems Documentation*, 24–26.

Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, et al. 2015. “Development of Phenotype Algorithms Using Electronic Medical Records and Incorporating Natural Language Processing.” *BMJ* 350 (11): 1885–1885.

Lin D. 1998. “An Information-Theoretic Definition of Similarity.” In *Fifteenth International Conference on Machine Learning, ICML*, 296–304.

Lindberg DA, Humphreys BL, and McCray AT. 1993. “The Unified Medical Language System.” *Methods of Information in Medicine* 32 (4): 281–91.

Lin, JH, and Haug PJ. 2008. “Exploiting Missing Clinical Data in Bayesian Network Modeling for Predicting Medical Problems.” *Journal of Biomedical Informatics* 41 (1): 1–14.

Lipsky-Gorman S, and Elhadad N. 2011. “ClinNote and HealthTermFinder: A Pipeline for Processing Clinical Notes.” Technical Report. Columbia University Technical Report.

Little RJA, and Rubin DB. 2002. “Statistical Analysis with Missing Data.” John Wiley & Sons.

Little RJA. 1993. “Pattern-Mixture Models for Multivariate Incomplete Data.” *Journal of the American Statistical Association* 88 (421): 125–34.

Liu H, and Friedman C. 2004. “CliniViewer: A Tool for Viewing Electronic Medical Records Based on Natural Language Processing and XML.” *Studies in Health Technology and Informatics* 107 (Pt 1): 639–43.

Li Y, Bandar ZA, and McLean D. 2003. “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.” *IEEE Transactions on Knowledge and Data Engineering* 15 (4): 871–82.

Li Y, Gorman SL, and Elhadad N. 2010. “Section Classification in Clinical Notes Using Supervised Hidden Markov Model,” In *ACM* 744–50.

Luhn HP. 1958. “The Automatic Creation of Literature Abstracts.” *IBM Journal of Research and Development* 2 (2): 159–65.

Lussier YA, and Liu Y. 2007. "Computational Approaches to Phenotyping: High-Throughput Phenomics." *Proceedings of the American Thoracic Society* 4 (1): 18–25.

Lyon AW, Higgins T, Wesenberg JC, Tran DV, and Cembrowski GS. 2009. "Variation in the Frequency of Hemoglobin A1c (HbA1c) Testing: Population Studies Used to Assess Compliance with Clinical Practice Guidelines and Use of HbA1c to Screen for Diabetes." *Journal of Diabetes Science and Technology* 3 (3): 411–17.

Manning CD, Raghavan P, and Schütze H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Manning CD, and Schütze H. 2003. *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcu D. 1997. "From Discourse Structures to Text Summaries" In *ACL*. 82–88.

Marlin, BM, Kale DC, Khemani RG, and Wetzel RC. 2012. "Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models," In *ACM*. 389–98.

Massey FJ. 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association* 46 (253): 68–78.

Matar Y, and Egyed-Zsigmond E. 2008. "KWSim: Concepts Similarity Measure." *Proceedings of Conférence En Recherche d'Information et Applications (CORIA08)*, August, 475–782.

McCallum, AK. 2002. "MALLET: A Machine Learning for Language Toolkit." www.mallet.cs.umass.edu.

McCarty, CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, et al. 2011. "The eMERGE Network: A Consortium of Biorepositories Linked to Electronic Medical Records Data for Conducting Genomic Studies." *BMC Medical Genomics* 4: 13.

McCray AT, Burgun A, and Bodenreider 0. 2001. "Aggregating UMLS Semantic Types for Reducing Conceptual Complexity." *Studies in Health Technology and Informatics*, 216–20.

McDonald CJ, Callaghan FM, Weissman A, Goodwin RM, Mundkur M, and Kuhn T. 2014. "Use of Internist's Free Time by Ambulatory Care Electronic Medical Record Systems." *JAMA Internal Medicine*, September. doi:10.1001/jamainternmed. 2014.

McDonald CJ. 1976. “Protocol-Based Computer Reminders, the Quality of Care and the Non-Perfectability of Man.” *New England Journal of Medicine* 295 (24): 1351.

McInnes BT, Pedersen T, and Pakhomov SVS. 2009. “UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity.” *AMIA Annual Symposium Proceedings* 2009: 431–36.

McPherson RA, and Pincus MR. 2011. *Henry’s Clinical Diagnosis and Management by Laboratory Methods*. Saunders.

Militello LG, Arbuckle NB, Saleem JJ, Patterson E, Flanagan M, Haggstrom D, and Doebbeling BN. 2014. “Sources of Variation in Primary Care Clinical Workflow: Implications for the Design of Cognitive Support.” *Health Informatics Journal* 20 (1): 35–49.

Miller RA, Pople HE, and Myers JD. 1982. “Internist-1, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine.” *The New England Journal of Medicine*.

Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, and Del Fiol G. 2014. “Text Summarization in the Biomedical Domain: A Systematic Review of Recent Research.” *Journal of Biomedical Informatics*, 52:457-67.

Mortensen JM, Horridge M, Musen MA, and Noy NF. 2012. “Applications of Ontology Design Patterns in Biomedical Ontologies.” *AMIA Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2012: 643–52.

Murray I, and Salakhutdinov R. 2009. “Evaluating Probabilities under High-Dimensional Latent Variable Models.” In *Advances in Neural Information Processing Systems 21*, 1137–44.

Nenkova A, and McKeown K. 2012. “A Survey of Text Summarization Techniques,” Mining Text Data. 43–76.

Nenkova A, and Passonneau RJ. 2004. “Evaluating Content Selection in Summarization: The Pyramid Method.” In *HLT/NAACL* 4:145–52.

Newman D, Lau JH, and Grieser K. 2010. “Automatic Evaluation of Topic Coherence.” In *ACL*. 100-108.

Newton KM, Peissig PL, and Kho AN. 2013. “Validation of Electronic Medical Record-Based Phenotyping Algorithms: Results and Lessons Learned from the eMERGE Network.” *JAMIA* 20(e1):e147-154.

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, et al. 2009. "BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse." *Nucleic Acids Research* 37 (Web Server issue): W170–73.

O'Keefe, QEW, and Simborg DW. 1980. "Summary Time Oriented Record (STOR)" Proc. Annu Symp Comput Appl Med Care. 2: 1175.

Pakhomov, S, McInnes, Adam T, Liu Y, Pedersen T, Melton GB. 2010. "Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study." *AMIA 2010 Annual Symposium*, July, 1–6.

Pakhomov, SVS, Pedersen T, McInnes B, GB, Ruggieri A, and Chute CG. 2011. "Towards a Framework for Developing Semantic Relatedness Reference Standards." *Journal of Biomedical Informatics* 44 (2): 251–65.

Patel VL, Arocha JF, and Kaufman DR. 2001. "A Primer on Aspects of Cognition for Medical Informatics." *AMIA* 8 (4): 324–43.

Patel VL, and Kushniruk AW. 1998. "Interface Design for Health Care Environments: The Role of Cognitive Science." *AMIA Symposium*, January, 29–37.

Pathak J, Kho AN, and Denny JC. 2013. "Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives." *Journal of the American Medical Informatics Association: JAMIA* 20 (e2): e206–11.

Patwardhan S, and Pedersen T. 2006. "Using WordNet-Based Context Vectors to Estimate the Semantic Relatedness of Concepts." In *EACL*, 1-8.

Payne, TH. 2000. "Computer Decision Support Systems." *Chest* 118 (2 Suppl): 47S – 52S.

Pedersen T, Pakhomov SVS, Patwardhan S, and Chute CG. 2007. "Measures of Semantic Similarity and Relatedness in the Biomedical Domain." *Journal of Biomedical Informatics* 40 (3): 288–99.

Perotte, AJ and Hripcsak G. 2013. "Temporal Properties of Diagnosis Code Time Series in Aggregate." *IEEE J Biomed Health Inform.* 17(2):477-83.

Perotte, AJ, Wood F, Elhadad N, and Bartlett N. 2011. "Hierarchically Supervised Latent Dirichlet Allocation," In *NIPS.* 2609–17.

Pesquita C, Faria D, Falcão AO, Lord P, and Couto FM. 2009. "Semantic Similarity in Biomedical Ontologies." *PLoS Computational Biology* 5 (7): e1000443.

Peters AL, Davidson MB, Schriger DL, and Hasselblad V. 1996. "A Clinical Approach for the Diagnosis of Diabetes Mellitus: An Analysis Using Glycosylated Hemoglobin Levels. Meta-Analysis Research Group on the Diagnosis of Diabetes Using Glycated Hemoglobin Levels." *JAMA: The Journal of the American Medical Association* 276 (15): 1246–52.

Petrakis EGM, Varelas G, Hiliaoutakis A, and Raftopoulou P. 2006. "Design and Evaluation of Semantic Similarity Measures for Concepts Stemming From the Same or Different Ontologies." *4th Workshop on Multimedia Semantics (WMS'06)*, 44–52.

Pirro G, and Seco N. 2008. "Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content." *Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II on On the Move to Meaningful Internet Systems*, August, 1–18.

Pivovarov R, Albers DJ, Sepulveda JL, and Elhadad N. 2014. "Identifying and Mitigating Biases in EHR Laboratory Tests." *Journal of Biomedical Informatics*,51:24-34.

Pivovarov R, and Elhadad N. 2012. "A Hybrid Knowledge-Based and Data-Driven Approach to Identifying Semantically Similar Concepts." *Journal of Biomedical Informatics* 45 (3): 471–81.

Plaisant C, Milash B, Rose A, Widoff S, and Shneiderman B. 1996. "LifeLines: Visualizing Personal Histories," In *SIGCHI* 221–27.

Plaisant C, Mushlin R, Snyder A, Li J, Heller D, and Shneiderman B. 1998. "LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records." *Proceedings of AMIA Annual Symposium. AMIA Symposium*, 76–80.

Poh N, and de Lusignan S. 2011a. "Data-Modelling and Visualisation in Chronic Kidney Disease (CKD): A Step towards Personalised Medicine." *Informatics in Primary Care* 19 (2): 57–63.

Poh N, and de Lusignan S. 2011b. "Modeling Rate of Change in Renal Function for Individual Patients: A Longitudinal Model Based on Routinely Collected Data." *NIPS*, August.

Popescu M, and Dong X. 2009. *Data Mining in Biomedicine Using Ontologies*. Artech House Publishers.

Powsner SM, and Tufte ER. 1994. "Graphical Summary of Patient Status." *The Lancet* 344 (8919): 386–89.

Powsner SM, and Tufte ER. 1997. “Summarizing Clinical Psychiatric Data.” *Psychiatric Services (Washington, D.C.)* 48 (11): 1458–61.

Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, and Savova G. 2014. “Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative.” *Journal of the American Medical Informatics Association : JAMIA*, August.

Prokosch HU, and Ganslandt T. 2009. “Perspectives for Medical Informatics. Reusing the Electronic Medical Record for Clinical Research.” *Methods of Information in Medicine* 48 (1): 38–44.

Rada R, Mili H, Bicknell E, and Blettner M. 1989. “Development and Application of a Metric on Semantic Nets.” *IEEE Transactions on Systems, Man and Cybernetics* 19 (1): 17–30.

Radev DR, Hovy E, and McKeown K. 2002. “Introduction to the Special Issue on Summarization.” *Computational Linguistics* 28 (4): 399–408.

Radev, DR, Jing H, and Budzikowska M. 2000. “Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies,” In *NAACL-ANLP* 21–30.

Raghavan P, Fosler-Lussier E, Elhadad N, and Lai A. 2014. “Cross-Narrative Temporal Ordering of Medical Events.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 998–1008.

Reichert D, Kaufman D, Bloxham B, Chase H, and Elhadad N. 2010. “Cognitive Analysis of the Summarization of Longitudinal Patient Records.” *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium* 2010 (July): 667–71.

Resnik P. 1995. “Using Information Content to Evaluate Semantic Similarity in a Taxonomy.” *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 448–53.

Rind A, Wang TD, Aigner W, Miksh S, Wongsuphasawat K, Plaisant C, and Shneiderman B. 2010. “Interactive Information Visualization for Exploring and Querying Electronic Health Records: A Systematic Review.” In *HCIL*. 207-298.

Rogers JL, Haring OM, and Watson RA. 1979. “Automating the Medical Record: Emerging Issues,” *Proc Annu Symp Comput Appl Med Care* 255-263.

Rogers J, Puleston C, and Rector A. 2006. “The CLEF Chronicle: Patient Histories Derived from Electronic Health Records,” In *ICDEW*. 109.

Rogers JL, and Haring OM. 1979. “The Impact of a Computerized Medical Record Summary System on Incidence and Length of Hospitalization.” *Medical Care* 17 (6): 618–30.

Roque, FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, et al. 2011. “Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts.” *PLoS Computational Biology* 7 (8): e1002141.

Roque FS, Slaughter L, and Tkatsenko A. 2010. “A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content,” In *NAACL HLT* 1–8.

Rosenbloom TS, and Shultz AW. 2012. “Managing the Flood of Codes: Maintaining Patient Problem Lists in the Era of Meaningful Use and ICD10” *Proceedings of AMIA* 2012: 8.

Rubin DB. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–92.

Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, and Parrott M. 2002. “Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus.” *Clinical Chemistry* 48 (3): 436–72.

Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, and Mark RG. 2011. “Multiparameter Intelligent Monitoring in Intensive Care II: A Public-Access Intensive Care Unit Database.” *Critical Care Medicine* 39 (5): 952–60.

Sagreiya H, and Altman RB. 2010. “The Utility of General Purpose versus Specialty Clinical Databases for Research: Warfarin Dose Estimation from Extracted Clinical Variables.” *Journal of Biomedical Informatics* 43 (5): 747–51.

Salvagno GL, Lippi G, Targher G, Montagnana M, and Guidi GC. 2007. “Monitoring Glycaemic Control: Is There Evidence for Appropriate Use of Routine Measurement of Glycated Haemoglobin?” *Clinical Chemical Laboratory Medicine* 45 (8).

Samal L, Wright A, Wong BT, Linder JA, and Bates DW. 2011. “Leveraging Electronic Health Records to Support Chronic Disease Management: The Need for Temporal Data Views.” *Informatics in Primary Care* 19 (2): 65–74.

Sammon, CJ, Miller A, Mahtani KR, Holt TA, McHugh NJ, Luqmani RA, and Nightingale AL. 2015. “Missing Laboratory Test Data in Electronic General Practice Records: Analysis of Rheumatoid Factor Recording in the Clinical Practice Research Datalink.” *Pharmacoepidemiology and Drug Safety* 24 (5): 504–9.

Saria S, Koller D, and Penn A. 2010. "Learning Individual and Population Level Traits from Clinical Temporal Data." In *NIPS*.

Saxena S, Anderson DW, Kaufman RL, Hannah JA, and Wong ET. 1993. "Quality Assurance Study of Cardiac Isoenzyme Utilization in a Large Teaching Hospital." *Archives of Pathology & Laboratory Medicine* 117 (2): 180–83.

Schafer JL, and Graham JW. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.

Schiff GD, and Bates DW. 2010. "Can Electronic Clinical Documentation Help Prevent Diagnostic Errors?" *The New England Journal of Medicine* 362 (12): 1066–69.

Schulam P, Wigley F, and Saria S. 2015. "Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery." In *AAAI.*

Selvin E, Steffes MW, Zhu H, Matsushita K, Wagenknecht L, Pankow J, Coresh J, and Brancati FL. 2010. "Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults." *The New England Journal of Medicine* 362 (9): 800–811.

Semeval-2015 task 14: Analysis of clinical text. 2015. http://alt.qcri.org/semeval2015/task14/.

Shahar Y, Goren-Bar D, Boaz D, and Tahan G. 2006. "Distributed, Intelligent, Interactive Visualization and Exploration of Time-Oriented Clinical Data and Their Abstractions." *Artificial Intelligence in Medicine* 38 (2): 115–35.

Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, and Cooper GF. 2005. "Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base." *Methods of Information in Medicine* 30 (4): 241–55.

Singer DE, Coley CM, Samet JH, and Nathan DM. 1989. "Tests of Glycemia in Diabetes Mellitus. Their Use in Establishing a Diagnosis and in Treatment." *Annals of Internal Medicine* 110 (2): 125–37.

Singh S, and Vajirkar P. 2003. "Context-Aware Data Mining Using Ontologies." *Proceedings of the 22nd International Conference on Conceptual Modeling*, August, 405–18.

Smith M, Saunders R, Stuckhardt L, and McGinnis JM. 2013. "Best Care at Lower Cost: The Path to Continuously Learning Health Care in America." *National Academic Press.*

Sonnenberg FA, Liu B, Feinberg JE, Kulikowski CA, and Johnson S. 2012. "Clinical Threading: Problem-Oriented Visual Summaries of Clinical Data." *AMIA 2012 Symposium* 353 (23): 2433–41.

Stead WW, and Lin HS. 2009. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. The National Academies Collection: Reports Funded by National Institutes of Health. Washington (DC): National Academies Press (US).

Stewart BA, Fernandes S, Rodriguez-Huertas E, and Landzberg M. 2010. "A Preliminary Look at Duplicate Testing Associated with Lack of Electronic Health Record Interoperability for Transferred Patients." *Journal of the American Medical Informatics Association : JAMIA* 17 (3): 341–44.

Styler W, Bethard S, and Finan S. 2014. "Temporal Annotation in the Clinical Domain." In *ACL*.

Suermondt HJ, Tang PC, Strong PC, Young CY, and Annevelink J. 1993. "Automated Identification of Relevant Patient Information in a Physician's Workstation." *Symposium on Computer Applications in Medical Care*, 229–32.

Sun W, Rumshisky A, and Uzuner Ö. 2013. "Temporal Reasoning over Clinical Text: The State of the Art." *Journal of the American Medical Inform. Assoc*. 20(5):814-819.

Tao C, Song D, Sharma D, and Chute CG. 2013. "Semantator: Semantic Annotator for Converting Biomedical Text to Linked Data." *Journal of Biomedical Informatics* 46 (5): 882–93.

Thornton DJ, Schold JD, Venkateshaiah L, and Lander B. 2013. "Prevalence of Copied Information by Attendings and Residents in Critical Care Progress Notes." *Critical Care Medicine* 41 (2): 382–88.

Thyvalikakath TP, Dziabiak MP, Johnson R, Torres-Urquidy MH, Acharya A, Yabes J, and Schleyer TK. 2014. "Advancing Cognitive Engineering Methods to Support User Interface Design for Electronic Health Records." *International Journal of Medical Informatics* 83 (4): 292–302.

Tversky A. 1977. "Features of Similarity." *Psychological Review* 84 (4): 327–52.

Unertl KM, Weinger MB, Johnson KB, and Lorenzi NM. 2009. "Describing and Modeling Workflow and Information Flow in Chronic Disease Care." *Journal of the American Medical Informatics Association* 16 (6): 826–36.

van der Meulen M, Logie RH, Freer Y, Sykes C, Mcintosh N, and Hunter J. 2010. "When a Graph Is Poorer than 100 Words: A Comparison of Computerised Natural Language Generation, Human Generated Descriptions and Graphical Displays in Neonatal Intensive Care." *Applied Cognitive Psychology* 24 (1): 77–89.

Van Vleck TT, and Elhadad N. 2010. "Corpus-Based Problem Selection for EHR Note Summarization." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium* 2010: 817–21.

Van Vleck TT, Wilcox A, Stetson PD, Johnson SB, and Elhadad N. 2008. "Content and Structure of Clinical Problem Lists: A Corpus Analysis." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, 753–57.

van Walraven, C. 2003. "Population-Based Study of Repeat Laboratory Testing." *Clinical Chemistry* 49 (12): 1997–2005.

Verspoor K, Dvorkin D, and Cohen KB. 2009. "Ontology Quality Assurance through Analysis of Term Transformations." *Bioinformatics (Oxford, England)* 25: 77–84.

Wallach HM, Murray I, Salakhutdinov R, and Mimno D. 2009. "Evaluation Methods for Topic Models," In *ICML*. 1105–12.

Wang X, Chused A, Elhadad N, Friedman C, and Markatou M. 2008. "Automated Knowledge Acquisition from Clinical Narrative Reports." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, 783–87.

Wang X, Sontag D, and Wang F. 2014. "Unsupervised Learning of Disease Progression Models." *Knowledge Discovery and Data Mining*.

Warner JL, and Alterovitz G. 2012. "Phenome Based Analysis as a Means for Discovering Context Dependent Clinical Reference Ranges" *AMIA Symposium AMIA Symposium*: 1441.

Weber GM, and Kohane IS. 2013. "Extracting Physician Group Intelligence from Electronic Health Records to Support Evidence Based Medicine." *PLoS ONE* 8 (5): e64933.

Wei WQ, and Denny JC. 2015. "Extracting Research-Quality Phenotypes from Electronic Health Records to Support Precision Medicine." *Genome Medicine* 7 (1): 41.

Wei WQ, Cui T, Jiang G, Chute CG. 2010. "A High Throughput Semantic Concept Frequency Based Approach for Patient Identification: A Case Study Using Type 2 Diabetes Mellitus Clinical Notes," *AMIA Symposium AMIA Symposium*, 857-861.

Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, and Denny JC. 2015. "Combining Billing Codes, Clinical Notes, and Medications from Electronic Health Records Provides Superior Phenotyping Performance." *Journal of the American Medical Informatics Association: JAMIA*.

Wells, BJ, Chagin KM, Nowacki AS, and Kattan MW. 2013. "Strategies for Handling Missing Data in Electronic Health Record Derived Data." In *EGEMS (Washington, DC)* 1 (3): 1035.

Were MC, Shen C, Bwana M, Emenyonu N, Musinguzi N, Nkuyahaga F, Kembabazi A, and Tierney WM. 2010. "Creation and Evaluation of EMR-Based Paper Clinical Summaries to Support HIV-Care in Uganda, Africa." *International Journal of Medical Informatics* 79 (2): 90–96.

West VL, Borland D, and Hammond WE. 2014. "Innovative Information Visualization of Electronic Health Record Data: A Systematic Review." *Journal of the American Medical Informatics Association : JAMIA*, 22(2):330-9.

Wilcox AB, Jones SS, Dorr DA, Cannon W, Burns L, Radican K, Christensen K, et al. 2005. "Use and Impact of a Computer-Generated Patient Summary Worksheet for Primary Care." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, 824–28.

Wrenn JO, Stein DM, Bakken S, and Stetson PD. 2010. "Quantifying Clinical Narrative Redundancy in an Electronic Health Record." *Journal of the American Medical Informatics Association: JAMIA* 17 (1): 49–53.

Wu G, Chang E, and Panda N. 2005. "Formulating Context-Dependent Similarity Functions." In *Proc. 13th Ann. ACM Int'l Conf. Multimedia (Multimedia '05)*.

Wu ST, Juhn YJ, Sohn S, and Liu H. 2014. "Patient-Level Temporal Aggregation for Text-Based Asthma Status Ascertainment." *Journal of the American Medical Informatics Association : JAMIA* 21 (5): 876–84.

Wu Z, and Palmer M. 1994. "Verbs Semantics and Lexical Selection." In *32ndAnnual Meeting of the Association for Computational Linguistics*, 133–38.

Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, and Cai T. 2015. "Toward High-Throughput Phenotyping: Unbiased Automated Feature Extraction and Selection from Knowledge Sources." *Journal of the American Medical Informatics Association : JAMIA*, April.

Zhang R. 2014. "Longitudinal Analysis of New Information Types in Clinical Notes," AMIA Jt Summits Transl Sci Proc; 232-237.

Zhang R, Pakhomov S, McInnes BT, and Melton GB. 2011. "Evaluating Measures of Redundancy in Clinical Texts." *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium* 2011 (July): 1612–20.

Zhang R, Pakhomov S, and Melton GB. 2012. "Automated Identification of Relevant New Information in Clinical Narrative," In *ACM IHI.* 837–42.

Zhi M, Ding EL, Theisen-Toupal J, Whelan J, and Arnaout R. 2013. "The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis." *PLoS ONE* 8 (11): e78962.

Zhou L, and Hripcsak G. 2007. "Temporal Reasoning with Medical Data–a Review with Emphasis on Medical Natural Language Processing." *Journal of Biomedical Informatics* 40 (2): 183–202.

Zhou L, Parsons S, and Hripcsak G. 2008. "The Evaluation of a Temporal Reasoning System in Processing Clinical Discharge Summaries." *Journal of the American Medical Informatics Association* 15 (1): 99.

Appendix A: List of Correlations between Laboratory Value and Measurement Gap

Table 1: The non-linear correlation between (i) measurement gap and numerical value and (ii) measurement gap and change in numerical value, for 70 laboratory tests. The measurement gap and numerical value correlations are calculated on three different time scales (full time scale, short time scale, long time scale). The following results are reported: the mutual information (MI), the confidence intervals for the MI results (CI of MI) and the MI total (which is the confidence interval subtracted from the MI). In addition, we calculated the ergodicity for every laboratory test measured via the Kolgomorov-Smirnov statistic and the result was 1, the same for all laboratory tests, and is therefore not shown.

Nonlinear Correlations between Laboratory Test Numerical Values and Time Between Measurement												
Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
25-OH Vitamin D	0.043	0.035	0.007	1.847	1.842	0.005	0.052	0.042	0.010	0.059	0.032	0.027
Albumin	0.154	0.001	0.153	0.045	0.004	0.041	0.051	0.001	0.050	0.020	0.001	0.019
Alkaline Phosphatase	0.038	0.001	0.037	0.012	0.005	0.007	0.023	0.002	0.021	0.010	0.001	0.008
ALT	0.037	0.001	0.036	0.012	0.005	0.007	0.014	0.002	0.012	0.009	0.001	0.008
Amylase	0.089	0.017	0.072	0.077	0.060	0.017	0.032	0.019	0.012	0.052	0.017	0.034

Continued on next page

Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
AST	0.062	0.001	0.061	0.016	0.005	0.011	0.019	0.002	0.018	0.023	0.001	0.022
Base Excess Arterial	0.029	0.017	0.012	0.025	0.019	0.007	0.102	0.085	0.017	0.058	0.020	0.038
Basophiles %	0.002	0.000	0.001	0.003	0.002	0.001	0.002	0.000	0.002	0.004	0.001	0.003
Bicarbonate Arterial	0.040	0.001	0.040	0.024	0.001	0.023	0.010	0.001	0.009	0.010	0.001	0.010
Bilirubin Direct	0.083	0.001	0.082	0.022	0.005	0.017	0.020	0.002	0.019	0.050	0.003	0.047
Bilirubin Total	0.063	0.001	0.061	0.019	0.006	0.013	0.013	0.002	0.011	0.021	0.002	0.019
Blood Protein	0.120	0.001	0.119	0.050	0.004	0.046	0.034	0.001	0.032	0.022	0.001	0.021
BNP	0.100	0.075	0.026	0.514	0.520	-0.005	0.097	0.082	0.015	0.071	0.066	0.005
BUN	0.057	0.001	0.056	0.016	0.001	0.015	0.035	0.001	0.034	0.031	0.001	0.030
Calcium	0.110	0.001	0.109	0.020	0.001	0.019	0.016	0.001	0.015	0.011	0.001	0.010
Calcium Ionized	0.023	0.012	0.011	0.024	0.014	0.010	0.048	0.036	0.012	0.025	0.011	0.014
CEA	0.093	0.066	0.027	0.870	0.897	-0.027	0.103	0.078	0.025	0.067	0.062	0.005
Chloride	0.039	0.000	0.038	0.011	0.001	0.010	0.012	0.001	0.012	0.006	0.000	0.006
Cholesterol	0.024	0.002	0.022	0.051	0.026	0.025	0.016	0.002	0.014	0.021	0.002	0.019
CK	0.034	0.004	0.030	0.023	0.014	0.008	0.011	0.006	0.005	0.019	0.005	0.014

Continued on next page

Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
CK-MB	0.106	0.112	-0.006	0.177	0.152	0.025	0.147	0.160	-0.013	0.102	0.098	0.005
Creatinine	0.060	0.001	0.059	0.025	0.002	0.024	0.033	0.001	0.032	0.045	0.001	0.043
CRP High Sensitivity	0.069	0.034	0.035	0.636	0.584	0.052	0.060	0.039	0.021	0.046	0.033	0.012
ESR	0.052	0.014	0.038	0.352	0.393	-0.041	0.048	0.015	0.033	0.020	0.014	0.006
Ferritin	0.066	0.016	0.050	0.400	0.390	0.010	0.066	0.019	0.048	0.037	0.016	0.021
Folic Acid	0.109	0.088	0.022	1.375	1.432	-0.056	0.123	0.096	0.028	0.096	0.077	0.019
Glucose Whole Blood	0.034	0.001	0.033	0.023	0.002	0.022	0.012	0.001	0.010	0.015	0.001	0.014
HCT	0.118	0.001	0.117	0.021	0.001	0.019	0.050	0.001	0.049	0.012	0.001	0.011
HDL	0.011	0.003	0.008	0.252	0.247	0.006	0.010	0.003	0.007	0.009	0.002	0.007
Hemoglobin	0.111	0.001	0.110	0.017	0.001	0.015	0.051	0.001	0.050	0.014	0.001	0.013
Hemoglobin A_{1c}	0.057	0.005	0.052	0.518	0.540	-0.022	0.058	0.005	0.053	0.023	0.004	0.019
Homocysteine	0.206	0.159	0.048	1.521	1.578	-0.058	0.246	0.192	0.054	0.111	0.115	-0.003
Iron	0.037	0.021	0.016	0.485	0.485	0.000	0.039	0.023	0.016	0.030	0.020	0.010
LDH	0.058	0.006	0.052	0.030	0.023	0.007	0.028	0.008	0.020	0.021	0.006	0.015
LDL	0.015	0.003	0.012	0.280	0.307	-0.026	0.015	0.003	0.012	0.015	0.003	0.013

Continued on next page

Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
Lipase	0.127	0.018	0.109	0.077	0.060	0.017	0.034	0.016	0.018	0.052	0.017	0.034
Lymphocytes %	0.062	0.002	0.060	0.013	0.009	0.004	0.029	0.002	0.027	0.015	0.002	0.013
Magnesium	0.020	0.002	0.018	0.016	0.002	0.013	0.031	0.005	0.025	0.015	0.002	0.013
MCH	0.009	0.001	0.008	0.003	0.002	0.002	0.011	0.001	0.010	0.014	0.000	0.013
MCHC	0.010	0.000	0.009	0.005	0.001	0.004	0.010	0.001	0.010	0.012	0.000	0.011
MCV	0.008	0.001	0.007	0.003	0.002	0.001	0.011	0.001	0.010	0.041	0.001	0.040
Microalbumin /Creatinine Ratio	0.034	0.022	0.011	1.319	1.249	0.070	0.039	0.030	0.009	0.029	0.021	0.008
Monocytes %	0.017	0.001	0.016	0.012	0.007	0.005	0.014	0.002	0.012	0.009	0.002	0.006
MPV	0.009	0.001	0.009	0.003	0.002	0.002	0.010	0.001	0.009	0.029	0.001	0.028
NRBC %	0.059	0.017	0.042	0.027	0.020	0.006	0.120	0.058	0.063	0.072	0.013	0.060
NRBC Absolute	0.022	0.014	0.008	0.018	0.016	0.001	0.078	0.063	0.015	0.016	0.013	0.003
PCO_2, Arterial	0.014	0.011	0.003	0.014	0.011	0.002	0.078	0.066	0.013	0.036	0.010	0.027
PCO_2, Venous	0.089	0.053	0.036	0.139	0.107	0.032	0.098	0.087	0.011	0.082	0.040	0.042
pH Arterial	0.016	0.006	0.010	0.016	0.007	0.009	0.056	0.048	0.008	0.027	0.004	0.023
pH Urine	0.006	0.001	0.005	0.007	0.008	-0.001	0.006	0.001	0.005	0.006	0.003	0.003

Continued on next page

Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
pH Venous	0.073	0.046	0.027	0.123	0.092	0.031	0.076	0.072	0.004	0.062	0.040	0.022
Phosphorus	0.026	0.002	0.024	0.012	0.003	0.009	0.019	0.004	0.015	0.015	0.002	0.013
Platelet Count	0.041	0.001	0.040	0.019	0.002	0.017	0.021	0.001	0.020	0.027	0.001	0.026
PO_2 Arterial	0.055	0.011	0.043	0.035	0.012	0.023	0.084	0.062	0.022	0.031	0.011	0.021
PO_2 Venous	0.197	0.056	0.140	0.128	0.096	0.032	0.121	0.098	0.023	0.211	0.054	0.157
Potassium	0.035	0.001	0.034	0.027	0.001	0.025	0.012	0.001	0.011	0.021	0.001	0.020
PSA Screening	0.124	0.022	0.103	0.745	0.773	-0.028	0.130	0.027	0.102	0.061	0.023	0.038
PTH	0.094	0.038	0.056	0.980	1.010	-0.030	0.098	0.044	0.053	0.052	0.033	0.018
RBC	0.104	0.001	0.103	0.020	0.002	0.018	0.049	0.001	0.047	0.012	0.001	0.011
RDW	0.051	0.001	0.050	0.005	0.002	0.003	0.042	0.001	0.040	0.019	0.001	0.018
Sodium	0.043	0.001	0.042	0.012	0.001	0.011	0.008	0.001	0.007	0.005	0.000	0.004
T4	0.061	0.016	0.044	0.560	0.600	-0.040	0.064	0.018	0.045	0.026	0.015	0.011
T4 Free	0.054	0.028	0.027	0.802	0.807	-0.005	0.057	0.031	0.025	0.042	0.026	0.016
TIBC	0.073	0.021	0.052	0.685	0.662	0.023	0.070	0.023	0.047	0.037	0.018	0.019
Triglycerides	0.017	0.003	0.014	0.226	0.215	0.011	0.016	0.003	0.013	0.008	0.003	0.005
Troponin	0.034	0.019	0.014	0.030	0.024	0.006	0.043	0.038	0.005	0.031	0.020	0.012
TSH	0.059	0.005	0.054	0.175	0.167	0.008	0.060	0.006	0.054	0.046	0.005	0.041
Uric Acid	0.029	0.006	0.023	0.045	0.025	0.020	0.023	0.008	0.015	0.031	0.006	0.025
Vitamin B_{12}	0.066	0.044	0.022	1.224	1.224	0.000	0.072	0.048	0.024	0.052	0.037	0.015

Continued on next page

Laboratory Test	MI (full)	CI of MI (full)	MI total (full)	MI (short)	CI of MI (short)	MI total (short)	MI (long)	CI of MI (long)	MI total (long)	MI (dv/dt)	CI of MI (dv/dt)	Total MI (dv/dt)
WBC	0.057	0.001	0.056	0.010	0.002	0.008	0.018	0.001	0.017	0.008	0.001	0.007

Table 2: The linear correlations (LC) and p-values of 70 laboratory tests, calculated on both the short and long time scale. A positive linear correlation indicates that a higher laboratory test value is correlated to a longer time until next measurement and a negative linear correlation indicates that a higher laboratory test value is correlated to a shorter time until next measurement. The linear correlation across the full time scale is not shown as it was very close to 0 for every laboratory test (all p-values for the full time scale linear correlations were statistically significant - except for the Vitamin B_{12} test.)

Linear Correlations between Laboratory Test Values and Measurement Gap				
Laboratory Test	LC (Short)	p-value of LC (Short)	LC (Long)	p-value of LC (Long)
25-OH Vitamin D	0.200	2.28E-01	0.053	2.63E-05
Albumin	-0.029	2.87E-11	0.296	0.00E+00
Alkaline Phosphatase	0.033	2.44E-14	-0.177	0.00E+00
ALT	-0.050	4.52E-31	-0.129	0.00E+00
Amylase	-0.083	3.58E-08	-0.111	3.65E-36
AST	-0.108	3.74E-135	-0.148	0.00E+00
Base Excess Arterial	0.066	3.58E-13	-0.123	3.06E-10
Basophiles %	0.020	1.15E-01	-0.048	2.32E-23
Bicarbonate Arterial	0.181	0.00E+00	0.103	0.00E+00
Bilirubin Direct	-0.050	4.30E-28	-0.169	0.00E+00
Bilirubin Total	-0.055	3.54E-32	-0.121	0.00E+00
Blood Protein	0.030	3.09E-11	0.195	0.00E+00
BNP	0.037	5.10E-01	-0.127	2.04E-12
BUN	-0.156	0.00E+00	-0.208	0.00E+00
Calcium	0.098	0.00E+00	0.115	0.00E+00
Calcium Ionized	0.057	3.87E-11	-0.071	4.57E-08
CEA	0.023	9.08E-01	-0.186	1.66E-16

Continued on next page

Laboratory Test	LC (Short)	p-value of LC (Short)	LC (Long)	p-value of LC (Long)
Chloride	0.033	2.75E-39	0.022	2.39E-24
Cholesterol	-0.160	7.80E-59	-0.005	1.01E-01
CK	0.011	1.28E-01	-0.027	2.94E-09
CK-MB	-0.003	9.14E-01	-0.101	2.26E-02
Creatinine	-0.208	0.00E+00	-0.186	0.00E+00
CRP High Sensitivity	-0.084	1.53E-01	-0.118	2.96E-24
ESR	0.088	2.69E-02	-0.237	8.92E-250
Ferritin	0.067	2.06E-01	-0.215	2.61E-150
Folic Acid	-0.138	1.86E-01	-0.072	7.22E-05
Glucose Whole Blood	-0.150	0.00E+00	-0.093	0.00E+00
HCT	0.142	0.00E+00	0.276	0.00E+00
HDL	-0.058	7.28E-02	0.087	1.04E-158
Hemoglobin	0.118	0.00E+00	0.282	0.00E+00
Hemoglobin A_{1c}	0.036	4.63E-01	-0.193	0.00E+00
Homocysteine	-0.163	3.90E-01	-0.275	8.87E-27
Iron	0.135	1.08E-02	0.106	1.71E-34
LDH	-0.004	6.34E-01	-0.140	6.73E-159
LDL	-0.001	9.70E-01	0.015	7.80E-06
Lipase	-0.068	9.19E-06	-0.152	1.47E-61
Lymphocytes %	0.022	1.44E-04	0.149	0.00E+00
Magnesium	-0.083	7.73E-149	0.167	1.80e-322
MCH	-0.009	2.70E-04	0.025	7.13E-32
MCHC	-0.053	1.91E-102	0.100	0.00E+00
MCV	0.018	1.75E-13	-0.022	1.55E-25
Microalbumin/Creatinine Ratio	0.004	9.84E-01	-0.074	9.10E-10
Monocytes %	0.061	7.61E-25	-0.091	1.01E-224
MPV	-0.003	3.15E-01	0.078	6.38E-257

Continued on next page

Laboratory Test	LC (Short)	p-value of LC (Short)	LC (Long)	p-value of LC (Long)
NRBC %	0.038	1.66E-05	-0.087	3.86E-10
NRBC Absolute	-0.023	1.14E-02	-0.044	8.37E-03
PCO_2 Arterial	-0.021	1.50E-03	0.061	4.57E-05
PCO_2, Venous	0.190	5.69E-20	0.091	3.18E-07
pH Arterial	0.100	3.74E-51	-0.109	3.64E-13
pH Urine	0.004	6.79E-01	-0.013	1.71E-04
pH Venous	0.239	5.04E-31	-0.065	2.63E-04
Phosphorus	-0.042	5.40E-36	-0.092	3.31E-136
Platelet Count	0.148	0.00E+00	0.029	1.53E-41
PO_2 Arterial	-0.153	3.44E-116	-0.094	3.24E-10
PO_2 Venous	-0.038	6.72E-02	-0.047	8.25E-03
Potassium	-0.071	8.56E-199	-0.044	6.33E-102
PSA Screening	0.174	2.95E-01	-0.341	2.29E-196
PTH	0.093	2.99E-01	-0.252	2.65E-92
RBC	0.145	0.00E+00	0.265	0.00E+00
RDW	-0.024	3.19E-20	-0.259	0.00E+00
Sodium	0.067	1.44E-182	0.102	0.00E+00
T4	-0.047	4.06E-01	-0.109	1.08E-44
T4 Free	0.039	6.30E-01	-0.107	2.43E-24
TIBC	0.034	5.49E-01	0.194	5.13E-104
Triglycerides	0.008	7.86E-01	-0.138	0.00E+00
Troponin	0.040	4.84E-05	-0.023	1.67E-01
TSH	0.026	4.27E-01	-0.226	0.00E+00
Uric Acid	-0.085	4.32E-18	-0.118	7.23E-111
Vitamin B_{12}	-0.150	8.43E-02	0.061	1.61E-06
WBC	-0.097	0.00E+00	-0.002	3.31E-01

Appendix B: Phenome Model Inference

Derivations of the Collapsed Gibbs Sampler Equations

$$\int\int\int\int\int p(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\epsilon},\boldsymbol{\zeta},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\iota},\boldsymbol{\kappa},\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y})d\beta_r d\eta d\theta d\iota d\kappa = \prod_{r=1}^{R}\int p(\beta_r \mid \alpha)\prod_{i=1}^{I_r} p(\gamma_{i,r} \mid \beta_r)\prod_{n=1}^{N_r} p(\delta_{n,r} \mid \beta_r)$$
$$\prod_{o=1}^{O_r} p(\epsilon_{o,r} \mid \beta_r)\prod_{m=1}^{M_r} p(\zeta_{m,r} \mid \beta_r)d\beta_r$$
$$\prod_{r=1}^{R}\int \prod_{i=1}^{I_r} p(v_{i,r} \mid \eta_{\gamma_{i,r}})\prod_{p=1}^{P} p(\eta)d\eta$$
$$\prod_{r=1}^{R}\int \prod_{n=1}^{N_r} p(w_{n,r} \mid \theta_{\delta_{n,r}})\prod_{p=1}^{P} p(\theta)d\theta$$
$$\prod_{r=1}^{R}\int \prod_{o=1}^{O_r} p(x_{o,r} \mid \iota_{\epsilon_{o,r}})\prod_{p=1}^{P} p(\iota)d\iota$$
$$\prod_{r=1}^{R}\int \prod_{m=1}^{M_r} p(y_{m,r} \mid \kappa_{\zeta_{m,r}})\prod_{p=1}^{P} p(\kappa)d\kappa$$

where $\boldsymbol{\beta}_{-\boldsymbol{r}}$ indicates all β variables except for that indexed by r. Here the notation $c^{p}_{r,I,(\cdot)}$ indicates the total count of observations from

record r of the subplate I have an assignment to phenotype p.

$$
\begin{aligned}
&\prod_{r=1}^{R}\int p(\beta_r \mid \alpha)\prod_{i=1}^{I_r}p(\gamma_{i,r}\mid\beta_r)\prod_{n=1}^{N_r}p(\delta_{n,r}\mid\beta_r)\prod_{o=1}^{O_r}p(\epsilon_{o,r}\mid\beta_r)\prod_{m=1}^{M_r}p(\zeta_{m,r}\mid\beta_r)d\beta_r\\
&=\prod_{r=1}^{R}\int\frac{\Gamma(\sum_{p=1}^{P}\alpha_p)}{\prod_{p=1}^{P}\Gamma(\alpha_p)}\prod_{p=1}^{P}\beta_{r,p}^{\alpha_p-1}\prod_{i=1}^{I_r}\prod_{p=1}^{P}\beta_{r,p}^{\gamma_{i,r,p}}\prod_{n=1}^{N_r}\prod_{p=1}^{P}\beta_{r,p}^{\delta_{i,r,p}}\prod_{o=1}^{O_r}\prod_{p=1}^{P}\beta_{r,p}^{\epsilon_{i,r,p}}\prod_{m=1}^{M_r}\prod_{p=1}^{P}\beta_{r,p}^{\zeta_{i,r,p}}d\beta_r\\
&=\prod_{r=1}^{R}\int\frac{\Gamma(\sum_{p=1}^{P}\alpha_p)}{\prod_{p=1}^{P}\Gamma(\alpha_p)}\prod_{p=1}^{P}\beta_{r,p}^{\alpha_p-1}\prod_{p=1}^{P}\beta_{r,p}^{C^p_{r,1:E,1:I_r}}\prod_{p=1}^{P}\beta_{r,p}^{C^p_{r,1:F,1:N_r}}\prod_{p=1}^{P}\beta_{r,p}^{C^p_{r,1:G,1:O_r}}\prod_{p=1}^{P}\beta_{r,p}^{C^p_{r,1:L,1:M_r}}d\beta_r\\
&=\prod_{r=1}^{R}\int\frac{\Gamma(\sum_{p=1}^{P}\alpha_p)}{\prod_{p=1}^{P}\Gamma(\alpha_p)}\prod_{p=1}^{P}\beta_{r,p}^{\alpha_p-1+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r}}d\beta_r\\
&=\frac{\Gamma(\sum_{p=1}^{P}\alpha_p)}{\prod_{p=1}^{P}\Gamma(\alpha_p)}\prod_{r=1}^{R}\frac{\prod_{p=1}^{P}\Gamma(\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}{\Gamma(\sum_{p=1}^{P}\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}\\
&\qquad\cdot\int\frac{\Gamma(\sum_{p=1}^{P}\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}{\prod_{p=1}^{P}\Gamma(\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}\\
&\qquad\cdot\prod_{p=1}^{P}\beta_{r,p}^{\alpha_p-1+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r}}d\beta_r\\
&=\frac{\Gamma(\sum_{p=1}^{P}\alpha_p)}{\prod_{p=1}^{P}\Gamma(\alpha_p)}\prod_{r=1}^{R}\frac{\prod_{p=1}^{P}\Gamma(\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}{\Gamma(\sum_{p=1}^{P}\alpha_p+C^p_{r,1:E,1:I_r}+C^p_{r,1:F,1:N_r}+C^p_{r,1:G,1:O_r}+C^p_{r,1:L,1:M_r})}
\end{aligned}
$$

η

$$
\begin{aligned}
&\textstyle\prod_{r=1}^{R}\int\prod_{i=1}^{I_r}p(v_{i,r}\mid\eta_{\gamma_{i,r}})\prod_{p=1}^{P}p(\eta_p\mid\mu)d\eta\\
&=\textstyle\int\prod_{r=1}^{R}\prod_{i=1}^{I}\prod_{e=1}^{E}\eta_{\gamma_{i,r,e}}^{v_{i,r}}\prod_{p=1}^{P}\frac{\Gamma(\sum_{e=1}^{E}\mu_e)}{\prod_{e=1}^{E}\Gamma(\mu_e)}\prod_{e=1}^{E}\eta_{p,e}^{\mu_e-1}d\eta\\
&=\textstyle\int\prod_{p=1}^{P}\prod_{e=1}^{E}\eta_{p,e}^{C^p_{1:R,e,1:I_r}}\prod_{p=1}^{P}\frac{\Gamma(\sum_{e=1}^{E}\mu_e)}{\prod_{e=1}^{E}\Gamma(\mu_e)}\prod_{e=1}^{E}\eta_{p,e}^{\mu_e-1}d\eta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{e=1}^{E}\mu_e)}{\prod_{e=1}^{E}\Gamma(\mu_e)}\int\prod_{e=1}^{E}\eta_{p,e}^{\mu_e-1+C^p_{1:R,e,1:I_r}}d\eta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{e=1}^{E}\mu_e)}{\prod_{e=1}^{E}\Gamma(\mu_e)}\int\frac{\prod_{e=1}^{E}\Gamma(\mu_e+C^p_{1:R,e,1:I_r})}{\Gamma(\sum_{e=1}^{E}\mu_e+C^p_{1:R,e,1:I_r})}\frac{\Gamma(\sum_{e=1}^{E}\mu_e+C^p_{1:R,e,1:I_r})}{\prod_{e=1}^{E}\Gamma(\mu_e+C^p_{1:R,e,1:I_r})}\prod_{e=1}^{E}\eta_{p,e}^{\mu_e-1+C^p_{1:R,e,1:I_r}}d\eta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{e=1}^{E}\mu_e)}{\prod_{e=1}^{E}\Gamma(\mu_e)}\frac{\prod_{e=1}^{E}\Gamma(\mu_e+C^p_{1:R,e,1:I_r})}{\Gamma(\sum_{e=1}^{E}\mu_e+C^p_{1:R,e,1:I_r})}
\end{aligned}
$$

θ

$$
\begin{aligned}
&\textstyle\prod_{r=1}^{R}\int\prod_{n=1}^{N_r}p(w_{n,r}\mid\theta_{\delta_{n,r}})\prod_{p=1}^{P}p(\theta_p\mid\nu)d\theta\\
&=\textstyle\int\prod_{r=1}^{R}\prod_{n=1}^{N}\prod_{f=1}^{F}\theta_{\delta_{n,r,f}}^{w_{n,r}}\prod_{p=1}^{P}\frac{\Gamma(\sum_{f=1}^{F}\nu_f)}{\prod_{f=1}^{F}\Gamma(\nu_f)}\prod_{f=1}^{F}\theta_{p,f}^{\nu_f-1}d\theta\\
&=\textstyle\int\prod_{p=1}^{P}\prod_{f=1}^{F}\theta_{p,f}^{C^p_{1:R,f,1:N_r}}\prod_{p=1}^{P}\frac{\Gamma(\sum_{f=1}^{F}\nu_f)}{\prod_{f=1}^{F}\Gamma(\nu_f)}\prod_{f=1}^{F}\theta_{p,f}^{\nu_f-1}d\theta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{f=1}^{F}\nu_f)}{\prod_{f=1}^{F}\Gamma(\nu_f)}\int\prod_{f=1}^{F}\theta_{p,f}^{\nu_f-1+C^p_{1:R,f,1:N_r}}d\theta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{f=1}^{F}\nu_f)}{\prod_{f=1}^{F}\Gamma(\nu_f)}\int\frac{\prod_{f=1}^{F}\Gamma(\nu_f+C^p_{1:R,f,1:N_r})}{\Gamma(\sum_{e=1}^{E}\nu_f+C^p_{1:R,f,1:N_r})}\frac{\Gamma(\sum_{f=1}^{F}\nu_f+C^p_{1:R,f,1:N_r})}{\prod_{f=1}^{F}\Gamma(\nu_f+C^p_{1:R,f,1:N_r})}\prod_{f=1}^{F}\theta_{p,f}^{\nu_f-1+C^p_{1:R,f,1:N_r}}d\theta\\
&=\textstyle\prod_{p=1}^{P}\frac{\Gamma(\sum_{f=1}^{F}\nu_f)}{\prod_{f=1}^{F}\Gamma(\nu_f)}\frac{\prod_{f=1}^{F}\Gamma(\nu_f+C^p_{1:R,f,1:N_r})}{\Gamma(\sum_{f=1}^{F}\nu_f+C^p_{1:R,f,1:N_r})}
\end{aligned}
$$

Equations for ι and κ $\prod_{p=1}^{P} \frac{\Gamma(\sum_{g=1}^{G} \xi_g)}{\prod_{g=1}^{G} \Gamma(\xi_g)} \frac{\prod_{g=1}^{G} \Gamma(\xi_g + C^p_{1:R,g,1:O_r})}{\Gamma(\sum_{g=1}^{G} \xi_g + C^p_{1:R,g,1:O_r})} \prod_{p=1}^{P} \frac{\Gamma(\sum_{l=1}^{L} \pi_l)}{\prod_{l=1}^{L} \Gamma(\pi_l)} \frac{\prod_{l=1}^{L} \Gamma(\pi_l + C^p_{1:R,l,1:M_r})}{\Gamma(\sum_{l=1}^{L} \pi_l + C^p_{1:R,l,1:M_r})}$

Collapsed Model Likelihood $\frac{\Gamma(\sum_{p=1}^{P} \alpha_p)}{\prod_{p=1}^{P} \Gamma(\alpha_p)} \prod_{r=1}^{R} \frac{\prod_{p=1}^{P} \Gamma(\alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r})}{\Gamma(\sum_{p=1}^{P} \alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r})}$
$\prod_{p=1}^{P} [\frac{\Gamma(\sum_{e=1}^{E} \mu_e)}{\prod_{e=1}^{E} \Gamma(\mu_e)} \frac{\prod_{e=1}^{E} \Gamma(\mu_e + C^p_{1:R,e,1:I_r})}{\Gamma(\sum_{e=1}^{E} \mu_e + C^p_{1:R,e,1:I_r})} \frac{\Gamma(\sum_{f=1}^{F} \nu_f)}{\prod_{f=1}^{F} \Gamma(\nu_f)} \frac{\prod_{f=1}^{F} \Gamma(\nu_f + C^p_{1:R,f,1:N_r})}{\Gamma(\sum_{f=1}^{F} \nu_f + C^p_{1:R,f,1:N_r})}$
$\frac{\Gamma(\sum_{g=1}^{G} \xi_g)}{\prod_{g=1}^{G} \Gamma(\xi_g)} \frac{\prod_{g=1}^{G} \Gamma(\xi_g + C^p_{1:R,g,1:O_r})}{\Gamma(\sum_{g=1}^{G} \xi_g + C^p_{1:R,g,1:O_r})} \frac{\Gamma(\sum_{l=1}^{L} \pi_l)}{\prod_{l=1}^{L} \Gamma(\pi_l)} \frac{\prod_{l=1}^{L} \Gamma(\pi_l + C^p_{1:R,l,1:M_r})}{\Gamma(\sum_{l=1}^{L} \pi_l + C^p_{1:R,l,1:M_r})}]$

Log-Likelihood Conditional probability of $\gamma_{i^*,r^*} \mid \gamma^{-(i^*,r^*)}, \ldots$

Isolate terms that contain γ_i, r

$$p(\gamma_{i^*,r^*} \mid \gamma^{-(i^*,r^*)}, \ldots) = \frac{\prod_{p=1}^{P} \Gamma(\alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r})}{\Gamma(\sum_{p=1}^{P} \alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r})} \prod_{p=1}^{P} \frac{\prod_{e=1}^{E} \Gamma(\mu_e + C^p_{1:R,e,1:I_r})}{\Gamma(\sum_{e=1}^{E} \mu_e + C^p_{1:R,e,1:I_r})}$$

Condition on p=p*

$p(\gamma_{i^*,r^*} = p^* \mid \gamma^{-(i^*,r^*)}, \ldots) =$
$\frac{\prod_{p \neq p^*} \Gamma(\alpha_p + C^{p,-(r^*,i^*)}_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r}) \Gamma(\alpha_{p^*} + C^{p^*,-(r^*,i^*)}_{r,1:E,1:I_r} + C^{P^*}_{r,1:F,1:N_r} + C^{P^*}_{r,1:G,1:O_r} + C^{P^*}_{r,1:L,1:M_r} + 1)}{\Gamma((\sum_{p=1}^{P} \alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r}) + 1)}$
$\cdot \frac{\prod_{p \neq p^*} \Gamma(\mu_{e^*} + C^{p,-(r^*,i^*)}_{1:R,e^*,1:I_r}) \Gamma(\mu_{e^*} + C^{p^*,-(r^*,i^*)}_{1:R,e^*,1:I_r} + 1)}{\prod_{p \neq p^*} \Gamma(\sum_{e=1}^{E} \mu_e + C^{p,-(r^*,i^*)}_{1:R,e,1:I_r}) \Gamma(1 + \sum_{e=1}^{E} \mu_e + C^{p^*,-(r^*,i^*)}_{1:R,e,1:I_r})}$

$$= \frac{(\alpha_{p^*} + C^{p^*}_{r,1:E,1:I_r} + C^{p^*,-(r^*,n^*)}_{r,1:F,1:N_r} + C^{p^*}_{r,1:G,1:O_r} + C^{p^*}_{r,1:L,1:M_r}) \prod_{p=1}^{P} \Gamma(\alpha_{p^*} + C^{p^*}_{r,1:E,1:I_r} + C^{p^*,-(n^*,r^*)}_{r,1:F,1:N_r} + C^{p^*}_{r,1:G,1:O_r} + C^{p^*}_{r,1:L,1:M_r})}{(\sum_{p=1}^{P} \alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r}) \Gamma(\sum \alpha_p + C^p_{r,1:E,1:I_r} + C^p_{r,1:F,1:N_r} + C^p_{r,1:G,1:O_r} + C^p_{r,1:L,1:M_r})}$$

$$\cdot \frac{(\nu_{f^*} + C^{p,-(n*,r*)}_{1:R:,f,*,1:Nr}) \prod_{p=1}^{P} \Gamma(\nu_{f^*} + C^{p,-(n*,r*)}_{1:R:,f,*,1:Nr})}{(\sum_{f=1}^{F} \nu_f + C^{p,-(n^*,r^*)}_{1:R,f,1:N_r}) \prod_{p=1}^{P} \Gamma(\sum_{f=1}^{F} \nu_f + C^{p^*,-(n^*,r^*)}_{1:R,f,1:N_r})}$$
$$= \frac{(\alpha_{p^*} + C^{p^*}_{r,1:E,1:I_r} + C^{p*,-(r^*,n^*)}_{r,1:F,1:N_r} + C^{p*}_{r,1:G,1:O_r} + C^{p*}_{r,1:L,1:M_r})(\nu_{f^*} + C^{p,-(n*,r*)}_{1:R:,f,*,1:Nr})}{(\sum_{f=1}^{F} \nu_f + C^{p,-(n^*,r^*)}_{1:R,f,1:N_r})}$$

Conditional probability of $\epsilon_{o^*,r^*} \mid \epsilon^{-(o^*,r^*)}$

$$p(\epsilon_{o^*,r^*} = p^* \mid \epsilon^{-(o^*,r^*)}, \ldots) = \frac{(\alpha_{p^*} + C^{p^*}_{r,1:E,1:I_r} + C^{p*}_{r,1:F,1:N_r} + C^{p*,-(o^*,r^*)}_{r,1:G,1:O_r} + C^{p*}_{r,1:L,1:M_r})(\xi_{g^*} + C^{p,-(o*,r*)}_{1:R:,g^*,1:Or})}{(\sum_{g=1}^{G} \xi_g + C^{p,-(o^*,r^*)}_{1:R,g,1:O_r})}$$

Conditional probability of $\zeta_{m^*,r^*} \mid \zeta^{-(m^*,r^*)}$

$$p(\zeta_{m^*,r^*} = p^* \mid \zeta^{-(m^*,r^*)}, \ldots) = \frac{(\alpha_{p^*} + C^{p^*}_{r,1:E,1:I_r} + C^{p*}_{r,1:F,1:N_r} + C^{p*}_{r,1:G,1:O_r} + C^{p*,-(m^*,r^*)}_{r,1:L,1:M_r})(\pi_{l^*} + C^{p,-(m*,r*)}_{1:R:,l^*,1:Mr})}{(\sum_{l=1}^{L} \pi_l + C^{p,-(m^*,r^*)}_{1:R,l,1:M_r})}$$

Now, knowing that $\Gamma(t+1) = t\Gamma(t)$...

$$\frac{(\alpha_{p^*} + C^{p^*,-(r*,i*)}_{r,1:E,1:I_r} + C^{P*}_{r,1:F,1:N_r} + C^{P*}_{r,1:G,1:O_r} + C^{P*}_{r,1:L,1:M_r}) \prod_{p=1}^{P} \Gamma(\alpha_{p^*} + C^{p^*,-(r*,i*)}_{r,1:E,1:I_r} + C^{P*}_{r,1:F,1:N_r} + C^{P*}_{r,1:G,1:O_r} + C^{P*}_{r,1:L,1:M_r})}{(\sum_{p=1}^{P} \alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r}) \Gamma(\sum_{p=1}^{P} \alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r})}$$

$$\cdot \frac{(\mu_{e^*} + C^{p,-(r*,i*)}_{1:R,e^*,1:I_r}) \prod_{p=1}^{P} \Gamma(\mu_{e^*} + C^{p^*,-(r^*,i^*)}_{1:R,e^*,1:I_r})}{(\sum_{e=1}^{E} \mu_e + C^{p,-(r^*,i^*)}_{1:R,e,1:I_r}) \Gamma(\sum_{e=1}^{E} \mu_e + C^{p,-(r^*,i^*)}_{1:R,e,1:I_r})}$$

We can remove everything that is a product over all Ps (because it's not specific to p*)

$$= \frac{(\alpha_{p^*} + C^{p^*,-(r*,i*)}_{r,1:E,1:I_r} + C^{p*}_{r,1:F,1:N_r} + C^{p*}_{r,1:G,1:O_r} + C^{p*}_{r,1:L,1:M_r})(\mu_{e^*} + C^{p,-(r*,i*)}_{1:R,e^*,1:I_r})}{(\sum_{e=1}^{E} \mu_e + C^{p^*,-(r*,i*)}_{1:R,e,1:I_r})}$$

Conditional probability of $\delta_{n*,r*} \mid \delta^{-(n^*,r^*)}$

$$p(\delta_{n*,r*} \mid \delta^{-(n^*,r^*)}, ...) = \frac{\prod_{p=1}^{P} \Gamma(\alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r})}{\Gamma(\sum_{p=1}^{P} \alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r})} \prod_{p=1}^{P} \frac{\prod_{f=1}^{F} \Gamma(\nu_f + C^{p}_{1:R,f,1:N_r})}{\Gamma(\sum_{f=1}^{F} \nu_f + C^{p}_{1:R,f,1:N_r})}$$

$$\mathrm{p}(\delta_{n^*,r^*} = p^* \mid \delta^{-(n^*,r^*)}, ...) = \frac{\prod_{p \neq p^*} \Gamma(\alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p,-(n^*,r^*)}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r}) \Gamma(\alpha_{p*} + C^{p*}_{r,1:E,1:I_r} + C^{p*,-(n^*,r^*)}_{r,1:F,1:N_r} + C^{p*}_{r,1:G,1:O_r} + C^{p*}_{r,1:L,1:M_r} + 1)}{\Gamma((\sum_{p=1}^{P} \alpha_p + C^{p}_{r,1:E,1:I_r} + C^{p}_{r,1:F,1:N_r} + C^{p}_{r,1:G,1:O_r} + C^{p}_{r,1:L,1:M_r}) + 1)}$$

$$\cdot \frac{\prod_{p \neq p*} \Gamma(\nu_{f^*} + C^{p,-(n*,r*)}_{1:R:,f,*,1:Nr}) \Gamma(\nu_{f^*} + C^{p,-(n*,r*)}_{1:R:,f,*,1:Nr} + 1)}{\prod_{p \neq p^*} \Gamma(\sum_{f=1}^{F} \nu_f + C^{p,-(n^*,r^*)}_{1:R,f,1:N_r}) \Gamma(1 + \sum_{f=1}^{F} \nu_f + C^{p^*,-(n^*,r^*)}_{1:R,f,1:N_r})}$$