Learning Diseases from Data - A Disease Space Odyssey

by

Bryan Haslam

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Signature redacted

Author . . . .
Department of Electrical Engineering and Computer Science
August 31, 2017

Signature redacted

Certified by . . . .
Dr. Luis Perez-Breva
Research Scientist
Thesis Supervisor

Signature redacted

Accepted by . . . .
Professor Leslie A. Kolodziejski
Chair of the Committee on Graduate Students

Learning Diseases from Data - A Disease Space Odyssey

by

Bryan Haslam

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Recent commitments to enhance the use of data for learning in medicine provide the opportunity to apply instruments and abstractions from computational learning theory to systematize learning in medicine. The hope is to accelerate the rate at which we incorporate knowledge and improve healthcare quality. In this thesis, we work to bring further clarity to the ways in which computational learning theory can be applied to update the collective knowledge about diseases.

Researchers continually study and learn about the complex nature of the human body. They summarize this knowledge with the best possible set of diseases and how those diseases relate to each other. We draw on computational learning theory to understand and broaden this form of collective learning. This mode of collective learning is regarded as unsupervised learning, as no disease labels are initially available. In unsupervised learning, variance is typically reduced to find an optimal function to organize the data. A significant challenge that remains is how to measure variance in the definition of diseases in a comprehensive way. Variance in the definition of a disease introduces a systematic error in both basic and clinical research. If measured, it would also be possible to use computers to efficiently minimize variance, providing a great opportunity for learning by utilizing medical data.

In this thesis, we demonstrate that it is possible to estimate variance in the disease taxonomy, effectively estimating an error bar for the current definitions of diseases. We do so using the history of the disease taxonomy and comparing it with a variety of external data sets that relate diseases to attributes such as symptoms, drugs and genes. We demonstrate that variance can be significant over relatively short time periods.

We further present methods for updating the disease taxonomy by reducing variance based on external disease data sets. This makes it possible to automatically incorporate information contained in disease data sets into the disease taxonomy. The approach also makes it possible to use expert information encoded in the taxonomy to systematically transfer knowledge and update other biomedical data sets that are often sparse (e.g. - symptoms associated with diseases).

A natural question stemming from these results is how granular does data need to be to make improvements? For instance, is patient-level data necessary to enable learning at the macro level of disease? Or are there strategies to extract information from other kinds of data to alleviate the need for very granular data. We show that detailed, patient-level data is not necessarily needed to extract detailed biological data. We do so by comparing disease relationships learned from clinical trial metadata to disease relationships learned from a detailed genetic database and show we can achieve similar results. This result shows that we can use currently available data and take advantage of computational learning to improve disease learning, which suggests a new avenue to improving patient outcomes.

By reducing variance within diseases using data available today, we can quickly update the space of diseases to be more precise. Precise diseases lead to better learning in other areas of medicine and ultimately improved healthcare quality.

Thesis Supervisor: Dr. Luis Perez-Breva
Title: Research Scientist

Acknowledgements

I would first like to acknowledge my exceedingly patient and supportive wife, Lis, who endured this exceptionally long thesis process. She also gave birth to our three children during my Ph.D. work and I acknowledge the personal sacrifices she has made during this time. I would also like to thank my other family members who inspired and taught me. Without my parents, brother and sisters I never would not have progressed this far.

I next would like to acknowledge my advisor, Dr. Luis Perez-Breva, who continually pushed me to greater exploration, creativity and precision. He was always generous with his time to help me along this journey.

I would also like to thank all of my previous mentors who encouraged me in my educational career and helped me get to this point: Zeynep Erim, Claire Gmachl, Andy Yun, Greg Nordin, Seung Kim, Mattijs de Groot, Johannes de Boer, Retsef Levi, Vivek Farias, George Verghese, Thomas Heldt and Pete Szolovits.

For this thesis I would also like to acknowledge Courntey Crummett and the other MIT librarians who helped me locate digital and hard copies of the history of the MeSH taxonomy from 1966 to 1996. Part of this work would not have been possible without those resources.

I also recognize my external funding sources: the McWhorter fellowship through the EECS department at MIT, the National Defense Science and Engineering Graduate fellowship through the Department of Defense, and the National Science Foundation Graduate Research Fellowship.

Lastly I would like to acknowledge the support of many friends across MIT. My neighbors at Westgate have been an excellent source of support during the PhD process. The MIT sailing pavilion was an excellent diversion for the stresses of my work. The MIT Addir Fellows provided an avenue to discuss unrelated, but important topics and make new friends. The EECS graduate office, especially Leslie Kolodziejski and Janet Fischer, provided me with support on many levels.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

There is an opportunity for faster learning in medicine. It requires three components and we focus on the one least explored.

Recent commitments to enhance the use of data for learning in medicine provide the opportunity to apply instruments and abstractions from computational learning theory to accelerate learning in medicine, thereby improving quality faster.[1, 2] There are three components related to seizing this opportunity: (1) collecting and sharing data, (2) learning how to map patients to diseases using predictive analytics, and (3) updating the collective knowledge of diseases.[3] The first two components have been addressed in a variety of ways in the literature, while the last component has received comparatively little attention. In this thesis, we work to bring further clarity to the ways in which computational learning theory can be applied to update the collective knowledge of diseases.

Learning in medicine already happens and revolves around diseases as discrete representations of continuous health states.

While this opportunity to introduce computational learning abstractions to medicine is relatively new, the concept of learning itself is not new to medicine. Researchers continually study and learn about the complex nature of the human body. There is a continuum of multidimensional health conditions that exist, as pointed out by Loscalzo, Kohane and Barabasi.[4] Because patients could lie anywhere along a continuum, diseases are used as discrete states[5] to categorize very similar health conditions, with a focus on conditions that are abnormal or undesired.[6] Learning about diseases is a central focus in medicine.[2]

There are two connected learning problems in medicine.

One way to view learning in medicine is through the dual learning hypothesis, which states that there are two intertwined learning problems that are actually distinct in nature. The first is to generate the "best" set of concepts, such as diseases. The second is to learn mappings between concepts, such as patient characteristics or features of diseases. The second learning problem relies on the concepts generated from the first learning problem, although the first learning problem can be updated based on new learning from the second problem. This hypothesis is consistent with existing literature on learning in medicine.[2]

The first is unsupervised learning to organize human biology features into diseases. The disease taxonomy is the most generally accepted summary of this learning.

The first mode of learning is to generate the "best" set of diseases and their relationships given everything known about human biology. This mode of learning is regarded as unsupervised learning, as no disease labels are initially available. The input data consists of human biology features (e.g., biochemical pathways, gene mutations), and the output includes groupings of features that can be labeled as diseases. While there is no single way to summarize the collective set of diseases and relationships, one formally accepted representation is a taxonomy.[7] A taxonomy defines a set of diseases as well as the hierarchical relationships between diseases. Some have pointed out that taxonomies have many imperfections, which may affect medical practice and outcomes.[8] Changing or updating the disease taxonomy is generally a laborious process that relies on expert consensus, and at best is a yearly endeavor.[9]

The second learning problem is supervised learning to map human biology features to diseases.

The second mode of learning, learning mappings to diseases, is considered supervised learning because the data contains disease labels. The input is human biology features and the labeled output data is diseases. For example, scientists learn mappings from biological models to diseases.[10] Researchers learn mappings from patient features to diseases based on new knowledge, methods or equipment.[11, 12] Clinicians learn to map patients more consistently to optimal diseases through education and experience.[13]

Computational learning has taught us how to approach these learning problems by minimizing empirical risk, loss or error.

Computer scientists have studied extensively how to use computers for supervised and unsupervised learning problems, which we refer to here as computational learning. Computational learning is based on minimizing an objective function, which Vapnik refers to as empirical risk minimization.[14] In the case of supervised learning, risk is quantified by a loss function to evaluate the difference between the output of the mapping and the labeled output data. The goal is to find a mapping that generalizes to any similar data set. Algorithms are then developed to find a minimum for the risk or loss function that performs similarly for different subsets of the data. In the case of unsupervised learning, the objective function is typically a measure of variance within or between groupings.

Measuring and reducing error has been a priority recently for supervised learning problems.

With the ability to quantify the objective function, subsequent improvements can be made. A landmark report in 2000 was the first to measure the amount of error in healthcare delivery.[15] A follow-on report singled out diagnostic errors, or where incorrect mappings from patients to diseases had occurred and were quantified.[12] Measuring these errors was the first step toward improved learning. In cases where data is readily available, such as ICU data[16, 17], machine learning has been used to develop models to predict diseases or future disease states.[18] These examples demonstrate the first type of learning described above. Computers are particularly useful because they can learn by efficiently minimizing risk and process large data sets; hence, there is a great opportunity for learning by utilizing medical data.

There are few examples of measuring or reducing error of the disease taxonomy.

To the best of our knowledge, there have only been a few examples of attempts to quantify or minimize the loss for the second type of learning. An absolute error cannot be calculated because no true labels are known; we can only make best approximations as diseases. However, we can measure uncertainty though variance. The most straightforward way to conduct this measurement would be to compare comprehensive longitudinal data on patients with detailed historical diagnostic criteria. Because the majority of this data cannot be accessed or does not exist, another approach is needed. One example of similar work was to reduce variance of information content in the Gene Ontology using several heuristics.[19] The heuristic and data used were specific to that taxonomy and would not apply to many other taxonomies.

If measuring and reducing loss was possible, we could improve patient outcomes faster.

If it was possible to quantify empirical risk of the disease taxonomy, it would also be possible to use computers to minimize this risk and learn better disease taxonomies. This would enable us to establish error bounds on supervised learning problems of diseases and better understand the impact of learning results such as clinical trials. It would also allow us to update the disease taxonomy more quickly, and in turn accelerate the pace of learning in medicine. A better set of diseases will enable improved diagnoses, and ultimately more effective treatments and better outcomes for patients.[12]

Reducing variance in disease would also achieve precision medicine.

Another way to characterize reducing variance is to increase precision. In particular, there has been a recent push for "precision medicine" to improve the quality of healthcare.[20, 21, 22] We define precision medicine in the context of this thesis as a reduced variance within diseases and reduced variance mapping patients to diseases. As will be explained later in this thesis, using data and computational learning may also bring us closer to precision medicine than the current trajectory.

What is going to happen in this thesis.

In this thesis, we demonstrate that it is possible to estimate variance in the disease taxonomy, effectively measuring an error bar on the current definitions of diseases. We further present one method for updating the disease taxonomy by reducing variance based on external disease data sets. A natural question stemming from these results is what data is needed to make improvements. We show that detailed, patient-level data is not necessarily needed to extract detailed biological data. Rather, we can use currently available data with different representations to extract latent information for comparing diseases. These results show that we can use currently available data and take advantage of computational learning to improve disease learning, thereby accelerating the improvement of patient outcomes.

In the next section, we provide a roadmap of the thesis, describing each chapter briefly. We then close the introduction with a summary of the contributions of the thesis.

1.2 Thesis Roadmap

In chapter 2, we provide a background of the literature on learning in medicine, disease definitions and a summary of computational learning. The following chapters (3-5) contain the setup, experiments and results which are described below. We then discuss the results in chapter 6 and provide conclusions in chapter 7.

In chapter 3, we start by quantifying how variance exists for a set of diseases. We do this by tracking the taxonomy of diseases over approximately 50 years. Our estimate is based on the assumption that the taxonomy changes in a relatively consistent manner over time. At the same time, we examine the history of the taxonomy and key changes that have taken place.

In chapter 4, we develop methods for connecting the disease space, as defined by the disease taxonomy, with other types of disease spaces based on different disease attributes such as genes, symptoms and drug interventions.

In chapter 5, we present a new disease representation we created for learning tasks. The representation utilizes the expert knowledge contained in clinical drug trials. Along with a distance metric, this representation creates a new space for diseases.

The appendices include two analogies for understanding variance in learning problems as well as supplemental material for chapters 3-5.

1.3 Contributions

The three main contributions in this thesis include:

1. We estimate quantitatively the variance in the accepted disease taxonomy. The taxonomy is used to define diseases, and up to this point measuring the

uncertainty or variance that the taxonomy introduces has remained a challenge. Our estimate enables researchers to put an error bar on learning results and use computational tools to learn an improved disease taxonomy.

2. We reduce the loss between the disease taxonomy and external disease data sets. This allows for a more precise update of the set of diseases based on available data, as well as for the discovery of new disease-feature associations.

3. We develop a new representation of diseases using clinical trial meta-data that contains latent information, similar to that found in a drug-gene networks. This result demonstrates that a surprising amount of learning is possible with superficial data sets. It further implies that diseases can be updated more precisely without having to obtain detailed patient-level data.

Chapter 2

Background

This thesis covers several fields including healthcare, machine learning and innovation. We use computational learning as a framework to unite these fields. Often terminology is ambiguous in these fields, so we provide specific definitions in Section A.1 in the Appendix for reference. For background on computational learning relevant to this work, see Section and for background on discrete versus continuous spaces, see Section , both found in the Appendix.

For any statistical learning to take place, multiple samples are needed from a consistent process. We assume that biology is consistent such that if a biological system is fully characterized and its environment controlled, an experiment will always produce the same result. In medicine, a patient's fully characterized state is their true condition, but is only partially observable and is unique. In order to have multiple data points for learning, we use the concept of disease to describe the overlap of true conditions of a population. Disease, as we use it in this thesis, is therefore a discrete approximation the true underlying state of many patients that we cannot fully observe.

2.1 Learning in Medicine

Learning in medicine involves updating knowledge to improve understanding of disease, treatment options available, and ultimately the health outcomes of patients.[12] Recently, an emphasis has been placed on using data more broadly to accelerate the pace of learning in medicine.[1] In 2012, the Institute of Medicine published an extensive report on their goals for a rapid and continuously learning health system.[2, 23] Though this has been a goal for decades,[24] learning in medicine has traditionally been very expensive, time consuming and often only batch processed.[2] The availability of data and computational tools offer the opportunity to realize the goal of accelerating the pace of learning in medicine.

The goal of healthcare is to help patients prevent or recover from sickness.[25, 26, 27]. Learning should translate to improving patient outcomes, and faster learning would lead to better outcomes sooner. To this end, learning occurs in many ways in medicine, from formal medical education and experiential training to published research. We distinguish between learning that is on an individual level compared to learning on aggregate. Individual learning involves individual clinicians updating their knowledge through continuing education. This process is primarily about disseminating new knowledge. Aggregate learning is the sum of all the medical knowledge gained. Aggregate knowledge is updated and then disseminated, whereby individual clinicians learn. In this thesis, we focus on aggregate learning.

To realize the goal of accelerated aggregate learning in medicine, three components are necessary: (1) collecting and sharing data, (2) mapping patients or features to diseases (3) updating the collective or aggregate knowledge of disease. We describe each of these these components in the following sections.

2.1.1 Collecting and Sharing Data

The health system generates vast quantities of data, though historically it has been inaccessible computationally by privacy concern, paper or propriety.[28] Medicine is now at the point where the valve is about to be released on this data. Government agencies have committed to make more data publicly available, such as the European Medicines Agency (EMA) making available clinical trial data [29] or the United States making a host of data sets public through healthdata.gov.[30] Industry is also releasing more, such as GlaxoSmithKline (GSK) and others making their clinical trial data public.[31] Increasingly, there is a trend to make data public or collect data in consortiums for broader use.[32, 33, 34, 35]

In many cases, data was collected for one learning task, as in the case of clinical trials, or none at all, as in the case of data collected in the course of medical practice.[36] There is great potential for new learning from this data.

2.1.2 Mapping Patients or Features to Diseases

Biological research is often the first step in learning. This can be performed *in vitro* or in animal models. Disease cell lines are created to model a disease.[37] New features which have been discovered are tentatively associated with the disease. They cannot conclusively be associated with the disease because they are not in a complete human system. Animal models allow for learning in a complete system, but not a human one.[38]

Clinical trials allow researchers to test a hypothesis in which some feature(s) are associated with a disease. Human subjects are sampled, and if the result is consistent across many people, it is concluded that there is an association between the feature and the disease. This could be a biological concept for research purposes or a diagnostic tool or therapeutic intervention for medical practice. Clinical trials are powerful because they involve real human subjects as opposed to models in animals or in vitro. However, they are limiting because they are require significant time and resources to make a single binary conclusion.[39]

The confidence of a trial result depends on the type of trial. In the past several decades, there has been an increased emphasis placed on using evidence from learning in medicine to inform medical practice.[40] A hierarchy of evidence has been constructed, which indicates the type of learning that is trusted relative to another.[41] For example, an observational trial is trusted over personal experience, and a randomized controlled trial (RCT) is trusted over an observational trial.

The learning process described in this section allows researchers to map biological features to diseases. This learning helps improve diagnostic criteria for diseases so patients can be mapped more precisely to diseases. The learning process also allows researchers to map treatments to diseases. This learning helps improve treatment options for diseases. Using data to enhance these efforts has been proposed and has shown promise. This type of learning is observational in nature, which is not preferred evidence, but is much less resource-intensive than forms of evidence higher in the hierarchy. Researchers have pointed out that much of medicine practiced is not based on evidence.[42, 43] This shows the need for the vast amount of learning that RCT's will not allow.[44]

Inherent in these learning problems is the assumption that the diseases are correct. In the next section. we discuss the problem of updating the knowledge about diseases so it is more correct.

2.1.3 Updating Collective Knowledge of Disease

Due to the large amount of knowledge accrued in past sets of diseases, the collective knowledge of diseases is not generated *de novo*, but is instead updated. There is no official or systematic way to do this; knowledge gets updated as new research or practice results are examined and shared within the research community. This may happen through conferences or journal publications. Official changes may happen when a group of experts agrees on a change to an official standard or guideline. For example, hypertension treatment guidelines for people over the age of 60 changes from 140 to 150 mmHg as per the American Heart Association (AHA) and the American College of Cardiology (ACC).[45]

Disease in this space is subject to a variety of limitations and biases.[10] For example, diseases are often defined by some knowledge base, such as anatomical location, symptoms observed, treatments, causative agents, biomolecular etiology, heredity and so forth.[46]

In addition to guideline changes, another currently prominent area of research to update knowledge about disease is the search for subpopulations of disease.[47] The premise behind these efforts is that diseases are accurate, but heterogeneous in nature. By finding sub-populations, it may be possible to find more heterogeneous subsets that possibly have differing etiologies. A classic example of this would be Diabetes being sub-categorized as type I and type II.

This type of learning has received less attention than the previous type. One reason may be that it is more subjective in nature - what makes a "good" disease can be a subjective assessment. To the best of our knowledge, nobody has undertaken to evaluate and update the collective knowledge of diseases using data.

2.2 Diseases

Underscoring the learning modes described above is the concept of disease. We describe in more depth here what the concept of disease entails.

Though widely used, the concept of disease has varying definitions, such as the opposite of health, conditions not biologically normal for humans or states of the body that carry a negative value judgement.[6, 48] The precise definition of sickness is debated, but typically is described as an internal state characterized by an impairment or deviation from a normal or more desired state.[48, 6] Diseases provide a way to describe the same or similar sickness in more than one patient, often through a consensus of healthcare professionals.[49] There is some ambiguity in the use of the term disease, so we clarify our use of terminology here. A health condition refers to the biological processes actually happening in a patient. There are billions of biological processes taking place in humans and there is no inherent or ground truth in the grouping of such processes. We refer to a disease, on the other hand, as a classification referring to a man-made grouping of symptoms, mechanisms or patients that hopefully reflects underlying biological processes. Diseases are used in medical practice as a means of identifying and describing a condition, and then prescribing a corresponding treatment.

The definitions of a diseases use rules that identify them, typically based on patient symptoms or biological mechanisms. For example, hypertension is blood pressure above a certain threshold. Such a designation allows clinicians to compress all of the information about a patient into one or more classifications. The rules for classifying operate on patient data, such as symptoms, signs and lab tests. Rule-based categorization is used in many situations because it is easy to apply and communicate, but is also known to have limits in computational learning.[50]

2.3 Dual Learning Hypothesis in Medicine

A hypothesis consistent with the learning in medicine described above is the dual learning hypothesis. The dual learning hypothesis states that categories or labels are learned in an unsupervised fashion, while the association of data points to specific labels is learned in a supervised fashion. Labels are necessary for the second type of learning, but each type of learning can have an effect on the other.

We describe each part of the dual learning hypothesis in the following sections.

2.3.1 Learning Diseases

Learning what is a disease is an unsupervised learning problem. There are no inherent labels for conditions and no absolute truths for what distinguishes one condition from another. Diseases are proposed based on sets of similar characteristics. There is an underlying joint distribution of all characteristics, but that distribution has a very large number of dimensions. It is not possible to comprehensively describe the distribution because of the high dimensionality, so clustering is used. With clustering, there are not necessarily inherent ways to divide up the distribution. It is therefore up to experts to describe what characteristics should be used, how many diseases should be identified and how to divide up parts of the condition spectrum into discrete diseases. This could be done based on anatomical location, organ systems involved, genetic or environmental causes and many more criteria.

It is not typical that disease are learned de novo, but rather updated with new knowledge. Updating could be the result of new characteristics or a new grouping of characteristics. Examples of new characteristics that are currently changing our understanding of diseases and how to group them include features of the microbiome[51] or mitochondrial defects.[52] Some characteristics are only know being appreciated in long-standing diseases, such as how metabolism affects cancer.[53] Still, while there are well-known dimensions such as diet and psychosocial variables which have long been associated with diseases, direct connections have been harder to draw.[54]

The set of diseases can have a dramatic impact on medical practice and learning. For example, billions of dollars have been spent searching for a treatment for Alzheimer disease (AD) with very little to show for it.[55] It may not be that the correct drug has not yet been found nor that some existing drug has not yet been tested on AD, but that AD is the wrong clustering of patients to use to search for treatments. In some cases, a drug in development has not been successful for the intended disease, but useful for a yet undefined condition. For example, erectile dysfunction was not a 'disease' before Viagra,[56] which also led to attempts to define more diseases such as Low-T [57] and female sexual dysfunction (FSD).[6]

The true continuous disease state space is relatively constant, changing on an evolutionary time scale. However, the approximated disease state space is constantly changing as biomedical research updates our understanding of disease. Furthermore, disease definitions change over time. Cancers have been defined by anatomical location (breast, brain, pancreatic) and then further refined to reflect tissue of origin (carcinoma, sarcoma, lymphoma). Recent work suggests that genetic similarity is important, often across different tissue types.[58] Definitions therefore change over time to incorporate new information. Osteoarthritis(OA) is another example; in 2007, the FDA requested clarification on the definition of OA, whether subcategories of OA could be combined and what the relation was to other specific diseases.[59] Yet another example is Myocardial Infarction.[60] Lastly, an example is Parkinsons disease.[61]

We provide two examples of diseases whose definition and relation to other diseases have changed over time. We can assume that the biology did not change, but rather our understanding changed to better approximate biology.

Example Case: Diabetes

Diabetes is believed to have been diagnosed anciently as frequent urination or sweet smelling urine. "Diabetes" means to pass through and "mellitus" means sweet like honey. It has been noted for a long time (1st recorded in 500 AD) that there was a difference between Diabetes in kids and Diabetes in overweight adults. In the late 1700s, diabetes mellitus was named to distinguish it from Diabetes Insipidus based on the sweet smell. In 1889, researchers found that a dog with its pancreas removed produced sugary urine. In 1901, someone first described Diabetes Mellitus as caused by the destruction of islets of Langerhans. People tried to make pancreatic extracts, and in 1922, insulin was successfully isolated and given to a diabetic patient. Insulin-dependent Diabetes is now known as type 1 Diabetes because it responds to insulin treatment. It is now believed that most cases are the result of an autoimmune reaction leading to the loss of beta cells. Some genes have been identified as risk factors, but environmental factors are also thought to be triggers.

The example shows how a disease first looked like a syndrome with symptoms of frequent urination, and became differentiated based on another symptom, sweet urine. This was further differentiated based on age and physical characteristics. After millennia of having seen symptoms, a crude mechanistic understanding developed, namely that the pancreas was linked to Diabetes. This became refined to show that beta cells were specifically the problem. There was then a race for a solution, and once found, the disease definition shifted to one that was based on the therapy that worked, insulin. Over time, the understanding of mechanisms has been refined, while the population has remained the same.

Example Case: Cystic Fibrosis

Cystic Fibrosis was possibly described in the 18th and 19th century as kids with salty sweat who died. The first pathological sign that was noticed was the fibrosing and cysts found in the pancreas. In the 1930s, researchers drew a connection between cystic fibrosis of the pancreas and lung function and intestinal disease (about 80% of deaths in CF are due to cardiorespiratory complications). At that point, it was hypothesized to be a recessive disease. It was not until 1988 that the first gene mutation was discovered. There is one gene (CFTR) that is affected, but more than 1,500 different mutations can have varying effects. Kalydeco (Ivacaftor) was a drug approved in 2012 to treat those with the G551D gene mutation, one that only represents about 4% of the population. In 2014, an NDA was filed for a Lumacaftor/Ivacaftor which is effective for patients with the F508del mutation for which approximately 60% of patients have.

The example shows a disease may have first been recognized by a symptom (salty sweat) but did not represent a unique way of identifying the disease. It really was defined by gross pathology of the pancreas. It was then linked to other pathophysiological problems which have been the cause of death. It has only been in the last 20-30 years that the genetic basis for the disease was found and then partitioned by type of mutation. Some therapies have subsequently been developed for specific mutations.

In the next section, we expand on the disease taxonomy, which is an accepted way of representing the aggregate knowledge about diseases.

Taxonomy

Taxonomies have been used for millennia to codify knowledge in many different disciplines.[62] A taxonomy delineates two things: (1) it defines what a class is in some space of knowledge, and (2) how those classes are related to each other in a structured hierarchy.[63] A taxonomy is an efficient way of compressing information and knowledge into a condensed format that is particularly useful for classification and communication. Taxonomies are useful for knowledge management, organizing information and sharing information, as in the case of teaching.[64]

Taxonomies are used in medicine in a variety of ways.[65] For example, the International Classification of Disease (ICD) is formally used for communicating about reimbursement and researching diseases.[66] The Medical Subject Headings (MeSH) is used for indexing medical literature and clinical trials.[67] The Systematic Nomenclature of Medicine (SNOMED) is a newer taxonomy increasingly used for communication among clinicians.[68] The Disease Ontology is a very recent taxonomy whose goal is to unify disease knowledge into one for biomedical research.[69, 70, 71] Beyond the formal delineated taxonomies, there are implicit taxonomies behind the structure of medical education, medical specialties, hospital departments, funding organizations such as the National Institutes of Health and so forth.

Medical taxonomies are also used for research and statistics by providing a controlled vocabulary and a set of hierarchical relationships.[65] Specifically in medicine, the taxonomy of diseases is useful in the context of medical training and licensing,[72] medical billing,[73] structuring research efforts,[74] and predicting outcomes.[75] For example, ICD codes are often used retrospectively to identify patient cohorts or as training data.[76] Research increasingly looks to automated techniques for learning, such as knowledge discovery in networks analysis or machine learning. Automated learning about concepts often relies on taxonomies to identify concepts and their classes as well as their relationships. This can be seen in work searching and comparing the corpus of medical literature, clinical trials or comparing disease-gene relationships.[77]

Though useful, taxonomies have their imperfections. In 1977, Engel wrote in Science "Thus taxonomy progresses from symptoms, to clusters of symptoms, to syndromes, and finally to diseases with specific pathogenesis and pathology. This sequence accurately describes the successful application of the scientific method to the elucidation and the classification into discrete entities of disease in its generic sense."[74] He argues there is much merit to this approach, but that "distortions" are also introduced. More recently, an IOM report in 2011 detailed how current medical understanding often does not fit well into the standard taxonomies used today, as well as the need for a new taxonomy, which should lead to faster and more precise learning and ultimately better health outcomes.[9]

One way to see how limiting a taxonomy might be is to look to the past. The first version of what is now known as the International Classification of Diseases included diseases such as "Want of breast milk," "Softening of brain," "Insanity," "Old age" and "Diarrhoea." These diseases, as classified more than 100 years ago, would not be considered diseases or causes of death today (the original ICD was created to track causes of death). It is easy now to see that these were likely not the best diseases, but this can be much harder to see in the current taxonomy.

Though many have pointed out there are limitations to a taxonomy structure and to our current disease taxonomy, the challenge remains to quantify how much error, uncertainty or variance there is in the disease taxonomy.

2.3.2 Mapping to Disease

Learning to map concepts or specific data points to diseases is a supervised learning problem. This learning problem assumes that diseases are fixed and the task is understanding what existing or new concepts are associated with a disease. Disease labels are provided by expert opinion or specific diagnostic criteria. Mappings are typically between biological features and diseases or between interventions and diseases. The two mappings often occur in that order, but that is not always the case. It is sometimes the case that searching for a treatment provides new insights about biological features. Both types of mappings are useful in medical practice because patients are mapped diseases based on features and then mapped to a treatment. The total process is referred to as "diagnosis" in the IOM report on Diagnosis in

2015.[12]

The process of diagnosis, as described in the IOM report, starts with a patient experience something they identify as a health problem and engaging with the health care system. Clinicians then gather patient information, integrate and interpret the information and come with a diagnosis. The diagnosis is to a "pre-existing set of categories agreed upon by the medical profession to designate a specific condition."[78] This can be summarized as healthcare professionals iteratively gathering patient information and mapping that to one or more disease categories. The information can be gathered through a clinical interview, clinical history, physical exam, diagnostic testing and more. The patient information may include clinician notes, images, test numbers and so forth.

The entire diagnostic process includes gathering information about a patient, determining the disease or diseases the patient has, recommending a treatment based on the disease and following up on the outcome.[12] The goal is an accurate diagnosis, meaning it is precise and complete, where precision means the disease classification corresponds to the patient's true condition. Errors in the process are unsafe or less effective, resulting in adverse events, death and billions of dollars wasted.[12] The diagnosis process is improved when these errors are reduced, such as when accurate information is acquired and shared.[15] The quality of their care is measured by safety and effectiveness of treatments to improve sickness.[79]

In the following two sections, we go into further details about the two general types of mapping and provides several examples.

Mapping Biological Features to Disease

For the learning problem of mapping biological features to disease, the input is human biology features and the labeled output is diseases or disease states. This represents much of biomedical research. Biological features could be genes, biochemical pathways, image characteristics or other diagnostic criteria. We typically assume that biology is consistent because evolution happens on such a long time scale. Much of basic biology research occurs in this way. In vitro or animal models are used to approximate a human disease. Characteristics of that model are studied by examining multiple samples. Samples may be modulated for further understanding.

Another area where features are mapped to disease states is related to clinical data. More specifically, this areas focuses on information collected in the normal course of care, but which can be insightful for learning. Clinicians use clinical experiences for learning, so it may be reasonable to use data that may summarize clinical experience for learning. Learning problems from clinical data are often set up to predict diseases or disease states, such as septic shock or death in the ICU.[80] Recent examples in machine learning have begun to realize the promise of making predictions comparable to doctors.[81]

Mapping Treatments to Disease

One of the most recognizable forms of learning is clinical trials to determine if a treatment is safe or effective. Effectiveness is a binary mapping to learn if the treatment will improve a specific disease pre-defined.[82] Safety is a set of binary mappings to learn if the treatment will lead to set of diseases. Ideally, a drug will improve the predefined disease and not lead to other diseases. There are other examples where a drug for one indication worked well for completely different indications, such as Thalidomide for leprosy and Isoniazid for tuberculosis. These discoveries were serendipitous, and many people have tried systematically finding such connections using a variety of computational tools and data.[83, 84, 85, 86, 87, 88, 89, 90, 91]

In the text boxes below, we give two examples of drug learning problems that had many steps with wide-ranging consequences. The first drug is Avandia where continued learning after approval reversed earlier findings. The second drug is Tecfidera where learning about one the drugs relationship with one disease led to new indications later on.

Example Case: Avandia

Avandia (rosiglitazone) was first approved in 1999 to control glucose levels for the management of type 2 Diabetes. Uncontrolled Diabetes can lead to heart disease and damage to blood vessels, nerves, kidneys and eyes. It quickly became a blockbuster drug with sales peaking in 2006 at $2.5 billion. In 2007, a meta-analysis associated rosiglitazone with increased risk of Myocardial infarction.[92] A congressional hearing was convened to investigate shortly after.[93] That same year, the FDA placed a black-box warning on the label of Avandia after an advisory committee of experts voted 22 to 3 that the overall benefit-risk profile of Avandia supported its continued marketing in the US.[94] In 2010, a new meeting of the advisory committee was divided between removing the drug from market or allowing continued albeit restricted marketing.[95] The FDA allowed continued marketing under a restricted access program that effectively reduced the patient population on rosiglitazone by 80%.[96, 97] GlaxoSmithKline, the manufacturing company, followed suit and decided to stop advertising Avandia. Rosen [98, 99] documents the evolution and lessons learned from the 2007 to the 2010 advisory committees. Though no new conclusive statistical evidence of the increased risk of myocardial infarction was gathered from 2007 until 2010, difficulties to draw associations from observational studies, disagreements with how adverse events were adjudicated in past clinical studies, and testimonials weighed in the evolution of the vote of the committee. In 2013, another advisory committee reviewed the independent analysis of the Rosiglitazone Evaluated for Cardiovascular Outcomes and Regulsation of Glycemia (RECORD) study conducted by GSK.[100] The conclusion of that study did not match Nissen's conclusion and the committee decided to remove certain restrictions on the drug.[101]

It is interesting to note that the association between rosiglitazone and cardiac events had been documented numerous times before Nissen's controversial meta-analysis.[102] comments on the limited clinical potential of thiazolidinediones (TZDs) because of the increased risk of cardiac events.[103] reports cases of heart failure associated with rosiglitazone. Yet [104] notes that TZDs have been documented to reduce the circulating markers of cardiovascular risk, and GlaxoSmithKline conducted inconclusive studies on rosiglitazone and cardiac failure as early as 2001. This disparity of opinions has spawned numerous publications to sort through the challenge of prescribing rosiglitazone.[105, 106]

From this example, we see that new learning can affect millions of people. We also see that data was reanalyzed in a very different way.

Example Case: Tecfidera

Dimenthyl fumarate (DMF) is an oral drug for treating relapsing multiple sclerosis. Biogen Idec began marketing DMF after successful phase 3 trials published in 2012.[107, 108] The story of DMF started in 1959 when a German chemist found that fumaric acid esters (FAE), one of which is DMF, helped treat his own psoriasis.[109] It was not widely used as a therapy though, but rather as a mold-inhibitor for bread, leather and furniture in Europe.[110, 111] It wasn't until the 1990s that trials were conducted in Germany that concluded FAE was efficacious in treating psoriasis.[112] In 1995 it was shown that Dimethylfumarate and Monoethylfumarate, branded under the name Fumaderm, were effective as an oral treatment for psoriasis and was approved in Germany.[113, 114] Biogen Idec partnered with the maker of Fumaderm, Fumapharm, because they were interested in finding a multiple sclerosis (MS) drug and knew MS and psoriasis has some similarities. In 2003 they acquired the worldwide license to Fumaderm and in 2006 acquired Fumapharm. In the second quarter of 2014, DMF, branded "Tecfidera," had sales of $700 million. There are several proposed mechanisms of action why DMF helps MS, but it is still not well-understood.[115]

From this example, we see that learning about a drug in the context of one disease can affect learning about that drug and other diseases. This makes learning much more efficient that a simple grid search.

2.4 Disease Representation

For computational learning, data must be structured in a way for a machine to compare different data points. Much of the data recorded in medicine is done for humans such as clinicians or patients to understand and make comparisons. Text data is an example that is very useful humans, but text must be modified for computers to use.[116] Computers can compare numbers, and to make any comparison greater than one dimension, a vector of numbers is used. Each number in the vector is called a feature, and choosing what the vector is composed of is known as feature selection. A single data point could be represented with many different vectors, and the design of the vector is the task of the user. The features used to represent data points is sometimes called a representation. At the end of the day, if one uses machine learning to learn about diseases, a disease needs to be represented by a vector of numbers. A representation affects what is learned, and therefore features can be engineered to suit a learning problem.

Much of the data we would like to analyze about humans is not available for a variety of reasons. One reason is that experiments are not performed on humans like they are on other models, so the types of experiments and sampling is less than desired. If humans exist with the condition, data might not be collected on them. Additionally, if the data exists, it may not be available for researchers to use. For these reasons, learning from data collected on humans can be much more challenging than typical biological studies. It appears to be a common first approach to try a representation that is similar to what humans would use. For example, a disease might be described by its diagnosis criteria or pathological pathways. These may be useful ways to represent disease, but to limit representations to those humans' understand would be over-constraining.

Though the data that many would desire for learning is not available, the same information may be available in existing data sources, albeit as latent information. Many problems in computational learning require uncovering latent information. More representations that can uncover latent information about disease will further our understanding with data that is currently available and help us prepare for the more detailed patient-level data that will become increasingly available in the coming years.

2.5 Precision Medicine

Precision medicine is a recent term used to describe an approach to healthcare research and delivery.[20, 22] It has garnered the attention of the National Institutes of Health (NIH), the White House and popular journals.[22] Definitions of what precision medicine entails vary; for example, "prevention and treatment strategies that take individual variability into account,"[9] "treatments targeted to the needs of individual patients on the basis of genetic, biomarker, phenotypic, or psychosocial characteristics that distinguish a given patient from other patients with similar clinical presentations,"[21] "coupling established clinical-pathological indexes with state-of-the-art molecular profiling to create diagnostic, prognostic, and therapeutic strategies precisely tailored to each patient's requirements"[20] and "precisely tailoring therapies to subcategories of disease, often defined by genomics."[117]

Common components associated with precision medicine are genomics, proteomics, metabolomics, epigenetics, cellular assays, biomarkers, mobile health, population-based research, big biomedical data, data sharing, data mining and decision-support tools.[118, 22, 21, 20] Still, a precise definition of precision medicine can be illusive. Precision in science is typically how close measurements are to each other; therefore, high precision implies a lack of variance. As a side note, precision is often confused with accuracy, the latter of which is how close something is to the truth or a standard. An important part of precision medicine is mapping patients precisely to diseases and treatments. This means that the same patient would be mapped to the same disease and receive the same treatment. The aforementioned components can be very useful for helping to map patients precisely to diseases.

There is another piece to precision medicine that is important in precisely defining diseases. Using diseases with low precision or high variability will not lead to precise treatments and predictable outcomes. Little attention has been paid to improving diseases on aggregate, such that they will lead to the more precise practice of medicine. Some have looked at refining individual diseases to be more precise, but to the best of our knowledge, nobody has looked at how to do so with the comprehensive set of diseases. A comprehensive approach is important because diseases are used for classification, so different parts of the space can affect each other.

2.6 Data Sets

We provide a list of all data sets used in this thesis with a short description. We use all of these data sets at different points in the thesis and include them here for reference.

- Aggregate Analysis of ClinicalTrials.gov (AACT) [119] - A database of information on ClinicalTrials.gov in downloadable format.
- Medical Subjects Headings (MeSH) [67] - A database of all terms in the MeSH vocabulary along with tree structures (hierarchical relationships) and additional information.
- Unified Medical Language System (UMLS) [120] - A database of terms and relationships for relating concepts from many different biomedical vocabularies.
- Online Mendelian Inheritance in Man (OMIM) [121] - A database of human genes and phenotypes regularly updated.
- Phenotype-Genotype Integrator Database [122] - The combination of several genome-wide association studies (GWAS) housed at the National Center for Biotechnology Information (NCMI).
- Pubmed - A database of journal citations indexed by MeSH terms.
- Pubmed Central (PMC) [123] - A subset of Pubmed of full-text articles.

- Human Symptoms Disease Network (HSDN) [124] - A database of diseases and symptoms connected in the PubMed literature.

- Human Disease Network (HDN) [125] - A database of diseases and associated genes derived from OMIM.

- Medication Indication (MEDI) [126] - A database of drugs and associated diseases.

- International Classification of Disease (ICD) [127, 128] - A database of disease terms and hierarchical relationships. Originally constructed for generating statistics on reasons for death and currently used for a wide range of purposed including medical billing.

- Systematized Nomenclature of Medicine - Clinical Terms (SnoMed CT) [129] - Standardized vocabulary of clinical terms and relationships.

- DiseaseOntology (DO) [71, 70] - A recent taxonomy of disease meant to unify biological concepts.

- Surveillance, Epidemiology, and End Results (SEER) [130, 131] - A history of cancer incidence compiled by the National Cancer Institute (NCI).

We also list data sets related to the thesis that were useful, but not explicitly used in the thesis.

- Side Effect Resource (SIDER) [132] - A database of drugs and side effects for marketed drugs.

- Off-label Side Effects [88] - A database of drug and side effects for non-marketed drugs.

- Search Tool for Interaction of Chemicals (STITCH) [133] - A database of relations among chemicals based on a variety of sources.

- Drubgank.ca [134, 135] - A database of drugs and characteristics such as drug targets.

Chapter 3

Estimating Disease Taxonomy Variance

It is possible to measure error, or conversely accuracy, in supervised learning because there is a gold standard or labeled data. One can think of defining diseases and how they relate to each other, such as a disease taxonomy, as an unsupervised task. It is therefore only possible to measure variance, or conversely precision, of diseases. In this chapter, we propose methods for measuring variance of the disease taxonomy, which can in turn be used to increase precision of the disease space.

3.1 Introduction

It is impossible to measure accuracy of diseases, but we might be able to measure precision.

We cannot assess the absolute error between diseases and true conditions, because true conditions are unknowns that diseases are trying to approximate. Uncertainty or variance within diseases might be assessed though, and it is recognized that uncertainty within diseases is a hurdle for learning.[136] To the best of our knowledge, nobody has ever quantified variance of diseases systematically. If we can estimate variance quantitatively, we can reduce the variance in the disease space and thereby increase the precision of diseases. We can also approximate uncertainty in supervised learning tasks that involve disease, enabling us to accelerate the rate of learning from data.[137]

Measuring precision of diseases in people is not straightforward.

In research, disease models (i.e., not in human subjects) can be used to produce multiple samples and record statistics to estimate variance. However, in human subjects this is not possible. Instead, we might use historical clinical data, but disease definitions change over time. The most straightforward way to measure variance within diseases is to require detailed diagnostic criteria and longitudinal patient data. Data on detailed diagnostic criteria over time does not exist, and longitudinal patient data is very difficult to obtain. Even if this data were available, there is the challenge of missing data because the patient data that exists depends on the diagnostic criteria. The question then arises as to whether there is another way to estimate variance in the set of diseases. But first, we need to understand the nature of true conditions versus diseases.

True health conditions have an infinite number of states.

The true health condition of any person exists in a complex, multidimensional and continuous space. For example, Atherosclerosis is the disease approximation for a condition where coronary arteries harden and is known to be a risk factor for heart attacks. Factors contributing to this condition may include diet,[138] physical activity,[139] genetics,[140] hormones in circulation,[141] other heart or vascular conditions and many more.[142] Such a condition include many different dimensions along which to understand, analyze or learn about the condition, indicating a complex and multidimensional space. Comparing two patients may require many different features such as how long the patient has experienced the condition, how thick the sclerosing is, how widespread it is throughout the coronary arteries, how stable any associated plaques are and so forth.[142] Values of these feature may lie on a continuum, which implies an infinite possibility of states.

We learn using discrete approximations of health conditions called diseases.

The goal of medical diagnosis is to assign a patient with a complex, continuous condition to one or more discrete disease states that will lead to their optimal treatment.[12] Diseases are discrete approximations of some portion of the health condition continuum. Learning about diseases involves searching for previously unknown associations. For example, to learn what genes[125] or symptoms[124] are associated with a given disease or what drugs might treat a given disease.[143, 144] All of the data we have on disease lives in the discrete disease space, which is subject to a variety of limitations and biases.[10] For example, diseases are often defined by only a few select attributes such as anatomical location, symptoms observed, treatments, causative agents, biomolecular etiology, heredity and so forth.[46]

Mappings of continuous conditions to discrete diseases changes over time.

The way diseases have been discretized and clustered has changed over time. Cancers have been defined by anatomical location (breast, brain, pancreatic...) and then further refined to reflect tissue of origin (carcinoma, sarcoma, lymphoma...). Recent work suggests that genetic similarity is important, often across different tissue types.[58] It may also be that cancer is the not necessarily the optimal discretization of the continuous condition as recent work has shown they are largely an immunological disease.[145] Osteoarthritis(OA) is a more specific example. In 2007 the FDA requested clarification on the definition of OA and whether subcategories of OA could be combined and what was the relation to other specific diseases.[59] Other examples include Myocardial Infarction[60], Parkinsons disease[61] and more. Definitions therefore change over time to incorporate new information.

We can measure variance of disease looking at diseases over time.

Distinguishing one disease or set of diseases from the rest is called diagnosis, and depends on the definitions of disease.[146] Disease classifications are often revised to inform medical diagnosis[147, 148, 149] Given that true conditions live in continuous space and diseases diagnosed live in discrete space, this begs the question: when trying to learn new insights from data where disease is a variable, how much can we trust results given that the diseases themselves are approximations with error? One way to study this would be to look at how disease descriptions and relationship defining the disease space have changed over time. If history is consistent, then it may be the best predictor of how much things will change in the future. The ability to measure variance or uncertainty has revolutionized entire fields, as in the case of Shannon's development of information theory for communication.[150]

Estimating variance over time may help promote precision medicine.

One reason estimating variance or uncertainty is important is that current research has sought to specify disease more precisely, such as collecting more detailed data on phenotypes associated with diseases[151] and further dividing disease into potential subclasses.[152] Some have pointed out the limitations of the current paradigm of diagnosis based on taxonomy, but little research has been done on how the disease taxonomy has evolved over time and how it should change to adapt to precision medicine.[8] To accelerate research to yield more precise medicine and clinical practice to implement precise medicine, disease definitions must be precise, must be classified precisely and correspond to precise treatments. Precision medicine therefore requires more precise diseases.

We can estimating variance using changes to a disease taxonomy over time.

With the abundance of data available that represent different bases of human health[153], we hypothesize that there are better way to represent disease than current taxonomies. In this chapter we explore how discretized disease space – the National Library of Medicine's Medical Subject Headings (MeSH) taxonomy – evolves over 50 years. We also compare this taxonomy to other taxonomies to show the variation in disease relationships. We further demonstrate a quantitative measure of variance to see how much error might be present in the disease taxonomy. This should allow for more precise research in those areas to define diseases and identify effective therapies.

3.2 Background

There is no official description of the space of all diseases, but unofficially it is the compendium of all medical knowledge. The most official, formal description of the disease space available is the disease taxonomy. A taxonomy is a particularly convenient way to represent the disease space because of its hierarchical structure, which makes for easy communication.[63]

A taxonomy is used for classification into ordered categories, where classification is grouping objects by common features. One common example of a taxonomy in science is the National Center for Biotechnology Information (NCBI) taxonomy of species.[154] Defining and classifying diseases has been a medical endeavor since the ancient Greeks.[155] From then to now diseases have been classified by signs and symptoms, anatomical location, organs affected, "chemical theory," pathology, biochemistry, genes and specific mutations.[155] Since the 15th century, many scientists and physicians have advocated for classification of disease such as Paracelsus, Syndenham, de la Croix, Morgagni, Bernard and others.[155] Taxonomies were recognized as an important way to practice medicine and also to study disease. They were also shown to predict outcomes as some have shown today.[75] In our data-rich world today, taxonomies are used for aggregating and sharing knowledge.[156, 64].

Official, comprehensive taxonomies have been developed for purposes of research,[67] statistics and later even reimbursement[157]. There is not one official taxonomy of disease, although the structure of medical education, medical specialties, hospital departments and research activities reflect general medical taxonomies engrained in the study and practice of medicine. Taxonomies are updated regularly and sometimes a completely new taxonomy is proposed. For example, at the peak of the human genome project it was proposed to create a "Gene Ontology" to unify all biology.[158]

The American Medical Association (AMA) in 2004 created a new concise taxonomy to simplify their categorization of research and practice.[7] The National Library of Medicine (NLM) also advocated for a new taxonomy based on molecular medicine in 2011.[9] A more recent example is creating a new taxonomy for inflammation specifically based on cytokines, given a modern molecular approach to taxonomy.[159] These examples show that medical taxonomies are constantly changing based on new understandings of health conditions and disease.

A precise disease taxonomy would be one that could precisely predict outcome and treatments. Precision medicine is a term used to characterize more accurate and precise diagnosis and treatment of patients. From the IOM report "Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease," precision medicine was defined as "tailoring of medical treatment to the individual characteristics of each patient... the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment."[9] In the New England Journal of Medicine the goal of precision medicine to "create diagnostic, prognostic, and therapeutic strategies precisely tailored to each patient's requirements."[20] This effort has led to increased research funding and support such as the White House's Precision Medicine Initiative.[22] Estimating precision, or conversely variance, of the taxonomy is an important step in advancing precision medicine.

3.3 Methods

3.3.1 Constructing the MeSH History Dataset

Various taxonomies have been constructed for medical use, such as the international classification of disease (ICD) which found its roots as the Bertillion Classification of Causes of Death in 1893 and was updated as the International Classification of Causes of Death in 1900.[157] This taxonomy was to be revised approximately every 10 years and became the International Statistical Classification of Diseases, Injuries and Causes of Death in its 6th revision. There have been 10 versions so far with the current version having 155,000 terms. The Systematized Nomenclature of Medicine (SNOMED)[68] is a collection of medical terms that have been arranged in a hierarchy. In 2002 it became SNOMED CT and now includes broader relationships. The Medical Subject Heading (MeSH) vocabulary started in 1960 for indexing literature.[67] Prior to this, the Index Medicus had a vocabulary for indexing medical literature, which started in 1879 by the Library of the Surgeon General's Office of the United States Army.[67] We chose to use the MeSH vocabulary and taxonomy because it has yearly updates, giving 50 continuous years of medical taxonomy. It is widely used today, especially for indexing and retrieving medical literature, which contains much of the medical learning of the last several decades.

MeSH is currently available online with digital versions going back to 1997. For versions prior to 1997, only print versions are available. MeSH was printed in a few different publication formats. The taxonomy is contained in what is called the Tree Structures publication. The tree structures did not start until approximately 1965 and were initially restricted to a maximum depth of four. The first year there was no restriction on depth was 1971. Each node of the tree was assigned a tree number associated with its location. One disease could have multiple tree numbers corresponding to different locations. We obtained PDF copies of the MeSH tree structures through the HathiTrust[160] and manual scanning ourselves. We then used optical character recognition to identify each character on each page. Each of the 35 years contained 50-200 pages with each page including approximately 30 terms and tree numbers. The optical character recognition (OCR) process was only approximately 70% accurate, making the data very noisy. Additionally there were some pages missing or unreadable, which added still more noise.

In order to compare disease terms exactly we needed to clean the data. To clean up terms and tree numbers, we took advantage of the smoothness in the data to infer the correct disease names and tree numbers. There were three types of smoothness. The first was that of tree numbers on a page. Terms on a page were organized by tree number. For example Arboviurs Infections (C02.081) is following by African Horse Sickness (C02.081.030) following by Bluetongue (C02.081.125), with tree numbers always ascending. Deviations from this pattern were not allowed. The second was alphabetical order of disease terms within the same section, as seen in the previous example. The third type of smoothness was between years. A disease that showed up in 1971 and 1973 was very likely to show up in 1972. The same was true for the tree numbers. It was also unlikely that one name or number only occurred for one year in isolation. We used the smoothness to automatically correct easy cases (such as when one year was missing) and produced a flag to manually revise disease-tree number pairs when the case was not clear. We then ran a series of checks to make sure the data set was consistent and manually reviewed errors comparing to the original PDFs. We checked for (1) tree numbers associated with more than one disease (2) tree numbers that did not have a parent tree number in a given year (3) all disease names that were not in neighboring years (4) irregularities in tree numbers over years for a given disease. A final check used the repetition of tree numbers. Each entry listed the tree number of the current location along with part of the tree numbers for other locations. This allowed us to ensure we had the correct number of tree numbers for each disease.

The next step was to match diseases to each other from one year to the next. The main challenge in doing so is that disease terms change over time. Sometimes the concept was the same, but the name changed slightly, for example Cryptorchidism was introduced in 1995, but was previously Cryptorchism. Others changed the disease. For example, Gonadal Dysgenesis was introduced in 1980 but previously there was Turner Syndrome and Sex Differentiation Disorders. Some of the history is captured in historical notes and previous indexings for diseases documented in recent versions of MeSH. We extracted these relationships to reconstruct the history of a given disease. Additionally we automatically went through each year of the taxonomy documenting the year of each disease and location. We then manually checked diseases where the historical relations from our data and the MeSH history did not align to find undocumented or incorrect historical relations.

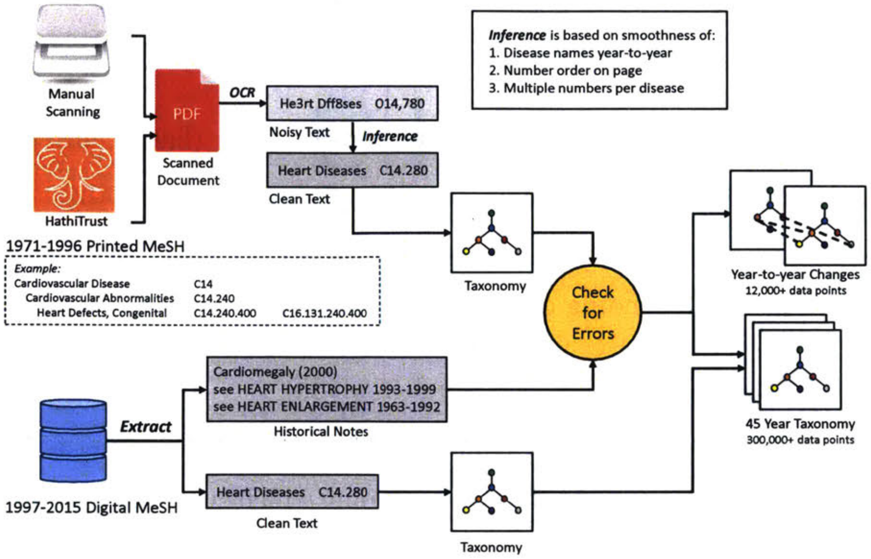


Figure 3-1: Schematic of how the data set was extracted and cleaned. Data was extracted from three data sources: manual scans of printed MeSH tree structures, scans of the MeSH tree structures from the HathiTrust and the digital versions of MeSH that started in 1997. Inference to determine correct terms and tree numbers was based on smoothness in the data and rules were set for error checking. The final data set includes a taxonomy for year each along with the relation of tree nodes between years.

3.3.2 Macro Characterization of MeSH Taxonomy History

To characterize how the taxonomy of disease has changed over time, we tracked four measurements over time. The first was to track the most detailed diseases from 50 years ago up to the present. The most detailed diseases are those that were not subdivided into further subtypes, or the leaves of the taxonomy tree. These may be viewed as the most specific diagnoses possible in the taxonomy. If a leaf changed to a branch of the tree, it would signify further subdividing of that disease into subclasses. We tracked this by finding the set of diseases that represented all leaves in the 1971 taxonomy. In all subsequent years we characterize each not node in one of three categories: (1) a disease that was a leaf in 1971, (2) a node that is a descendant of a node of type (1), and (3) all other nodes.

The second measurement was the number of diseases that occurred in a given number of locations in the taxonomy tree. For all years we categorized each disease by the number of nodes labeled with that disease name. There were between 1 and 19 nodes per disease depending on the disease and the year.

The third measurement was the number of diseases that occurred in a given number of general categories in the taxonomy tree. We define general categories as the highest level categories of the disease taxonomy which currently include categories such as Virus Diseases, Neoplasms, Eye Diseases, and Cardiovascular Diseases. For all years we categorized each disease by the number of unique general categories it belonged to. The set of categories for a given disease is found by identifying all nodes for a given disease and moving up the tree and finding the label of the ancestor that is a child of the root node.

The final measurement was the number of new diseases in the taxonomy. This was done by (1) looking at where new diseases occur in the taxonomy by year and (2) plotting the number of diseases in each level of the taxonomy tree by year. New diseases were identified as disease terms that did not occur in the previous year and did not have an alternate name in the previous year.

The taxonomy has several thousand nodes, with an increasing number over the 50 years. To plot such a large graph we use a toolkit developed for visualizing phylogenetic trees.[161] We color the tree node and the link connecting the node to its parent to highlight groups of nodes. We found a circular layout with nodes and the edge directly above each node colored was easiest to visualize and interpret.

Diseases with largest number of locations

We found that particular types of diseases had more node locations than other diseases. In Table 3.3 we examine all diseases in 2015 that have more than 10 node locations. We noted that many diseases seemed to either be (1) genetic, (2) be cancer or (3) have the word syndrome. We went through and categorized each in the following way: (1) if there was a genetic component detailed in the MeSH description such as a specific gene or mutation, (2) included under the Neoplasms category and (3) if the word 'syndrome' is included in the name or its MeSH description. For diseases that did not fit one of the categories we suggested a categorization based on the MeSH description.

We then reversed the analysis and counted the number of locations per disease for those categorized described above. We calculate this number for three categories for each year. To generate this data we need a method for automatically categorizing diseases instead of the manual method we used in the previous paragraph. To do this we determined that any diseases including the word 'syndrome' in the name could be categorized as a syndrome. To determine if a disease was genetic we looked to see if it was in the OMIM database. We plotted the mean number of locations for these two categories compared to all diseases in Figure 3-5.

Disease overlap

To visualize the overlap of diseases within different disease categories, defined as the highest level of the disease taxonomy, we used a plot of connected arcs. The 360 degrees of the arc corresponds to the number of diseases in 2015. Each disease that only occurred in one category was counted as a self-arc. For diseases shared between multiple categories, an arc was created between each pair of categories inversely proportional to the number of categories. To illustrate, if a disease was found to occur in only one category, it would add 1/N width to the self-arc of that category. If the disease occurred in two categories, an arc between the two categories would be added with width 1/2N so that each end of the arc added together was 1/N or the width of one disease. If the disease occurred in three categories, three arc would be added, each with 1/6N width so the total would add to 1/N again. In this way the width of the arcs between categories is actually underestimated. This approach was necessary because only pairwise interactions could be visualized, but diseases with more than two categories was common. The graph of 1971 and 2015 is shown in Figure 3-6.

3.3.3 Micro Characterization of MeSH Taxonomy History

To characterize how the taxonomy changes over time at the level of individual disease locations in the tree, we traced each change from one year to the next. We chose to document changes in diseases or disease node locations instead of a single metric like tree edit distance (see Appendix, Section B.4). The tree operations we identified are shown in Table 3.1 with an explanation for identifying them. Operations are identified on a disease-by-disease basis and may affect the entire disease (e.g. - new disease, deleted disease, or name change operations) or only a specific location of a disease (e.g. - location change or split operations). We plot the number of different types of changes over time.

We studied what factors might determine what types of taxonomy operations occurred. One factor of interest was how long had the disease been considered a disease. Records rarely exist stating when a disease name precisely came into use, but we could search historical medical texts for disease names. The National Library

Operation	Description
New Disease	A disease exists in a year that could not be traced to any diseases in the previous year.
Deleted Disease	A disease exists in a year that could not be traced to any diseases in the following year.
Name Change	A disease exists in a year that can be traced to one disease in the previous year that does not exist in the current year.
Location Change	A node is added or deleted for a given disease relative to the previous year.
Split	A disease has a child disease that was not a child disease in the previous year.

Table 3.1: Description of taxonomy operations from one year to the next.

of Medicine (NLM) has a freely available corpus of historical medical books, mainly from 1800's. The data is based on optical character recognition (OCR) of scanned books. We split the set of diseases from the first year of our data set (1971) into two sets: (1) disease names that occurred in the historical books and (2) those that did not. We then tracked the number of taxonomy operations that occurred for disease in each of the two sets. The data is plotted in Figure 3-7.

3.3.4 Estimating Variance Among Taxonomies

In addition to analyzing the MeSH taxonomy over time, we compare it to other taxonomies currently used. The International Classification of Disease (ICD) was originally developed to track statistics of diseases and is now additionally used for reimbursement purposes.[157] The Systematized Nomenclature of Medicine (SNOMED)[68] is a collection of medical terms that have been arranged in a hierarchy. To compare terms in different taxonomies we used the United Medical Language System (UMLS) to match terms to each other.[120]

The Disease Ontology (DO) is a more recent taxonomy developed to be a comprehensive knowledge base of human disease including inherited, developmental and acquired.[71] It has served as a tool for linking biomedical data by human diseases.[69, 70] We did not compare to this taxonomy because UMLS does not have DO mappings and the mappings provided in DO do not appear to preserve generality of terms.

We compare ancestor/descendant relationships in the different taxonomies to examine variation across taxonomies. To do this, we enumerate all disease ancestor→descendant relationships in a taxonomy. For example, 'Heart Defects, Congenital' shows up in three locations:

1. Cardiovascular Diseases→Heart Diseases→Heart Defects, Congenital
2. Cardiovascular Diseases→Cardiovascular Abnormalities→Heart Defects, Congenital
3. Congenital, Hereditary, and Neonatal Diseases and Abnormalities→Congenital Abnormalities→Cardiovascular Abnormalities→Heart Defects, Congenital

would result in the following ancestor→descendant relations:

1. Heart Diseases→Heart Defects, Congenital
2. Cardiovascular Diseases→Heart Defects, Congenital
3. Cardiovascular Abnormalities→Heart Defects, Congenital
4. Congenital Abnormalities→Heart Defects, Congenital
5. Congenital, Hereditary, and Neonatal Diseases and Abnormalities→Heart Defects, Congenital

We then convert the relations into the UMLS unique identifiers (CUI's). To compare the relations, we calculate the percentage of relations preserved between two taxonomies for which both terms exist in both taxonomies. For example, if 'Heart Disease' and 'Cardiovascular Disease' are both in MeSH and ICD9, we might expect 'Heart Disease' to be a subset of 'Cardiovascular Disease' in each. More precisely this can be written as

$$\frac{\sum_{d_1,d_2 \in D} \mathbb{1}_{T_A}(r_{12}) * \mathbb{1}_{T_B}(r_{12}) + \mathbb{1}_{T_A}(r_{21}) * \mathbb{1}_{T_B}(r_{21})}{|D|}$$

where D is the set of all disease pairs that occur in both taxonomies T_A and T_B, r_{ij} is the relation $d_i \rightarrow d_j$, and $\mathbb{1}_{T_X}(r)$ is an indicator function for whether or not a relation r occurs in a taxonomy T_X, given by:

$$\mathbb{1}_{T_X}(r) := \begin{cases} 1 & \text{if } r \in T_X, \\ 0 & \text{if } r \notin T_X. \end{cases}$$

3.3.5 Estimating Disease Taxonomy Variance

A taxonomy is used for learning to aggregate data by utilizing inferred relations. These inferred relations are a function of the taxonomy and the taxonomy is a function of time. Given a data set of diseases-feature associations, the inferred disease-feature data could change due to the changing taxonomy. Figure 3-2 depicts visually how the measure of variance is calculated. Taxonomies are commonly used to aggregate features of concepts. For example, in Figure 3-2 concept A has no features (represented by colored boxes) associated with it, but inherits the features associated with all of all concepts that are its descendants. In the disease taxonomy, each disease can be viewed as a collection of features, shown in a bucket in the figure. With a fixed data set of disease-feature association, each collection of features or bucket may change over time due to the changing taxonomy.

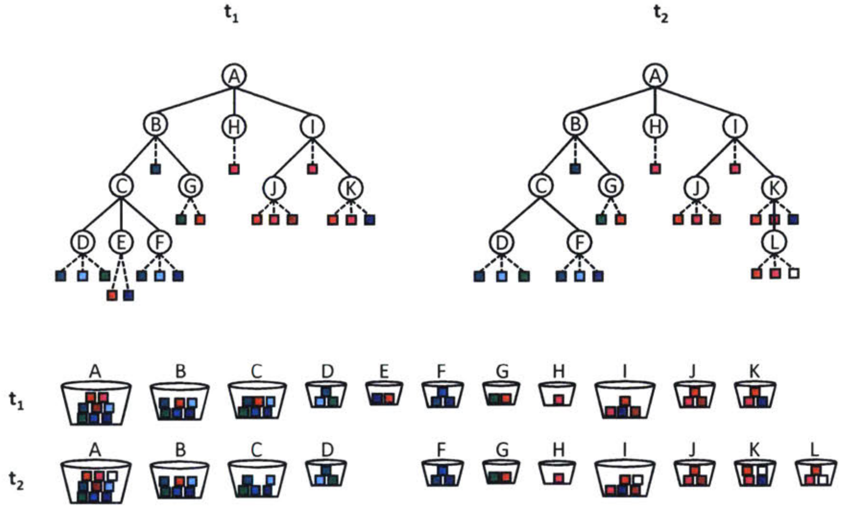


Figure 3-2: A graphical depiction of how variance can be measured in the taxonomy over time. On top each taxonomy has a set of concepts labeled with different letters and the same set of features associated with concepts represented by colored squares. On the bottom are two sets of buckets corresponding to the taxonomies showing how concept-features associations have changed.

We quantify the difference between the set of features association with a disease in the taxonomy at two different time points using Jaccard similarity. The Jaccard similarity of two sets A and B is given by the following equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

For diseases, we define the set of features associated with a disease d and its descendants at time t as $F(d, t$. We can then write the Jaccard similarity as:

$$J(d, t, \Delta t) = \frac{|F(d, t) \cap F(d, t + \Delta t)|}{|F(d, t) \cup F(d, t + \Delta t)|}$$

which gives a similarity measure for one disease at two time points. To generate statistics over the entire taxonomy we can generate a distribution over all diseases and all time periods between taxonomies possible. This can be written:

$$p(d, \Delta t) = \frac{1}{45 - \Delta t} \sum_{t=1971}^{2015-\Delta t} J(d, t, \Delta t)$$

We use four distinct data sets of features including genes from OMIM and PheGenI, journal articles from PubMed and clinical trials from clinicaltrials.gov. For each of these data sets, researchers have used MeSH to aggregate data.

For each year we find the inferred disease-features by passing features of one node to all its parents. For each disease we then trace over time for which features are associated with that disease.

3.3.6 Estimating Variance within a Disease

The variance we measure in the previous section is based on the observation that diseases are combined into discrete higher level disease concepts. it is not based on mapping continuous variables into the discretized disease space because data does not exist to do so as previously noted. One way approximate this though is to find a set of diseases where the incidence should not change over some time period because the genetic or environmental factors causing the disease would not likely change over that given time period. Many cancers would fall under that category for a time period of a several years. We make this assumption because the known environmental factor that cause cancer change slowly and the disease develops over the course of years. The changes in incidence on short time scales would then be due to changing diagnostic criteria or screening practices.

We use the SEER data set [162, 163] to plot the incidence of difference cancers over time. The incidence is given in the data by year from 1973 to 2011 for 128 different cancers. We then estimate what the anticipated variance might be for a randomly chosen type of cancer. We calculate a distribution over Δt where t is measured in years. The formula for calculating the distribution is given by:

$$p(\Delta t) = \frac{1}{|D|} \sum_{d \in D} \sum_{t=1973}^{2011-\Delta t} \frac{|I(d, t + \Delta t) - I(d, t)|}{I(d, t)}$$

where $I(d, t)$ is the incidence of disease d at time t.

3.4 Results

3.4.1 Macro Taxonomy Changes over Time

We show how the taxonomy structure changes as a whole by plotting the entire tree structure over time. Figure 3-3 shows the taxonomy at three different time points: 1971, 1990 and 2015. Our data set starts in 1971, the final year of the data set is 2015 and an approximate midpoint is 1990. The center of the circle represents the root of the tree and the farther away from the root radially represents greater depth in the tree. We note that the depth of different parts of the tree increases over time. We plotted each taxonomy with three different node/edge color schemes in the three columns of Figure 3-3.

The first column shows the leaf nodes and the corresponding edges in 1971 colored blue with all other nodes/edges in 1971 colored black. Leaf nodes are diseases that represent the most specific diseases or ones that are not broken down into more specific diseases. In subsequent years we plot descendants of the 1971 leaves in red. The diseases in red represent diseases that were newly introduced after 1971 that

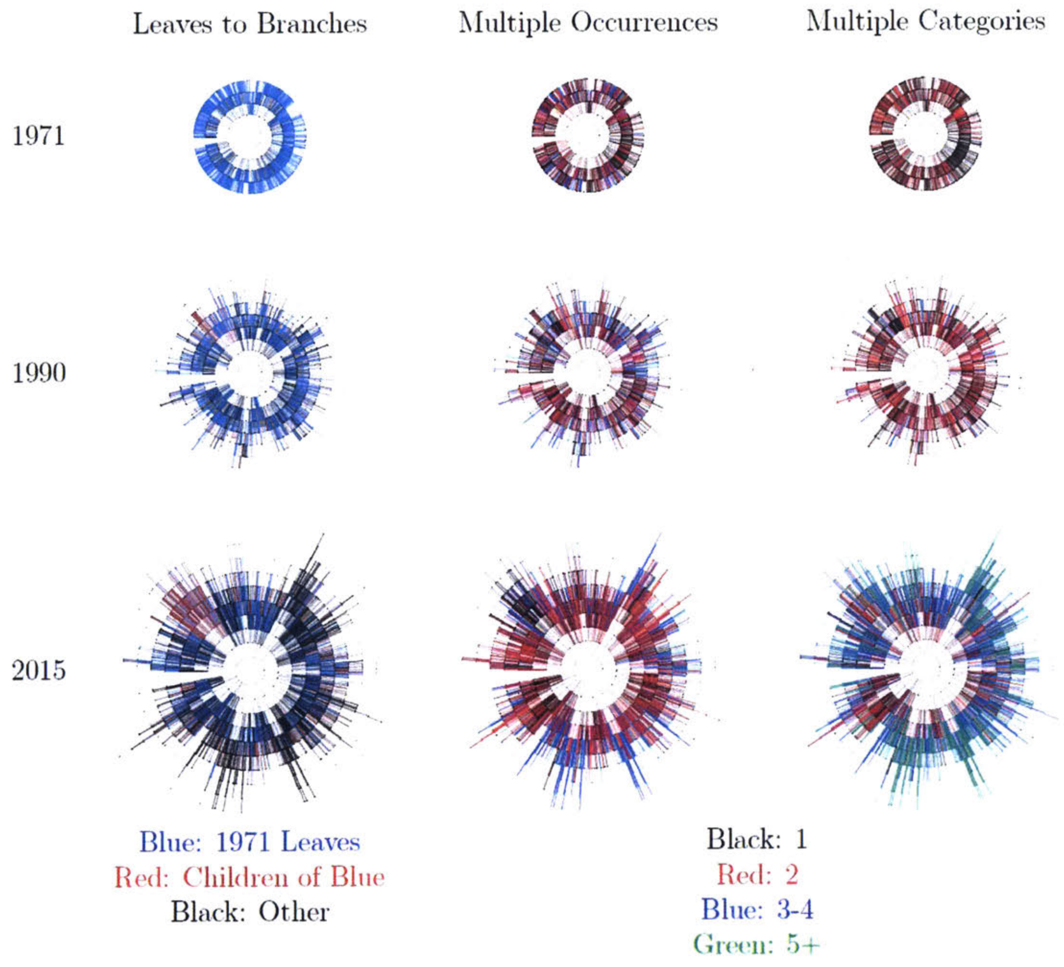


Figure 3-3: The entire MeSH disease taxonomy is plotted for three different years as rows: 1971, 1990, and 2015. In the first column, taxonomy leaves are plotted in blue in 1971 and descendants of those leaves are plotted in red while all other diseases are plotted in black. In the second column diseases occurring in multiple parts of the tree are plotted according to the following color scheme for number of tree locations: 1 (black), 2 (red), 3 times (blue), 4 times (orange) or 5+ times (green). In the third column are diseases occurring in multiple categories of disease, which correspond to the top level of disease classification in the MeSH taxonomy such as 'Neoplasms' and 'Cardiovascular Diseases.' The same color scheme from the second column is used.

subdivided what had previously been a leaf disease. We found that the subdividing process was non-uniform with some diseases subdividing several times while others do not subdivide at all. This may give us an indication of how much our current diseases may be broken into more specific diseases in the future.

The second column of Figure 3-3 shows all diseases colored by the number of tree nodes per disease. We found that over time the number of locations per disease on average increased. If the taxonomy partitions diseases neatly, we would expect diseases to show up in few places. At the opposite extreme, if all diseases occurred in all possible locations, the tree would contain no information. To see if diseases with multiple locations were restricted to the same disease category (again defined as the highest level category in the taxonomy), we plotted how many categories each disease occurred in, seen in column three. We observe that many diseases occur in many (5+) categories. If categories nicely partitioned diseases we would expect diseases to have one location and most diseases would show up black. Instead we see diseases that show up in multiple locations showing the overlap in the taxonomy. We also see that the tree overlaps more and more over time. By 2015 the majority of the diseases show up in several disease categories.

We quantify these changes in the Figure 3-4. The tree and diagram on the top left and matrix on the top right show how each measure is calculated. The depth of a node is defined as the minimum number of edges away from the root seen in the top left plot. The number of nodes at depths 1-3 remains relatively constant over time while the number at depths 4-6 represent the vast majority of increase in number of nodes over time. There are also a few nodes that occur at extreme depths (8+). In the top right plot we see the number of nodes by height, defined as the minimum number of edges to a leaf node. It can be seen that the vast majority of nodes added have height 0 or in other words are leaves of the tree.

The bottom plots show the number of diseases with a given number of nodes (bottom left) or number of categories (bottom right). We were surprised to see that

Measure	1971	2015	2065 (modeled)
Number of Diseases	2267	4662	7430
Number of Nodes	3130	11398	24175
Average Nodes per Disease	1.38	2.44	3.25
% of diseases with more than 3 locations	10.1%	19.5%	50.4%
% of diseases with more than 2 categories	2.2%	15.1%	31.7%
% of category pairs with greater than 5% overlap	3.7%	9.5%	26.6%

Table 3.2: Past, current and projected MeSH tree statistics. Regression plots can be found in Figure B-1 in the Appendix. The complexity of the taxonomy could increase by approximately 2-3 fold over 50 years.

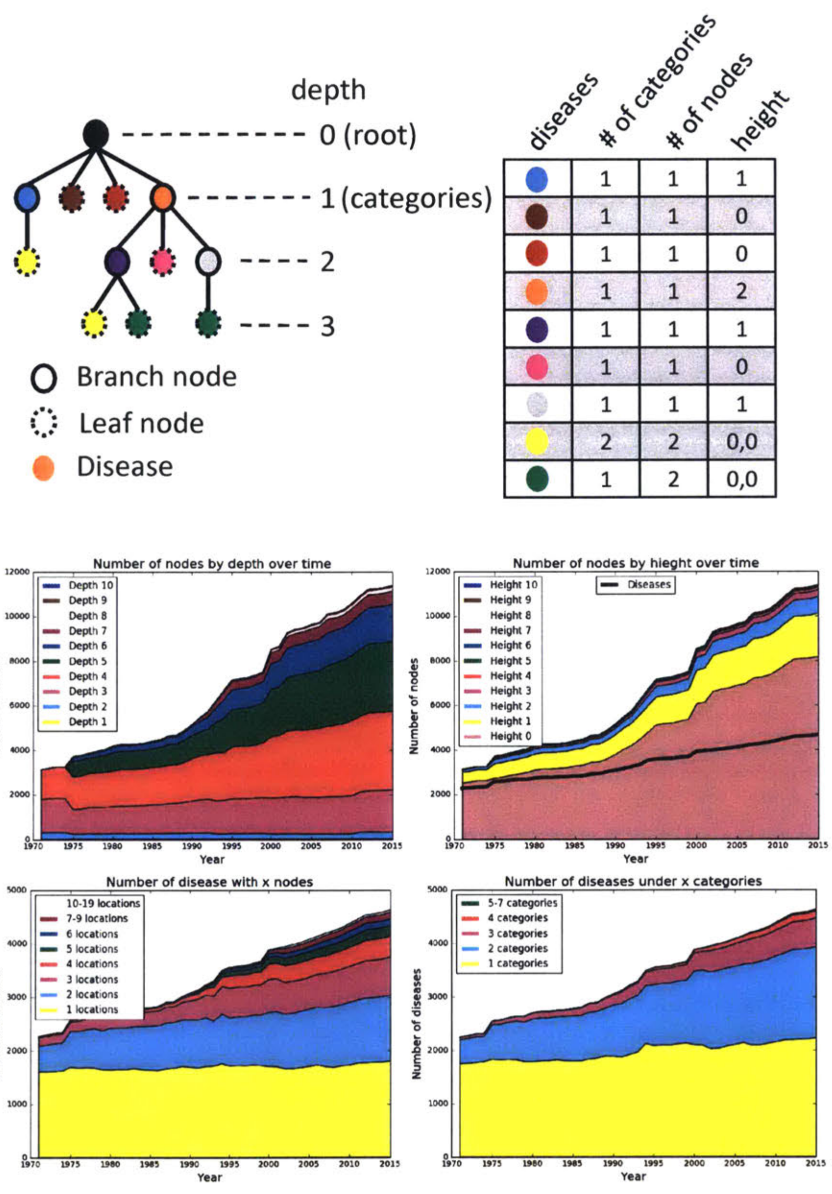
the number of diseases that occur in one location or occur in one category remains relatively constant. Explanations consistent with this finding could be that those are the diseases that we have not learned much about and therefore have not changed or that the taxonomy is constructed so those diseases are classified neatly and they only require one location.

One characterization of the disease taxonomy is that it has become too large and complex for physicians to have a handle on. We plotted the measures we have explored and used linear regression to fit a line to the data as seen in Figure B-1 in the Appendix. We report in Table 3.2 the statistics from 1971 and 2015 along with projected measures for 50 years into the future if the trends continue. If the trend does continue, we will have a taxonomy in 50 years that increases in complexity 2-3 fold depending on the metric used.

To understand why some diseases have so many tree node locations, we looked for patterns in the diseases with the most diseases or categories. We did not find consistency over time in diseases with the most nodes. In other words, the diseases in 1971 with the most tree nodes did not have the most nodes in 2015, nor vice versa. Neither did we find any observable patterns in the disease categories. Further details can be found in Figure B-2 in the Appendix. We did find a noticeable pattern for diseases considered syndromes or genetic diseases. Figure 3-5 shows the mean number of tree node locations per disease for all diseases, for syndromes and for genetic diseases. The syndromes and genetic diseases have significantly more tree node locations on average.

We also examined all of the diseases with the most tree node locations setting a threshold of 10 tree locations in 2015 and classified each disease as genetic, syndrome or other as seen in Table 3.3. We considered a disease a syndrome if syndrome was in the name or the MeSH disease description. We considered a disease genetic if there was a reference to a genetic cause in the MeSH description. We classified all other diseases into general categories and found a large number of cancers and only a few diseases that would fit in other categories. It was clear that the majority of diseases were genetic diseases, syndrome or both. It appears that these types of diseases do not fit neatly into the categories in the taxonomy. We describe potential reasons why in the Discussion.

Our final macro characterization is to examine the data from the opposite perspective, not from diseases but from categories. In Figure 3-6 we plot the amount of overlap between disease categories in 1971 and 2015. The circumference of each circle represents the total number of diseases in 2015. Each arc length with a different MeSH tree number represents a diseases category. The same colors between 1971 and 2015 represent the same categories at the two different time points. Sometimes a category divides into multiple categories such as 'Infectious Diseases (C1.)' changing to 'Bacterial Infections and Mycoses (C01),' 'Virus Diseases (C02),' and 'Parasitic Diseases (C03).' The arcs within the circle represent the amount of overlap between categories as the number of diseases that occur in each category. We observe that most diseases in 1971 have a majority of diseases unique with sparse connections to other diseases. In 2015 the number of unique diseases often represents a small minority and the connections are dense. It can be seen that in the past the disease categories did a much better job at neatly partitioning diseases, but that is not the case today given increased the understanding of disease resulting in multiple categories.



diseases	# of categories	# of nodes	height
●	1	1	1
●	1	1	0
●	1	1	0
●	1	1	2
●	1	1	1
●	1	1	0
●	1	1	1
●	2	2	0,0
●	1	2	0,0

Figure 3-4: The tree diagram and matrix on the top visually show the tree measures in the plots. The top left plot shows the number of nodes by depth demonstrating that the majority of growth in the tree is at depths 4-6. The top right plot shows the number of nodes by height demonstrating that the vast majority of nodes are leaves. The bottom left plot shows the number of diseases stacked and colored by the number of tree node locations. The bottom right plot shows the number of diseases stacked and colored by the number of disease categories.

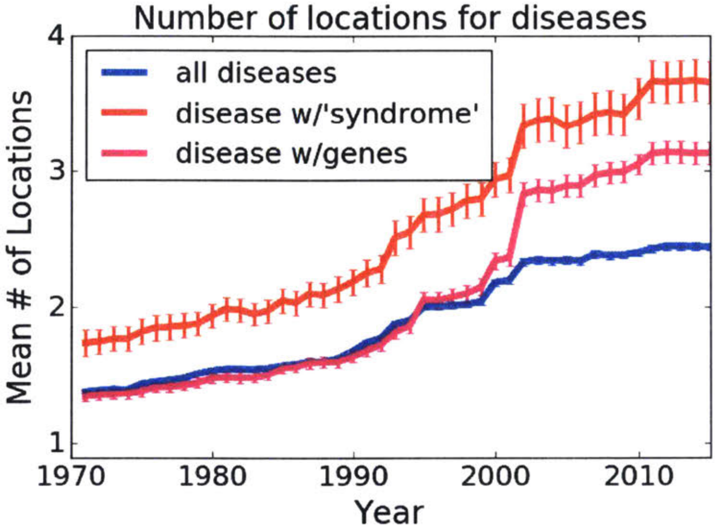
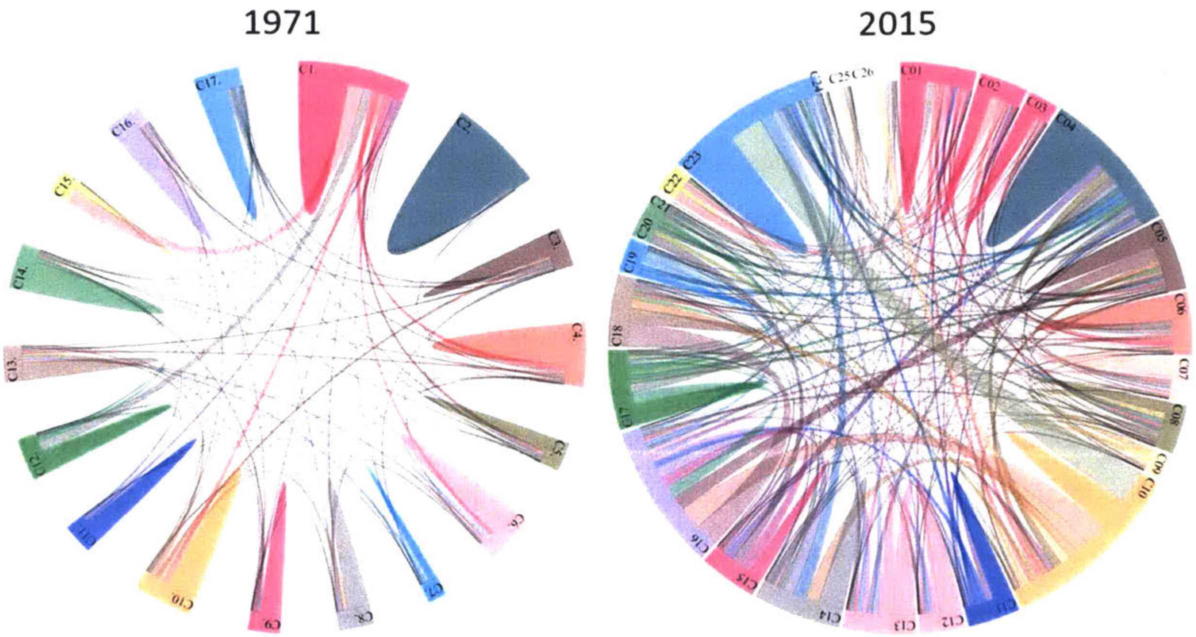


Figure 3-5: Number of tree node locations per disease for all diseases, syndromes and those with associated genes. The number of locations for syndromes and genetic diseases is significantly higher.

Our Categorization	MeSH Disease
Genetic Syndrome	Adrenogenital Syndrome
	DiGeorge Syndrome
	Oculocerebrorenal Syndrome
	Gardner Syndrome
	Lesch-Nyhan Syndrome
	WAGR Syndrome
	Kinky Hair Syndrome
	Wolfram Syndrome
	Lesch Nyhan Syndrome
	Usher Syndromes
	Denys-Drash Syndrome
	Turner Syndrome
	Adenomatous Polyposis Coli
	Menkes Kinky Hair Syndrome
	22q11 Deletion Syndrome
	Polyposis Syndrome, Familial
Genetic	Tay-Sachs Disease
	Sphingolipidoses
	Familial Hypophosphatemic Rickets
	Gangliosidoses
	Sphingolipidosis
	Leukodystrophy, Metachromatic
	Neurofibromatosis
	Tay-Sachs Disease, AB Variant
	Gangliosidoses GM2
	Refsum Disease
	Niemann-Pick Diseases
	Adrenoleukodystrophy
	Angiokeratoma Corporis Diffusum
	Sandhoff Disease
	Leukodystrophy, Globoid Cell
	Neurofibromatosis 2
	Gaucher Disease
	Fabry Disease
	Cerebral Amyloid Angiopathy, Familial
	Sertoli-Leydig Cell Tumor
Cancer	Lymphoma, Mixed-Cell, Follicular
	Lymphoma, Small Cleaved-Cell, Diffuse
	Lymphoma, Large-Cell, Follicular
	Optic Nerve Glioma
	Lymphoma, Lymphoblastic
	Androblastoma
	Lymphoma, Histiocytic
	Lymphoma, Large-Cell, Diffuse
	Arrhenoblastoma
	Lymphoma, Mixed-Cell, Diffuse
	Lymphoma, Small Noncleaved-Cell
	Lymphoma, Small Cleaved-Cell, Follicular
	Nephroblastoma
	Burkitt Lymphoma
Syndrome	Neuromyelitis Optica
	Hermanski-Pudlak Syndrome
Trauma	Cerebral Hemorrhage, Traumatic
Pre-Cancer	Colonic Polyps
Autoinflammatory Syndrome	Reiter Syndrome
Inflammatory	Lupus Vasculitis, Central Nervous System

Table 3.3: Diseases that have more than 10 nodes in 2015. We categorized each disease using the term description in MeSH. Syndromes include the word 'syndrome' in the name or description. Genetic disease have a genetic basis in the description. We classified all others according to the MeSH description.



1971	2015
C1. - Infectious Diseases	C01 - Bacterial Infections and Mycoses
	C02 - Virus Diseases
	C03 - Parasitic Diseases
C2.- Neoplasms, Cysts and Polyps	C04 - Neoplasms
C3. - Musculoskeletal Diseases	C05 - Musculoskeletal Diseases
C4. - Digestive System Diseases	C06 - Digestive System Diseases
N/A	C07 - Stomatognathic Diseases
C5. - Respiratory Tract Diseases	C08 - Respiratory Tract Diseases
N/A	C09 - Otorhinolaryngologic Diseases
C6. - Urogenital System Diseases	C12 - 'Male Urogenital Diseases
	C13 - Female Urogenital Diseases and Pregnancy Complications
C7. - Endocrine Diseases	C19 - Endocrine System Diseases
C8. - Cardiovascular Diseases	C14 - Cardiovascular Diseases
C9. - Hemic and Lymphatic Diseases	C15 - Hemic and Lymphatic Diseases
C10. - Nervous System Diseases	C10 - Nervous System Diseases
C11. - Sense Organ Diseases	C11 - Eye Diseases
C12. - Skin Diseases	C17 - Skin and Connective Tissue Diseases
C13. - Diseases of Nutrition and Metabolism	C18 - Nutritional and Metabolic Diseases
C14. - Injury, Immune Disease, Poisoning	C20 - Immune System Diseases
	C21 - Disorders of Environmental Origin
C15. - Diseases Exclusively of Animals	C22 - Animal Diseases
C16. - Neonatal Diseases and Abnormalities	C16 - Congenital, Hereditary, and Neonatal Diseases and Abnormalities
C17. - Symptoms and General Pathology	C23 - Pathological Conditions, Signs and Symptoms
	C24 - Occupational Diseases
	C25 - Chemically-Induced Disorders
	C26 - Wounds and Injuries

Figure 3-6: Overlap of diseases categories at two time points. The 360 degree arc length represents the number of diseases in 2015. The arc length for each disease category represents the number of diseases in that category. The width of the lines between categories represents the number of overlapping diseases. The table shows the names of the disease categories in 1971 and 2015.

3.4.2 Micro Characterization of MeSH Taxonomy History

There is a prevalent assumption that diseases change relatively little over time, but split into more refined disease definitions.[9] To test this assumption we characterize all tree operations, which are described in Table 3.1. These operations plotted over time are found in Figure B-4 in the Appendix. The majority of operations were tree node location changes. New diseases were the the next prevalent with the majority of new diseases showing up in leaves. Subdivides were less common than re-arrangements of tree nodes for a disease. It appears that diseases are less likely to be refined into more specific diseases. It is more likely that new aspects of a disease are learned, which changes where it shows up in the taxonomy.

We found that the disease operations do not appear to be more prevalent for certain categories of disease. We did find a relationship between the number of operations and how long a disease had been considered a disease. To determine if diseases had been considered diseases for a long period of time before our data set starts, we looked at the set of medical books from the 1700-1800's available on the National Library of Medicine website.[164] There are hundreds of books available with accompanying optical character recognition (OCR) text data. We searched through all books for the terms found in 1971 to determine if the diseases existed in the previous centuries. Due to the noise in the OCR text, our searching can be viewed as a lower bound on possible occurrences.

In Figure 3-7 we plot the percent of diseases that have an operation depending separately for whether they occur in the historical medical books or not. We observe that diseases present in historical medical books are approximately half as likely to undergo a name change or be deleted. They are also twice as likely to subdivide, while there is no noticeable difference in the rate at which they change locations.

It appears that diseases that have been considered diseases for a long time are less likely to change in name, but just as likely to change in terms of locations in the tree. Two possible explanations are that (1) diseases that have been called by a term for a long time are less likely to change name because of convention or (2) there is a process of refining diseases and those that survive the process more closely resemble a consistently identifiable condition. We expand on the second explanation in the discussion.

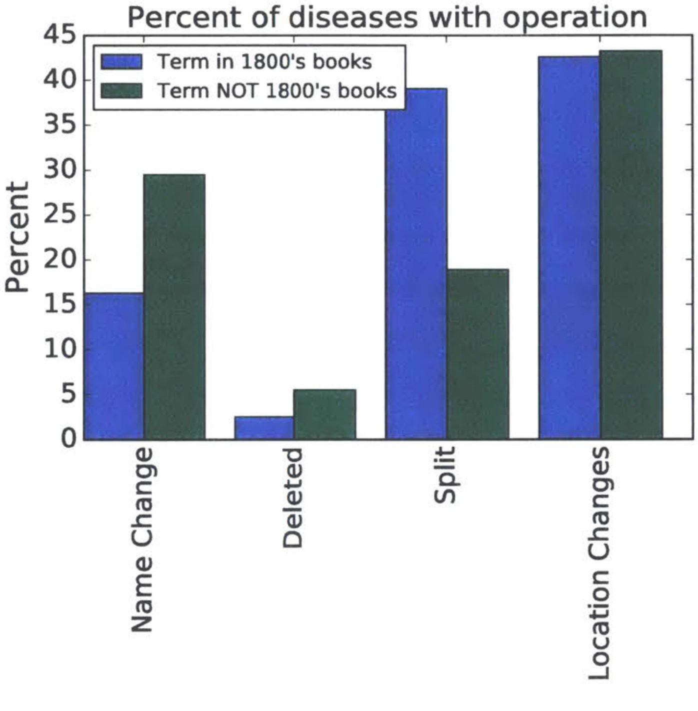


Figure 3-7: Percent of diseases with operations determined by two disease sets. One set is diseases from 1971 MeSH that can be found named in medical books from the 1700-1800s and the other set is diseases found in 1971 MeSH not found in those books.

3.4.3 Estimating Taxonomy Variance

Up to this point we have studied how the structure of the taxonomy has changed over time and characterized how much variance there is with internal measures. We now look at how the disease taxonomy is used for learning to characterize what variance a changing taxonomy introduces to the learning problem.

One of the primary uses of any space, and a taxonomy in particular, is to make inferences. A taxonomy is structured so that everything below a node shares something in common and is also distinct in some way from all other nodes in the taxonomy. MeSH has been used for retrospective statistical learning to aggregate data using this assumption. For example, if a disease is associated with a symptom or a gene or a patient, it may also be assumed that any ancestors of the disease also have the same relationship even if not explicitly stated.

Using the structure of the taxonomy, we developed a way for quantifying variance of the taxonomy over time based on external data sets. For any disease-feature associations, ancestor disease-features can be inferred. We identified how inferred ancestor-disease features change over time as the taxonomy changes. This can be quantified as a percentage of associations that change over a given time period. If disease-disease hierarchical relationships were constant over time, the percent change would be zero.

Figure 3-8 shows the variance measure plotted for gene data including OMIM and PheGenI (GWAS). Figure 3-9 shows the variance measure plotted for PubMed papers (Pmid) and Trials from clinicaltrials.gov. The left column for each figure shows the number of diseases while the right column shows the percent of diseases. The different colors represent different time periods the change was measured over (1, 5, 10, 20 and 44 years) and the x-axis is the Jaccard similarity, or percent similarity (intersection over union of sets).

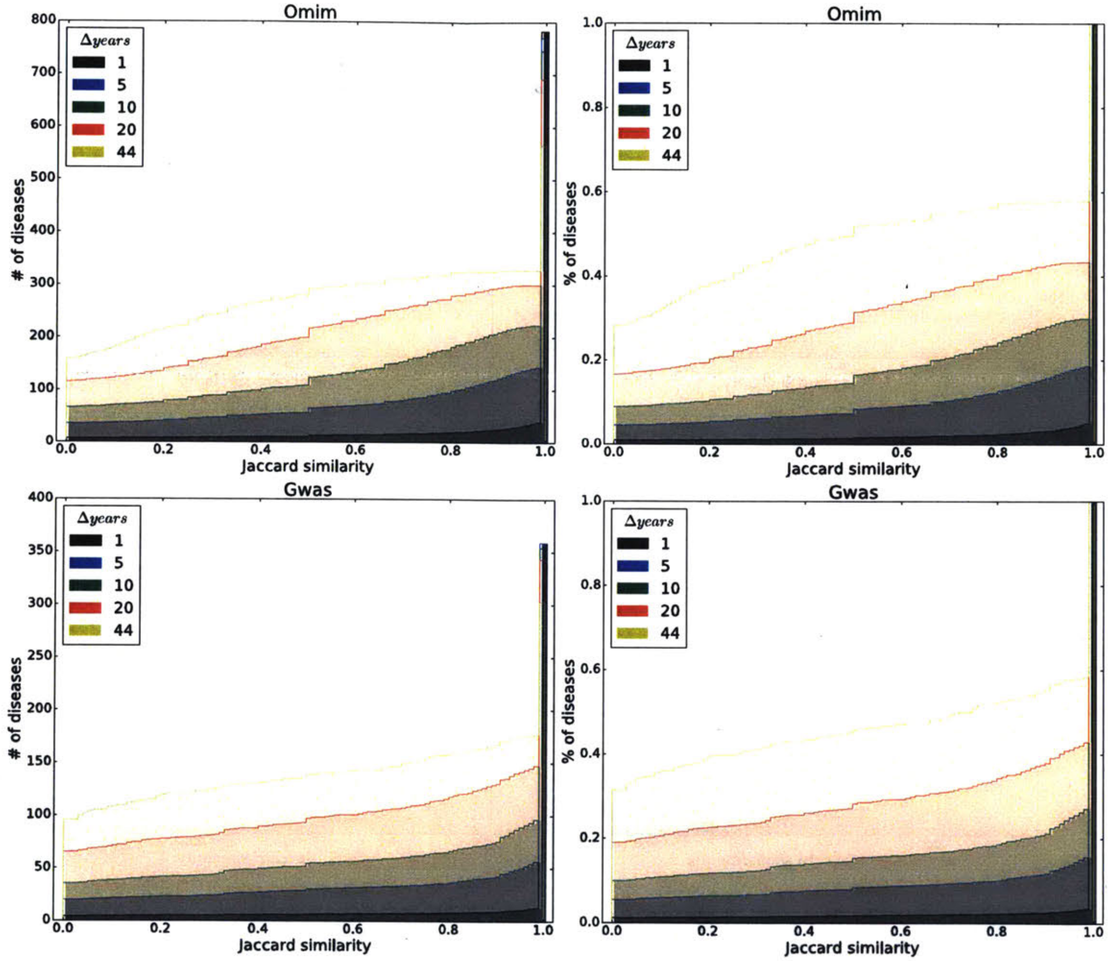


Figure 3-8: Distribution of variance for diseases for five different year differences. The OMIM and PheGenI (GWAS) databases used to create the distributions are based on disease-gene relations. Approximately 30% of diseases change by more than 0.5 over the course of 20 years.

For the gene data, we were surprised to see that over a time difference of 10 years, the similarity was only 0.5 for 20% of the diseases. This indicates that for randomly chosen diseases and a randomly chosen starting date, over the next 10 years, 20% of diseases would have 50% or more different genes. The medical literature and trials

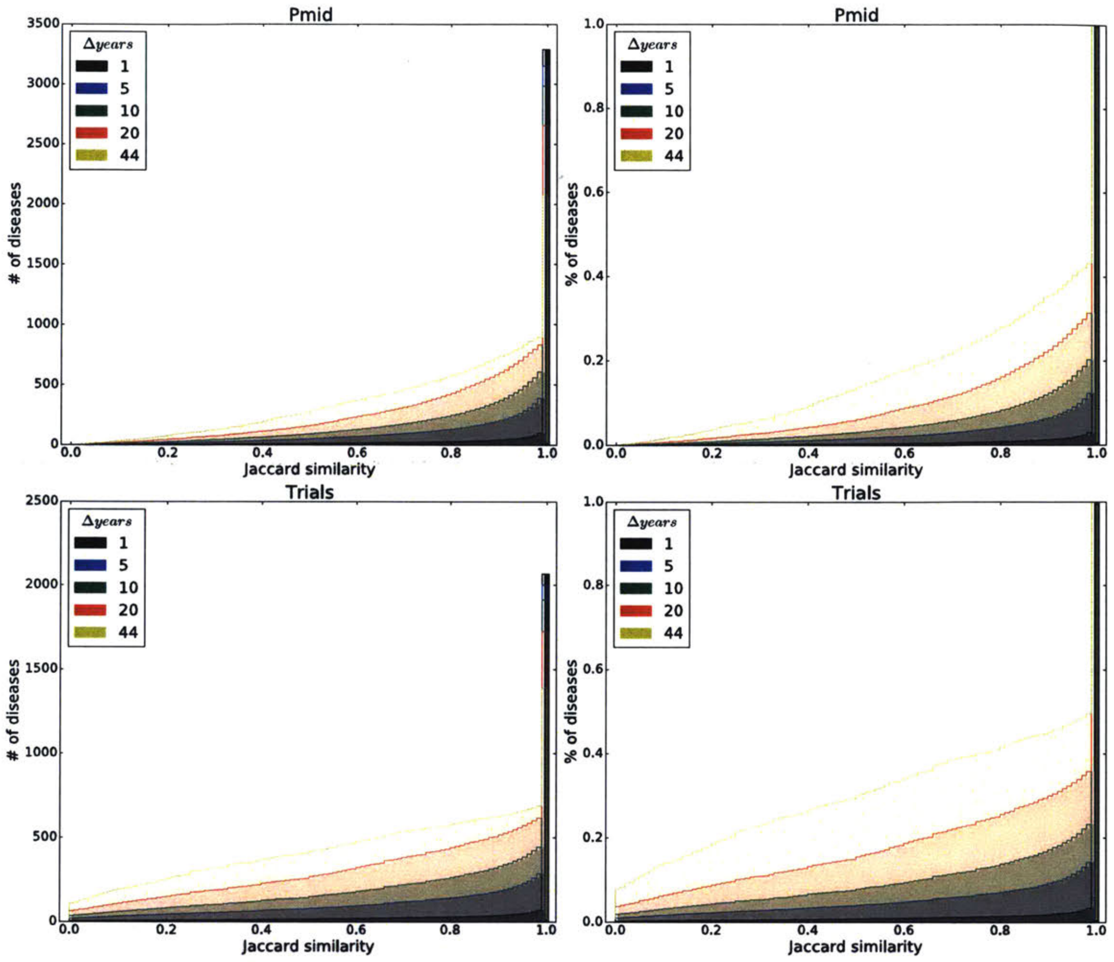


Figure 3-9: Distribution of variance for disease for five different year differences using two databases: PubMed papers and clinicaltrials.gov Trials.

were less pronounced, but still showed significant changes. The entire distribution for each can be found in Figure B-6 in the Appendix.

To reiterate, Figures 3-8 and 3-9 were made with a fixed data set of disease-feature relations where the features were genes, papers and clinical trials respectively. The only thing that changed over time was the taxonomy which led to different inferred disease-feature associations. This result suggests that using the same disease-feature data sets with a future taxonomy would also generate different results. These plots give us a measure of how much variation there is in the taxonomy with respect to a given external data set. For example, our results would indicate that using the taxonomy to infer relations would be more justified for papers and clinical trials than genes.

Our estimate also allows us to to put an estimate on the variance introduced by using the taxonomy in learning. For example, if we use the MeSH taxonomy to study clinical trials and infer diseases that are associated with trials, using our method we could put an error bar on the results. This should allow researchers to put an error bar on their learning problems that use the taxonomy. We have not seen work using a taxonomy for inferred disease-feature ever estimate an error bar.

We additionally found that the changes are not necessarily monotonic. Approximately 10% of disease-gene changes reversed, whether it was an addition or removal. An example is shown in Figure 3-10 showing how diseases associated with Alzheimer Disease would have changed over time due to the changing taxonomy. We then show in Figure 3-11 how the Jaccard similarity would be calculated for Alzheimer Disease at different time points. Another example, Tay-Sachs Disease, is seen in Figure B-5 in the Appendix.

We now provide two benchmarks for how much variance there is in the taxonomy. In the first we benchmark how much change the taxonomy can introduce with the disease-symptom associations. These can be extracted from PubMed literature as done in [124]. We determined how much a change in the taxonomy would change the results. We also determined how much the results would change if PubMed papers were only used up to a certain date. We found that the changes in similarities were approximately same. Details of the methods and plots can be found in Section B.10 in the Appendix. This implies that the changing taxonomy has as much of an impact

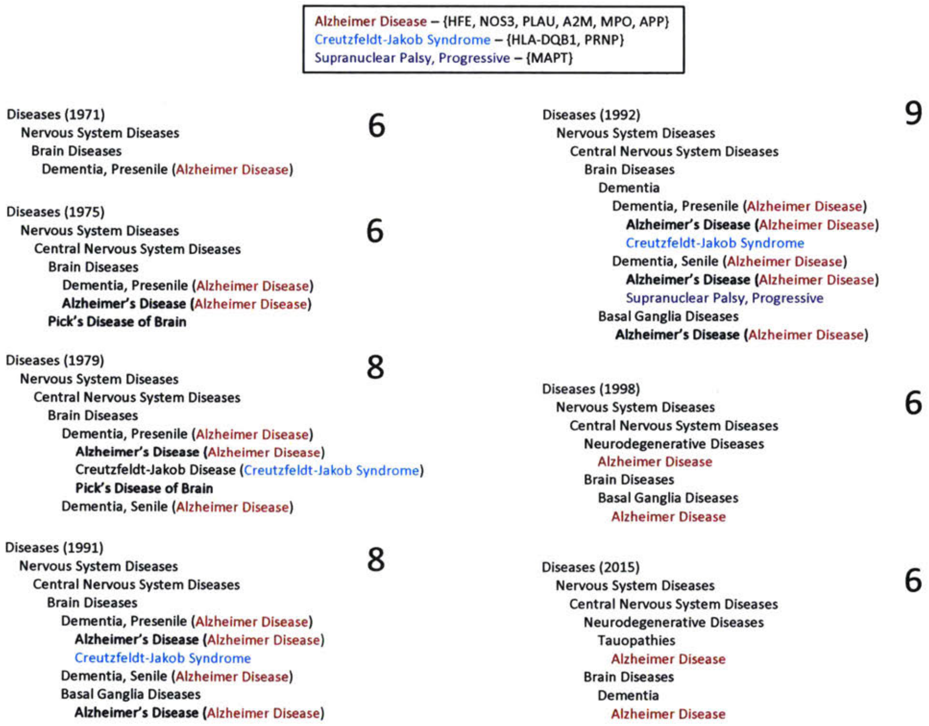

Alzheimer Disease – {HFE, NOS3, PLAU, A2M, MPO, APP}
Creutzfeldt-Jakob Syndrome – {HLA-DQB1, PRNP}
Supranuclear Palsy, Progressive – {MAPT}
Diseases (1971)
Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
6
Diseases (1975)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Pick's Disease of Brain
6
Diseases (1979)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Creutzfeldt-Jakob Disease (Creutzfeldt-Jakob Syndrome)
Pick's Disease of Brain
Dementia, Senile (Alzheimer Disease)
8
Diseases (1991)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Creutzfeldt-Jakob Syndrome
Dementia, Senile (Alzheimer Disease)
Basal Ganglia Diseases
Alzheimer's Disease (Alzheimer Disease)
8
Diseases (1992)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Creutzfeldt-Jakob Syndrome
Dementia, Senile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Supranuclear Palsy, Progressive
Basal Ganglia Diseases
Alzheimer's Disease (Alzheimer Disease)
9
Diseases (1998)
Nervous System Diseases
Central Nervous System Diseases
Neurodegenerative Diseases
Alzheimer Disease
Brain Diseases
Basal Ganglia Diseases
Alzheimer Disease
6
Diseases (2015)
Nervous System Diseases
Central Nervous System Diseases
Neurodegenerative Diseases
Tauopathies
Alzheimer Disease
Brain Diseases
Dementia
Alzheimer Disease
6

Figure 3-10: Example of disease where the number of genes associated with it changed because of changes to the taxonomy structure even though the actual disease-gene associations are constant as seen at the top. The number of genes associated with Alzheimer Disease would vary between 6-9 genes depending on the taxonomy.

as the changing knowledge found in new PubMed papers.

Another way to benchmark the taxonomy change is to compare the similarity of a disease at a time point 1 to the same disease at a future time point 2 compared to similarity with its closest neighbors at time point 1. To do this, we search through all diseases for examples that satisfy:

$$S_{jaccard}(d_{i;1971}, d_{i;2015}) < S_{jaccard}(d_{i;1971}, d_{j;1971}) \tag{3.1}$$

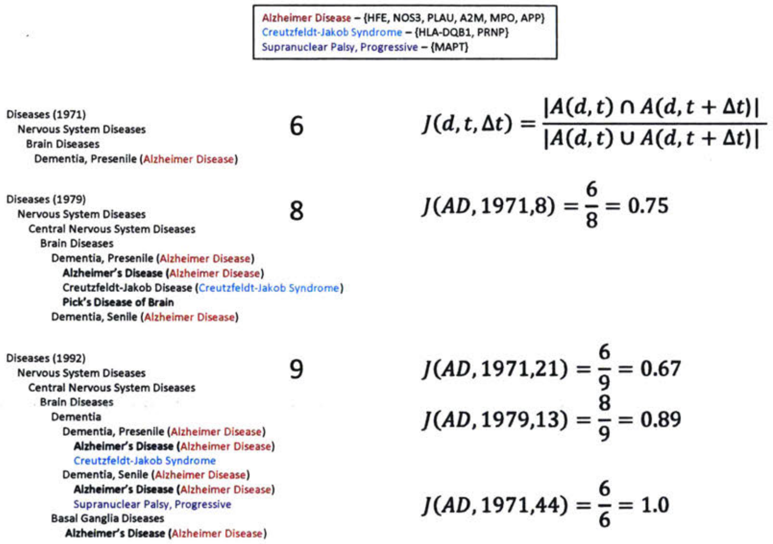

Alzheimer Disease – {HFE, NOS3, PLAU, A2M, MPO, APP}
Creutzfeldt-Jakob Syndrome – {HLA-DQB1, PRNP}
Supranuclear Palsy, Progressive – {MAPT}
Diseases (1971)
Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
6
$J(d,t,\Delta t) = \frac{|A(d,t) \cap A(d,t+\Delta t)|}{|A(d,t) \cup A(d,t+\Delta t)|}$
Diseases (1979)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Creutzfeldt-Jakob Disease (Creutzfeldt-Jakob Syndrome)
Pick's Disease of Brain
Dementia, Senile (Alzheimer Disease)
8
$J(AD,1971,8) = \frac{6}{8} = 0.75$
Diseases (1992)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Dementia
Dementia, Presenile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Creutzfeldt-Jakob Syndrome
Dementia, Senile (Alzheimer Disease)
Alzheimer's Disease (Alzheimer Disease)
Supranuclear Palsy, Progressive
Basal Ganglia Diseases
Alzheimer's Disease (Alzheimer Disease)
9
$J(AD,1971,21) = \frac{6}{9} = 0.67$
$J(AD,1979,13) = \frac{8}{9} = 0.89$
$J(AD,1971,44) = \frac{6}{6} = 1.0$

Figure 3-11: Example calculation of the Jaccard similarity for Alzheimer Disease using the taxonomy at three different years.

Restated, disease i in 1971 looks less like disease i at a future year than it does to some other disease j in 1971. We use the PubMed data set for this example. The number of diseases where such cases is shown in Table 3.4. Approximately 5-10% of diseases had at least one other disease more similar in 1971 than the disease in question at a later date. We provide three example in Figure B-7 in the Appendix. Looking forward, this suggests that 5-10% of current diseases may not be distinguishable from other diseases in the taxonomy.

start year	1971
# diseases with at least one diseases with more similarity	140
# diseases eligible	2049
% of diseases	6.8%
mean # years	15.2
std dev time	10.6

Table 3.4: Diseases that were less similar to themselves in time than they were to other diseases in the same year.

3.4.4 Estimating Variance Among Current Taxonomies

Thus far we have compared one taxonomy over time to examine variations. We also compare the MeSH taxonomy to two other widely used taxonomies, SnoMed CT and ICD9. We would expect that if two diseases are found in two different taxonomies, that the relationship between the two diseases would also be preserved. To test this assumption we identify all terms possible in all taxonomies with a unique UMLS concept id. We then look at all taxonomy pairs and find pairs of diseases that occur in both taxonomies. Finally, we determine what percent of those disease pairs have a relationship in one, but not the other compared to how many are in both as seen in Table 3.5

We found that the smallest overlap was 17% and the largest overlap was 63%. The smallest percent overlaps into ICD9. This is likely because ICD9 is a strict taxonomy with only one location per disease, so it has significantly fewer relations in general – approximately half the number in MeSH and approximately 5% of SnoMed. The relatively small degree of agreement is surprising since different taxonomies are used for learning problems in similar ways.

	MeSH	ICD9	Snomed CT
Num diseases	4662	17523	69158
Num disease pairs	31752	68634	1199677
Num concepts	8399	16499	68532
Num concept relations (CR)	97200	47368	1171048
Num CR w/both in MeSH	–	1044	26851
Num CR w/both in ICD9	3838	–	22539
Num CR w/both in Snomed CT	30797	9085	–
Overlap in MeSH/ICD9	660	660	–
% CR in other taxonomy	17.2%	63.2%	–
Overlap in MeSH/Snomed CT	14778	–	14778
% CR in other taxonomy	48.0%	–	55.0%
Overlap in ICD/Snomed CT	–	5272	5272
% CR in other taxonomy	–	58.0%	23.4%

Table 3.5: Overlap of disease relations between taxonomies. 'Concepts' refer to the UMLS concepts, all of which are diseases in our case. There may be zero, one or multiple UMLS concepts for any given disease in one of the three taxonomies. Concept relations (CR) are the pairs of UMLS concepts that have an ancestor-descendant relationship in the taxonomy specified by the column.

3.4.5 Estimating Variance within a Disease

Up to this point we have estimated variance within the taxonomy and among taxonomies as it pertains to learning. Another type of variance is how data points map to diseases. To make a proper estimate we would require diagnostic criteria for all diseases at each point in time. A change in diagnostic criteria implies a change in population associated with the disease. We are not aware of any historical diagnostic criteria data that would allow for a reasonable estimate.

Another way to estimate this variance is to find diseases where the environmental or genetic factors are unlikely to have changed for the course of several years. We could then assume the change in incidence of the disease would be due to changes in diagnostic criteria or patterns. Cancers fit this description. We plot the incidence of disease for cancer which we assume should not change over short time periods due

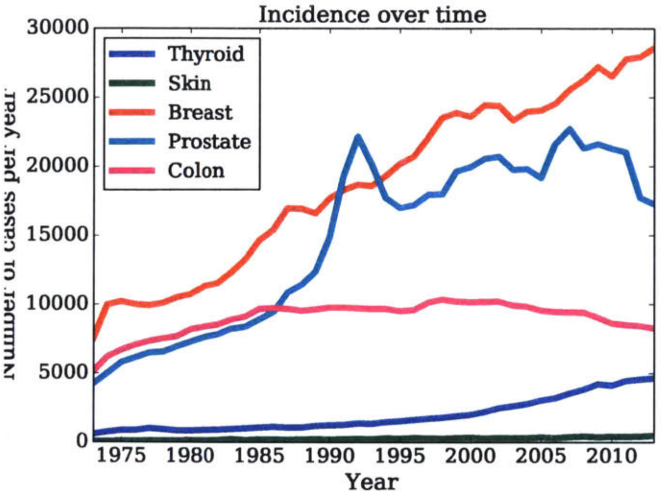


Figure 3-12: Incidence of several types of cancer over time. Some types of cancer have highly a variable incidence even though there is no evidence that causes would have changed.

to causal factors, but more likely due to changes in disease definition or diagnostic criteria. Figure 3-12 shows several examples of cancers that change significantly or minimally over only a few years. Though some are relatively constant, some change dramatically over only a few years. The most dramatic case is Prostate Cancer which at one point started using prostate-specific antigen (PSA) as a diagnostic.[165] For the purposes of learning, this changed dramatically the population being studied.

From all of the cancer data we plot in Figure 3-13 an estimate of how incidence might change over a given time period for a randomly chosen disease.

This estimate appears to work well for cancer, but we might apply the estimate to a other diseases where causal factors would not change over a time period of several years. Diseases that would not be included would be infectious diseases, diet-based diseases or diseases that have a strong environmental factor like lung diseases from asbestos.

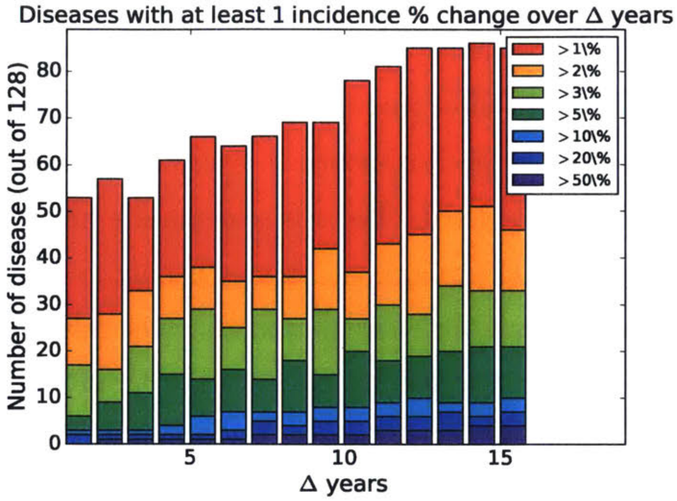


Figure 3-13: Changes of incidence over Δ years for all cancers in the SEER data set.

3.5 Discussion

To the best of our knowledge, this work provides the first look at how the taxonomy of diseases has evolved over the last 50 years. Others have studied how diseases or disease categories have changed over time, but no work has looked at the comprehensive set of diseases in the taxonomy. We observed that the taxonomy grows rapidly in the leaves, while the branches change relatively little. The highest level categories of the tree tend not to change. At the same time, diseases have shown up in multiple locations more and more over time. One explanation is that the disease categories in the taxonomy that once neatly classified diseases no longer do so as our understanding of diseases has been updated. For example, cancers have recently been mixed with immunological, which mirrors much of recent cancer research focusing on immunology.[166, 145] It may be that diseases need more descriptors or tree node locations for accurately fit into the taxonomy given the legacy disease categories.

It is also clear from the history of the taxonomy that the greatest amount of change is from adding and deleting tree node locations of existing diseases. Diseases also subdivide to add newer, more specific diseases, but that is not the primary type of learning incorporated into the taxonomy. It is also common that diseases change name, but rarely are they deleted. There is sometimes a perception that diseases are inherently correct and the process of learning is only to refine the current diseases into sub-diseases. Our work shows that much of the new learning is redefining and reclassifying diseases that may not have been accurately described given the taxonomy structure, resulting in tree node locations added or deleted.

The past changes in the taxonomy may be the best predictor of the future taxonomy, which indicates an increasingly complex taxonomy. As each disease can be placed in more and more locations, the amount of information in the tree structure goes down and the less useful the taxonomy is. Increasingly we see that diseases previously not thought to be related share commonalities and systemic views of disease show interconnections. One way to address this problem is to change the taxonomy categories. In computational learning there are a few ways to control for complexity in classification trees. One approach is to prune branches when the information content decreases. This ensures that the upper levels of the tree lead to the most information in the tree below. An intuitive example of this can be explained using the taxonomy of life. If a new organism was found that was classified as a reptile, mammal and marsupial - scientists might question the validity of those categorizations.

One way to characterize the process by the taxonomy is updated is as an iterative process. A disease is described by the other diseases in the taxonomy, but once more diseases are added to the taxonomy, each disease will have relationships that have change and consequently diseases may need to change position. This is similar to the cycle of learning diseases from populations. A population is defined by some features and is classified to have a specific disease. Then patients diagnosed with the diseases are studied to find associated features, which may change the population to study and the disease definition. This explanation would explain why some diseases appear to diverge or converge as would be expected with an iterative process. For example, Dropsy is a disease that disease that was widely diagnosed in the 1800s and the last patient diagnosed was in the mid 20th century.[167] On the other hand, Cystic Fibrosis was a disease observed based on observation of fibrosis and cysts in the lungs and turns out that chloride ion channels were not working properly.[168] Current diseases that may diverge may be Diabetes or Alzheimer Disease. As we showed, it appears that diseases that have been around for a longer time have survived the process and are more likely to stay.

Some of the variation in the process may be inherent to the task of placing diseases in a discrete space when the true health conditions live in a continuous space. We demonstrated one way to estimate that variance by using incidence data to show that a continuous health condition can be approximated by a discrete disease differently over time. The taxonomy also provides a discrete set of diseases that is used to compare diseases to each other. If one part of the taxonomy changes, other parts of the taxonomy may need to change to be correct.

To measure how much variance there is in the taxonomy, we developed a way to quantify variance over time that may be used to put an error bar on the taxonomy. Our estimates may be the best predictor of how different the taxonomy changes will affect learning in the coming years. This is important since many retrospective learning tasks rely on disease relations in the disease taxonomy. Our method allows researchers to demonstrate that their results are greater than the noise in the taxonomy. Furthermore, it will allow researchers to demonstrate how consistent the taxonomy is for use with the data set they are exploring. For example, we showed that disease-gene relations changed significantly over time and inferred relations with a gene data set may be too noisy for analysis.

There are cases where this variance could have significant implications. For example, in drug approval a disease must be named as the primary indication. The goal is to show that there is a statistically significant result demonstrating the treatment modulates the targeted condition. If the disease has variance that is significant, it is like looking for a signal in noise. We are not aware of any previously proposed methods to systematically correct for the error bar around diseases for drug approval.

One concern from this work is that the definition of disease and current use of taxonomy may lead the scientific community farther away from precision medicine. Certainly, the disease categories appear less precise over time. There are many diseases, particularly genetic diseases, syndromes and cancers that appear not to be classified precisely and perhaps another classification scheme would allow for more treatments. We need to be able to learn about individuals and precise conditions without the need to rely on the discrete disease bins defined in the taxonomy. Any learning requires aggregating and the taxonomy provides a framework to do that, but there may be better ways to create a manifold of disease relationships for doing that. Future work will examine ways to compare diseases for more accurate learning and possibly better ways to update the disease taxonomy to align with data and contain more information. Future work should also focus on incorporating different data sets and combining them for a more comprehensive measure of variance.

One limitation of this work is that we only examined the history of one disease taxonomy. This was due to the fact that we could only obtain the regular history of the MeSH taxonomy. However, we were able to compare MeSH to ICD9 and SnoMed CT in the present finding, concluding that the variance among taxonomies appears to be high.

We conclude from this work that taxonomies have been and continue to be useful for many aspects of medicine, particularly as a tool for communication. Taxonomies may not be the best tool for learning from data, especially computational learning. Taxonomies allow information to be compressed in an efficient way to partition a space. This is a convenient representation for humans, but may result in significant information loss that would not be required for a representation designed for computing. Computational techniques can handle high-dimensional data that is uncompressed. One area of future research is how to codify learning about diseases in a way that can be used for officially comparing diseases without having to rely on a taxonomy.

Chapter 4

Updating the Disease Space from Data

In the previous chapter, we showed how to measure uncertainty in the taxonomy by estimating variance. We made such estimates using the history of the disease taxonomy and external data sets. In this chapter, we look at how to reduce variance when comparing external data sets to the disease taxonomy. Using the concept of reducing uncertainty allows us to setup several learning problems.

4.1 Introduction

In order to alleviate, cure and prevent diseases, there is a vast, ongoing research effort to learn more about diseases.[9] A recently growing approach to learning is retrospective data analysis, which uses novel approaches (new data aggregation or analysis) to use previously collected data to glean new information.[2] This approach has gained greater acceptance and hope in recent years due to the vast amount of untapped data [32, 169, 29, 170] and the recent progress in machine learning and artificial intelligence algorithms.[171]

In order to aggregate data for statistical learning on diseases, a structured vocabulary is used to define what is considered a disease.[62, 172] Diseases are then associated with attributes of interest, such as genes, symptoms or drugs. This data can be used to find further disease-attribute associations or discover new disease-disease associations. It is often the case that structured vocabularies of diseases are organized into a taxonomy,[62] which represents the output of experts to help determine how diseases are related to each other.[173]

It is not unusual for a taxonomy structure to aid in retrospective statistical learning.[174] What has not been studied in depth though is how well a taxonomy reflects the attributes of given diseases. Furthermore, there has been little work showing how disease attributes can be used to update disease taxonomies that better align with the data, nor how the expert knowledge captured in taxonomies could be used to expand disease-attribute data sets. In this work, we show several ways to analyze how well the taxonomy and the disease-attribute data are aligned, as well as how one data set could be used to update another. This work creates a way to map between the disease taxonomy space and the disease attribute space, enabling us to reduce the variance between the two. These tools will be valuable for computational learning, where data is sparse and a manifold in disease space is needed.

4.2 Background

Efforts to learn associations between diseases and disease attributes include attributes such as genetic causes,[175, 125, 176] meta-genomics, environmental factors, symptoms,[124] organs or tissues affected, comorbid diseases,[177, 152] cellular pathways,[178] treatments,[144] outcomes and many more. Each attribute has a fixed set of features that spans the attribute. This spanning creates a mathematical space in which diseases can be compared. The ability to relate diseases to each other has several significant uses. For example, finding a differential diagnosis[8, 12] in medical practice or drawing inferences about similar diseases in retrospective learning.[77] Developing methods for comparing similarity or distance between diseases in a given feature space is an ongoing area of research. Often these efforts focus on one disease attribute in isolation. More recent work has attempted to combine different attributes.[90, 143, 90]

Another way to define a disease space is through a curated taxonomy as discussed in the previous chapter. To reiterate, we briefly describe what a taxonomy is and how it is used in medicine.

A curated taxonomy is constructed by biomedical experts and uses a nested structure to relate diseases to each other.[7, 65] Typically there is no explicit reasoning provided for why diseases are grouped in a particular way, but it is assumed that all diseases underneath a given disease category share some features in common. A taxonomy is generally accepted as a formal organization of diseases, though each taxonomy is constructed and/or curated for a different purpose. For example, the Medical Subject Headings (MeSH) is used to index medical literature.[67] The International Classification of Disease (ICD) was originally created for performing statistics on causes of death and more recently is widely used for medical billing in the United States.[179] The Systematic Nomenclature of Clinical Terms (SnoMed CT) is used to standardize medical notes.[68]

Each taxonomy is also a structured vocabulary, which is a discrete set of terms used to codify concepts that may inherently be continuous. For example, cancers are often broken down into stages and any patient can be matched to a term in the vocabulary even though patients may present across a continuum. Standards are created such as stage I-IV to create discrete disease concepts.[180] When data sets matching diseases to attributes are created they rely on a structured vocabulary such as MeSH, ICD or SnoMed CT.

Recently, there have been calls to restructure the disease taxonomy as a whole to better reflect up-to-date understanding of diseases.[9, 58] In the previous chapter, our work showed that taxonomies change substantially over time and inferences derived by using a taxonomy to augment learning problems can have a significant impact in only a few years. There has also been a push for more continuous updating of the taxonomy to avoid time delays between the generation of knowledge and the formal codification of knowledge.[23] A question these points raise is if a taxonomy can be automatically updated given new data? The reverse is also a valid question, can expert knowledge in taxonomies be used to generate hypotheses to update disease-attribute data sets?

4.3 Data and Methods

4.3.1 Data

The data used includes two disease taxonomies: MeSH and ICD9, and three sets of attributes associated with diseases: genes, symptoms and drugs.

To represent each disease in the taxonomy as a vector, each disease i is represented as a binary vector of all other diseases where a 1 in position j indicates that disease j is an ancestor of disease i. The disease attribute data includes disease-gene, disease-symptom and disease-drug associations. We refer to genes, symptoms and drugs as attributes. Each attribute includes a set of features. For example, features of the 'symptoms' attributes include 'pain,' 'vomiting,' 'amnesia' and so on. Each disease is represented as a binary vector for each association between a given disease i and each feature k of a given attribute.

The genetic data came from the Online Mendelian Inheritance in Man (OMIM) database,[121] the symptoms data was from work extracting relationships between diseases and symptoms in the PubMed database,[124] and the Medical Indication Resource - High Precision (MEDI-HPS) data set.[126] The symptoms data set was constructed using the MeSH vocabulary, so all diseases were included as MeSH diseases. To convert phenotypes in the OMIM database to diseases in MeSH we used the UniProt Knolwedgebase's "Controlled vocabulary of human diseases,"[181] which includes identifiers used by different databases including OMIM and MeSH. The drugs data set was constructed using the ICD-9 taxonomy, so we used the Unified Medical Language System (UMLS) to map ICD-9 concepts to unique UMLS concepts to MeSH diseases.

We created a binary vector for each disease of length N, where N was the cardinality of the set of unique features in each data set. For example, the symptoms data set included 322 different unique symptoms. For each element in the vector, a zero indicates no association in the database and a one indicates there is one.

We then created a set of vector to represent the disease taxonomy. We created a binary vector for each disease of length N, where N was the total number of diseases in the taxonomy. For each element in the vector, a zero indicates the disease representing the vector was not a descendent of the disease represented by the element and a one indicates it was a descendant. A visual depiction of the two types of disease vectors can be seen in Figure 4-1.

4.3.2 Similarity Metrics

To define a space, there must be a way to measure distance, similarity or dissimilarity between data points. The typical way a structured vocabulary is used is that all data points with the same label have a distance of 0 and all else have a distance of 0. In a taxonomy it may be assumed that all descendants of a given node share something in common and therefore have a distance of 0, while all else have a distance of 1 according to the function:

$$d(n_i, n_j) = \begin{cases} 0, & \text{if } n_j < n_i \text{ or } n_i < n_j \\ 1, & \text{otherwise} \end{cases} \tag{4.1}$$

where $n_j < n_i$ indicates that node j is a descendant of node i. This binary metric does not allow for very much granularity. Other distances have been proposed including least-common subsumer [173, 182]:

$$\begin{aligned} \max_k \quad & depth(n_k) \\ \text{s.t.} \quad & n_i < n_k, \\ & n_j < n_k. \end{aligned} \tag{4.2}$$

where depth is the number of edges from the root node. Wu and Palmer[183] derive a metric that uses the same principle as equation 4.1 that all descendant nodes are considered equal, but seeks to find the most specific node where that is true. The metric suffers from the same nonlinearity as equation 4.1. Another basic metric that takes into account distance along links is shortest path [184]:

$$\begin{aligned} \min_k \quad & depth(n_i) + depth(n_j) - 2 * depth(n_k) \\ \text{s.t.} \quad & n_i < n_k, \\ & n_j < n_k. \end{aligned} \tag{4.3}$$

which is commonly used in network structures. The metric counts the number of edges between nodes, but does not take into account the depth of the least common subsumer. There are many more metrics proposed [185, 186, 173], but they are generally based on these two basic metrics.

There are a number of ways to measure distance between two binary vectors for the attributes. We choose to use an information-based similarity metric [187]:

$$s(w_1, w_2) = \frac{2 * I(F(w_1) \bigcap F(w_2))}{I(F(w_1)) + I(F(w_2))} \tag{4.4}$$

where $F(w)$ is the set of features associated with w.

The metric is essentially the Jaccard similarity score, but weighted according information content of each vector element. Some features are more common and are therefore less informative in distinguishing diseases, so those are weighted less. For our data sets, there are cases where generic features like 'pain' are less informative in comparing diseases.

4.3.3 Information Measures

Mutual information (MI) is a measure of how much information is known about one random variable given another random variable.[188] Mutual information is used for many tasks such as feature selection[189] or clustering[190] for applications from Natural Language Processing[191] to Image Analysis.[192] MI for discrete variables is calculated:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)ln\left[\frac{p(x,y)}{p(x)p(y)}\right] \tag{4.5}$$

To calculate mutual information between a taxonomy and a disease attribute data set, the joint probability $p(d, f)$ that a randomly chosen disease has disease d as an ancestor and is associated with feature f. MI is the summation of all point-wise mutual information (PMI) over all elements in each distribution. It can be used to assess the shared information between elements from the two distributions.

4.3.4 Predictions

The premise of this chapter is that there may be information in data sets that are not in a formal disease space like the disease taxonomy and vice versa. We would like to learn to update one data space using the other. The learning problem is the following: given two set of data vectors, X and Y respectively, create a mapping from X to Y in order to predict missing values. Specifically, this could be mapping from a taxonomy T_a to a disease attribute A_z or vice versa, mapping A_z to T_a. For example, a mapping could be made from disease-symptom associations to relationships in the MeSH taxonomy. If the taxonomy has some basis in symptoms of diseases and if some of those associations have not been considered by the experts updating the taxonomy, then we will learn new taxonomy updates based on data. The problem can be expanded to include sets of data such as mapping T_a, T_b to A_x, A_y, A_z. Figure 4-1 shows the vector representation of the prediction problem.

We first give details on the implementation of the approach described above. We also discovered that the prediction problem could be reversed and we could predict symptoms from the taxonomy structure. This could be useful for imputing values in

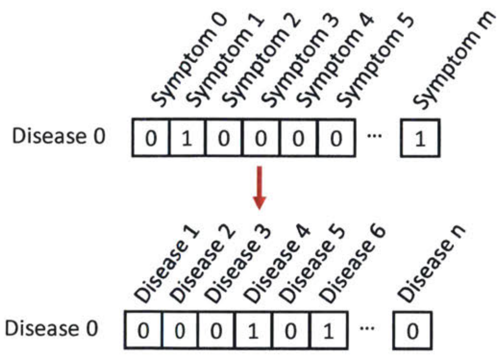


Figure 4-1: Vector representations of disease-attribute data sets and the disease taxonomy. We start by finding a mapping from symptoms to the taxonomy with the goal of predicting where a new disease would be located in the taxonomy.

the disease-symptoms database.

Taxonomy Predictions

To provide a baseline prediction, we used a maximum likelihood approach by summing the number of descendant of each disease. Some categories have more diseases, making it more likely that a randomly selected disease would occur underneath one of those diseases.

The most straight-forward prediction we performed was logistic regression without any penalty term.

We also tested a highly-nonlinear and potentially complex model using a deep neural network (DNN).[193] We tried three different DNN architectures. The first was a multilayer perceptron with the symptom for the input vector and the taxonomy as the output vector. The input layer had 322 symptoms and the output layer had 4662 diseases. There were two hidden layers, the first with 322 nodes and the second with 3000 nodes. All layers were connected in a dense manner with the first three

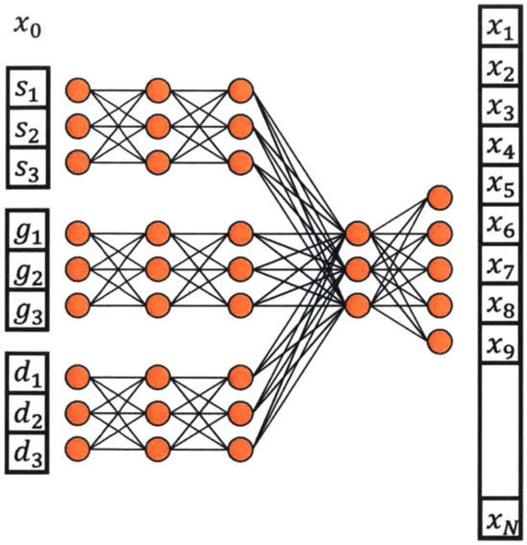


Figure 4-2: Diagram of deep neural network combining multiple disease attributes in a multi-view structure to learn mapping to taxonomy.

layers using a rectified linear activation function and the final layer using a sigmoid activation for binary classification. The loss function used to train the network was the binary cross entropy function and a dropout rate of 0.25 is used for each hidden layer to prevent overfitting.[194]

We also used a multi-view architecture to incorporate different data sets. To do this we created two more input vectors, one for genes and one for drugs. There were two hidden layers as described above, but each independent of the others. We then combine the three hidden layers together into one combined hidden layer. We tried several merge functions including element-wise addition, multiplication and maximum. The final hidden layer was then connected to the output representing the taxonomy as described above. The same activation functions and dropout rates were also used. We implement the DNN using the keras library for python with TensorFlow as a backend enginer.chollet2015keras,tensorflow2015-whitepaper A visualization of this architecture can be seen in Figure 4-2.

We performed two slightly different prediction tasks with the methods described. One was predict where a new node would go for a disease that already exists in the taxonomy and the second is to predict node locations for a new disease. In the first case there already exist some locations and we are trying to find another one. In the second case we have a completely new disease that does not yet occur in the taxonomy. To create a test set in the first we randomly remove 10% of leaf nodes from the tree and in the second case we remove 10% of the diseases that have at least one leaf node. To evaluate performance of this prediction we plot a precision-recall curve and extract the f-max score. We do this because the problem can be viewed as information retrieval trying to identify a few diseases out of 4662 and precision-recall is typically used for information retrieval.[195] We run the prediction/evaluation 100 times with different training and test sets.

Attribute Prediction

To provide a baseline prediction, we used a maximum likelihood approach by summing all symptoms vectors. Some symptoms occur more frequently than others, making it more likely that a randomly selected disease would have those symptoms.

As done in the previous section, the most-straightforward prediction was a logistic regression without any penalty term. We also created a multilayer perceptron as done in the previous section. The first layer included the diseases in the taxonomy with length 4662, the next hidden layer had 4662 nodes followed by a hidden layer with 1000 nodes and the final output layer of symptoms with 322 nodes. The activation functions were rectified linear functions with 25% dropout for the first three layers and the final layer had a sigmoid activation function.

In this case we only performed one prediction problem by removing 10% of the symptoms randomly for the test set. This prediction problem can be viewed as a multi-class classification problem and we evaluate performance using the area under the receiver operation curve (AUC). We similarly run the prediction 100 times with different training and test sets.

Combined Prediction

A DNN architecture that combines the two prediction problems is an autoencoder. Autoencoders are a popular deep learning structure that has the same input and output layers with a hidden layer(s) that is smaller.[171] Corrupt data is used at the input to train the network to ensure that a smooth relationship is learned and noise is averaged out. We build such an autoencoder structure, but make the hidden smallest layer also an output being the disease attributes. We train the network by leaving out 10% of the disease node locations and corrupting the input data. We then predict the output taxonomy using the uncorrupted input and predict where new node locations will be and validate on the 10% holdout. Figure 4-3 shows the structure of the architecture. Each hidden layer has 1000 nodes. Each hidden layer has a rectified linear activation function and the two output layers have sigmoid activation functions.

4.4 Results

4.4.1 Similarity in Different Disease Spaces

To visualize the similarity between diseases and symptoms, we plotted the taxonomy as a treemap and outlined specific diseases associated with a given symptom. We give two examples, Transient Global Amnesia in Figure 4-4 and Pain in Figure 4-

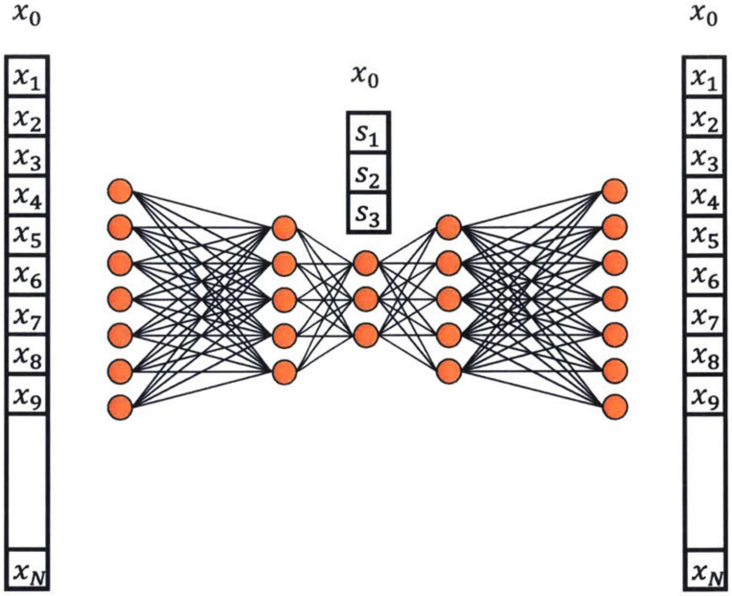


Figure 4-3: Deep neural network structure encoding the taxonomy as a set of symptoms and decoding back to the taxonomy structure. When the input data is corrupted and the network trained, a denoised mapping is learned between each disease space.

5. Transient Global Amnesia shows specific locations which correspond to specific etiologies such as neurological disease, stroke or head injuries. This shows that a symptom could be used to help divide up the taxonomy. Pain, on the other hand, shows up in many different locations, with the exception of Animal Disease. It appears that different symptoms may have different weight for distinguishing diseases in the taxonomy.

We found visual evidence there is similarity between symptoms and the taxonomy for an example; now we look more comprehensively and quantitatively. We compare similarity in different diseases spaces to test if there is shared information between the different data sets. We plotted two tree distance measures, minimum path length (MPL) and least common subsumer (LCS), compared to similarity based on information in different data sets of diseases and attributes according to Eq. 4.4 seen in Figure 4-6. We observe that there is a general relation between the tree distances and

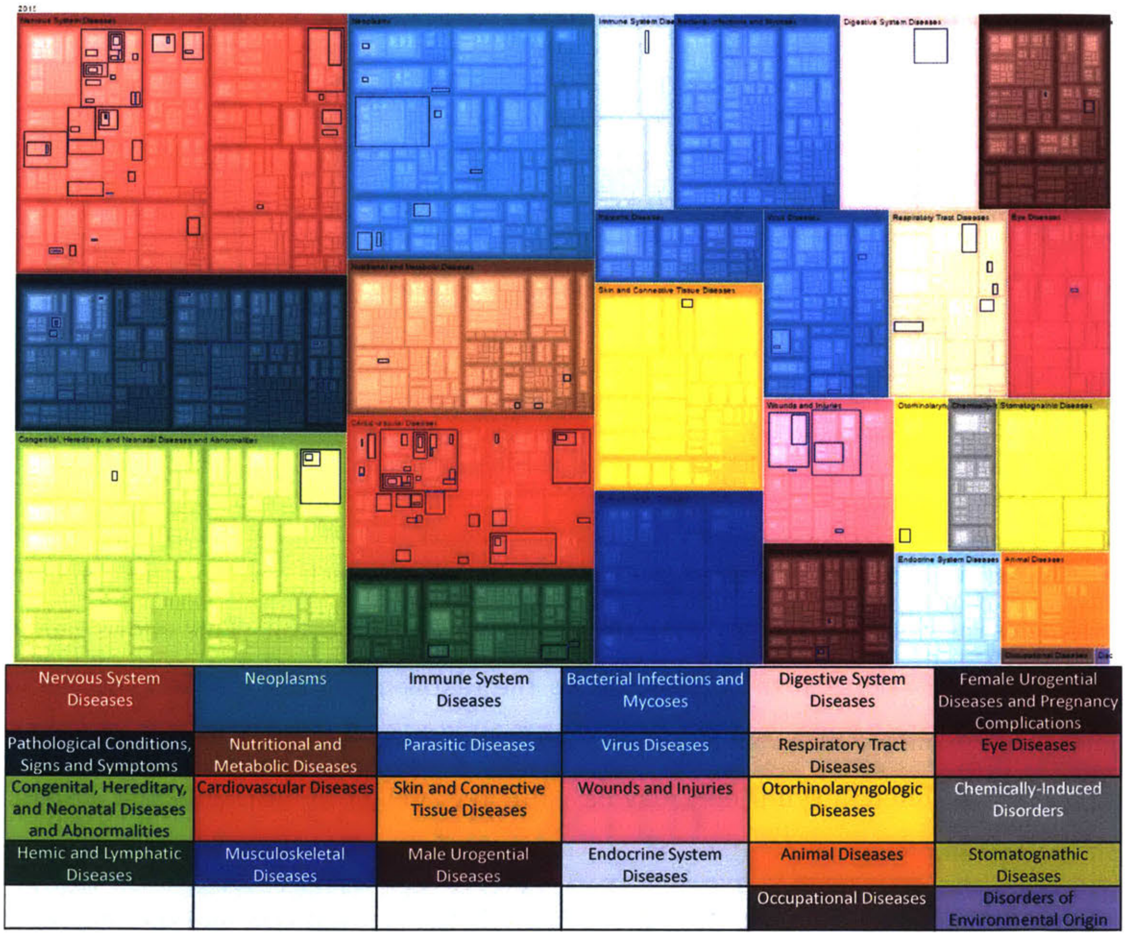


Figure 4-4: Diseases associated with the symptom 'Transient Global Amnesia.' The three main categories of disease they show up in are 'Nervous System Diseases,' 'Cardiovascular Diseases,' and 'Wounds and Injuries.'

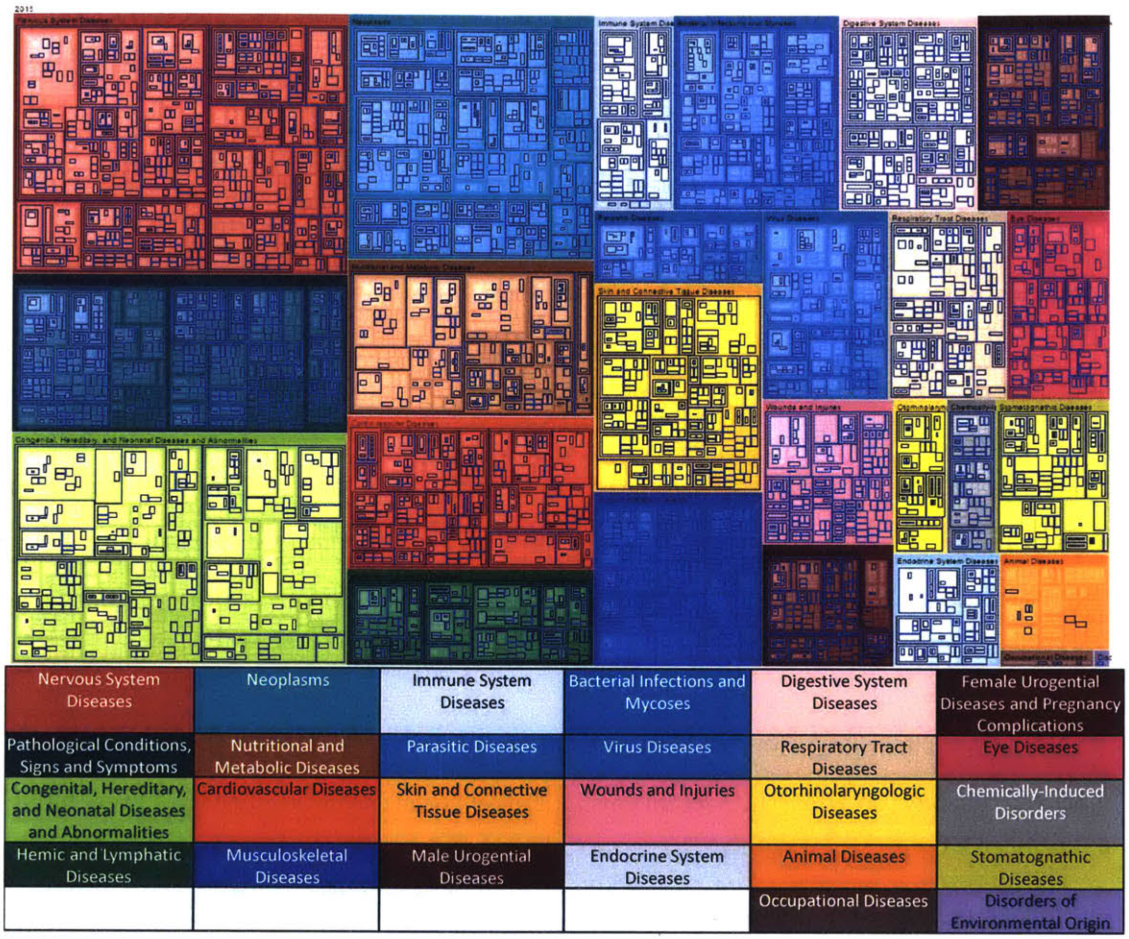


Figure 4-5: Diseases associated with the symptom 'Pain.' This symptom is prevalent in all categories except 'Animal Diseases' where pain may not be as much of a priority.

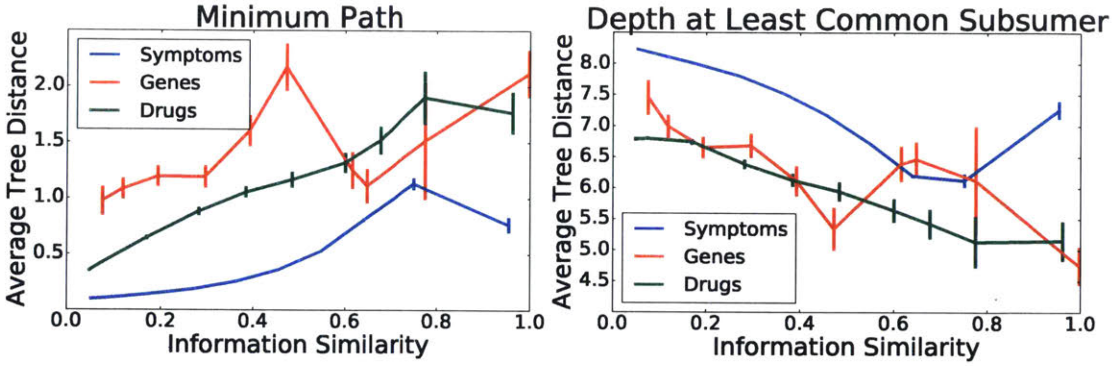


Figure 4-6: Comparing tree-based distance measures to similarity measures based on disease attributes. The similarity metric used is information similarity from Eq. 4.4. Error bars are standard error.

the attribute-based measures. None of the relationships are strictly linear or even monotonic. These two observations suggest that overall the tree measures capture the attribute information generally, but one cannot be used as a surrogate for the other.

We were surprised that drugs had the most linear relationship with the tree metrics. It may be that the taxonomy as currently structured captures treatments better because disease classification reflects mechanisms that drugs treat more than any other metric. The taxonomy is widely described as capturing symptoms [9], so we expected that attribute to most follow the tree measures, but the two do not appear to match when the similarity is close to 1.

In this chapter we primarily use symptoms as an attribute. We also show in Section C.0.1 in the Appendix another example of overlap in the taxonomy using drugs from clinical trials.

	Symptoms	Drugs	Genes
0%	0.277	2.11	2.139
1%	0.274 ± 0.0007	2.10 ± 0.0085	2.138 ± 0.0051
5%	0.263 ± 0.0015	2.090 ± 0.0143	2.132 ± 0.0105
10%	0.250 ± 0.0026	2.073 ± 0.0179	2.130 ± 0.0154
25%	0.216 ± 0.0026	2.025 ± 0.0275	2.116 ± 0.0247

Table 4.1: Mutual information for different attributes with variance measures. The left column is the percent of data that has been randomly shuffled. The mean and standard deviation of 100 trials is given in each cell.

4.4.2 Information Shared between Disease Spaces

Another way to measure how aligned two diseases spaces are is to measure how much information is contained in one space when another space is known. This can be measured with mutual information. We calculate the average MI over all diseases for a given attribute shown in Table 4.1. To measure variance, we randomly shuffle a percentage of the disease-attribute data and calculate the mutual information. We repeated 100 times each and report the mean and standard deviation.

We then look at specific cases of the point-wise mutual information (PMI) between specific diseases and specific symptoms. The largest PMI's are given in Table 4.2. We found that most of the diseases were high level categories in the taxonomy and associated with symptoms that were often associated with descendants of the disease, even though the symptom is not associated with the disease category itself in the data set. Such a relationship opens the opportunity to relate disease categories to disease attributes without simply percolating relationships up the tree.

PMI is used in natural language processing to find word pairs that occur exclusively together or infrequently together.[196] A negative value indicates concepts occur together less than random. Some examples of disease-symptom pairs with negative values include: Central Nervous System Diseases, Edema; Nervous System

Disease	Symptom
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Mental Retardation
Skin Diseases	Pruritus
Skin and Connective Tissue Diseases	Pruritus
Eye Diseases	Amblyopia
Eye Diseases	Vision, Low
Eye Diseases	Pseudophakia
Stomatognathic Diseases	Toothache
Genetic Diseases, Inborn	Mental Retardation
Eye Diseases	Vision Disorders
Digestive System Diseases	Abdominal Pain
Eye Diseases	Eye Hemorrhage
Nervous System Diseases	Perceptual Disorders
Eye Diseases	Blindness
Cardiovascular Diseases	Angina Pectoris
Cardiovascular Diseases	Heart Murmurs
Cardiovascular Diseases	Acute Coronary Syndrome
Eye Diseases	Pupil Disorders
Digestive System Diseases	Dyspepsia
Congenital Abnormalities	Mental Retardation
Eye Diseases	Scotoma

Table 4.2: Pairs of diseases and symptoms with the largest PMI

Most General	Num diseases	Most Specific	Num diseases
Seizures	1811	Aphasia, Primary Progressive	42
Neurologic Manifestations	114	Pseudophakia	99
Muscle Weakness	1562	Virilism	246
Paralysis	1515	Primary Progressive Nonfluent Aphasia	14
Fatigue	1348	Echolalia	48
Hemiplegia	1156	Feminization	98
Headache	1740	Urinary Incontinence, Urge	68
Quadriplegia	1073	Hypercalciuria	104
Dizziness	1102	Vomiting, Anticipatory	45

Table 4.3: The most general and most specific symptoms calculated using point-wise mutual information.

Diseases, Diarrhea; Neoplasms, Dystonia; and Brain Diseases, Chest Pain.

We can also look at average PMI for a given disease or symptom to see which ones are least/most informative on average. We list the most general and the most specific symptoms in Table 4.3. We note that pain is not in the top 10 even though it occurs with 2592 diseases, while "Nuerologic Manifestations" has only 114 and is in the top 10. Specificity doesn't necessarily correspond to number of diseases. In Figure 4-7 we show an example of one of the general symptoms, Neurologic Manifestations. In Figure 4-8 we show an example of one of the most specific symptoms, Virilism. It can be seen that the first is distributed throughout the tree, while the second is concentrated in several locations. We might suspect that symptoms associated with more diseases would be more general, but we selected these examples because Virilism actually has more diseases associated with it than Neurological Manifestations.

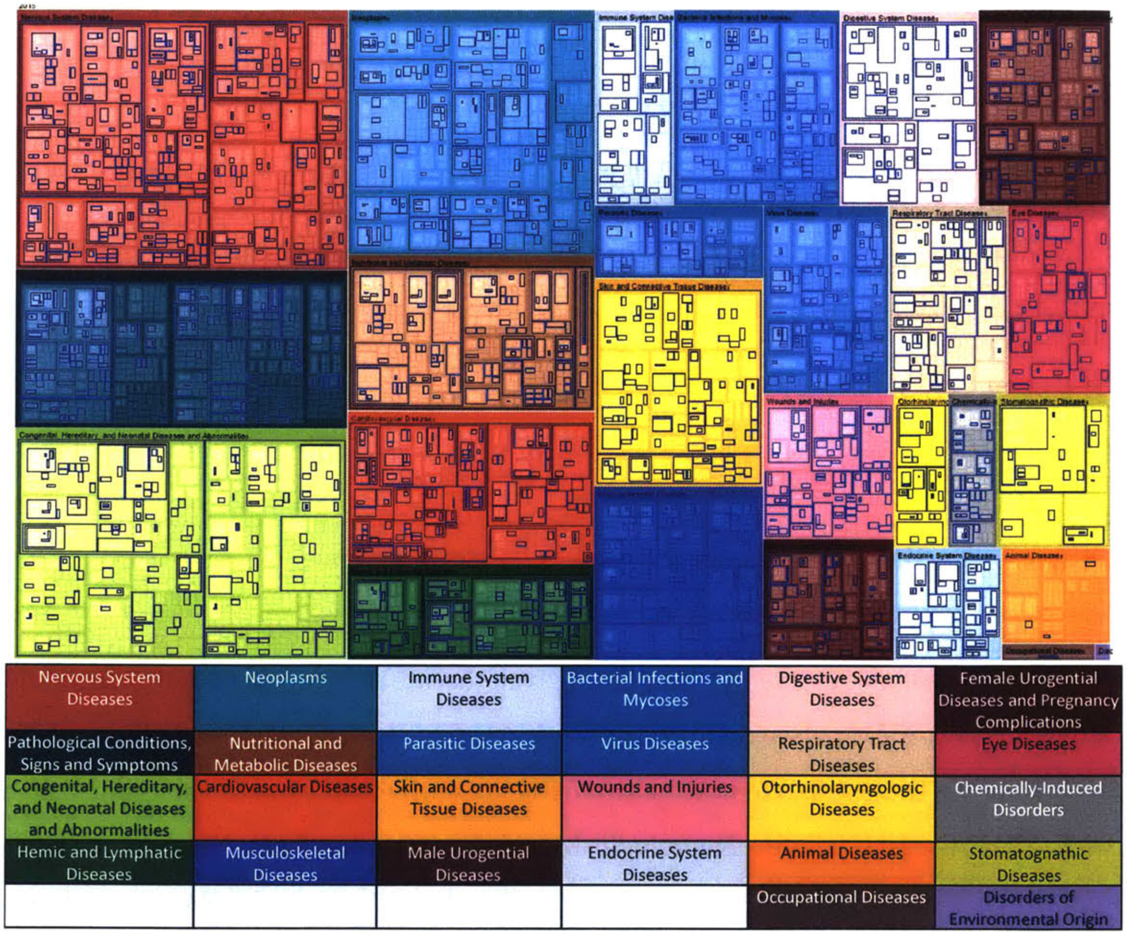


Figure 4-7: Diseases associated with Neurological Manifestations are outlined in blue. The associated diseases are distributed throughout the taxonomy, showing the symptom is very general.

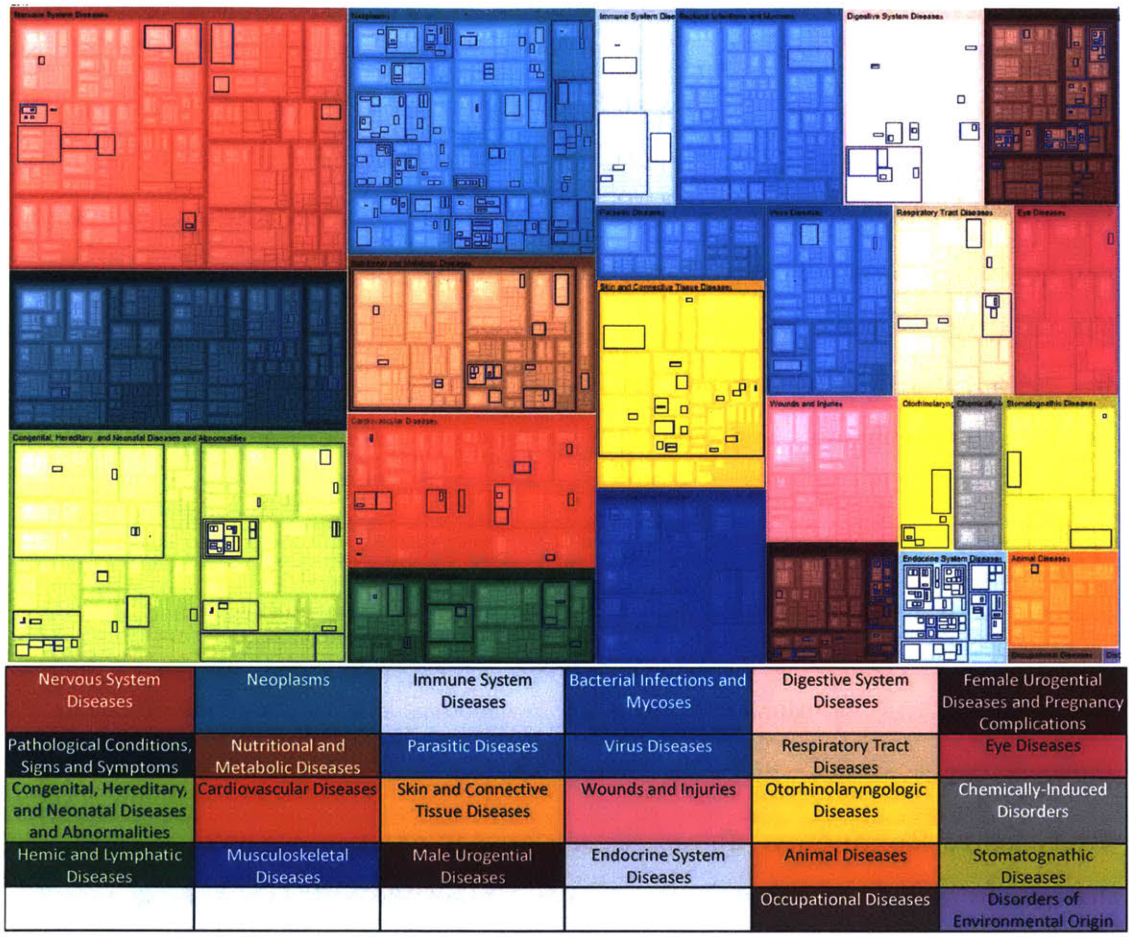


Figure 4-8: Diseases associated with Virilism are outlined in blue. The associated diseases are found in many specific locations in the taxonomy, showing the symptom is very specific.

4.4.3 Learning Between Disease Spaces

Model Selection

Taxonomy to Features In our first learning problem, we predict disease features from taxonomies. We remove 10% of the disease-feature associations in each data set, train the model, make predictions and validate on the 10% test set. We repeat 100 times and generate an ROC curve and compute the AUC for each disease. We report the overall average AUC for each technique in Table 4.4. We note that a mapping can be found that significantly improves over a naïve guess and that deep learning is the algorithm with the best performance. We also found that combining taxonomies did not significantly improve the results. This may be because the overlap of data between the taxonomies is sparse.

Method	AUC
Most Common	0.825
Logistic Regression	0.891
Deep Learning	0.900
Deep Learning w/Multiple Taxonomies	0.900

Table 4.4: AUC for mapping the taxonomy to symptoms and validating with prediction.

Features to Taxonomy In our second learning problem, we predict taxonomies updates from disease features. We remove 10% of the leaf disease node locations in the MeSH taxonomy, train the model, make predictions and validate on the 10% test set (not removing any disease completely), repeating 100 times for results. We perform prediction using most descendants, collaborative filtering, logistic regression and deep learning. This problem is much sparser than the problem above and can be cast as an information retrieval problem, so we use precision-recall curves to analyze the prediction results. We report the f-max score in Table 4.5. We show for MeSH

Method	F-Max
Most Descendants	0.181
Logistic Regression	0.364
Deep Learning	0.48
Deep Learning w/Multiple Features	0.48

Table 4.5: F-score for mapping symptoms to the taxonomy and validating with prediction on randomly removed nodes.

Method	F-Max
Most Descendants	0.240
Logistic Regression	0.486
Deep Learning	0.49
Deep Learning w/Multiple Features	0.49

Table 4.6: Mapping symptoms to the taxonomy and validating with prediction on randomly removed diseases.

with symptoms alone and with all three attributes. We found similar results as in the previous case, with the exception that deep learning was a more significant improvement over logistic regression.

We perform the same task of mapping attributes to a taxonomy, but validate by removing all nodes for randomly chosen diseases. We remove 10% of the diseases and make predictions. We do not perform collaborative filtering here because of the cold-start problem where there is no data yet for the predictions we are interested in -- there are no nodes for the diseases we are trying to predict. The results are reported in Table 4.6. We were surprised that the most descendants and logistic regression results improved over the previous results. It may be because each disease now has more node locations to predict allowing for a more robust prediction.

Combined problem A final way to combine the mappings between taxonomy and disease attributes is by incorporating an encoder. Autoencoders are a popular deep learning structure that has the same input and output layers with a hidden layer(s) that is smaller. Corrupt data is used at the input to train the network. We build such a structure, but make the hidden smallest layer also an output being the disease attributes. We train the network by leaving out 10% of the disease node locations and corrupting the input data. We then predict the output taxonomy using the uncorrupted input and predict where new node locations will be and validate on the 10% holdout. The network performs similar to the results in Table 2 with an f-max of 0.49. This structure allows us to move back and forth between taxonomy space and attribute space. It also would allow us to expand further research into discovering new attributes or disease categories as hidden nodes.

4.4.4 Predicting New Associations

We examined the symptom predictions with the highest prediction scores. Each prediction fell into one of three categories: (1) associations that were straightforward (2) associations that were surprising and we could find evidence for and (3) associations that were surprising and we could not find evidence for.

Straightforward associations are found in Table 4.7. It may be that such associations are unsurprising or even known, but did not appear in the data set. Identifying such associations could be helpful to complete a sparse data set. For example, Lyme Disease has many neurological symptoms in the data set, but did not include the specific symptoms "Neurobehavioral Manifestations" nor "Synkinesis." It is unlikely that is because the association is biologically inaccurate, but due to the manner in which the data set was compiled that was not comprehensive.

Surprising associations for which evidence exists are found in Table 4.8. These associations did not appear obvious, but the literature included reports that sug-

Disease	Symptom
Lyme Disease	Neurobehavioral Manifestations
Lyme Disease	Synkinesis
Tobacco Use Disorder	Cheyne-Stokes Respiration
Meningism	Primary Progressive Nonfluent Aphasia
Tuberculosis, Cardiovascular	Systolic Murmurs

Table 4.7: Predicted disease-symptom associations that were straightforward.

gested they are possible. Bernard-Soulier Syndrome is a rare genetic disease affecting platelets resulting in prolonged bleeding. This would appear not to be related to Morning Sickness, but in fact causes several problem in pregnancy.[197] In this case the symptom does not appear to be causal in either direction, but associated with a particular subpopulation. Tauopathies are neurodegenerative diseases characterized by aggregation of tau proteins in the brain, whereas Systolic Murmurs are heart murmurs that can be heard during systole. The two appear unrelated, but a case report of one of the first cases of a tauopathy not in a human was a chimpanzee who was relatively healthy except for a history of systolic heart murmur, moderate obesity and high serum cholesterol.[198] This could be random chance or an unexplored association.

Kaposi Sarcoma is a tumor caused by a herpesvirus infection while Eructation is what may be commonly termed belching. Case reports have identified patients with Kaposi Sarcoma noting weight loss, abdominal distention, frequent eructation, fever and weakness.[199] Though Kaposi Sarcoma is often found on the skin, it can also be found in the gastrointestinal tract, particularly in patients with AIDS, and may explain the excess gas and eructation.[200] Chemotherapy-Induced Febrile Neutropenia is a low number of neutrophil granulocytes in the blood accompanied by fever. Myokmyia is involuntary twitching of the eyelid. Though seemingly unrelated, there have been case reports of patients with myokmyia after chemotherapy.[201]

Disease	Symptom
Bernard-Soulier Syndrome	Morning Sickness
Tauopathies	Systolic Murmurs
Sarcoma, Kaposi	Eructation
Chemotherapy-Induced Febrile Neutropenia	Myokymia

Table 4.8: Predicted disease-symptom associations that were surprising and supporting evidence could be found.

Disease	Symptom
Fever of Unknown Origin	Metatarsalgia
Contrecoup Injury	Halitosis
Invasive Pulmonary Aspergillosis	Urinoma
Liver Diseases, Parasitic	Alien Hand Syndrome
Ehrlichiosis	Respiratory Aspiration
Mandibular Diseases	Prosopagnosia

Table 4.9: Predicted disease-symptom associations that were surprising and no supporting evidence could be found.

These examples are not conclusive evidence that an association is present, but they do suggest the predictive algorithm may be able to find non-obvious associations. For surprising associations like these, it would be useful to find and explore them in ranked order as provided by our algorithm.

Surprising associations where no evidence could be located are found in Table 4.9. These associations did not appear obvious and a literature search did not provide any evidence that they actually exist. That does imply that these are not associations, but perhaps less likely than the ones above for which some evidence could be found. There were a few more of these examples than the examples given above.

These predictions in this section could be used to complete a sparse data set. For example, our data set had 4661 diseases and 322 symptoms. That is more than 1.5 million potential associations, which would be difficult to manually review. It could also be used to find unknown associations for which we were able to find evidence for in a significant number of cases. It may not be likely these associations would be used to immediately augment a data set, but could be use to inform future research.

4.5 Discussion

This work has demonstrated the potential for using taxonomies to augment disease-attribute data sets and vice versa. The immediate application is to fill in missing values in the disease-attribute data sets which may be sparse. For example, the disease-symptoms data was derived from the corpus of medical literature and is not meant to be comprehensive. If there is no paper that includes a disease and symptom together, it does not imply that there is no association. It could be that the association is known, but does not have any literature. It could be that the association is not yet known. It could also be that the association is known, but the literatures is not indexed with the given terms. For the genes data set, it may be that the experiments have not yet been performed to know if the association exists.

The taxonomy may similarly be sparse. Our previous work showed that the number of locations in the MeSH taxonomy that a disease occurs in has increased dramatically in the last decade. This may indicate that experts believe the previous locations were not sufficient to characterize the disease in the taxonomy. This work demonstrates one way to make suggestions to experts on how the taxonomy could be made less sparse based on data.

Beyond the immediate applications described above, this work shows the ability to move between disease spaces. Many learning problems have used the disease taxonomy to infer relations between diseases when a data set is sparse. This is particularly true for disease concepts that are higher in the tree. For example, in the disease-symptom data set, there are no associations of symptoms with cardiovascular disease because cardiovascular disease is too general of a term to be indexed with in the literature. Similarly, there are no directly associated genes or drugs. In order to infer relationships between cardiovascular disease and these attributes, many have proposed using the attributes of all descendants. The challenge with this approach is that one weak association with one descendant would be treated the same as many strong associations with many descendants. It also constrains the relationships to hierarchical ones. A more general disease space could allow for inference more broadly to provide a scaffold for filling in missing data when a data set may be sparse.

For cases where the taxonomy is used to fill in data, the work here provides a framework for deciding if there is enough information in the taxonomy about the attribute to use the taxonomy structure. For example, if a taxonomy did not contain much of the information that the disease-symptom data set contained, it would not be reasonable to use the taxonomy for making inferences about disease-symptom associations.

Finally, we have provided a way to map between disease spaces. Taxonomies are useful for communication between humans and provide a compact and sparse way to represent disease associations. For machine learning though, these attributes are not necessary. It may be that a more complex and dense representation may reflect the true disease space better. This work may provide the basis for moving away from using the taxonomy as the accepted disease space to other representations that can be used in machine learning as formal disease spaces.

Chapter 5

Disease-Drug Representation from Clinical Trials

5.1 Introduction

There is a wealth of biological, medical or health data available to use computational methods to learn new insights. Such learning may dramatically save time or resources compared to basic biology research or clinical trials. Gleaning new insights from data is not just about what data is used or the algorithm applied for learning, but also the representation of the data setup for the specific learning problem. There have been strong and laudable efforts to make more medical data available for learning, but we suspect there may be unexplored opportunities for new representations using data that is already available. One such opportunity is using clinical trial data.

The collection of all clinical trials represents research carried out on millions of humans[202] and represents thousands of decisions made by researchers, clinicians, and executives managing billions of dollars.[203] Each trial teaches us something about the specific relationship between a disease and an intervention like a drug that can be summarized in terms of safety and efficacy. However, the question of how to learn from the collection of clinical trials remains open. Little is known about the information clinical trials reveal as a whole.

We hypothesize that, at present, different trials are only connected indirectly through the expertise accumulated by the scientific, clinical, regulatory and executive decision-makers that deployed their expertise in several trials. We believe this expertise is latent in the collection of clinical trials and can be accessed via inference. Viewing the collection of clinical trials as observables emerging from this latent knowledge, we show how to leverage clinical trial meta-data to arrive at unique insights into the relationships between diseases. We further show how these insights may be useful in linking clinical data back to biology by generating hypotheses for future biological research.

Collections of clinical trials are now available in clinical trials registries, enabling work like ours to learn from multiple clinical trials together.[82] The desire to increase learning from multiple trials and possibly further inform scientific research has led many researchers to push for greater access to patient-level clinical trial data and even electronic health records.[204, 205, 206] The prevailing assumption within the field is that the more granular the data the better the learning.

In this chapter, we explore whether relationships among diseases can be learned from superficial data, like the kind of meta-data available in clinical trial registries, and how the relationships uncovered compare to relationships learned from molecular and biological data. If true, this would provide evidence contrary to the prevailing assumption.

The notion that new knowledge may be generated from the structured aggregation of databases has been used before to explore relationships between diseases or between drugs. For example, Goh et al. and Butte et al. compared diseases and phenotypes by shared gene correlations based on the Online Mendelian Inheritance in Man database[125] and the Gene Expression Omnibus database.[175] Zhou et al. compared diseases by shared symptoms from text in PubMed.[124] Hidalgo et al. compared diseases by shared comorbid conditions.[207] Yildirim et al. compared drugs based on shared targets.[208] Campillos et al. and Tatonetti et al. compared drugs based on shared side-effects.[209, 88] These examples provide a strategy to access disease similarities that would not have been apparent from individual experiments.[210]

New insights into disease similarities or dissimilarities can have dramatic results. For example, the similarity between psoriasis and multiple sclerosis led to the blockbuster drug Tecfidera (dimethyl fumarate) for treating relapsing-remitting multiple sclerosis.[108] As another example, the discovery of genetic dissimilarity in breast cancer corresponding to dissimilar prognoses led to new and improved diagnosis and treatment options.[211] Furthermore, similarity metrics between drugs have been used to predict indications, targets and drug interactions.[212, 213, 89] Such predictions have enabled repurposing of existing drugs and avoiding adverse drug reactions.[87]

To some degree, all of these examples may be viewed as accessing latent information about underlying biology. In this work, we follow an approach similar in spirit, but introduce the hypothesis that the underlying biology is contained in the expert knowledge used to decide to conduct the trials. Expert knowledge may come from literature reviews, proprietary biological experiments, other clinical trials and the summary of expert opinions. We further hypothesize that such expert knowledge may be accessed as latent information from clinical trial meta-data. By uncovering that latent information, it is possible to extract what experts collectively know about the relationships between diseases tested. Using this information, one may explore the relationships between diseases and generate hypotheses to guide future research. In this work, we use aggregated clinical trial meta-data to construct relationships among diseases and show that similar relationships can be learned compared to analyses based on detailed biological data.

To the best of our knowledge, studying disease relationships by using the entire set of clinical drug trials along with the premise of expert knowledge latent in the trial meta-data is novel and a key contribution of this work. Our approach and subsequent results imply that there is much more knowledge to be gained from existing clinical trial data. It also suggests a path to compare patient-level data across different trials when such data becomes more widely available.

The chapter is organized as follows. We explain how we used free text meta-data from drug trials on ClinicalTrails.gov to construct a model of the diseasome. We then explore the connectivity between diseases and drugs and visualize the data in a network layout. We report on the validation of the disease-disease network (DDN) against a standard disease taxonomy and a diseasome built from genetic data. The relationships derived from our network show surprising agreement with relationships based on genetic data or medical taxonomies and show promise for informing future scientific research.

5.2 Methods

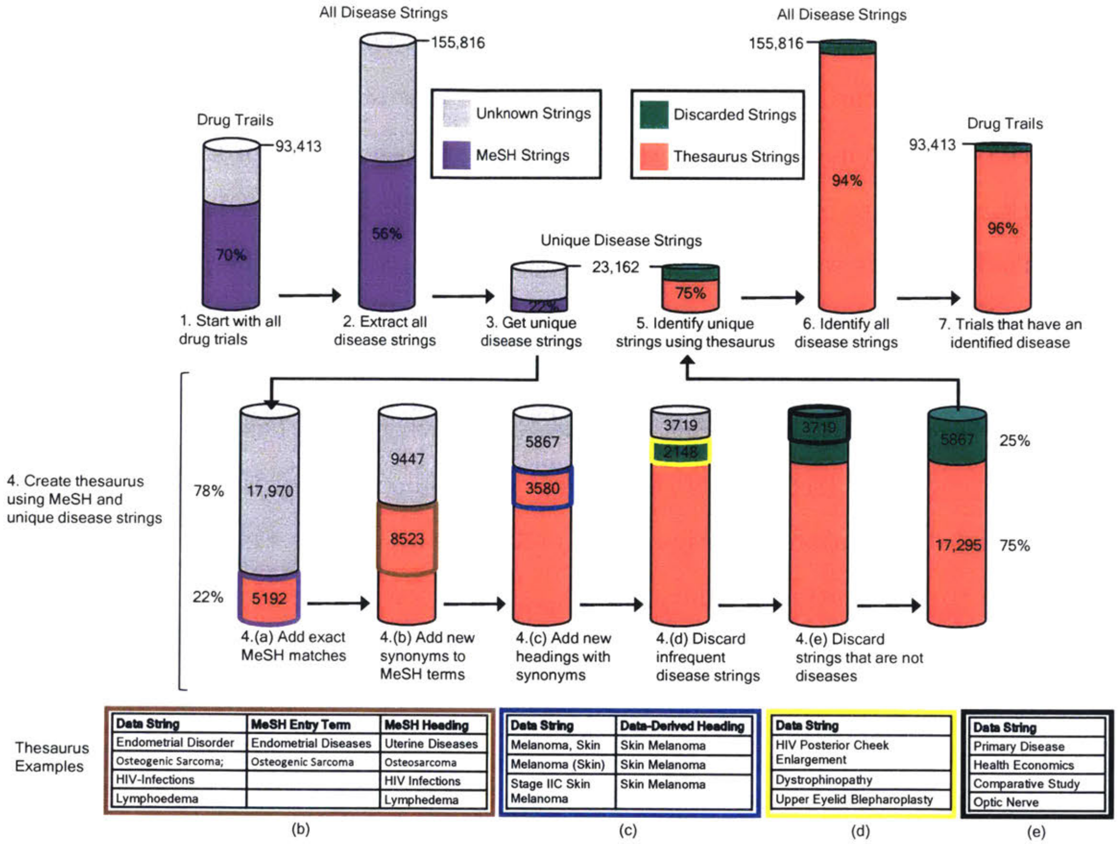
5.2.1 Construction of the Disease-Disease Network

We extracted meta-data from 93,654 clinical drug trials from ClinicalTrials.gov (see Appendix, Section D.1). The meta-data included a list of one or more free text strings for conditions (diseases) and a similar list for drugs. Comparing diseases or drugs from different trials was sometimes ambiguous because two different text strings could represent the same concept. For example, there are 73 different strings that represent the single concept Type 2 Diabetes Mellitus (see Appendix, Table D.1).

We found some resources in use to disambiguate disease and drugs on ClinicalTrials.gov such as browsing trials by condition or drug intervention on ClinicalTrials.gov[214] or using the AACT database.[215] These resources included errors though, such as "Imidacloprid" being a drug for 129 trials when browsing by intervention or in 6 trials in the AACT database. Imidacloprid is actually an insecticide that does not show up in ClinicalTrials.gov. We traced this false positive to the trade name "Advantage," which was a synonym for Imidacloprid in the 2013 version of MeSH. These false positives occur because both resources rely on an automatic algorithm for finding MeSH terms.[215] Such false positives would erroneously connect diseases in our analysis. We also found that these resources have built in inferences based on the NLM algorithm or annotations by clinicians. Our goal was to use the raw data as much as possible, without introducing layers of inference.

Much of the work that analyses large sets of clinical trials is based on the AACT database.[216, 217, 218] Other works focus on specific aspects of the trials such as drug combinations[219] or participation criteria,[220, 221] with algorithms tailored to the specific fields used to automatically extract their data sets. Due to the nature of our approach, it more important for us to reduce false positives and we therefore used manual curating of the data. To enable comparison we built a thesaurus of terms, starting with diseases. The process is outlined in Figure 5-1. We started with the medical subject headings vocabulary (MeSH)[222] as a base thesaurus. Another option would have been UMLS, [120] which is a more comprehensive thesaurus based on many databases and can be accessed with enhanced tools such as MetaMap.[223] We chose MeSH because contributors to ClinicalTrials.gov are encourage to use its vocabulary [224] and by using one database we avoid inferences as described above. MeSH only identified 22% of the unique disease strings listed in ClinicalTrials.gov, accounting for 56% of all disease strings and 70% of the trials. We manually reviewed the 17,970 disease strings not identified by MeSH, comparing each one to the 20 closest terms in MeSH generated by fuzzy string matching (see Appendix, Section D.2). If a match could not be made and the disease string occurred repeatedly, we created a new "data-derived" term in the thesaurus. The number of MeSH disease terms used over time peaked in 2008 and the percentage of terms we added to our thesaurus increased linearly over time (see Appendix, Figure D-1), suggesting that the 2014 MeSH vocabulary did not include terms currently used in research, hence the need for our enhanced thesaurus. Using our thesaurus, we identified 94% of the disease strings for 96% of the drug trials listed in ClinicalTrials.gov.

Drugs were more challenging to disambiguate because one drug string often included multiple drugs and because experimental drugs are often not found in MeSH. Of 63,066 unique drug strings, we generated 503,270 possible substrings that could be drugs and used automated and manual filtering to identify all drugs and construct a thesaurus (see Appendix, Section D.3 and Figure D-2). We analyzed the drug thesaurus and found similar patterns to that of the disease thesaurus (see Appendix, Figures D-3 and D-4). We then used the drug thesaurus to identify individual drugs in drug strings (see Appendix, Table D.5). To assess accuracy, we took a sampling of 100 random trials with 216 drugs and found that 98% were identified (see Appendix, Table D.6). We also sampled 100 drug strings that were not mapped to any drugs and found that 92% were correctly excluded (see Appendix, Table D.7). The resulting disease-drug data set accounts for 93,069 trials and includes 132,822 diseases


All Disease Strings
155,816
Drug Trails
93,413
70%
56%
Unknown Strings
MeSH Strings
Discarded Strings
Thesaurus Strings
All Disease Strings
155,816
94%
Drug Trails
93,413
96%
Unique Disease Strings
23,162
22%
75%
1. Start with all drug trials
2. Extract all disease strings
3. Get unique disease strings
5. Identify unique strings using thesaurus
6. Identify all disease strings
7. Trials that have an identified disease
4. Create thesaurus using MeSH and unique disease strings
78%
17,970
22%
5192
9447
8523
5867
3580
3719
2148
3719
5867
25%
17,295
75%
4.(a) Add exact MeSH matches
4.(b) Add new synonyms to MeSH terms
4.(c) Add new headings with synonyms
4.(d) Discard infrequent disease strings
4.(e) Discard strings that are not diseases
Thesaurus Examples
(b)
(c)
(d)
(e)

Data String	MeSH Entry Term	MeSH Heading
Endometrial Disorder	Endometrial Diseases	Uterine Diseases
Osteogenic Sarcoma;	Osteogenic Sarcoma	Osteosarcoma
HIV-Infections		HIV Infections
Lymphoedema		Lymphedema

Data String	Data-Derived Heading
Melanoma, Skin	Skin Melanoma
Melanoma (Skin)	Skin Melanoma
Stage IIC Skin Melanoma	Skin Melanoma

Data String
HIV Posterior Cheek Enlargement
Dystrophinopathy
Upper Eyelid Blepharoplasty

Data String
Primary Disease
Health Economics
Comparative Study
Optic Nerve

Figure 5-1: Creating a thesaurus maximizes the data that can be used. To compare disease terms to each other, we needed a standard vocabulary with synonyms. We start with the Medical Subjects Headings (MeSH), but only 70% of drug trials on ClinicalTrials.gov and 56% of the diseases listed in those trials can be found in MeSH. We augment MeSH by looking at every unique disease string, of which only 22% are in MeSH. Going through the remaining 78% manually we either add another synonym to a MeSH term (4b), create new terms from the data with accompanying synonyms (4c) or discard infrequent or irrelevant strings (4d and 4e). Every unique string has been reviewed and either included in our thesaurus or discarded. From our thesaurus we can identify 94% of all disease strings, which mean we can compare data from 96% of the trials.

(3,663 unique) and 175,584 drugs (7,349 unique). "HIV Infections" was the most tested disease with 2,772 different trials involving 700 different drugs. There were

1,211 unique diseases tested in more than 10 trials with at least one drug and 1,784 unique drugs tested in more than 10 trials with at least one disease. The variety of trials involving different diseases and drugs provides the connectivity for a network.

To construct the network we start with the disease-drug network, a bipartite network of disease nodes and drug nodes. Figure 5-2 shows the process creating the network and Figure 5-3 shows descriptive statistics of the data. Diseases and drugs are linked by trials with the width of the edge proportional to the number of trials with both the disease and drug. The Disease-Disease network (DDN) is constructed using only diseases as nodes and the edges representing the number of different drugs tested at least once on both diseases. The connectivity of the Disease-Drug network follows a power law (see Figure D-5), while the connectivity of the DDN follows an exponential distribution (see Figure D-5).[225] We assign the weight of all edges in the graph to 1 indicating a binary connection between diseases for simplicity in this work.

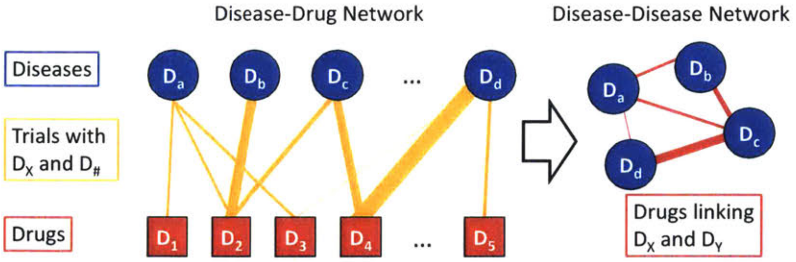


Figure 5-2: Construction of the Disease-Disease Network with descriptive statistics. The disease-disease network (DDN) is constructed from a bipartite network of diseases and drugs linked by trials. In the bipartite network, the thickness of edges corresponds to the number of trial that have both the disease and drug the edge connects. In the DDN, diseases are linked by drugs with the edges proportional to the number of different drugs tested in trials with both diseases the edge connects

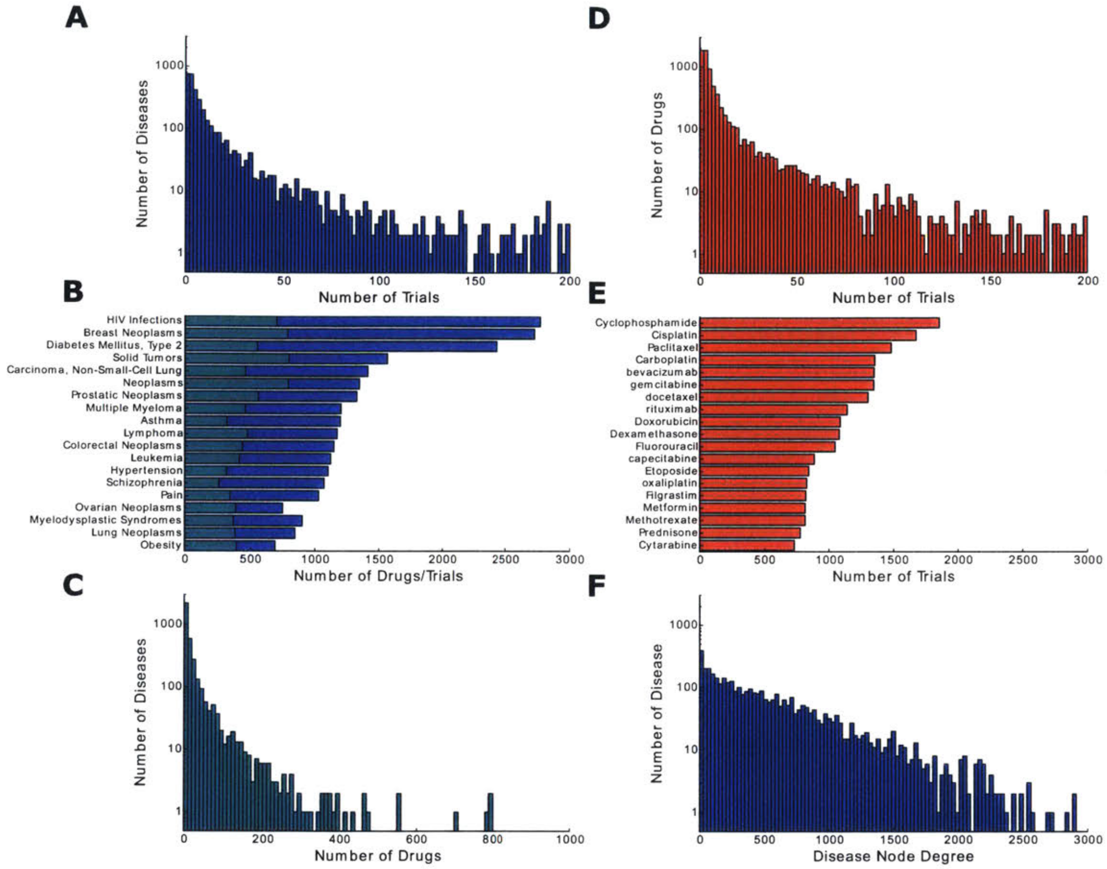


Figure 5-3: Descriptive statistics of the disease-drug data set. (A) The number of diseases that occur in a given number of trials. There are diseases appearing in thousands of trial, but for visualization purposes the plot is truncated at 200 trials. (B) The top 15 diseases by number of trials (dark blue) and number of drugs (cyan). (C) The number of diseases tested with a given number of unique drugs. (D) The number of drugs that occur in a given number of trials. (E) The top 15 drugs by number of trials. (F) The number of diseases associated with how a given number of other diseases by the criteria that both diseases were tested with a particular drug. The x-axis can be viewed as the degree of each node in a network of diseases linked to each other.

5.2.2 Visualization of the Disease-Disease Network

Visualizing the network is difficult because there are 3,663 disease nodes with hundreds of thousands of edges between them. To reduce the number of edges, we filter edges to keep ones with strong relationships. We define an edge as strong if one of the two disease nodes it connects is frequently associated with the drug the edge represents. A frequently associated disease for a given drug is one that shows up in a significant percent of all trials for that drug. To determine significance we used a binary test with a cutoff p-value of 0.001 and used Bonferonni correction for comparing multiple diseases. We selected this method for filtering because of a pattern we found in drug trials. New diseases a drug is tested on are either tested in conjunction with an established indication for that drug or tested on completely new diseases. We hypothesized that the first case suggests a deeper characterization of the drug and diseases, while the latter case suggests an exploration of possible new indications. We treat the diseases in the first case as having a strong relationship. Our filtering for strong relationships between diseases is only one way to examine the data. In the discussion we explain why subtle weak relationships are potentially more interesting.

The network graph was laid out using the Fruchterman and Reingold method, though other force-directed layout algorithms gave similar results (see Appendix, Section D.7 and Figure D-6). Node size is proportional to the number of drugs tested on the disease and edge width proportional to the number of drugs tested on both diseases. Nodes are colored according to MeSH categories of diseases (see Appendix, Section D.8 and Table D.8).

5.2.3 Validation with the MeSH Taxonomy

The MeSH disease taxonomy is constructed by experts based on their biological and clinical understanding. If the DDN can reproduce the MeSH taxonomy, it would suggest the DDN captures the same level of information implicitly that experts explicitly outlined when constructing MeSH. To explore similarity between the DDN and MeSH we quantitatively evaluate clustering of diseases in the DDN by MeSH category. First we evaluated internal consistency of clusters in our network visualization using the nearest neighbor index.[226] Second we evaluated how distinct clusters are based on graph theoretic distance. Third we evaluated how consistent the DDN and MeSH are compared to a randomly constructed network using a binomial test.

5.2.4 Validation with the Human Disease Network (HDN)

We also validated the DDN by comparing it to the Human Disease Network (HDN), which was constructed using a database of disease-gene associations.[125] The HDN was validated by examining clustering by disease categories that match MeSH categories (see Appendix, Table D.8). First we evaluate the internal consistency of clusters of nodes within the same category by measuring the fraction of edges connecting nodes within that category. Second we use the ratio of shortest paths within versus without a category to derive a graph theoretic measure of clustering for each disease category. Lastly we evaluate how much overlap there is relative to a randomly constructed network using the binomial test.

5.2.5 Prediction Potential

We test the potential of the DDN for prediction by building a rudimentary recommender engine for clinical trials. We used the entire unfiltered DDN, rather than the filtered version that we used for visualization purposes, to capture the subtle relationships among diseases and not just the strongest ones. Our training set contained trials starting before 2011 and our test set contained trials starting in 2011 or later. The data set contains 2,160 diseases that were tested with at least one drug in the training set and one drug in the test set. There are 7,349 possible drugs to predict with 54,509 disease-drug pairs in the training set and 19,157 disease-drug pairs in the test set. Each disease is represented by a vector of drug variables with a 1 indicating that the drug was tested in a trial with the disease and a 0 otherwise. The purpose of the recommender engine is to suggest drugs that had not previously been tested on a given disease, but may be relevant to a disease based on data in the training set. For a given disease, we made predictions about each drug using collaborative filtering[227] with a cosine similarity metric, or the normalized inner product between disease vectors. We evaluated the performance of the recommender engine looking at the area under the ROC curve[228] for each disease in the 3.5 year period after 2011.

5.3 Results

5.3.1 The Disease-Disease Network

The network graph resulting from our layout is plotted in Figure 5-4, which contains a giant component with 1,101 nodes and 6,972 edges. The distance between nodes in the graph represents similarity based on shared drugs directly or through other disease nodes. By visual inspection there is clustering of nodes of the same color or MeSH category. Many nodes in close proximity are expected such as Crohn's Disease and Ulcerative Colitis or Parkinson's and Alzheimer's. At the same time there are surprises, such as Hypertension and Parkinson's being close together. We also observe similarity of disease categories, such as Psychiatric and Nervous System diseases next to each other or Cardiovascular and Metabolic diseases mixed together.

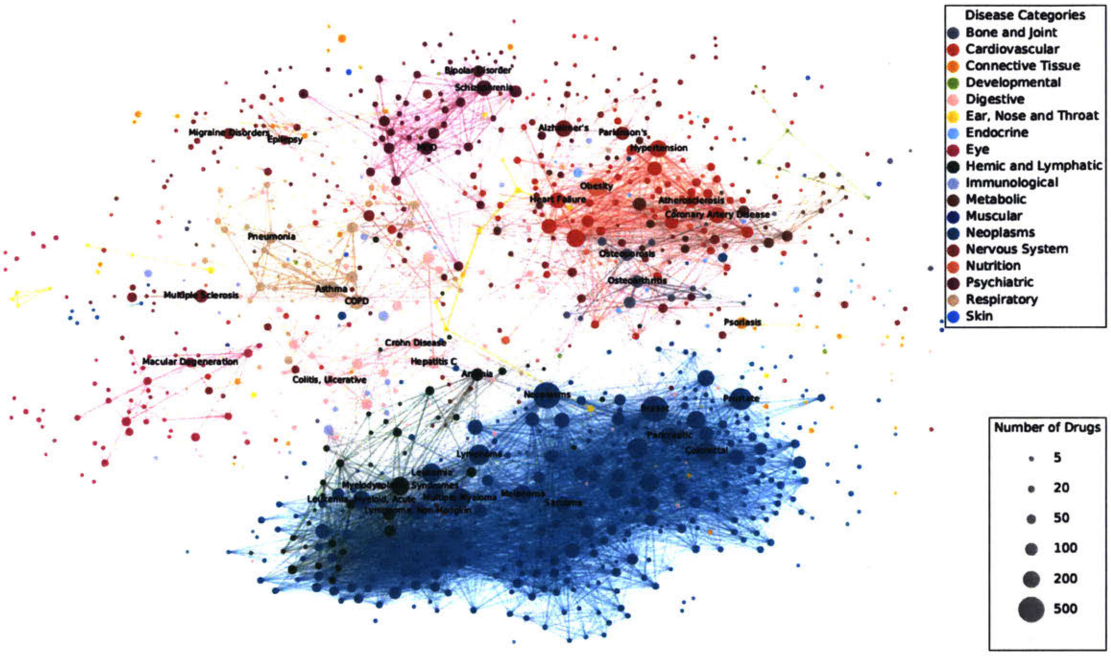


Figure 5-4: Visualization of the Disease-Disease Network. In the disease-disease network, diseases are represented as nodes in the network with the size of the node proportional to the number of different drugs tested on that disease. Edges between nodes are drugs tested on both diseases the edge is connected to. Thickness of the edge is proportional to the number of drugs tested on the two diseases. For ease of comparison with MeSH we colored nodes according to MeSH disease subtrees, though that information was not used by the visualization algorithm. A cluster of nodes of one color indicates the disease-disease database captures information about the relationships between diseases that can be found in the MeSH taxonomy.

5.3.2 Validation with MeSH

Validation of the DDN against the MeSH taxonomy is shown in Figure 5-5. The nearest neighbor index (NNI) for a group of data points in a plane indicates whether the points are randomly spaced (an index of 1), non-randomly clustered (an index smaller than 1) or non-randomly spaced apart (an index larger than 1). The NNI for each MeSH category (Figure 5-5A) is less than 1 for all categories except "Skin." Compared to randomly connected networks, 13 of the 15 disease categories with 10 or more nodes have a significantly smaller ($p<.05$) NNI (see Appendix, Table D.9). All three disease categories with fewer than 10 nodes were not significant. All p-values were calculated empirically using Monte Carlo simulations (see Appendix, Table D.9).[229]

Figure 5-5B shows how close nodes of the same MeSH category are (colored bars) compared to how close nodes of different MeSH categories are (gray bars) using shortest path distance. Distinct clusters have a significantly shorter colored bar than gray bar, which is the case for all 15 disease categories with 10 or more nodes. Compared to randomly connected networks, 14 of these categories have a significantly ($p<.05$) shorter colored bar (see Appendix, Table D.9). Only 1 of the 3 disease categories with fewer than 10 nodes had a shorter colored bar, which was also significant compared to randomly connected networks. Figure 5-5C shows the binomial distribution of the number of edges connecting nodes of the same category if they were randomly placed in the network and the red arrow indicates how many correct links we observe in the DDN. For the binomial test the p-value cannot be calculated using double floating point precision (see Appendix, Section D.9).

The three evaluations show that the connectivity of the network as a whole significantly reflects categories in MeSH. In addition to the clustering within categories, we note related diseases of different categories that are close such as "AIDS" (Endocrine) and "Hepatitis C" (Digestive), "Myelodysplastic Syndromes" (Hemic and Lymphatic) and "Leukemia" (Neoplasms) or "Hypercholesterolemia" (Metabolic) and several Cardiovascular diseases.

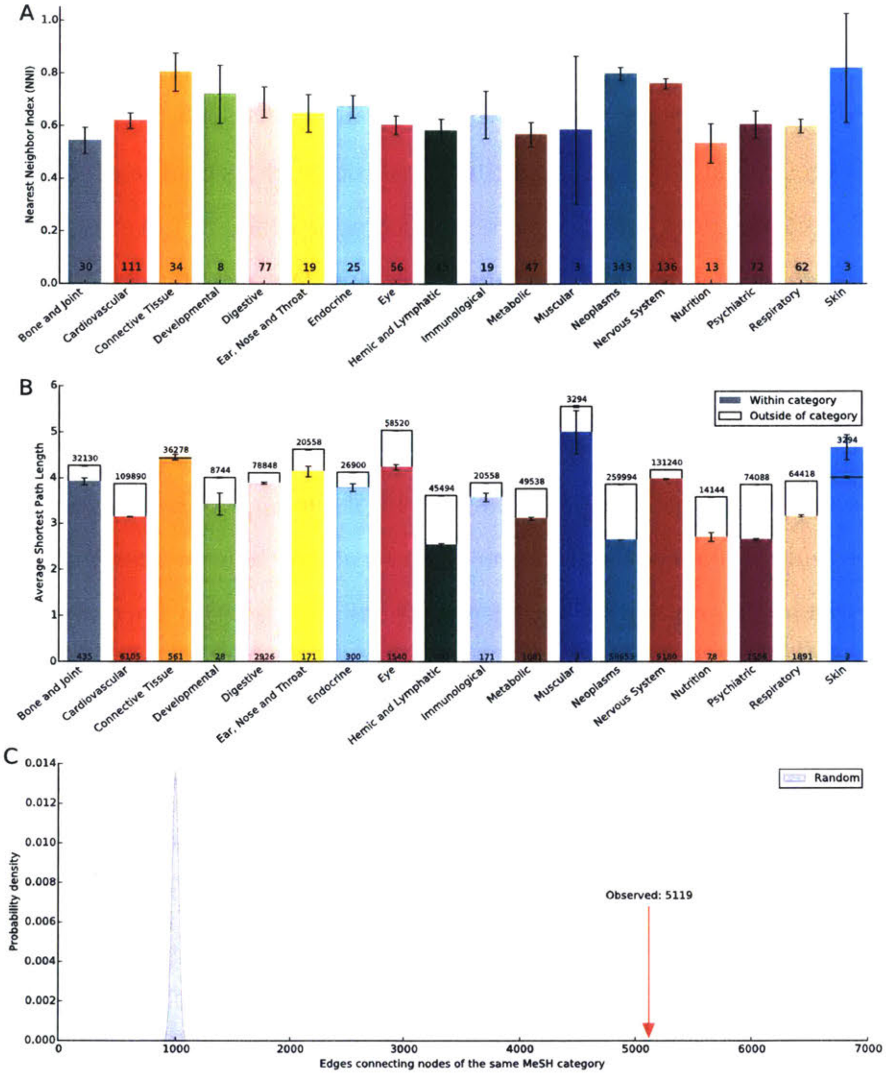


Figure 5-5: The DDN shows clustering by MeSH category. (A) Nearest neighbor index (NNI) for each MeSH disease category. **Values significantly less than 1 indicate clustering in our visualization.** Error bars are standard error and numbers on bars are the number of nodes in the category. (B) Average shortest path length between nodes in the same category (colored bars) compared to the average shortest path between a node within a given category and all nodes outside the category (white bars). **Colored bars significantly lower than white bars indicate more distinct clusters in** the network. Error bars are standard error, the numbers on the bottom indicate how many edges are within the category and the numbers on top indicate how many edges leave the category. (C) Binomial distribution of edges between nodes of the same category if they were randomly placed on the graph (purple shading) compared to the observed number of edges in the disease-disease network (red arrow).

5.3.3 Validation with the HDN

There are visual similarities between the DDN and HDN such as Neoplasms/Cancer being the largest cluster. There are differences too, such as deafness being prominent in the HDN but absent in our plot, which reflects the different data sources. Deafness may be strongly associated with certain genes, but does not currently have any pharmaceutical treatment options. Quantitatively, the average degrees fraction within a disease category indicates how connected nodes in the category are to each other. Figure 5-6A shows average degree fraction is similar or larger for the DDN (color bars) compared to the HDN (gray bars) for 13 of the 15 MeSH categories with 10 or more nodes and for 0 of the 3 with fewer than 10 nodes. The two categories with more than 10 nodes and a smaller average degree fraction in the DDN are "Metabolic" and "Hemic and Lymphatic." "Hemic and Lymphatic" is an interesting case where the DDN has a smaller degree fraction than the HDN. This is true because it is mixed with Neoplasms in the DDN, which may reflect the similarity in treatment in hematology and oncology while the genetic basis may be more distinct. The difference between the DDN and HDN compared to the difference between randomly connected networks and the HDN is significant ($p<.01$) for all categories except "Muscular," and "Skin," which do not have any directly connected nodes (see Appendix, Table D.10).

Taking into account indirect connections between nodes, Figure 5-6B shows the mean of the ratio of shortest path within versus without each category, where a smaller ratio indicates tighter clustering within the category. The DDN has a similar or smaller ratio for 10 of the 15 categories with 10 or more nodes and for 1 of the 3 with fewer than 10 nodes. The difference between the DDN and HDN compared to the difference between randomly connected networks and the HDN is significant (p<.01) for all categories except “Connective Tissue,” “Muscular” and “Skin.” (see Appendix, Table D.10).

Comparing the two networks directly we found 181 common nodes with 764 edges among those nodes in the DDN and 192 edges among the same nodes in the HDN. The expected number of overlapping edges is a binomial distribution as seen in Figure 5-6C. We observed 73 edges were the same giving a p-value of 9×10^{-42} (see Appendix, Section D.9). There is significant overlap in disease relationships found in the DDN compared to the HDN even though the two networks are constructed using very different datasets.

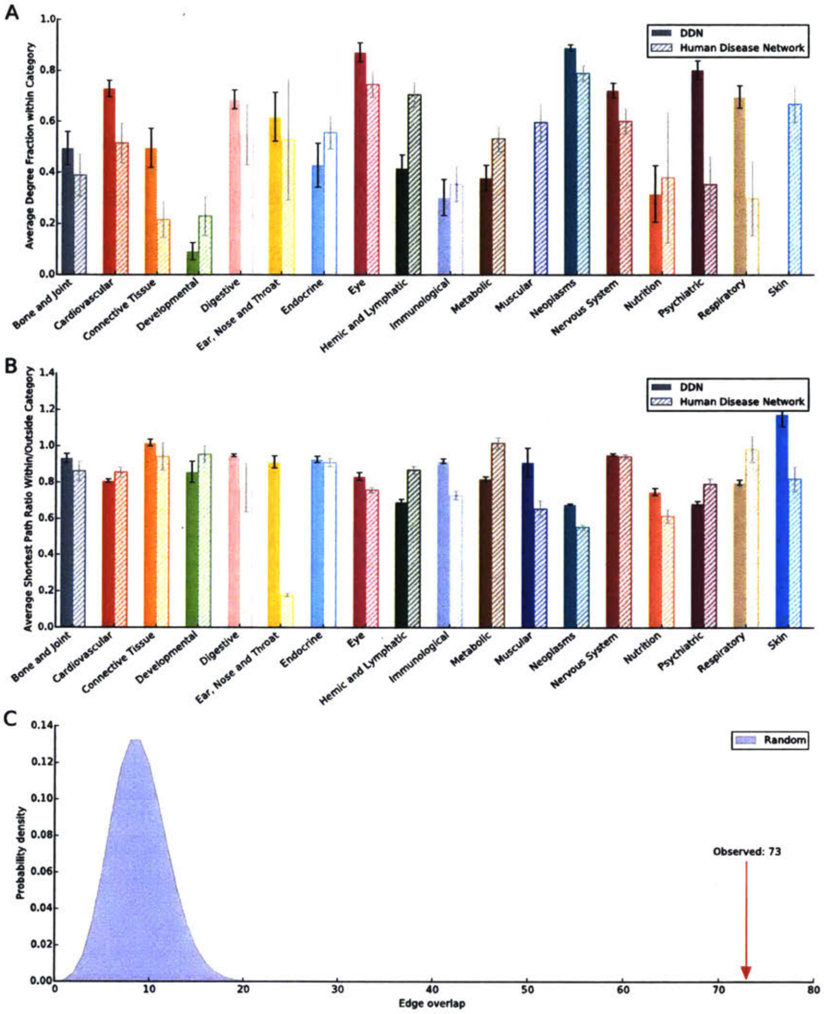


Figure 5-6: The DDN shows similarity in clustering of MeSH categories compared to the Human Disease Network (HDN), which is based on a very different data set. (A) The average degree fraction ratio for nodes within the same category. The ratio is the number of edges extending to nodes in the same category to the number of edges extending to nodes in different categories. The DDN is shown in color for each category with the HDN in gray next to it. In general the DDN shows a similar or greater ratio than the HDN demonstrating similar or even tighter clustering. (B) The average shortest path ratio for nodes within the same category. The shortest path ratio for a node is the mean of the shortest path to every node in the same category to the mean of the shortest path to every node outside the category. As in (A) the color bars correspond to the DDN and the gray to the HDN. In general the DDN shows a similar or smaller ratio than the HDN demonstrating similar or possibly better clustering. (C) Binomial distribution of overlapping edges between the disease-disease network and the HDN if edges were randomly placed between nodes with the observed number shown with the arrow. The comparison is only made for nodes that are identical in the two graphs.

5.3.4 Prediction Potential

Figure 5-7 shows the AUC for all diseases as a scatter plot also showing the disease category, the number of drugs in the training set and the number of drugs in the test set. Random predictions would give an AUC around 0.5 while the majority of our predictions have an AUC much larger. The average AUC for diseases was 0.845. The histogram of all AUC's compared to random predictions benchmarks the predictive ability of the network (see Appendix, Figure D-7). Using the Shapiro-Wilks test for normality of the AUC scores, the p-value is 6×10^{-39}. We note that if a trial did not occur in the 3.5 year test set time period it does not indicate that a trial will not happen in the future or that there is no connection between the drug and disease, so this result represents a conservative estimate. Examples of diseases with an AUC of more than 0.95 are provided in Table D.11 in the Appendix along with references to recent literature supporting the connection between the disease and the predicted drug.

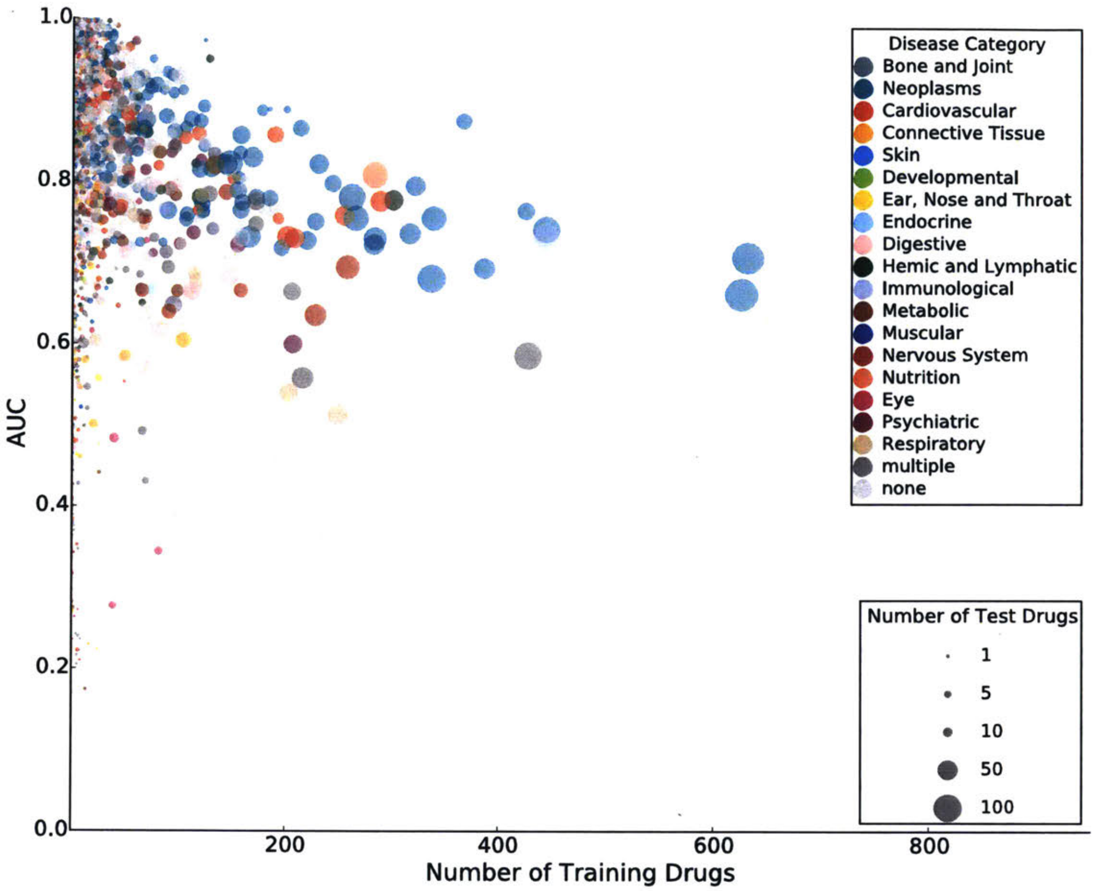


Figure 5-7: Area under the curve for drug predictions for individual diseases. The AUC for every disease is shown on the y-axis with the number of drugs in the training set on the x-axis and number of drugs in the test set for each disease as the size of each circle. The disease category is also shown as the color of the circle. Intuitively, more data should lead to more accurate predictions, but as the number of drugs in the data set increases, the AUC actually decreases. This may be because diseases that have many drugs (more than 50-100) may be more prevalent and result in more random exploration of drugs than those with just a few.

5.4 Discussion

Our results demonstrate that clinical trial meta-data can be used to infer disease relationships found in genetic data or medical taxonomies. Such agreement is surprising given that the meta-data used contains no explicit information about biology. The agreement suggests that our method succeeded in leveraging information latent in the collection of clinical trials to draw conclusions beyond what any single trial could reveal. The aggregate expert knowledge may reveal disease relationships one group of experts may not have identified by themselves. This result opens the possibility that such latent information in clinical trial data or other types of clinical data can be used to uncover biological relationships that otherwise might only be found by using detailed biological data, by gaining access to large amounts of clinical data or by resource-intensive research. There are several promising avenues to use the disease-disease network (DDN) as a resource to generate new hypotheses for biological and medical research.

The visualization of the DDN we presented provides a quick global reference of the therapeutic links among diseases that conveys the underlying similarity between diseases. This similarity is based on the decisions to run specific clinical trials as observed in ClinicalTrials.gov. The decision to conduct a trial is based on the summation of biological and medical knowledge, such as published research, proprietary in vitro or animal study results, clinical observations, results from previous clinical trials, and economic considerations. This represents a significant body of cross-disciplinary information leading to the decision to run any single trial. At present, lessons learned from this cross-discipline endeavor are shared in part through publications, reviews and conference proceedings; collating this information for the entire body of clinical trials to derive lessons about human biology would be very time intensive. Instead, our representation of clinical trial meta-data allows us to access that cross-discipline information implicitly to derive conclusions and lessons learned. The DDN we built demonstrates one way in which access to that implicit or latent information can be used to draw conclusions beyond the information contained in any single trial - such as similarities between diseases.

As an example, asthma and inflammatory bowel diseases (Crohn's disease and ulcerative colitis) are closely related in the DDN map even though the MeSH taxonomy classifies them differently and there is little direct connection between the diseases in the data set. They are each inflammatory diseases though and recent research suggests that patients with asthma are at higher risk for inflammatory bowel disease.[230] Such similarities could stimulate hypotheses about related biological pathways, epidemiological connections, comorbidities in patients or new indications for drugs.

As a computational tool, we showed that the DDN may be used to recommend drugs to test on a given disease. Similarly, for a given drug, one might predict which set of diseases would most likely benefit. Others have tried this approach by relating drugs using aggregated datasets as described earlier.[208, 83, 88] The DDN may be most useful in combination with other data sources. For example, overlap of gene expression in two diseases in the Human Disease Network (HDN) [125] seemed to indicate the genes might be involved in the same biochemical pathways in the two diseases. The DDN also provides information about shared biochemical pathways, but from the perspective of drugs that could modulate those pathways. Together, the HDN and DDN could point more precisely to pathways or help distinguish between a genetic or environmental etiology. The DDN could be used in conjunction with data sets such as epigenetics, metabolomics, environmental factors, symptoms and others that could be layered together to support inference.

In this work, we limited ourselves to trial header information, based on our desire to focus on what we called superficial information and our intent to explore the limits of learning imposed by availability of data. Our work shows that there is potential to extract more information from clinical trial data.

Now that we have established this as a benchmark, future work could make use of additional information about each trial such as inclusion/exclusion criteria or trial results. The thesaurus we built and the cleaned trial data will be extremely useful for expanding this work and may prove useful for other research in related areas such as meta-analysis.

For visualization and validation, we limited our exploration to a subset of the network. We filtered out some of the nodes for easier comparison with the human disease network (HDN), such as infectious diseases that do not occur in the HDN. We filtered out edges for clearer visualization and to demonstrate the structure present in our data. Having validated a large subset of the network, future work could explore visualization methods on the unfiltered network, other filtering techniques to extract different meaning from the network, and the use of other similarity metrics and inference on graphs such as in [231].

There is other information in the clinical trial data such as economic potential of drugs, special interest in orphan diseases or prevalence of diseases in developed countries. Such information may bias an attempt to draw inferences about biological relationships. In this work, we focused on capturing all learning from trials. Future work using this learning for inference should account for bias depending on the specific inference problem. One example where bias would not need to be removed would be predicting which sets of diseases are least explored. Such predictions would be useful for determining what future trials would lead to the greatest increase in understanding of diseases. The amount of learning indicated by the topology of the DDN does not always match a straightforward measure such as the number of trials or drugs tested on a disease (see Appendix, Figure D-8). Beyond the DDN as a resource, the approach we demonstrated may prove useful in other areas where latent information is contained in seemingly superficial data. Though it is anticipated that floodgates of medical data will be released in the future, there is much more that can be done with the seemingly superficial data that is currently available. For example, “off-label” prescription data could be used in the same manner to uncover aggregated learning implicitly taking place by physicians in clinical practice.

Our results provide an example of using experimental data on humans, which is rare and valuable, to extract biologically useful information. This approach is different from the typical approach of learning biology mechanisms in models and testing to see if they also hold in humans. Here, we have shown how relationships can be derived from testing in humans and then explored to see if those relationships can improve understanding of biological mechanisms. As more clinical data does become available, it will be important to have tools like these in place to more rapidly uncover biological insights and discover effective treatments. For example, patient-level clinical trial data is becoming more available to researchers, but it is not clear how to compare two patients from two different trials that were constructed for different purposes. As patient-level data becomes available, we see opportunities to extend this work to provide a structure for making such comparisons and posing research questions that do not depend on clinical end-points.

We demonstrated that clinical trial meta-data can be used to derive biologically meaningful disease relationships as tested using other disease networks and taxonomies. We therefore conclude that there is latent expert knowledge in the meta-data. Our disease-disease network (DDN) shows a way to access that knowledge and to leverage the collective expert understanding of diseases. The relationships unique to the network can be used to generate new hypotheses for future biological and clinical research. This demonstrates a new strategy to leverage research data on humans to advance our understanding of biological mechanisms. Furthering this approach of translating clinical data back to biological research will become even more important as more granular clinical data becomes available.

Chapter 6

Discussion

In this discussion we will recap the central premise we started with and the main results. We then go through each of the results chapters individually, providing a brief summary of the results along with a discussion of the impact of the results. Additionally, for each chapter we enumerate the limitations and avenues for future work to build towards. We conclude the discussion by relating this work to the current movement for advancing precision medicine.

6.1 Summary of Central Premise

This thesis started with the premise that more precise diseases lead to better healthcare and that measuring and reducing variance will lead to more precise diseases. Researchers and practitioners aspire to learn what causes diseases, how to identify diseases, what treatments will affectively alter a disease and so forth.[9] Using computers to aid learning is a relatively young field of study and provides a useful perspective from which to understand learning in medicine.[232, 233] In the introduction, we defined computational learning as selecting a mapping or function from a set of possible functions to explain a relationship using data.[14] This definition implies that the mapping identified is optimal for any intended data set – not just the limited training set. One of the biggest and most fundamental challenges in computational learning is that data is limited and often sparse, requiring prior knowledge to allow for models to generalize.

Prior knowledge can be incorporated in a variety of ways. All computational learning problems require a representation, typically a vector, that a computer can interpret. To compare vectors, a distance or similarity function is required. Together, the representation and distance create a space in which the data exists, where each piece of data is a point in space as seen in Figure 6-1. Prior knowledge can be incorporated into the design of the representation, the distance functions or the regularization technique. One of the most accepted and formalized knowledge bases of diseases is the disease taxonomy. The disease taxonomy has been used as prior information for learning tasks to help cope with the challenge of data sparsity.[174] Essentially, it partitions the space of diseases in a nested manner, but it has not yet been proven that this is the best way to incorporate prior knowledge for computational learning.

In this work, we explored how computational learning may be affected by variance inherent in the disease space, particularly the disease space defined by the disease taxonomy. To do so, we (1) quantified how much variance exists in the disease taxonomy, (2) propose new methods for updating the disease space using the taxonomy as a starting point, and (3) explored new representations of diseases. We explored the history of the disease taxonomy and developed a method for measuring variance of the disease taxonomy in chapter 3. We then looked at ways to reduce variance in the disease space by using external data sets in chapter 4. We also found that we may

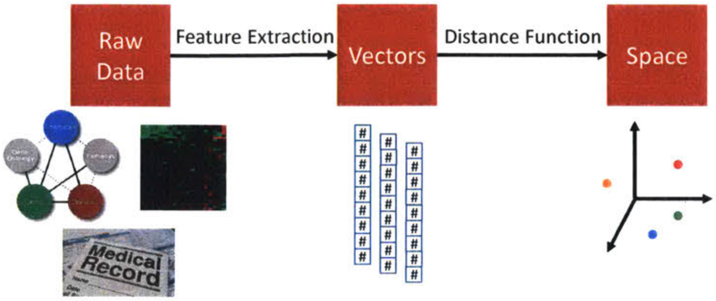


Figure 6-1: Diagram showing how raw data is mapped to a space where learning takes place. Features are extracted from the data to create a vector of numbers representing each data point. A distance function relating vectors to each other indicates how far points are away from each other, creating a mathematical space.

use the disease space to predict gaps in external data sets. Finally, in chapter 5 we showed that external data sets do not need to be detailed, patient-level data sets, but can be superficial data that is currently available. We discuss each of these results below, including limitations of the work and future directions. We then discuss the results of the current efforts to achieve more precise medical practice.

6.2 Variance in the Disease Space

We first used the history of the disease taxonomy over approximately 50 years to estimate variance within the disease taxonomy. We compared several data sets of diseases and other associated variables such as genes, clinical trials and medical literature to quantify the extent to which inferred relations in the disease space could vary over time. We found the amount of variance surprising; in the case of gene associations, we observed that approximately 20% of diseases changed by 50% over a period of 10 years. We cannot definitively say that one taxonomy is better than another over that 10-year span, but the variance over that time allows us to quantify an error bar. If current trends continue as they have over the past 50 years, our estimate may be the best predictor of, how much inferred associations could change 10 years in the future. This technique provides a tool for quantifying noise by deriving an estimate in contrast to assuming there is no noise.

One implication of this work is that the confidence of the conclusion of learning problems about diseases depends on the confidence of disease definitions, not just the data. For example, if the taxonomy is used to infer disease relationships with other variables, the degree to which the taxonomy changes should be used to estimate variance in the output to determine confidence.

We made the assumption that biology is consistent, but human understanding of biology is not. True underlying conditions are continuous concepts that remain constant over years or decades, but diseases used to describe those conditions are discrete and change over time. Sometimes, they change so much that they cease to exist. The conclusion of a learning problem can only be as good as the data is consistent. For example, "Cardiovascular Disease" may mean something different in 2015 than it did in 1971. In this thesis, we validated this assertion by examining the diseases that are underneath another disease in the taxonomy.

Other variables, such as genes, biochemical pathways, RNA sequences, cellular processes, physiological processes, organs affected, external effects, structural changes and interventions, can be associated with a disease. If the disease changes – such as being split or removed, as we saw in the history of the taxonomy – the associated variables would be re-allocated. The same is true if relationships in the taxonomy change. The key idea is that diseases are not only human constructs created to approximate true conditions, but are ultimately a variable themselves.

We further uncovered evidence that diseases converge or diverge in the updating

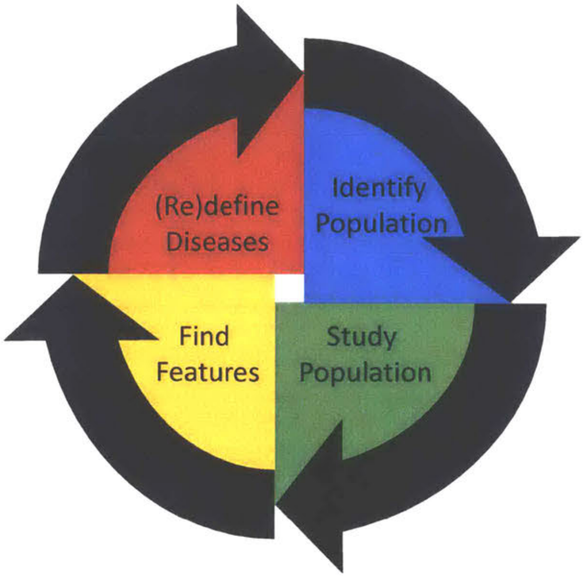


Figure 6-2: The cycle of the disease learning hypothesis. Once a disease is defined, it implies a given population, which can be sampled for learning. New features can be found in that population, such as certain symptoms, genetic markers or drugs, that may be effective. New learning from these features may lead to a redefinition of the disease or associated diseases.

process just described. If a disease is precisely defined, it can likely withstand the test of time. We established that many diseases change over time – some do not change very much, while some change so much that they eventually go away. This notion indicates a process of convergence or divergence. One reason we found that a disease may diverge is because there may be a way to break it into very different diseases. On the other hand, a converging disease may be one that can be identified and where new learning is consistent with its original definition. Over time, our understanding of the biological processes will be updated.

One way to describe the process by which diseases are described relative to other diseases, as done in a taxonomy, is as an iterative process, as shown in Figure 6-2. A disease is described by other diseases in the taxonomy, but once more diseases are added, the taxonomy has been changed, meaning that each disease may need to change position. This is similar to the cycle of learning diseases from populations. A population is defined by some features and is classified as having a specific disease. Then, patients diagnosed with the disease are studied to find associated features, which may change the study population and the disease definition.

This hypothesis would explain why some diseases appear to diverge or converge, as would be expected with an iterative process. For example, Dropsy is a disease that was widely diagnosed in the 1800s and removed around the middle the 20th century.[234] On the other hand, Cystic Fibrosis is a disease based on the observation of fibrosis and cysts in the lungs, and only later was the cause discovered to be chloride ion channels were not working properly.[235, 236] Current diseases that may diverge include Diabetes or Alzheimer disease.[237, 238] As we have shown, it appears that diseases which have been around for a longer time have survived the process and are more likely to remain. Understanding this cycle may help future researchers as they define and redefine diseases, both for medical practice and learning.

6.2.1 Limitations

The first limitation of our result is that we only had one taxonomy history available for testing. The MeSH taxonomy is the only taxonomy that has been consistently and regularly updated over the last 50 years, and we selected it for this reason. We did compare the 2015 version of MeSH to current versions of the ICD9 and SnoMed CT, finding high variances among the taxonomies. However, we did not have historical versions of ICD or SnoMed to perform our analysis.

The second limitation is that we were only able to examine changes due to the taxonomy. We would have liked to examine changes that resulted exclusively from moving continuous information to a discrete space by examining how disease-feature associations change over time. For example, when 'Dropsy' was removed from the disease taxonomy, were features of Dropsy re-allocated to new diseases? Such an analysis would require historical data on disease-feature association, which does not currently exist. We used one data set of cancer incidence to approximate this.

Finally, our work was setup to treat all data sets independently. However, it is possible that there was an overlap in information between data sets, such as genes and symptoms if certain gene mutations resulted in certain symptoms. If these two attributes have features that are coupled, they should not be treated independently. It may be possible to categorize variance based on attributes such as causes, effects or interventions. As has been pointed out in previous work, the disease taxonomy may focus more on symptoms and capture that information better.[9] We can expand the data sets used to evaluate variance, and then combine them for a more comprehensive measure of variance along multiple dimensions.

6.2.2 Future Work

One example where our work could have a significant impact is in drug testing and approval. A drug must be tested on a specific indication, and the starting point for such an indication is a disease from the given set. However, if the variance within that disease is high, trying to learn whether a treatment is effective on the disease may not be feasible. The disease itself should be treated as a variable and modified to result in a consistent treatment response. One way to further develop precision medicine would be to reduce variance in diseases by redefining diseases. To test the validity of this approach, it may be possible to look at trials that were successful versus trials that were not successful and compare the variance of diseases within those groups. We would expect unsuccessful trials to be more likely to have higher variance overall.

Estimating precision of diseases in the context of clinical trials could open the possibility of multiple tiers of evidence required for trials. Currently there is a standard measure of evidence required by the FDA and a somewhat different bar of evidence for orphan diseases.[239] This has lead to a steady increase in the number of drugs approved for orphan diseases.[240] Many of these diseases are more precisely defined because they have a specific genetic cause.[241] An alternative paradigm to this two-tiered approach would be to have the bar of evidence inversely proportional to the disease variance. Thus a precise orphan disease would require less evidence compared to a highly variable or less precise broadly-defined disease, but instead of only two options the bar of evidence would match the disease. This would encourage development of drugs not only for orphan diseases and diseases that correspond to a blockbuster, but everything in between.

6.3 Updating Disease Spaces

Our work in the previous chapter showed that a disease space can be defined in many ways. Formally it could defined as a disease taxonomy, but could also be defined by features in a data set connecting diseases to an attribute such as genes or symptoms. We demonstrated several ways for using one approach to augment the other. This provides us with a technique to make hypotheses about missing data, either in the taxonomy or for a given data set. The implication is that we can update the taxonomy automatically from data or update a sparse data set using the expert knowledge contained in the disease taxonomy. We gave examples of predictions for imputing values in the disease-symptom data set that appeared surprising, but had support in the literature.

The literature on taxonomies show they are useful for communication between humans, and provide a compact way to represent disease associations. However, for machine learning, these attributes are not necessary; it may be that a more complicated and dense representation reflects the disease space better. This work may provide the basis for moving away from using the taxonomy as the accepted disease space to other representations that can be used in machine learning as formal disease spaces. Our predictions would form the basis for using the current disease taxonomy as a starting point and automatically changing the structure to create a new disease space.

6.3.1 Limitations

The main limitation of this work is that we only used one taxonomy. There are several available taxonomies, but they do not overlap completely (i.e. terms that occur in one, but not another taxonomy). There is no straightforward way to combine all taxonomies and vocabularies, so we only used one. However, one taxonomy only captures one set of experts and perhaps one set of applications. It would also be useful to estimate confidence levels in predictions. In this work we demonstrated that algorithms could provide scores with a reasonable ordering of predictions, but it would also be important to know confidence to understand where a threshold should be set for predictions.

6.3.2 Future Work

Our work has three main areas of application. The first is updating the given taxonomy. For example, we could make predictions about taxonomy updates, provide them to the National Library of Medicine, and see how many of the proposed changes would be adopted. The advantages of systematically updating the taxonomy are that it can include a broad range of data that experts may not have mentally present and allow for continuous updating based on new data points as they are produced. An improved taxonomy would be less likely to change over time if it better reflects underlying conditions, so in the future, the rate of change of the taxonomy could be a proxy for how well previous taxonomies approximated disease.

The second area of application is merging data from different taxonomies. Many data sets are constructed using a controlled vocabulary from a taxonomy. For example, our symptoms data set was constructed using MeSH terms for diseases and symptoms. The MEDI-HPS drugs data set we used was constructed using ICD-9 codes. The only way to compare diseases from one data set to another is using a thesaurus, such as the United Medical Language System (UMLS). Many terms do not have an equivalent in another vocabulary and therefore cannot be used. The methods we developed would allow us to predict where a term might go in another taxonomy. This ability would allow us to merge data sets and taxonomies in a way not currently possible.

The third area of application is imputing values in sparse data sets based on the taxonomy. The disease taxonomy was constructed by experts and contains a wealth of information. This expert information could be used to predict what values might be missing in a sparse data set. In our work we gave the example of 322 symptoms for 4662 diseases resulting in a possible 1.5 million associations. It would not be feasible to manually assess each association to ensure the data set is complete. If a list of predictions were available, perhaps the top 10 symptoms for each disease, it might be possible to manually review them. Up to this point, taxonomies have typically been used to impute values only by inheritance, meaning that a higher level concept inherits all associations of lower level concepts. This approach may not properly account for patterns of diseases that occur in many different locations. It has been shown in previous machine learning literature that the taxonomy can be very useful for augmenting a data set and results in improve predictive performance.[174] Our method may be able to further improve performance.

6.4 Disease Representation

In the previous chapters, we found that the variance of diseases could be measured using features of diseases and that the taxonomy could be updated using a variety of external data sets. We next wanted to answer the question of how granular does data need to be to improve our understanding of diseases? For example, is patient-level data necessary? We developed a representation of diseases based on clinical drug trials that contains latent information about expert knowledge connecting diseases and drugs (we note that the data set of disease and drugs we created for this chapter is distinct from the MEDI-HPS data set of diseases and drugs that we used in the previous chapter).

Our representation produced results with surprising similarities to those obtained using a genetic database, even though our representation uses only trial meta-data and contains no explicit information about biology. We demonstrated that this new representation could be used for predicting new clinical trials. It is also possible that the latent information uncovered can also be used to reveal biological relationships that otherwise might only be found by using detailed biological data, gaining access to large amounts of clinical data or conducting resource-intensive research. At present, lessons learned from this cross-discipline endeavor are shared in part through publications, reviews and conference proceedings; collating this information for the entire body of clinical trials to derive lessons about human biology would be very time intensive. In contrast, our representation of clinical trial meta-data allows us to access cross-discipline information implicitly and derive conclusions and lessons learned.

For many retrospective learning tasks, data is fixed and therefore new experiments cannot be engineered to collect the data of most interest for the learning problem. For this reason, the information of interest is often latent. If the representation is engineered correctly, it can represent the desired information and allow that information to be extracted. There are many examples of data used for one purpose, where researchers later determined that latent information in the data could be used for another learning problem. For example, drug side effects are documented carefully to warn patients and inform clinicians what to watch out for. However, researchers later realized that the latent information was biochemical pathways.[83] With the assumption that the same pathway leads to the same observed effect, (whether a primary effect or side effect) the researchers set up a learning problem to find similar drugs based on the pathways they affect and the observed side-effects. We believe there is much more to be learned from latent information in the data that is currently available.

Currently, there is a push for more detailed data in health care, which due to privacy concerns is hard to obtain. Though we feel this is a productive development, it may result in efforts only concentrated on fighting for more access to data[242, 243] or only result in work completed with the detailed data currently available, typically using ICU data.[244] While excellent work is being undertaken in these areas, our results also show more could be done with the data available today. Not only will this help us learn more now, but it will also help future researchers design adaptive learning problems for when more detailed is made available.

Diseases are connected to other variables in many data sets and will become connected to patients once more patient-level data becomes available. There are still many pairwise relationships which can be learned and combined for further research. Humans are very good at learning, but they can only handle so much data and complexity, which are both aspects that computers can handle much better. For example, finding relationships between genes, biochemical pathways, mRNA, cellular processes, physiological processes, organs affected, external effects, structural changes and interventions will likely require the use of computers.

6.4.1 Limitations

There is other information in the clinical trial data besides biological information, such as economic potential of drugs, special interest in orphan diseases and prevalence of diseases in developed countries. If the goal of using the representation is learning new biology, then these other sources of information could bias results. For learning problems where this information is of interest, such as determining which diseases are least explored, different biases would need to be accounted for. In this work, we did not try to remove bias because defining what bias is depends on the learning problem itself.

Another limitation to our approach is that we treated all trials equally. It might be that later stage trials (e.g. stage-3 versus stage-1) or trials with more participants should be weighted more heavily. We also limited ourselves to trial header information, but there may be more information to be gleaned using trial stages, study design, available results and funding sources to augment our approach. We also propose exploring new filtering techniques for comparing diseases as an area that may need more exploration. Another recommendation to expand this work is to explore similarity metrics that define the disease-drug space and inference on graphs for prediction. Our work only touched on each of these areas.

6.4.2 Future Work

This work may have significance going forward in a few different areas. The first is generating hypotheses for future biomedical research. Our representation allowed us to make comparisons between diseases and find some surprising connections between diseases. We do not know if these connections are accurate, but there were enough accurate connections to believe they could be. It would be possible to select several overlapping diseases and design biology experiments to see if there is indeed biological overlap.

Another area of impact is using our disease representation as a scaffold to interpret other data sets. Our representation is entirely empirical and does not make any biological assumptions. Because it was constructed independent of any other data set, it can be used as a prior to compare other data sets with attributes of disease. For example, one could compare diseases based on genes associated with disease and infer further relations using our representation as the underlying space. Another example could be comparing patient level data in clinical trials. There is a current push to make more patient-level data available from clinical trials, but comparing patients from different trials is not straight-forward given the setup and assumptions behind different trials may be substantially different. Our representation would provide a yard stick for comparing patients from different trials. Such an effort would help bridge the gap between 'micro' patient-level research and 'macro' disease-level research.

An advantage that our representation has is it allows for continuous updates, making a real-time intelligent database of disease relations possible. As each new trial is added to clinicaltrials.gov, the representation can automatically update and one could track how our understanding of disease is updating with each new data point. This would provide a model for the future of continuous learning in medicine and could translate to patient-level data when it become more readily available.

In addition to the representation we created, we believe our approach could be applied to other areas as well. We believe there is more to be done with the superficial data that is currently available. One example that is very similar to our approach would be to use "off-label" prescription data to uncover aggregated learning implicitly taking place by physicians in clinical practice.

Finally, an area for further exploration and impact is the visualization and data cleaning methods from this work. Data cleaning in medicine is time-consuming and our results may help speed up further efforts. In a similar way, visualization is key to interpreting data and what we have provided in this thesis will allow for more exploration of the data by a broader group of researchers and clinicians.

6.5 Precision Medicine

The work done in this thesis ties in to the movement of precision medicine. Precision medicine has gained momentum and there is much hope it can help progress medicine rapidly in the coming decade.[22, 21, 20] Ironically, precision medicine is rarely defined precisely. In this thesis, we define precision medicine as being able to precisely map a patient to a particular treatment, which currently happens through diseases. A disease-centric approach focuses on diseases and what they map to. On the other hand, a patient-centered approach focuses on the treatment patients receive. There is an underlying assumption in the disease-centric approach that learning diseases is the best way to treat people. In the patient-centric approach, diseases are simply variables that can be manipulated and are convenient for communication.

One explanation for why patient-centric precision medicine is so elusive is that diseases often get stuck in local optima. In optimization theory, there are strict conditions under which a global optimum can be achieved. Nonlinear spaces are almost never convex, meaning that one can only find local minima.[245, 246] It appears that the way to break out of a disease local optima is for evidence to accrue that meets some threshold. As there is generally inertia or tradition that keeps diseases in a status quo, the evidence has to be compelling. In nonlinear optimization, one way to deal with significant nonlinearity is through simulated annealing or stochastic gradient descent, which make random jumps that allow one to get out of a local optima.[247, 248] Larger jumps may be needed to break away from local optima when refining the disease space. Another way is to run the algorithm again with a different random starting point or make significant changes, such as pruning high up in a decision tree.[249, 250] Such an overhaul may be challenging in medicine, but we would recommend a regular overhaul of the disease space, include the disease taxonomy. It may be difficult to come to an agreement, so automated methods like those we developed for updating the taxonomy with expert tweaking could be a preferred approach.

We also concluded that quantifying variance may be one of the keys to precision medicine. The primary aims are to identify, quantify and then reduce variance to achieve precision. These factors have been identified as a cause for success in a number of other industries as well.[251] Attention to measuring variance in medicine has been lacking, however, and this thesis should draw attention to this.

Chapter 7

Conclusions and Contributions

7.1 Conclusions

We conclude with the premise we started with, that there is an opportunity for faster learning in medicine. In this thesis we focused on learning diseases as a space of discrete concepts that are related to each other. Using computational learning theory as a guide, we characterize learning the disease space as an unsupervised learning problem. We conclude that the best way to measure precision of disease learning is by estimating variance within diseases. This is challenging due to the continuous nature of underlying health conditions and the discrete nature of disease, but we developed a method for estimating variance based on the historical relationships between diseases, using the disease taxonomy as a model disease space. We found that the variance within diseases was often significant (greater than 25%) even on short time scales (10 years). We conclude that disease variance should be considered when drawing conclusions from other types of learning problems. We also see the potential for reducing variance to achieve precision medicine.

We further showed how to reduce variance between external data sets that associate diseases with attributes and the disease taxonomy. We conclude that expert knowledge encoded in the disease taxonomy can be used to hypothesize data points that are sparse in data sets. Similarly, information in external data sets can be used to the disease taxonomy or any disease space. Taxonomies have been and continue to be useful for many aspects of medicine, particularly as a tool for communication, but may not be the best tool for representing the disease space using computational learning. One area of future research is how to codify learning about diseases in a way that can be used for officially comparing diseases without having to rely on a taxonomy.

Finally, we point out that a movement in the field of machine learning in medicine is a push to make more detailed, patient-level data available. We support these efforts, but also are interested in limits of what can be learned with data currently available. In this thesis we show that it is possible to get similar results from clinical trial meta-data compared to results from detailed genetic data. This results is surprising and encouraging. We conclude that highly granular or patient-level data is not necessary to achieve significant learning results. Such work with the data available now will prepare the way for more detailed patient-level data when it becomes available.

7.2 Main Contributions

The three main contributions of this thesis were the following:

1. We estimated the variance quantitatively in the accepted disease taxonomy. The disease taxonomy is used to define diseases and the a significant challenge

has been how to measure the uncertainty or variance that the taxonomy introduces. Our estimate enables researchers to put an error bar on learning results and use computational tools to learn an improved disease taxonomy.

2. We developed methods to reduce the variance between the disease taxonomy and external disease data sets. This reduction allows for a more precise update of the set of diseases based on available data, as well as for the discovery of new disease-feature associations.

3. We developed a new representation of diseases using clinical trial meta-data that contains latent information, similar to that found in a drug-gene networks. This result demonstrates that a surprising amount of learning is possible with superficial data sets. It further implies that diseases can be updated more precisely without having to obtain detailed patient-level data.

7.3 Additional Contributions

In order to address the central premise undergirding this thesis we found it necessary to explore multiple data sources and strategies, which is a central part of all data science endeavors. The data science required to explore the central premise of this thesis yielded numerous additional contributions in the form of visualization tools, methods and data sets that can become the basis for future work. We enumerate the most salient contributions emerging from our data science efforts below:

1. The disease-drug data set constructed from meta-data on clinicaltrials.gov. This data set required significant automated processing and manual review to ensure clean data. We published this data set as part of the supplemental

materials in [144]. The thesaurus we compiled to augment MeSH and our data cleaning methods may also have application for cleaning other data sets.

2. The history of the Medical Subject Headings (MeSH) disease taxonomy from 1971-2015, including the taxonomy for each year as well as how concepts in one year correspond to concepts in the next year. To get digital files required extracting digital data from manually scanned PDF files. This data set also required significant automated processing and manual review to ensure clean data. At the time of writing this thesis we are preparing to publish this data set along with a journal article.

3. A suite of web-based visualization tools for exploring the history of the MeSH taxonomy and how it relates to associated variables such as genes, drugs and symptoms. These tools include tool to track diseases in the taxonomy over time, a tool to compare disease categories in different years (see Figure 3-6) and a tool to see where disease features, such as a specific symptom, occur in the taxonomy in a given year. These tools make it easy to explore the data for researchers and clinicians alike. We will make these tools available online.

Appendix A

Background - Supplemental Materials

A.1 Terminology

Terminology in medicine can be ambiguous, which may lead to confusion in communication. We define important terms here, as we use them in this thesis, that may have different definitions in other contexts.

- Continuous - Unbroken or uninterrupted. An infinite set of possibilities over a range exist.
- Discrete - Individually distinct and separable. A finite set of possibilities exist.
- Condition - The state of the human body characterized by all biological processes occurring at a given time. Conditions lie on a continuum because the possible conditions are near-infinite.

- Disease - A discrete, agreed-upon approximation of part of the spectrum of conditions.
- Disease State - A discrete subcategory of a disease. It could be a more specific classification than what is generally accepted as a disease or a disease at a particular point in a progression.
- Disease Attribute - A class of characteristics associated with many diseases. Characteristics within an attribute are either disjoint or are subsets of each other. Examples include genes and symptoms.
- Disease Feature - A single characteristic associated with a disease. This could be a specific gene, symptom or process. Typically, these are discrete biological concepts.

A.2 Computational Learning

We start by giving an overview of computational learning for those unfamiliar with the topic. We also expound on terminology as a reference for the rest of this chapter.

Learning can be an ambiguous term. One dictionary definition is "the acquisition of knowledge or skills through experience, study or being taught."[252] Experiences or studies provide new information or data from which new knowledge can be gleaned. For the past several decades, researchers have studied how computers can be used to acquire new knowledge from data with promising results.[50] We refer to this field as computational learning, though aspects of it can also be described as machine learning, artificial intelligence or statistical learning.[253]

Computers require explicit instruction and we therefore need a more precise definition for learning than one used to describe human learning. We adopt a definition by Vapnik of the learning problem as "a problem of finding a desired dependence using a limited number of observations" [254], where the dependence can be described by a mathematical function. Broadly speaking, there are two types of learning: supervised and unsupervised. In computational learning theory, these modes of learning have important similarities and differences. Each are described in the context of computational learning below.

A.2.1 Supervised Learning

Vapnik defined supervised learning precisely as "that of choosing from the given set of functions the one which best approximates the supervisor's response." [254] Supervised learning assumes there is some fixed, but unknown function that acts on variables.[254] Though the true function is unknown, we have input and output data and can infer what function may approximate the true function. For example, one could learn the relationship between the velocity of falling items (output) and height (input) to model gravity (true function). The force of gravity is constant on the surface of the earth, so many samples could be used to learn a function to approximate gravity. If the true function is not consistent over the data points, an approximate function could not be accurately learned. For example, if measurements were taken from the surface of the Earth, the moon and Jupiter, a single constant for the force of gravity could not be learned.

Two main goals of supervised learning are: (1) to understand the relationship between inputs and outputs explicitly, as in the case of modeling gravity, and (2) to make predictions of output values based on inputs that are not in the data set used for learning. However, the second goal does not require the first to be evident. A complicated function may be approximated that does not reflect the true function, but is still useful for making predictions.

A.2.2 Unsupervised Learning

Unsupervised learning likewise assumes that while there is a fixed function generating the data, there is no output data. The goal is therefore to find the relationship among the data points. The relationships can be described comprehensively as a probabilistic density, or less complex models such as clustering can be used. As an example, we might be given a group of animals and be tasked with organizing them by similar characteristics. We could create a joint distribution of the characteristics shared by all animals, but that would be a very complicated model (N^{n_c}), where N is the number of animals and n_c is the total number of characteristics. Such a density would be hard for people to work with. A simpler model would be to cluster animals by similar characteristics into groups. Unsupervised learning can be more subjective in nature because there is no absolute standard for what these groupings should be. This is generally a free parameter in the learning algorithm that is pre-defined or trained according to some pre-defined criteria.

Like supervised learning, there are two main goals of unsupervised learning. One is to understand which data points are clustered together. The other is to identify which points are close to a new data point. There is no explicit output to predict or label to identify, but there is often an assumption that similar data points share similar features. Identifying points that are near each other can help predict other features.

Unsupervised learning is related to Hebbian learning,[255] a hypothesis proposed in the 1950's that is often summarized as "neurons that fire together are wired together."[256] This hypothesis suggests that unsupervised learning is similar to natural neurological patterns used to organize information in the brain.

A.2.3 Other Learning

Additional types of learning have been proposed which are variations of the above two. For example, hybrid supervised/unsupervised learning is where only a subset of data points are labeled.[257] Reinforcement learning starts with a reward function, and through unsupervised exploration, data points receive a label according to the reward function.[258] Reinforcement learning is particularly useful in robotics.[259] Active learning is where some data points are labeled and the goal is to find the next data point that provides the most information for learning a supervised relationships.[260]

A.2.4 Empirical Risk Minimization Principle

In all of the above cases, there is the concept of what Vapnik terms the "risk functional"[261] The risk functional represents the amount of error or how different the outputs of the model or function are from a ground truth or pre-defined function. The risk is quantified by a loss function over a set of functions.[14] Minimizing the risk makes the model as close to the truth as possible. This principle makes the learning problem well suited for computers because methods have been developed for optimizing (minimizing) functions with many variables. Algorithms have been developed to efficiently solve optimization problems. Due to the complicated nature of the optimization problems, there is no guarantee that the solution found can be considered a global optimum, but local optima have been useful in practice. There are several design considerations when setting up a computational learning problem that the user must define. We describe several of these in the following subsections.

Space of Functions

A necessary constraint for computational learning problems is to define a set of functions to search over. As there are an infinite number of types of functions, it would not be possible to search exhaustively in a finite time period. A set of functions and parameter ranges must be pre-defined to constrain the problem to one that can be solved in a reasonable amount of time. A straightforward example is linear regression where the set of functions is only one ($y = \beta x + \alpha$) and the search is designed to find the best values for α and β, which could be any real numbers.[262] In this case, an analytic solution is available for the optimal solution. In many cases, specific algorithms are developed to solve an optimization problem for a given function or set of functions.

The power of computational learning is that many functions can be quickly evaluated. Traditional hypothesis-based approaches are forward-looking, proposing a single hypothesis and using statistics to refute or confirm the hypothesis.[263] Machine learning can rapidly sort through a large number of functions to find the best fit to the data. The set of functions to search over is the choice of the user who is responsible for setting up the learning problem.

With a loss function and set of functions to search over, it is possible to minimize the risk. However, there is one more design consideration to ensure that results extend beyond the data set used for learning, which is known as generalization.[137]

Rule-based functions: a history We pause here to give a small piece of history in computational learning regarding functions to search over. One of the most natural functions for humans to consider is rules. A rule is a binary function, with one output if a condition is met and another output if it is not met. Rules are generally easy to interpret and are familiar in many contexts. Early efforts in computational learning focused on rule-based learning.[264] One example is machine translation where early efforts focused on inducing rules for how a word or group of words would translate to a similar word or words in another language. Progress hit a wall, which was broken through by enabling a more probabilistic approach. This expanded the set of functions to search over. One proposed explanation is that rules are considered the most nonlinear function, and a smoother set of functions is better.[265]

Generalization

Identifying possible functions for a single data set is a relatively simple task, as an arbitrarily complicated function can fit the data. For the learned relationship to be useful, it must hold for any data set that exhibits that relationship. The challenge of any learning task is doing so with a limited data set. However, because learning requires a limited number of examples to identify a mapping, a strategy must be developed regarding the most efficient use of the data to ensure generalization and to include prior knowledge about the learning problem if needed. The first is known as validation and the second as regularization.[266]

The method to ensure generalization is a choice for the user. Validation involves dividing the data set into multiple sets and performing training and validating on different data subsets.[261] A validation data set is one that the model does not see for training. The data subsets can be fixed in advance, or many disjoint data subsets can be generated using cross-validation.[266] Validation does not use external information; rather, it only uses the data set more carefully. Regularization uses external information to help make a decision about the function selected. An example could be Occam's Razor, which says that a simpler explanation is more likely the correct explanation, so a penalty would be placed on functions that are more complex.

The opposite of generalization is overfitting. Overfitting refers to fitting a model to a data set that is specific only to that data set.[267] It is easy to find an arbitrarily complex function that will yield very little loss for a given data set, but much of what has been learned would be specific to that data set. In general, we do not want to learn a function that fits all facets of the data set, but rather some particular aspect that is of interest. For example, with respect to learning gravity, the data is affected by every other force, such as wind resistance, as well as measurement error. If we developed a model of gravity on data using feathers, it would not generalize to other objects. It is also worth noting that to learn a function, the data must contain information about that function. For example, it would not be possible to learn about the gravitational force of the object itself because its affect would be too small to be found in the data. Even if the data allowed us to draw some conclusions, we know it would not generalize.

For unsupervised learning, there is no absolute truth, so a complexity parameter is often pre-defined. For example, in clustering, one may set the number of clusters, as in k-means, or set the number of neighbors to compare to, as in nearest neighbors.[268] These parameters would be based on prior domain knowledge.

A.2.5 Data Space

Up to this point, we have assumed that the data given to the user for setting up a learning problem is already matched to the problem. For some problems, the data measured can be used directly for the learning problem, as in the gravity example. In many cases though, features must be extracted from the data to create a set of numbers to represent a data point. For example, text cannot be directly used by a computer – there must be a way to represent text or compare it quantitatively. This is both a challenge and an opportunity for users.

There is inherent bias in machine learning which is manifested in how the user chooses to features that represent concepts and the hypothesis space to explore.[269] It has been shown that if machine learning "tasks are known to possess a common internal representation or preprocessing then the number of examples required per task for good generalization" goes down.[270] The "Physical Grounding Hypothesis," which states that an intelligent system should have its representations grounded in the physical world as opposed to derived symbolic representations, suggests that representations try to make use of measured data when possible as opposed to data derived from measurements.[271] According to this hypothesis, a good representation for computational learning will minimize the number of biases imposed.

The user therefore needs to engineer a data space by defining features to represent each data point and develop a way to compare similarities or distances between data points. It is also desirable to avoid bias and inferences in constructing the space.

Continuous vs Discrete

Feature values (either input or output data) can be continuous or discrete. When the output data is discrete, the task is classification, and when the output data is continuous, the task is regression.[137] The task does not depend on the nature of the input values. Computers can only work with discrete values, but can discretize continuous values with a high level of precision.

In all cases, each data point must consist of a discrete set of values. It is not possible to represent all values along a continuous dimension. For example, a physical picture must be discretized in space. Data that is best represented on a continuous spectrum or across many dimensions must be converted into a discrete set of values.

A.3 Analogies for Variance in Learning

It is not immediately obvious why estimating variance or uncertainty is important for analyzing the disease taxonomy. Two reasons why this is the case is diseases are often studied one at a time narrowing our focus when learning and because the taxonomy changes so slowly many may view it as constant. Yet, measuring uncertainty can revolutionize a field as Shannon did for communication.[150]

To aid in framing the problem, we propose two analogies that may help clarify the objective and approach of this thesis. These analogies frame the problem in more familiar domains and show where the benefits of the approach can be seen more clearly.

A.3.1 Organizing Toys

In explaining the thesis to my 6-year old daughter, I found that our family uses the same principles when organizing toys.

With three kids we have a variety of toys, which sometimes end up strewn across our apartment. We have several containers to store the toys, where our kids can put toys after they have finished playing with them. It is our desire to organize the toys so we can find specific toys when someone wants one. Sometimes, we decide to change our organization system.

We could have one big bucket with all toys, but that would not really help us find specific toys. We could also buy a bucket for every toy, but that would be equally ineffective. Generally, our approach is to group toys together and put them together in a bucket. This is typically based on some subjective intuition or general set of rules. We then identify attributes for toys in a given container, such as color, size, age group, electronic component and so on. We then teach the rules to our kids so they know where any given toy should go. Some rules we have are that animals go in one basket, things that make music go in another, small toys that could be a choking hazard go in a bucket up high and so forth.

At some point, we decide that our rules for what goes where need to be updated. We also find we may need to redefine what goes into a bucket. We therefore have two tasks that depend on each other: (1) how do we divide up toys into groups, and (2) how do we identify which toys go where? The two tasks may appear very similar, but are actually distinct. The rules allow us to communicate to our kids what goes where and to generalize this to new toys. As an example of the former, we had an animal bucket but also had a lot of other similarly sized toys that often got played with alongside the animals. We changed the bucket to include "creatures," which consisted of animals, dinosaurs, toy people, snowmen figures and similar toys.

Why it is useful

This process is surprisingly similar to the process of learning about diseases. There are many people and it would not be helpful to have one disease to categorize all people under. It would be similarly ineffective to say every individual has a different disease. In either case, we could not learn from other cases.

In medicine, there are also two related learning tasks, one that is clustering (i.e. putting biological features in a bucket) and the other is determining how to map data points to diseases. The data points mapped could represent patients or other concepts like treatments. Mapping rules allow clinicians to place new patients into buckets.

The analogy is particularly useful because there is no absolute truth to organizing diseases. In medicine, there is similarly no absolute truth to disease organizations, but we could cluster diseases by organ systems affected, symptoms or genes. There is no way to determine which one is more accurate.

Where it breaks down

The analogy is dissimilar from medicine in several ways. One is that in the toy example, clustering and mapping are often done very close together. This is because features are generally readily apparent. In medicine, the process is often much slower. A disease may be defined initially by symptoms, and over time, other diagnostic criteria may be used to map patients to a disease. Moreover, with new features, the disease itself gets updated. This could take decades.

The analogy may also seem unrelated because the primary purpose of organizing toys is to find them later. We generally think that the primary purpose of diseases is to provide treatments, but they too are used for quickly finding examples of patients for future learning.

A.3.2 Exploring North America

This analogy is to help make the concept of measuring precision more concrete.

Exploration of diseases is not dissimilar to physical exploration of the earth. The goal of explorers is to map out some physical area for future use, such as charting routes. One particularly famous case is the exploration of what eventually became the United States. Prior to 1492, there were two entire continents virtually unknown. In that year, Columbus discovered that there was something there, but by no means had he mapped out the continents. Over the following centuries, explorers chartered what the surface of the land looked like.

This process resulted in numerous data points that cartographers could examine to create a model of physical space. These models are called maps. Many maps were made over the years with varying degrees of accuracy. We may assume that geography has not changed dramatically over the last several centuries, so geographical features like coastlines, rivers, mountains and so forth should be in the same location they were then. We can therefore compare current maps to earlier versions and see whether any inaccuracies exist.

It is easy to see where cartographers were not accurate. If we were to go back in time though, there would be no absolute reference. It would not be possible to look at a set of maps and determine accuracy, but it would be possible to determine precision. This could be done by tracking variance among maps over time. For features that have high variance, we may assume that those features are not precisely identified.

Why it is useful

Diseases are similar in that we are currently exploring what diseases are. Biology does not change significantly over a short time scale, so we may attribute variance to the exploration process. If we go back in time, we would find that there is no absolute truth in this exploration analogy. It may be useful to have a measure of precision about physical features though, just like case of diseases.

Where it breaks down

The analogy is clearly different in that the physical world has a ground truth. This makes it easy to see inaccuracies in our example since the world has now been

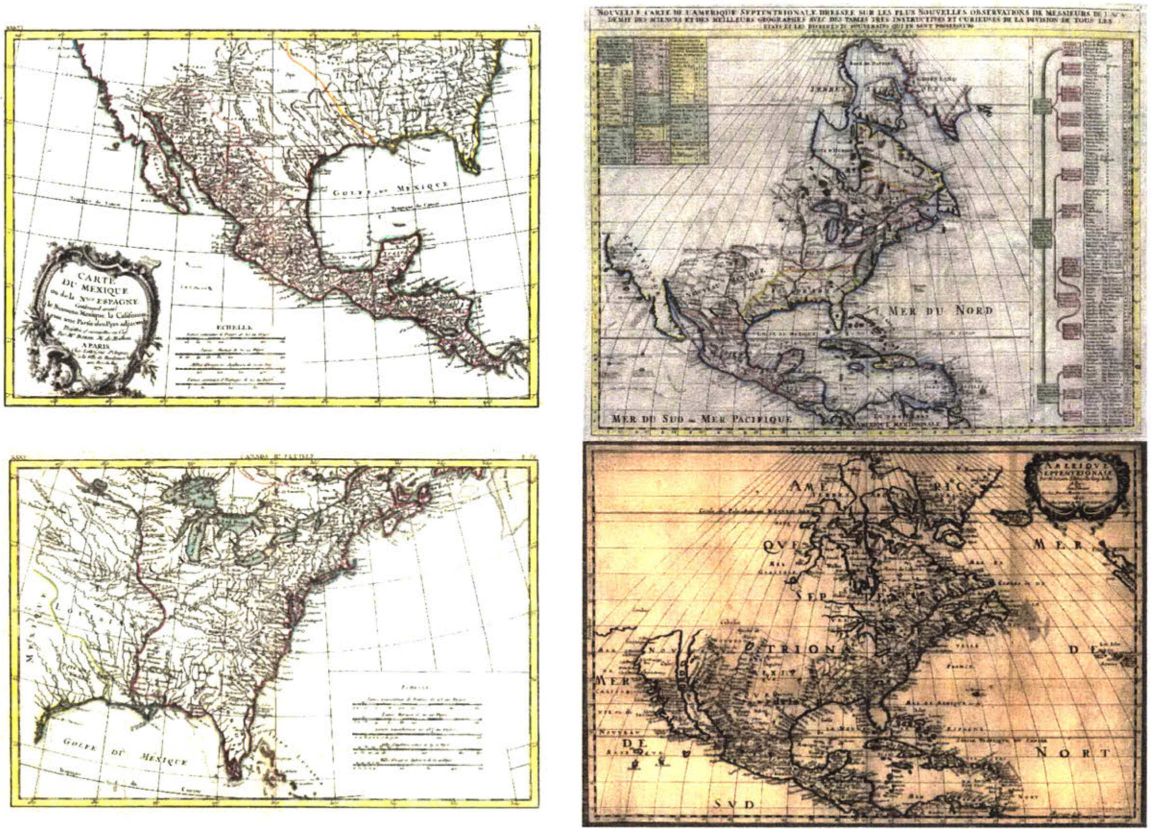


Figure A-1: Old maps of the America continent. There are inaccuracies that are apparent compared to current maps. From these maps imprecision can also be seen.

mapped accurately in detail, but that is not the case for current disease understanding. However, it is easy to look back and see where we got some things wrong about disease. The physical world is also very constrained in dimensions; typically maps are two or three dimensions – perhaps a few more if information about soil or water is given. Exploring diseases involves many more dimensions. The exploration process is therefore much more sparse.

Appendix B

Estimating Disease Taxonomy Variance - Supplemental Materials

B.1 Taxonomy statistics prediction

We plotted several taxonomy statistics over time and fit a line to the data using linear regression. We used these linear approximations to predict what the statistics might be 50 years into the future.

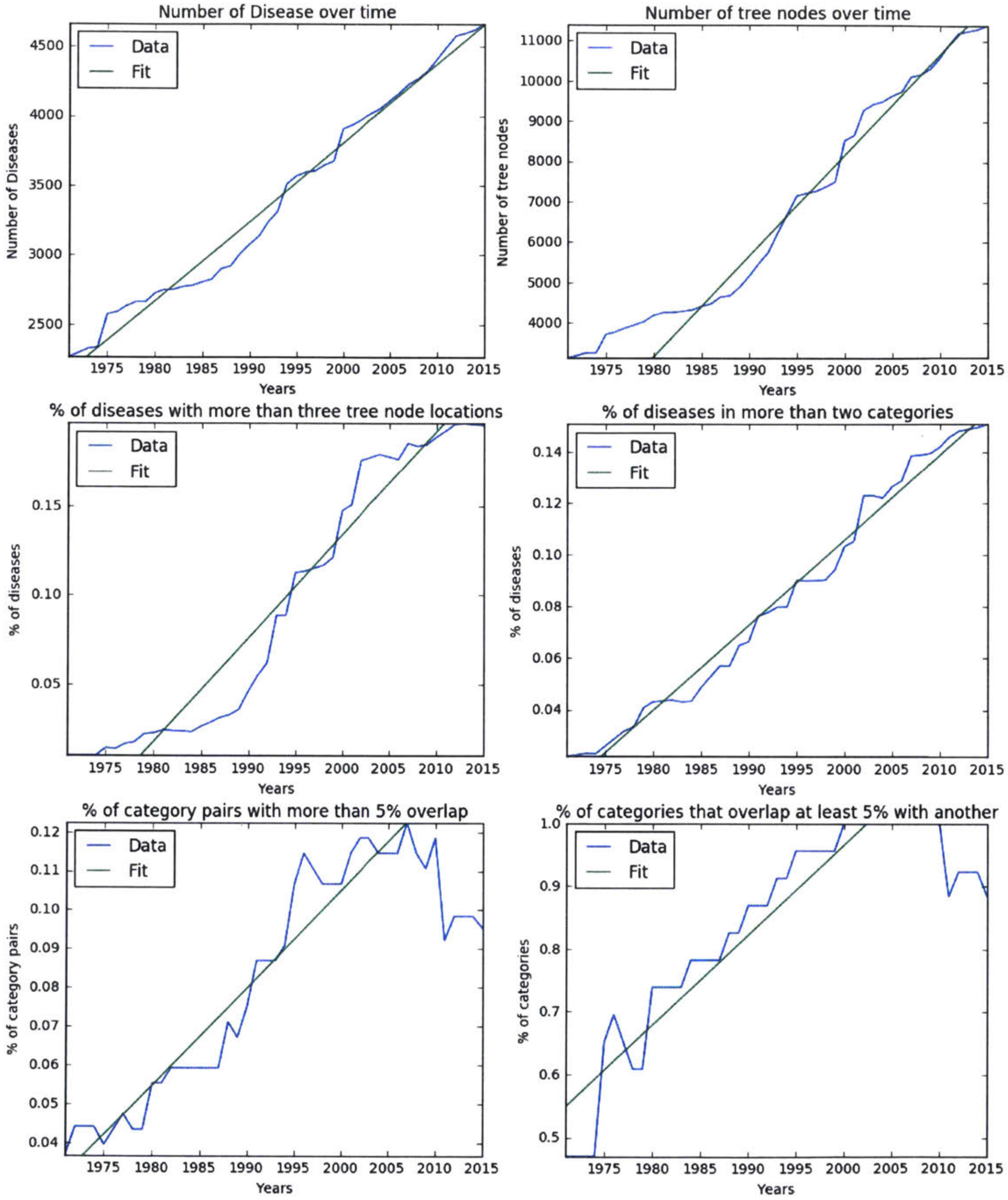


Figure B-1: Plots of data and linear regression used to make future predictions about taxonomy statistics.

B.2 Diseases with the most tree nodes

To understand why some diseases do not appear to fit neatly into the taxonomy, we explored the diseases that occur in many different locations. In Figure B-2 we plot the average number of locations for diseases over time. In the top left plot is the number of nodes per disease for four sets. The four sets are based on the number of tree node locations in 1971. If diseases were known to be complex then and the number of nodes simply increased over time we would expect to see the set of diseases with most locations over time result in the set with the most locations in the future. Instead we see only a small increase over time. In the top left plot the sets of diseases are fixed by the number of tree node locations in 2015. We arrive at the same conclusion. It therefore appears that the number of locations in the past cannot be used to predict which diseases would have the largest number of locations in the future.

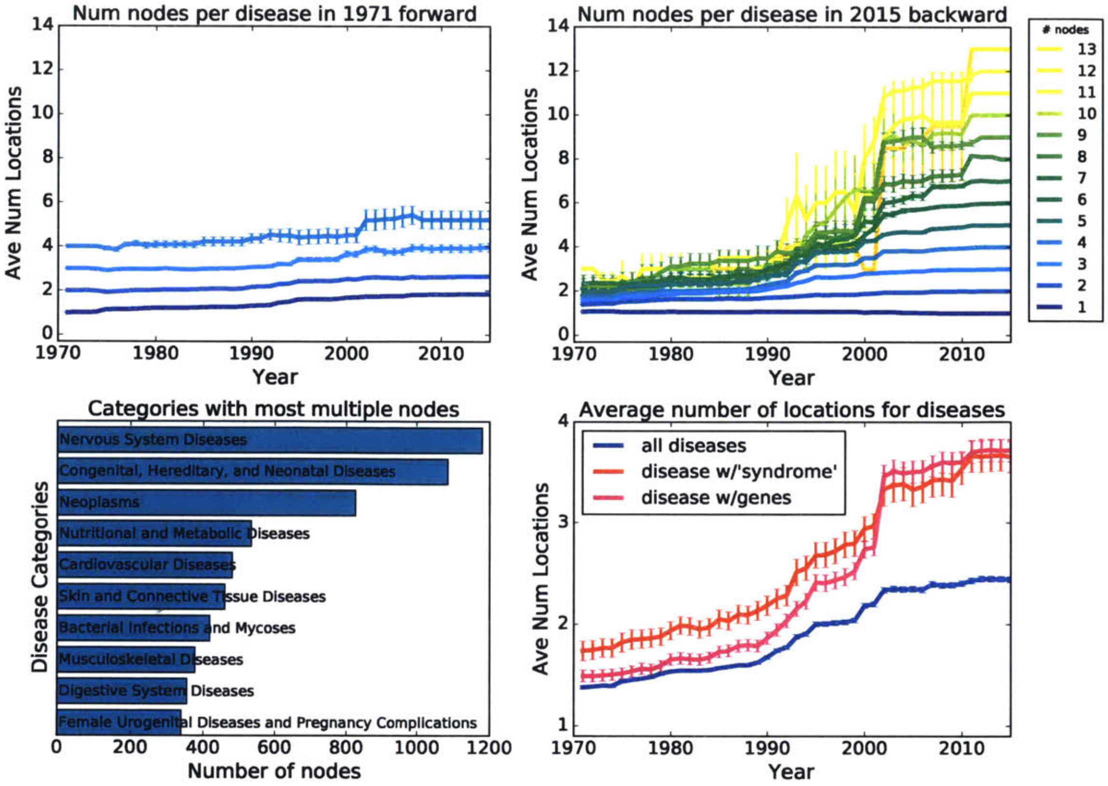


Figure B-2: Number of nodes plotted over time. The top two plots show that number of nodes in the past could not be used to predict which diseases would have the largest number of nodes in the future. The bottom left plot shows which categories have the most diseases with multiple nodes. The bottom right shows that syndromes and genetic diseases have significantly more locations that the average.

B.3 Taxonomy operations over time

To approximate the amount of time between taxonomy operations, we plotted the inter-operation times.Approximate chance of an event is 5%. The distributions approximately match suggesting that for any given disease at any give time point, the chance of a taxonomy change is 5%. From this estimate we may assume that a given disease may change once every 20 years or for a given year, one in every 20 diseases will change. This rate of change is more frequent that we would have anticipated.

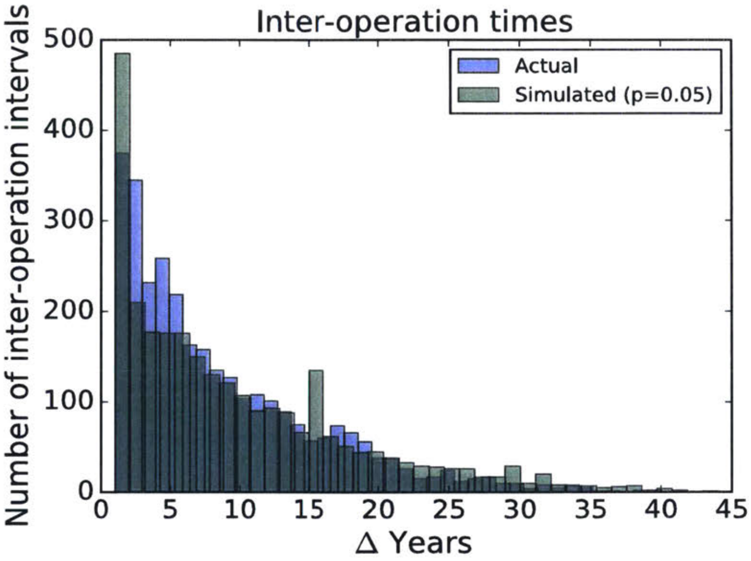


Figure B-3: Inter-operation times, actual and simulation for a probability of 0.05.

B.4 Tree Edit Distance

Tree edit distance is one way to estimate how similar/dissimilar two trees are.[272] The basic edits or operations are node deletions and node adds and perhaps a node relocation. The distance is a minimization over the number of edits required to to change one tree to the other and unordered trees, like the disease taxonomy, are computationally much more expensive than ordered trees. [273, 274] Trees where the children of a node have an inherit ordering have efficient algorithms to calculate distance.[275] Our trees are so large that the optimization problem is could not be calculated and there is no intuitive meaning to the distance metric.

The reason we did not use it is that the taxonomy trees are updated in a deliberate way that does not necessarily match a least number of edits framework. Also, it is assumed that taxonomy changes are made on a disease-by-disease basis or by group of diseases, but on a node-by-node basis as tree edit distance assumes. For these reasons we chose to document the disease changes and not use tree edit distance.

B.5 Taxonomy operations

On the top left we see the total number of operations, which is dominated by location changes, then new diseases added and very few diseases deleted. There are a couple years where the whole taxonomy was overhauled, including 1975 and 2000, where spikes in activity can be found.

The other three plots we show what happens to diseases when the disease name is removed (top right), when new diseases are added (bottom left), and when changes occur where the disease name does not change (bottom right). When a disease name is removed, it is most likely because the name of the disease has changed and not because the disease was deleted. There were very few cases where the disease went away and was replaced by multiple diseases, representing a split in the disease. When new diseases were added, most were new leaves, but in 1996 and 2000 there was a significant number of new branches added. This may suggest that in general new diseases are leaves or very specific diseases, but sometimes the taxonomy is overhauled to change the branch structure. Finally, when a disease stays and there is a change, most of the time the change is a subdivide (new diseases are added below) or a location added or removed.

If diseases simply subdivided, we would expect to see few location changes with many splits or subdivides. Instead we see new diseases throughout the tree and many location changes, both deletions and insertions.

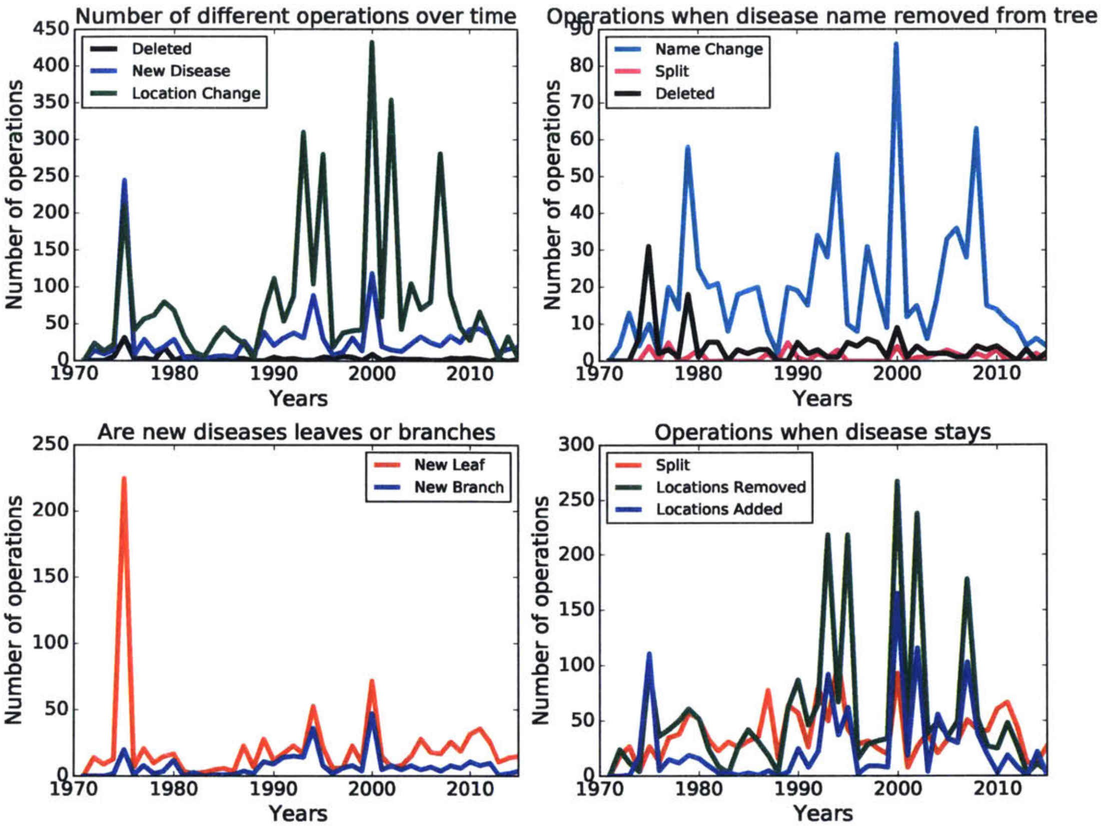
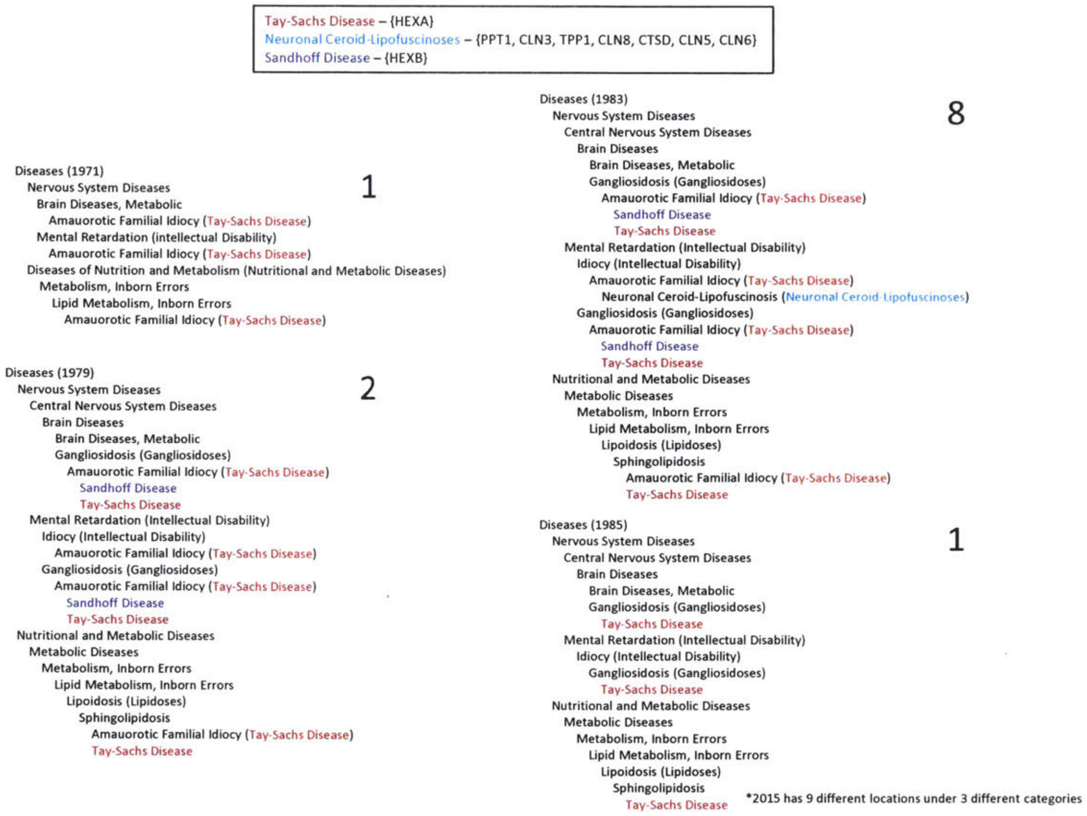


Figure B-4: Taxonomy operations over time. On the top it is seen that most taxonomy operations are location changes compared to new diseases and very few diseases are actually deleted. On the top left it can be seen that when a disease name is removed from the tree, it is most likely from a name change and diseases being split into multiple diseases is uncommon. On the bottom left it can be seen that most new diseases are leaves, but in some years there is a spike in new branch diseases or disease categories suggesting an overhaul to the taxonomy structure. On the bottom left we see that when a disease changes in the taxonomy without changing the disease name, it is approximately equally split between a subdivision, a node deletion or a node addition.

B.6 Example of inferred disease-feature associations changing: Tay Sachs


Tay-Sachs Disease – {HEXA}
Neuronal Ceroid-Lipofuscinoses – {PPT1, CLN3, TPP1, CLN8, CTSD, CLN5, CLN6}
Sandhoff Disease – {HEXB}
Diseases (1971)
Nervous System Diseases
Brain Diseases, Metabolic
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Mental Retardation (intellectual Disability)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Diseases of Nutrition and Metabolism (Nutritional and Metabolic Diseases)
Metabolism, Inborn Errors
Lipid Metabolism, Inborn Errors
Amauorotic Familial Idiocy (Tay-Sachs Disease)
1
Diseases (1979)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Brain Diseases, Metabolic
Gangliosidosis (Gangliosidoses)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Sandhoff Disease
Tay-Sachs Disease
Mental Retardation (Intellectual Disability)
Idiocy (Intellectual Disability)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Gangliosidosis (Gangliosidoses)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Sandhoff Disease
Tay-Sachs Disease
Nutritional and Metabolic Diseases
Metabolic Diseases
Metabolism, Inborn Errors
Lipid Metabolism, Inborn Errors
Lipoidosis (Lipidoses)
Sphingolipidosis
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Tay-Sachs Disease
2
Diseases (1983)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Brain Diseases, Metabolic
Gangliosidosis (Gangliosidoses)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Sandhoff Disease
Tay-Sachs Disease
Mental Retardation (Intellectual Disability)
Idiocy (Intellectual Disability)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Neuronal Ceroid-Lipofuscinosis (Neuronal Ceroid-Lipofuscinoses)
Gangliosidosis (Gangliosidoses)
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Sandhoff Disease
Tay-Sachs Disease
Nutritional and Metabolic Diseases
Metabolic Diseases
Metabolism, Inborn Errors
Lipid Metabolism, Inborn Errors
Lipoidosis (Lipidoses)
Sphingolipidosis
Amauorotic Familial Idiocy (Tay-Sachs Disease)
Tay-Sachs Disease
8
Diseases (1985)
Nervous System Diseases
Central Nervous System Diseases
Brain Diseases
Brain Diseases, Metabolic
Gangliosidosis (Gangliosidoses)
Tay-Sachs Disease
Mental Retardation (Intellectual Disability)
Idiocy (Intellectual Disability)
Gangliosidosis (Gangliosidoses)
Tay-Sachs Disease
Nutritional and Metabolic Diseases
Metabolic Diseases
Metabolism, Inborn Errors
Lipid Metabolism, Inborn Errors
Lipoidosis (Lipidoses)
Sphingolipidosis
Tay-Sachs Disease
1

*2015 has 9 different locations under 3 different categories

Figure B-5: Example of disease (Alzheimer Disease) that changed the number of genes associated with it because of changes to the taxonomy structure even though the actual disease-gene associations are constant as seen at the top. The number of genes associated with Tay-Sachs Disease would vary between 1-8 genes depending on the taxonomy.

B.7 Error Bar complete

Figure B-6 shows the complete 2-dimensional distribution over years whereas Figures 3-8 and 3-9 only showed 5 cases of the 1-dimensional (for 1, 5, 10, 20 and 44 years).

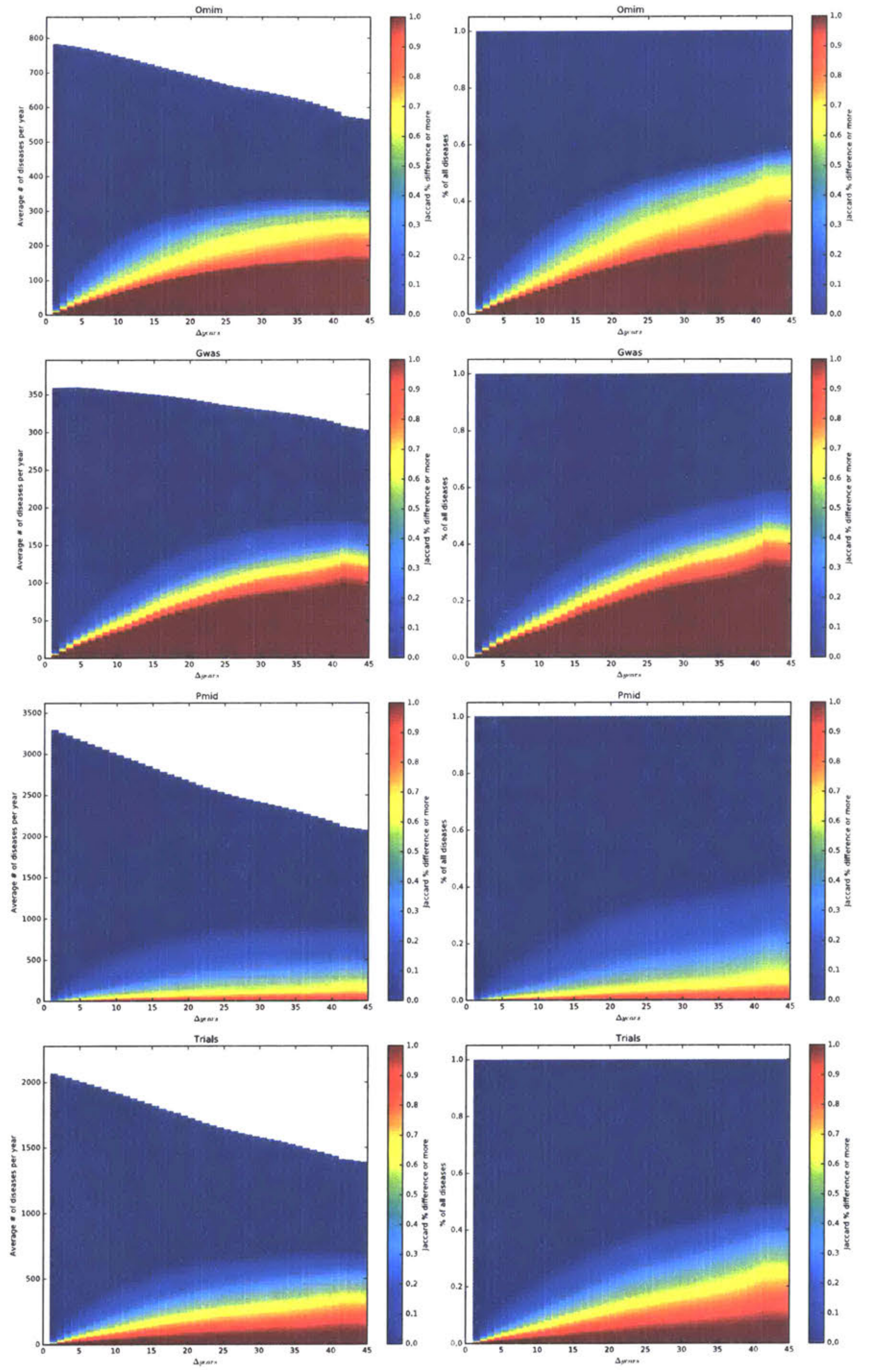


Figure B-6: Error bar imparted by inferred relations.

B.8 Examples of diseases less like themselves

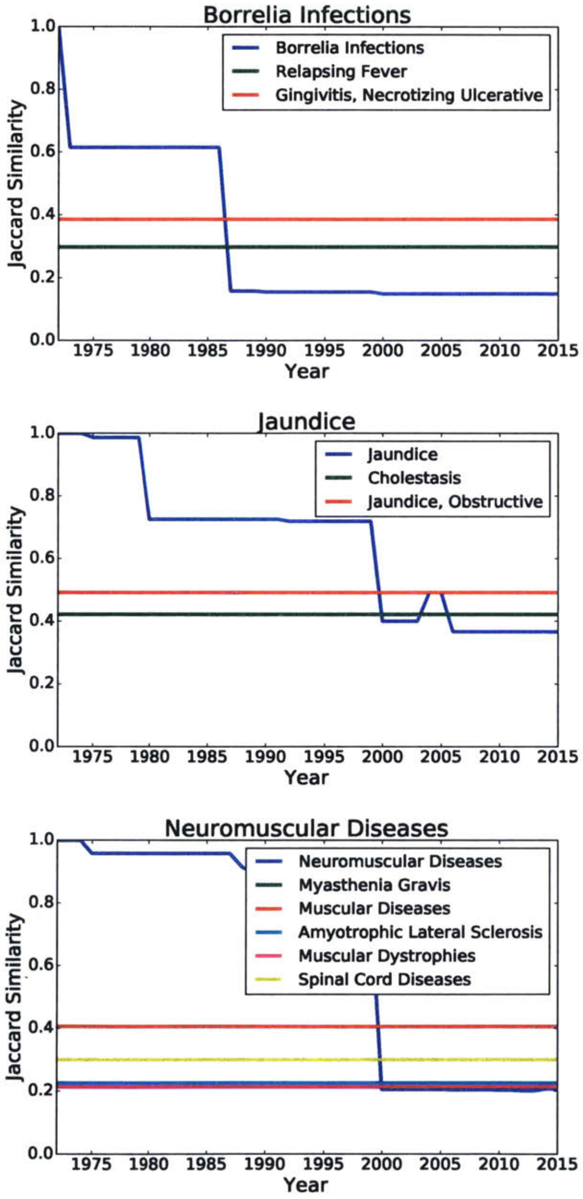


Figure B-7: Examples of a disease and other diseases that are less like themselves than another over time.

B.9 Systematic Reviews

One area the inferred disease-feature relations could have a significant impact is for systematic reviews. All systematic reviews start with a structured literature search. The majority of systematic reviews use the MEDLINE search which is an advanced version of a standard PubMed search. MeSH terms are‘exploded’ to search for everything under a term. The explode operation is a function of time since the taxonomy changes over time. It is therefore possible that a taxonomy change could result in a different set of papers/trials to start looking for studies to include.

We examined 20 different systematic reviews in detail to see if a taxonomy changes would result in a different starting set of studies that could affect the conclusion of the systematic review. We found that the study search strategies use expanded strategies beyond exploding MeSH terms. As an example, one systematic review on barbiturates for traumatic brain injury [276] used a search term to explode ‘Barbiturates’ and ‘Craniocereberal Trauma.’ The systematic review was published in 2000 with 8 studies included. The change in taxonomy between 1999 and 2000 would have resulted in one study not being include, except that the systematic review used additional search terms including: “pentobarb* or phenobarb* or methohexital* or thiamyl* or thiopental* or amobarb* or mephobarb* or barbital* or hexobarb* or murexide* or primidone* or secobarb* or thiobarb*” and “((injur* or trauma* or lesion* or damage* or wound* or destruction* oedema* or edema* or fracture* or contusion* or concus* or commotion* or pressur*) and (head or crani* or capitis or brain* or forebrain* or skull* or hemisphere or intracran* or orbit* or cerebr*))” It appears that the experts conducting the systematic review were aware of limitations that MeSH has and knew terms that could be used specifically for this systematic review overcome those limitations.

B.10 Changes in taxonomy vs literature for disease-symptom associations

$$sim(d_i, d_j) = f(\bar{s}_i, \bar{s}_j)$$

where $\bar{s}_i$ is a vector of all symptoms used to describe disease i. It is a binary vector with a 1 in cases where there is an association between disease i and a given symptom and 0 elsewhere. We could change the year in which the symptoms were selected from or a year up to which PubMed papers are examined. This new similarity takes the form

$$sim(d_i, d_j, y_s, y_p) = f(\bar{s}_i, \bar{s}_j)$$

where y_s is the year used to get the symptoms and y_p is the year up to which papers are used.

In Figure B-8 we plot $sim(d_i, d_j, y_s, 2011) = f(\bar{s}_i, \bar{s}_j)$ in solid lines with y_s on the x-axis and $sim(d_i, d_j, 2011, y_p) = f(\bar{s}_i, \bar{s}_j)$ in dashed lines with y_p on the x-axis. For two examples the change in similarity is approximately the same.

We also plot the room mean squared error (RMSE) of all similarities between the year 2011 and all other years according to the following two equations:

$$\sqrt{\sum_i \sum_j [sim(d_i, d_j, y_s, 2011) - sim(d_i, d_j, 2011, 2011)]^2} \tag{B.1}$$

$$\sqrt{\sum_i \sum_j [sim(d_i, d_j, 2011, y_p) - sim(d_i, d_j, 2011, 2011)]^2} \tag{B.2}$$

The blue line in Figure B-9 corresponds to Equation B.1 and the green line corresponds to Equation B.2.

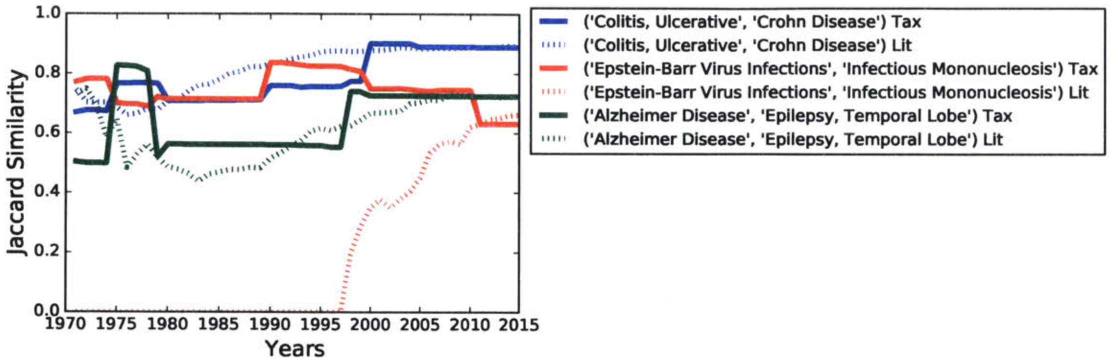


Figure B-8: Examples of three disease pairs whose similarity changes over time. The solid lines represent similarity changes due to the changing taxonomy while dotted lines represent similarity changes due to changes in the literature. For these examples the change in similarity is approximately with one exception.

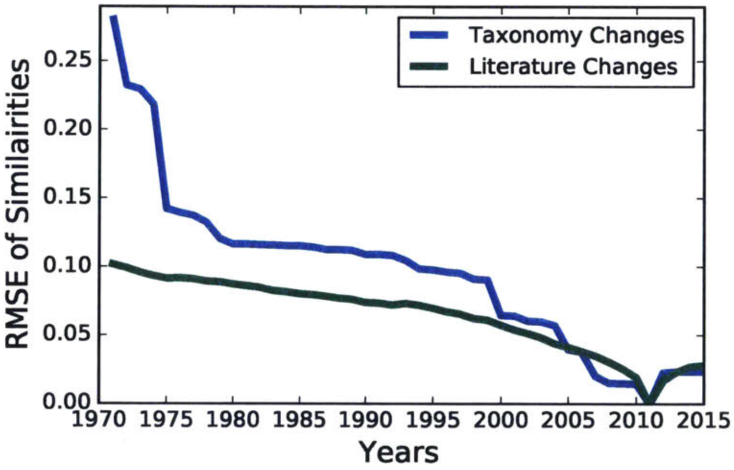


Figure B-9: The root mean squared error (RMSE) of the difference in similarity between the year 2011 and other years. The blue line represents the change due to taxonomy changes and the green line represents changes due to changes in the literature.

In this case, the taxonomy changes were only the symptoms. One section of the disease tree contains symptoms. For disease-symptoms relations we used diseases in 2011 and varied the symptoms by taxonomy from different years.

B.11 Examples of Symptoms in Taxonomy

We provide two examples of where symptoms show up in the taxonomy according to the diseases they are associated with. Pain is a general symptom that is associated with many diseases throughout the taxonomy and therefore provides little information about diseases it is associated with. Transient Global Amnesia on the other hand is much more specific and shows up with a limited set of diseases in a much more limited set of the taxonomy.

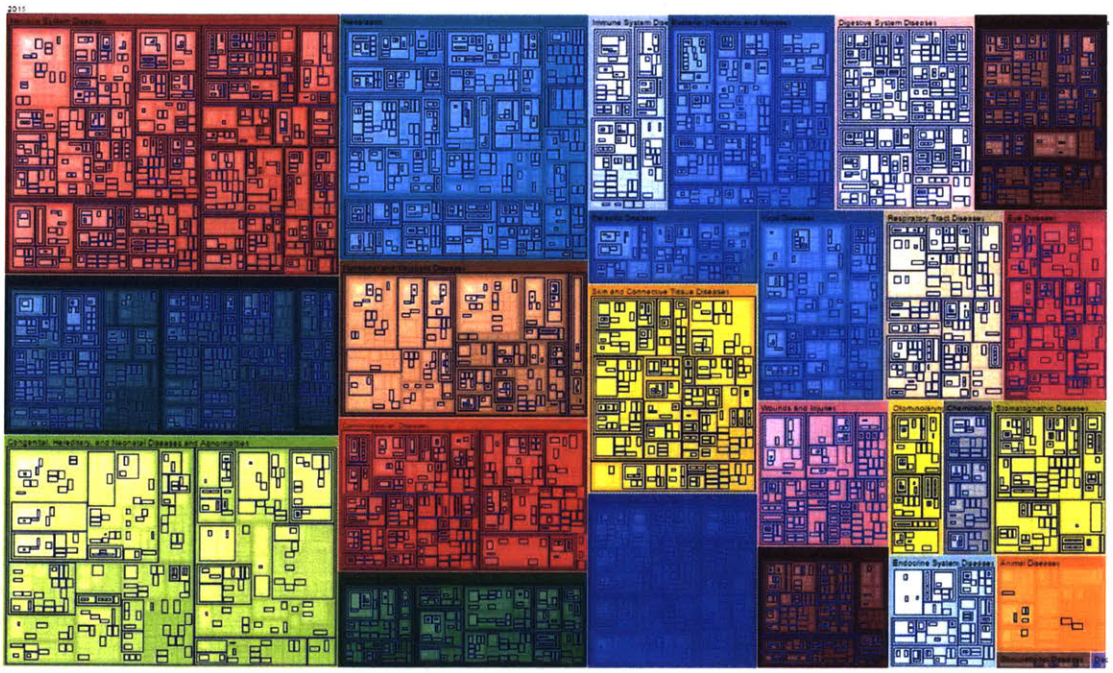


Figure B-10: Diseases associated with the symptoms 'Pain' are highlighted in blue in this tree map. This symptom is dense with little information content.

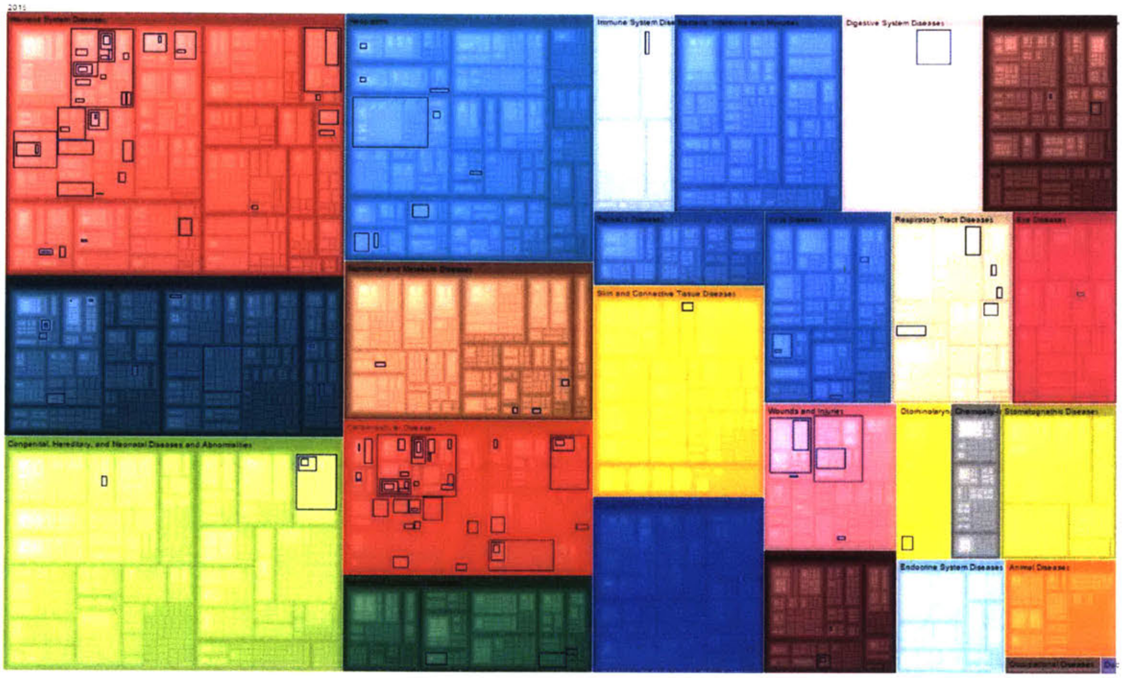


Figure B-11: Diseases associated with the symptoms 'Transient Global Amnesia' are highlighted in blue in this tree map. This symptom is more sparse with a higher information content.

Appendix C

Updating the Disease Space from Data - Supplementary Materials

C.0.1 Example of Drug Variance in Taxonomy

We give one example of variance that results in apparent inconsistencies in the disease taxonomy. The data set we use is drugs used in clinical trials. To achieve precision medicine, disease definitions and classifications must map to treatments. Drug therapies represent a large subset of possible disease treatments. Drugs used in clinical trials have previously been used to show disease relationships. Here we use all drugs tested in clinical trials to explore variance in the disease taxonomy. Each drug can be mapped to one or more disease and each disease can be mapped to one or more locations in the disease taxonomy. We use drugs from clinical trials for this example only because it is a much larger data set than approved drugs.

We first find drugs with high variance in the taxonomy. We describe a high variance drug as one that has been tested on diseases that show up in many locations in the disease taxonomy or in ones that show up in many unique locations. To find such drugs we use the metric of average shortest path length between diseases in the taxonomy that correspond to a drug. A short path length suggests clustering of diseases while a longer path length suggests diseases are distributed throughout the taxonomy. To restrict the scope of the paths, we only examine the k-nearest-neighbors paths.

We next find drugs in parent/child nodes. If the child nodes partition the parent node we would expect attributes of the parent disease to be divided among the child diseases. If child diseases overlap with each other mechanistically it would be expected that the same drugs would be used to treat them. We find the parent/child relationships with the largest discrepancies of drugs used. We expect to find areas of the taxonomy that do not reflect well the treatments for those diseases. We then show which areas of the taxonomy would be most changed by incorporating the data from drug trials in a modified taxonomy.

Figure C-1 shows the 2015 taxonomy with diseases colored on which the drug "Cyclophosphamide" was tested in clinical trials. It can be seen that this drug was tested on many different diseases that are found throughout the taxonomy. Other drugs are only tested on a few diseases which may be clustered together or may be dispersed throughout the taxonomy. There are two interpretations to the dispersion. One is that the drug affects many different disease mechanisms or that the disease definitions and classifications are not precisely defined.

Figure C-2 shows specific parent/child disease relationships in the MeSH taxonomy along with the number of drugs tested in clinical trials on that disease. The pie charts for each disease show the number of overlapping drugs with sister nodes, or with child nodes in the case of the one parent node. If diseases were partitioned nicely we might expect to find most drugs for a disease were specific to that disease

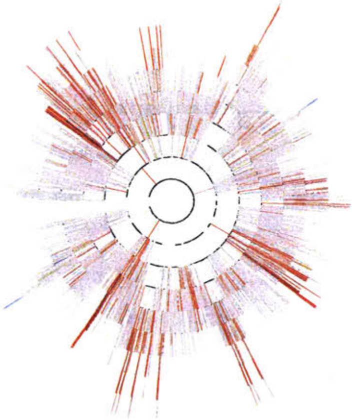


Figure C-1: The drug "Cyclophosphamide used in clinical trials mapped to the diseases it is tested on in clinical trials. Those diseases are colored and dispersed throughout the taxonomy.

or child node. At the same time, if we see that a disease shares few drugs with other diseases it may indicate they are not related in a therapeutic sense. The lack of overlap of some disease and the extent of overlap for others suggest the taxonomy may not capture the therapeutically relevant relationship in parts of the taxonomy. This suggests the opportunity to modify the MeSH disease taxonomy according to the clinical trial drug data.

We provide several additional examples of drug variance among diseases very close together. We selected several parent diseases that were distinct from each other and exhibit a variety of patterns of overlap. The color scale corresponding to overlap shown in Figure C-2 is the same for all other plots.

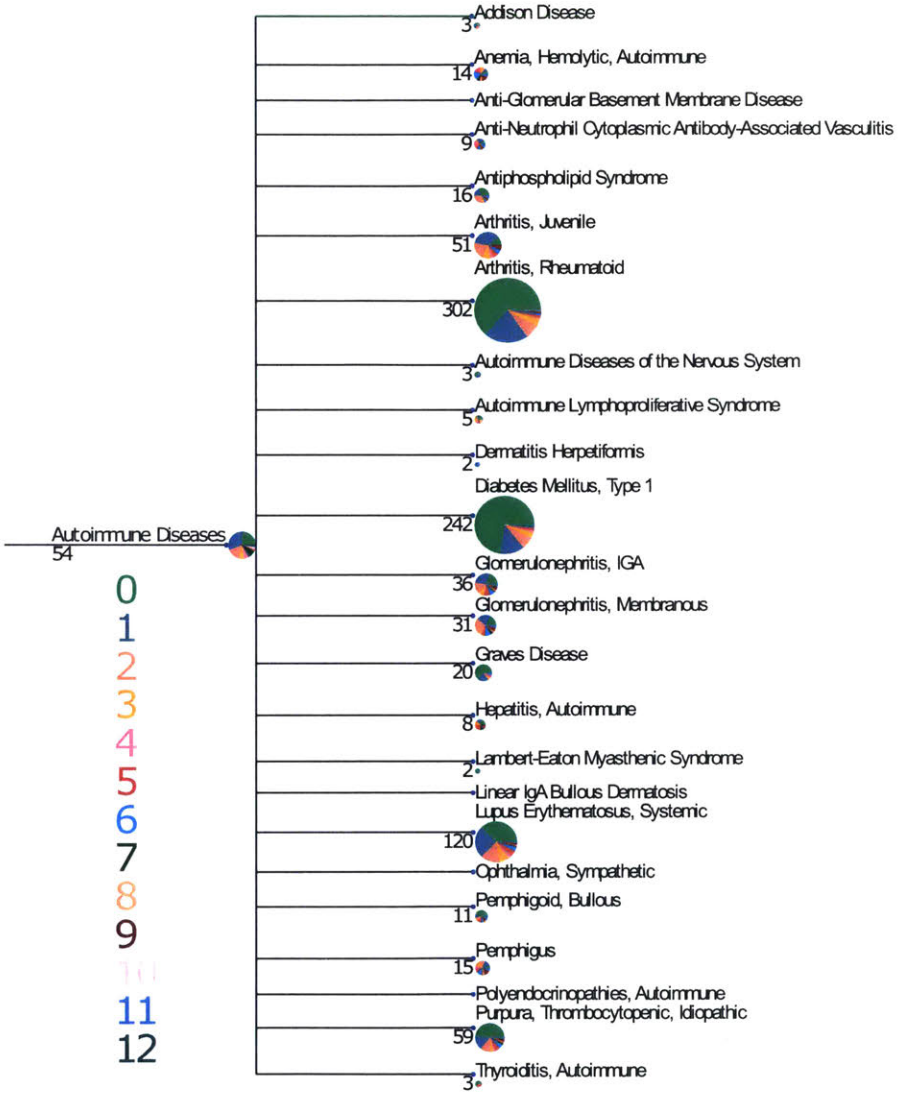


Figure C-2: Overlap of drugs tested on diseases in close proximity in the disease taxonomy. The parent disease is Autoimmune Diseases.

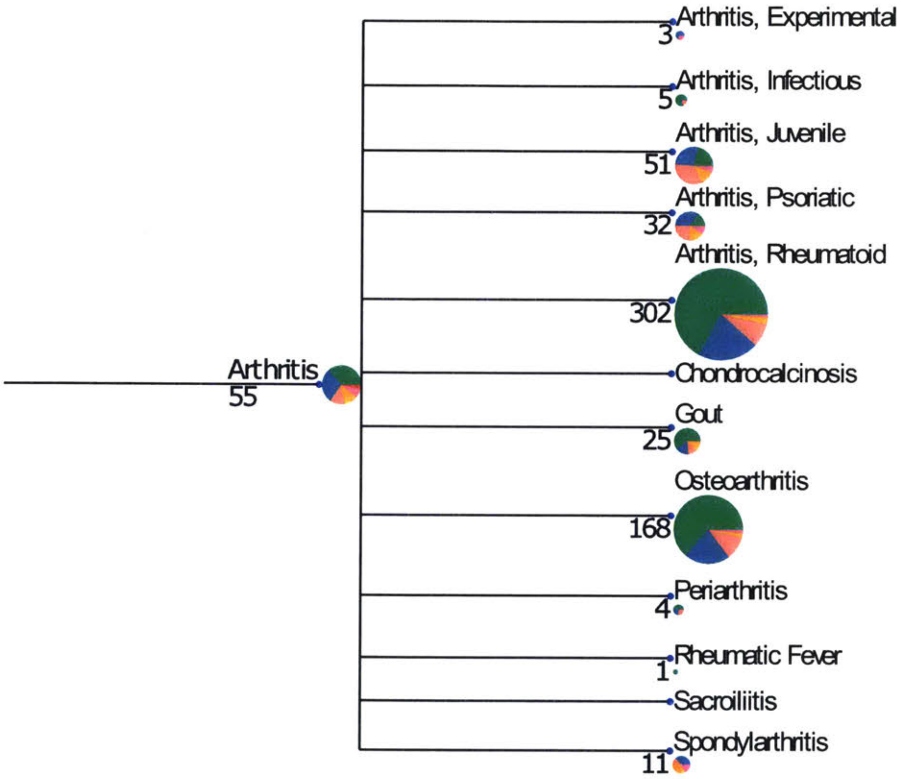


Figure C-3: The parent disease is Arthritis.

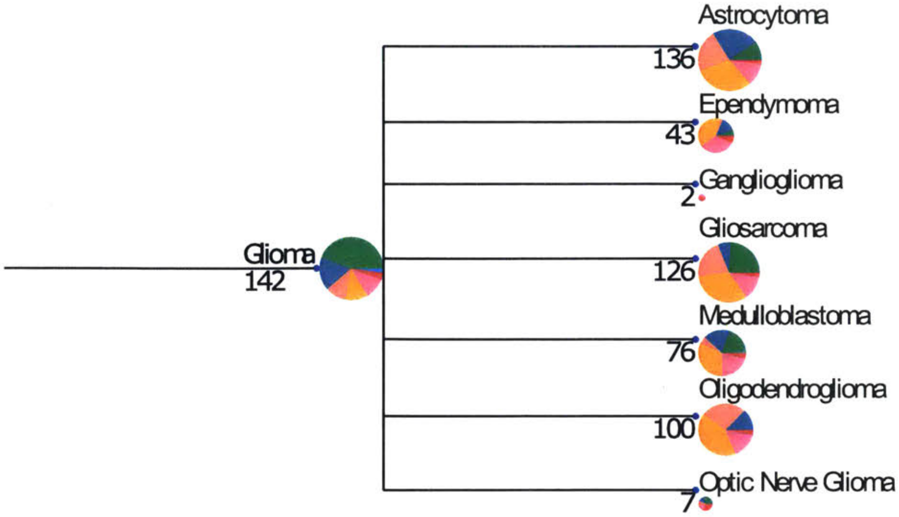


Figure C-4: The parent disease is Glioma.

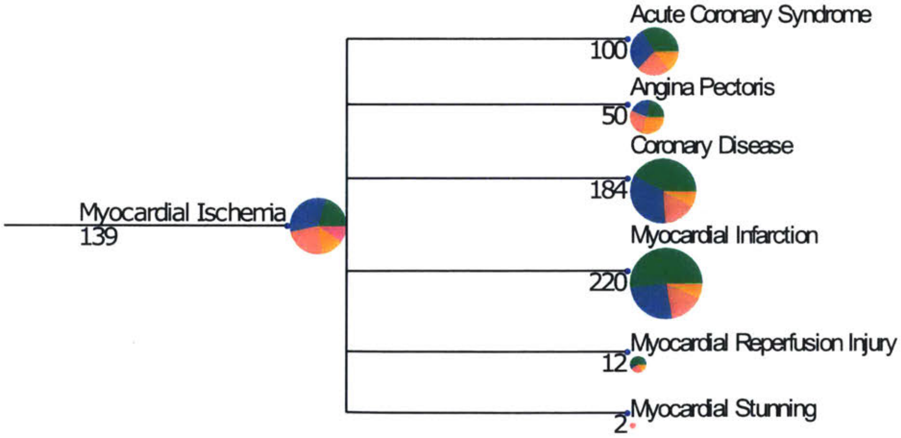


Figure C-5: The parent disease is Myocardial Ischemia.

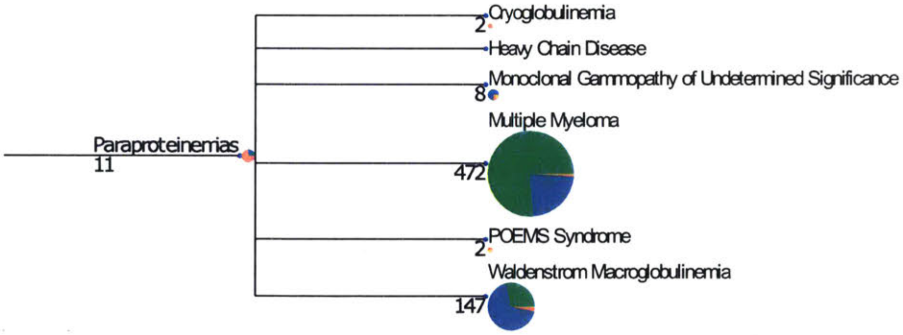


Figure C-6: The parent disease is Paraproteinemias.

Appendix D

Disease-Drug Representation from Clinical Trials - Supplemental Materials

D.1 Clinical Trials Database

We examined the largest clinical trial registries including CclinicalTtrials.gov, the EU Clinical Trials Registry, ISRCTN, the Japan Primary Registries Network and the Australian New Zealand Clinical Trials Registry;[277] as well as the GlaxoSmithKline company registry and the International Clinical Trial Registry Platform. We chose to use CclinicalTtrials.gov because it is the largest and most consistent source of trials. We explored the option of combining CclinicalTtrials.gov with other registries such as the EU Clinical Trials Register or the International Clinical Trials Registry Platform, but found it difficult to match trials from one database to another. We also found in a small sample that CclinicalTtrials.gov had the vast majority of trials found in the other registries. We used the March 27, 2014 release of the Clinical Trials Transformation Initiative (CTTI) Database for Aggregate Analysis of ClinicalTrials.gov (AACT available on www.ctti-clinicaltrials.org), comprised of 163,764 trials.

D.2 Disease Disambiguation

Many diseases are named in multiple ways, but our analysis requires each concept to be identified by the same string. An example of one disease with many different strings found for it on clinicaltrials.gov is seen in Table D.1. This section describes how we mapped on all terms to a common vocabulary of diseases.

Of the 163,764 available trials, 93,654 have at least one intervention labelled "Drug," "Dietary Supplement," or "Biologic." For those 93,654 trials there are 155,816 diseases listed, but only 23,162 of them are unique strings ignoring case. Of those, 5,192 are in the MeSH vocabulary under the Diseases[C] or Pyschiatry and Psychology[F] subtrees. To make use of the remaining 18,420 strings we augmented the MeSH thesaurus with new terms and synonyms. The steps to create the thesaurus are shown in step 4 of Fig. 1.

The first step (4a) was to remove all exact matches to the MeSH vocabulary as previously mentioned. The second step (4b) was to compare each disease string to every disease term in MeSH (approximately 75,000) using fuzzy string matching. For fuzzy string matching we considered Levenshtein distance[278] and Ratcliff/Obershelp pattern matching[279] for similarity metrics used to compare strings. We selected a modified version of the R/O algorithm because it looks for the longest matching substrings and can therefore account for word order changes. We identified the closest 20 fuzzy matches and manually selected one or more if it was the same

1. t2dm	38. non insulin dependent diabetes
2. non-insulin-dependent diabetes mellitus	39. adolescent type 2 diabetes
3. patient with type 2 diabetes treated with insulin using a baseline/bolus strategy	40. diabetes mellitus non-insulin-dependent
4. diabetes mellitus type 2 (t2dm),	41. insulin-requiring type 2 diabetes mellitus
5. diabetes mellitus type 2	42. foot dryness in patients with niddm
6. type 2 diabetes on medication	43. type-2 diabetes mellitus
7. diabetes mellitus type 2 irc or nir	44. impaired glucose or type 2 diabetes
8. type 2 diabetes	45. type2 diabetes
9. diabetes mellitus, type 2 and metformin	46. patients with type 2 diabetes
10. type 2 diabetes (t2d)	47. diabetes mellitus, adult-onset
11. type 2 diabetes melitus	48. patients with type 2 diabetes mellitus who have been examined at a medical institution
12. diabetes type 2	49. newly diagnosed type 2 diabetes (during the last 12 months)
13. type 2 diabetes mellitus without insulin treatment	50. diabetes mellitus, type 2
14. insulin dependent diabetes mellitus (type ii diabetes)	51. type 2 diabetes (treated with exenatide or other oral antidiabetic therapies)
15. diabetes, type 2	52. type 2 diabetes treated with insulin
16. type 2 diabetes mellitus(t2dm)	53. gad ab positive clinically type 2 diabetic patients
17. type 2 diabets mellitus	54. type-ii diabetes mellitus
18. type 2-diabetes	55. diabetes mellitus, adult onset
19. diabetes mellitus type ii non insulin dependent	56. type 2 diabetes mellitus, non insulin dependent.
20. subjects with type 2 diabetes mellitus.	57. type ii diabetes
21. diabetes mellitus type 2;	58. diabetes mellitus type ii
22. diabetes mellitus type-2	59. diabetes, type ii
23. type 2 diabetes mellitus related endothelial dysfunction	60. niddm
24. diabetes mellitus ii	61. type 2 diabetes mellitus (t2dm)
25. type 2 diabetic patients with ihd	62. newly diagnosed type 2 diabetes
26. type 2 diabetic patients	63. foot transepidermal water loss in patients in niddm
27. type two diabetes mellitus	64. diabetes mellitus non insulin dependent oral agent therapy
28. diabetes mellitus, non-insulin dependant	65. type ii diabetes in the not so obese
29. diabetes mellitus, non insulin dependent	66. non-insulin dependent diabetes mellitus
30. diabetes mellitus, non-insulin dependent	67. type 2 diabetes with nephropathy
31. diabetes mellitus type 2 not well controlled	68. diabetes mellitus - type 2
32. diabetes mellitus, non-insulin-dependent	69. type ii diabetes mellitus
33. diabetes mellitis type 2	70. adult type diabetes mellitus
34. type 2 diabetes mellitus (t2d)	71. diabetes mellitus, type ii
35. type-2 diabetes	72. type 2 diabetes mellitus
36. type-2-diabetes mellitus	73. type 2 diabetes patients'
37. antihyperglycemic effect in type 2 diabetic patients with secondary failure to oral hypoglycemic agents	

Table D.1: List of strings in data corresponding to Type 2 Diabetes Mellitus. Each of the 73 is a different way the free text in the disease field on ClinicalTrials.gov represents the same concept Type 2 Diabetes Mellitus.

concept as the disease string.

The next step (4c) was to create new terms to include in our vocabulary based on their occurrence in the data set. Two strings were merged if they represented the same concept. If terms that were merged together still appeared in 2 or fewer trials or the trials had 2 or fewer different diseases total, then the term was discarded in the next step (4d). Our motivation is to show relationships between diseases and drugs, so if one is only connected to the other once or twice it provides very little insight. We found that diseases occurred infrequently either because they were too specific (e.g. - "cancer with transdermal accessible tumour," "muscle sensitivity to pressure," or "ny-eso-1-expressing tumors") or too general (e.g. - "spinal disorder," "intestinal inflammation," or "chemical injuries").

The remaining disease strings were set aside in the next step (4e) because they were not diseases. These strings fell into one of the following categories: drugs (e.g. - "pharmacokinetics of asp015k and midazolam," "beta blocker," "desmopressin"), procedures (e.g. - "mitral valve surgery," "surgery of the pancreatic head," "sedated for cardiac catheterization"), measurements (e.g. - "detection rate," "blood markers," "effects of 2 mu-opiates on gastrointestinal transit"), body parts or processes (e.g. - "middle ear gas exchange," "inflammatory status," "abdominal") and other ("randomized clinical trial," "for recipients:," "health care quality"). When mapping to MeSH terms, we only used the portions of the vocabulary matching diseases, "Diseases[C]" and "Pyschiatry and Psychology[F]." For strings mapped to MeSH terms, we convert any synonyms or entry terms to the MeSH heading. Entry terms for a heading are "synonyms, alternate forms, and other closely related terms in a given MeSH record that are generally used interchangeably with the preferred term for the purposes of indexing and retrieval."[279]

Figure D-1 shows how many terms were in MeSH versus data-derived terms as well as how many mappings were "exact" vs "approximate" synonyms.

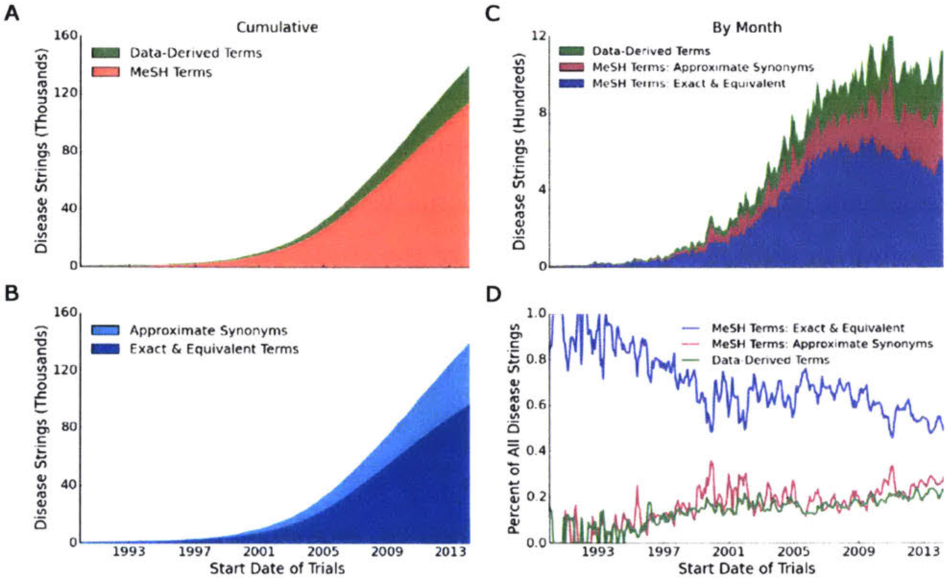


Figure D-1: Categorization of disease strings by date show a fewer percentage of strings were either direct synonyms or MeSH terms over time. (A) The cumulative number of disease strings that were mapped to MeSH terms or Data-Derived Terms. By 2014 more than 75% of strings were mapped to MeSH terms, though at a decreasing rate. (B) The cumulative number of strings that were either approximate synonyms (approximately the same as the term in the thesaurus) compared to strings that were exactly the same or equivalent synonyms. By 2014 approximately 67% of the disease strings were exact or equivalent, though at a decreasing rate. (C) The number of diseases strings by month with MeSH terms distinguished by approximate and equivalent. The number of exact and equivalent MeSH terms peaks around 2008. (D) The percentage of disease strings by month. The percentage of approximate MeSH synonyms and data-derived terms increases linearly starting around 1995.

D.3 Drug Disambiguation

Drugs are different from diseases because they are more discrete in nature, do not often change, are often shorter in length and less variable in spelling. There are additional challenges specific to our data set including new drug terms, strings that include multiple drugs and strings with additional information. New drugs occur often in our dataset and do not have official names (identified only by a code name) or have not yet been included in the MeSH vocabulary. The other challenge is that each drug string may contain more than one drug or contain information such as dose or administration route, which makes it hard to automatically extract drug names. It often takes a careful reading of a trial to determine if drugs are used as a combination or in different treatment arms of the study. We therefore assume that all drugs may individually have an effect on the disease.

To identify new drugs or find different spellings of drugs in MeSH, we start with the assumption that every possible drug name is a word or group of consecutive words in the drug string, which we will refer to as substrings. Substrings are identified as every possible set of characters between non-alphanumeric characters in a string. Each substring must have alphanumeric characters, but is not restricted to them because drug strings often contain spaces and commas or in the case of chemical names, dashes and parentheses. As an example of possible substrings, “Fenofibric Acid (Fibricor) 105 mg Tablet” has five spaces, “(”, “”, and “)” as non-alphanumeric characters which leads to 42 different substrings (“Fenofibric”, “Fenofibric Acid”, “Fenofibric Acid (”, “Fenofibric Acid (Fibricor”, “Fenofibric Acid (Fibricor” and so on). Two of the substrings are “Fenofibric Acid” and “Fibricor” which we want to identify as drugs. The rest of the substrings are not drug strings because they do not contain a drug (e.g. - “Tablet,” “105 mg,” or “)”) or contain additional information with the drug (e.g. - “Fibricor”).

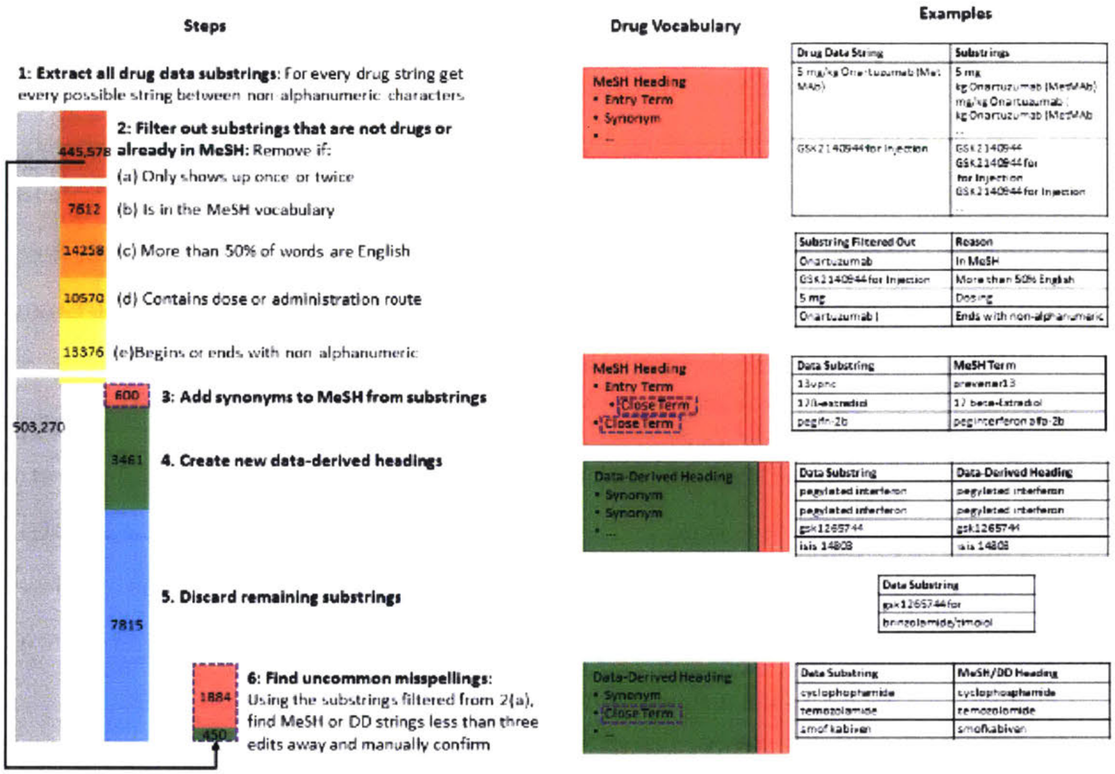
Figure D-2 in the Appendix shows the processing which starts by extracting 503,270 substrings from the original 66,066 unique intervention strings. The next step is to automatically filter out substrings that are already in MeSH or are not drug terms. We constructed several different filters after examining the different types of substrings. The first is to remove any substring that only shows up 2 or fewer times because we are only interested in drugs that show up multiple times. Misspellings that only occurs once or twice may be removed at this step, so we return to them later. The second filer is when more than 50% of the words are English words, which we identified using a basic dictionary available online with 109,582 English words (SIL International). Many of the substrings in the data are combinations of English words that are used to describe something about the intervention and this filter will discard those.

The next filter was to remove substrings that contained dosing or administration information, because those could not be drug strings. It is possible such a substring could contain a drug name, but that drug name would be found by itself as a different substring. Examples of specific words are "and," "tablet," "mg" and "experimental." A complete list of all 121 words are found in Table D.2.

The last filter removed substrings that began or ended with a non-alphanumeric character or had unbalanced parentheses. From our previous example "Fenofibric Acid (Fibricor) 105 mg Tablet", the following substrings would be filtered: "Fenofibric Acid (Fibricor" - unbalanced parentheses, "Fenofibric Acid (Fibricor$^{\mathrm{TM}}$) 105 mg" - dosing information, and "Tablet" - more than 50% English words.

After filtering, we were left with 11,876 substrings. We performed fuzzy string matching to find close MeSH terms and manually made matches if the drug was the same. These steps are described above under disease disambiguation. We also used fuzzy matching to identify terms that could be merged. Lastly, we set aside substrings that were not drug strings.

Once we had a preliminary thesaurus, we returned to the substrings that were filtered because they only showed up once or twice. We then computed the Levenshtein distance between each of those substrings and all terms in MeSH and our thesaurus.



1: Extract all drug data substrings: For every drug string get every possible string between non alphanumeric characters

2: Filter out substrings that are not drugs or already in MeSH: Remove if:

(a) Only shows up once or twice

(b) Is in the MeSH vocabulary

(c) More than 50% of words are English

(d) Contains dose or administration route

(e)Begins or ends with non alphanumeric

3: Add synonyms to MeSH from substrings

4. Create new data-derived headings

5. Discard remaining substrings

6: Find uncommon misspellings: Using the substrings filtered from 2(a), find MeSH or DD strings less than three edits away and manually confirm

Drug Data String	Substrings
5 mg/kg Onartuzumab (MetMAb)	5 mg kg Onartuzumab (MetMAb) mg/kg Onartuzumab (kg Onartuzumab (MetMAb ..
GSK2140944 for Injection	GSK2140944 GSK2140944 for for Injection GSK2140944 for Injection ..

Substring Filtered Out	Reason
Onartuzumab	In MeSH
GSK2140944 for Injection	More than 50% English
5 mg	Dosing
Onartuzumab (	Ends with non-alphanumeric

Data Substring	MeSH Term
13vpnc	prevenar13
17β-estradiol	17 beta-Estradiol
pegifn-2b	peginterferon alfa-2b

Data Substring	Data-Derived Heading
pegylated interferon	pegylated interferon
pegylated interferon	pegylated interferon
gsk1265744	gsk1265744
isis 14803	isis 14803

Data Substring
gsk1265744 for
brinzolamide/timolol

Data Substring	MeSH/DD Heading
cyclophophamide	cyclophosphamide
temozolamide	temozolomide
smof kabiven	smofkabiven

Figure D-2: Process to create a thesaurus of drug terms to maximize data that can be used. The process to disambiguate drugs was similar to disambiguation of disease strings, but had to be modified because multiple drugs often occur in one string and many drugs still being researched are not in MeSH. The steps are shown on the left. The first is to find every possible substring of the drug strings. Those are then filtered automatically in step 2 to remove substrings that are clearly not drugs or only occur once or twice. In steps 3-5 we manually review the remaining substrings. Step 6 is used to go back through all the substrings and find any substrings that are infrequent, but very close to one of the MeSH or Data-Derived terms. The output of the process is a thesaurus of MeSH and Data-Derived terms along with synonyms. Each drug string is then matched to one or more terms in the in the vocabulary if possible.

Levenshtein distance is the minimum number of character insertions, deletions or substitutions to go from one string to another,[278] which we chose specifically for finding misspellings. We only looked at string pairs that had a distance of less than or equal to 2 for substrings with more than 4 characters. We then manually went through those remaining terms and created a match if the drug was the same.

and
plus
tablet
tablets
with
dose
doses
mg
mcg
g
kg
+
ml
treatment
treatments
months
month
year
years
cycle
cycles
day
days
level
placebo
of
or
in
ug
every

for
week
weeks
pharmaceuticals
drug
comparator
formulation
oral
period
high
low
transdermal
intramuscularly
usp
hour
hours
patch
iu
biologic
ophthalmic
diseases
subcutaneous
maintenance
period
inhaled
inhalation
microspheres
system
solution
sustained

lot
spray
sprays
seasonal
cohort
cohorts
group
cream
release
infusion
infused
cohort
phase
combined
air
injectable
suspension
experimental
chemotherapy
via
adjuvant
extended
children
nanoparticle
intraperitoneal
mist
treatment
solution
formulation
prodrug

used
panel
gel
transdermal
nasal
hourly
capsules
weekly
coated
tumor
candidate
over
24hr
24hrs
48hr
48hrs
adjusted
prolonged
from
containing
powder
inhaler
er
diet
dose
excipient
subconjunctival
prefilled
label
transdermal
administered

Table D.2: Words used to filter potential drug substrings. These words appear frequently in the intervention strings for drug trials. Most of the words relate to dosing or administration. If one of these words occurs in a drug substring we filter out the substring because it cannot be a drug by itself.

Figure D-2 shows how many strings were filtered or manually assigned at each step. The majority of substrings are filtered because they only show up once or twice, leaving approximately 40,000 substrings. Through the additional filtering steps we reach a more manageable number of 11,876. From these we were able to map 4,061 substrings and then go back to the misspellings and map 2,334 more substrings.

The final step was to review all mappings manually. For mappings that were not straightforward we searched ClinicalTrials.gov to ensure that the terms referred to the same drug. This often occurred when there were other names listed for a drug on ClinicalTrials.gov that were not listed as synonyms in MeSH. We documented at least one of the trials for each of these mappings. We also checked the MeSH synonym mappings to ensure that very general terms and non-drug terms were excluded as well as synonyms that might not refer to a drug. We included these lists of terms in Tables D.3 and D.4.

After completing our drugs thesaurus with MeSH and supplemented terms, we went through the drug strings and identified terms in the thesaurus. We only kept the longest substring if one substring was the complete subset of another. We also searched in the strings in the "intervention arm groups" and "intervention other names" fields which are sometimes included on CclinicalTtrials.gov. Table D.5 in the Appendix shows 10 randomly selected intervention strings with the drug names that were automatically extracted using our thesaurus.

We plot in Figure D-3 how many strings are MeSH terms vs data-derived terms vs those that could not be mapped. Figure D-4 shows the temporal patterns of the same data.

Acids
Adhesives
Adjuvants, Immunologic
Alloys
Amino Acids
Anabolic Agents
Analgesics
Anti-Anxiety Agents
Anti-Arrhythmia Agents
Anti-Asthmatic Agents
Anti-Bacterial Agents
Anti-HIV Agents
Anti-Obesity Agents
Antibodies
Anticoagulants
Anticonvulsants
Antidepressive Agents
Antiemetics
Antifibrinolytic Agents
Antigens
Antigens, Surface
Antihypertensive Agents
Antiparasitic Agents
Antiperspirants
Antipsychotic Agents
Antirheumatic Agents
Antiviral Agents
Bacterial Vaccines
Biological Factors
Biological Markers
Biological Products
Bronchodilator Agents
Capsules
Carbohydrates
Carbon
Cardiovascular Agents

Ceramics
Chewing Gum
Cholinergic Agents
Colloids
Coloring Agents
Contraceptive Agents
Contrast Media
Cosmetics
Dentifrices
Deodorants
Detergents
DNA
Drug Carriers
Dust
Emetics
Emulsions
Enzymes
Food Additives
Gels
Genetic Markers
Hormones
Hydrogel
Hypoglycemic Agents
Hypnotics and Sedatives
Ice
Immunosuppressive Agents
Indicators and Reagents
Keratolytic Agents
Lipoproteins
Minerals
Nasal Sprays
New-Fill
Nucleoproteins
Oils
Ophthalmic Solutions
Peptides

Pharmaceutical Preparations
Phytochemicals
Plant Extracts
Plant Nectar
Plant Oils
Plastics
Polymers
Powders
Prescription Drugs
Proteins
Proton Pumps
Protons
Receptor, Adenosine A1
Receptor, Cannabinoid, CB1
Receptor, Cholecystokinin B
Receptor, Endothelin A
Receptor, Epidermal Growth Factor
Receptors, AMPA
Receptors, Androgen
Receptors, Angiotensin
Receptors, Antigen
Receptors, Antigen, T-Cell
Receptors, Bradykinin
Receptors, Calcium-Sensing
Receptors, Cholinergic
Receptors, Endothelin
Receptors, Erythropoietin
Receptors, Ghrelin
Receptors, Glutamate
Receptors, Interleukin-1
Receptors, Interleukin-4
Receptors, Leukotriene
Receptors, Lysophosphatidic Acid
Receptors, Mineralocorticoid
Receptors, Muscarinic
Receptors, Nicotinic

Receptors, Opioid
Receptors, Platelet-Derived Growth Factor
Receptors, Progesterone
Receptors, Prolactin
Receptors, Purinergic P2
Receptors, Serotonin, 5-HT3
Receptors, Somatostatin
Receptors, Tumor Necrosis Factor
Receptors, Vascular Endothelial Growth Factor
Receptors, Vasopressin
Salts
Silicates
Silicones
Soaps
Smoke
Solutions
Starch
Steam
Surface-Active Agents
Sweetening Agents
Tablets
Tablets, Enteric-Coated
Tissue Adhesives
Toothpastes
Tumor Markers, Biological
Vaccines
Vaccines, Conjugate
Vaccines, DNA
Vaccines, Inactivated
Vaccines, Subunit
Vaginal Creams, Foams, and Jellies
Vasoconstrictor Agents
Vasodilator Agents
Venoms
Vitamins

Table D.3: MeSH drug terms that are too general or do not refer to a drug. These terms or one of their MeSH synonyms were found in the ClinicalTrials.gov drug interventions fields. They were removed because they were too general to be useful, e.g. - Vaccines, or because they were not actually drugs, e.g. - Capsules.

dtpa
pep
alum
merlin
monitor
imp
ado
advantage
mph
amp
luminal
sentinel
toto
prod
peek
complement
ors
carob
crystalloid
simplex
ppd
manna
sol
vessel
int
brand
retinal
quad
smokeless
cit
trim
tan
mil
cave
foxy
nota
sham
lam
lab
mad
icon
cpi
dom
counter
mimic

Table D.4: MeSH drug synonyms that are ambiguous. These MeSH synonyms were found in the ClinicalTrials.gov drug interventions fields, but did not always refer to the MeSH term they were a synonym for. They were removed to prevent false positives.

Original Drug	**Extracted Terms**
diclofenac + calcitriol	diclofenac, calcitriol
placebo (plb)	placebo
candesartan cilexetil (atacand)	atacand, candesartan cilexetil
intra-amniotic injection of digoxin	digoxin
loperamide/simeticone 2 mg/125 mg chewable tablets	loperamide
recombinant luteinizing hormone (r-lh)	luteinizing hormone
supplements of l-methionine, betaine and folate	folate, betaine, l-methionine
dabigatran with asa	dabigatran
buspirone hcl	buspirone
mp-470 + carboplatin/etoposide	mp-470, carboplatin, etoposide

Table D.5: Original drug strings and extracted drug terms. These 10 strings were randomly selected from the data in the intervention field on ClinicalTrials.gov. The terms on the right are those that were extracted from the original data using the thesaurus we constructed.

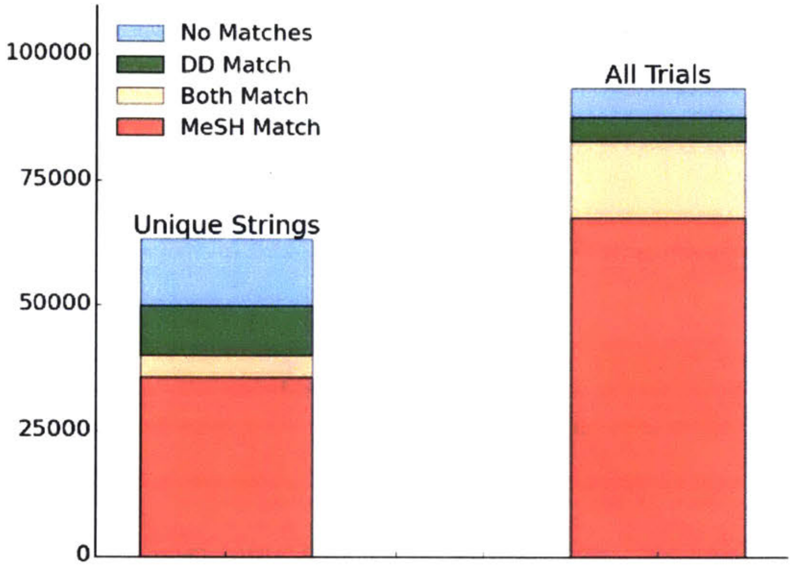


Figure D-3: Categorization of drug strings shows most trials have a MeSH drug term. Each unique drug string was categorized as having a Data-Derived (DD) match, a MeSH match, both or neither. We did the same categorization for trials on the right where each trial could have one or more unique strings. Even though there are many unique drug string that could not be matched to a drug in the thesaurus, they are infrequent and make up a small percent of all trials. Drug strings matched to MeSH terms are the opposite, even though they account for approximately 60% of unique strings, they account for approximately 85% of all trials.

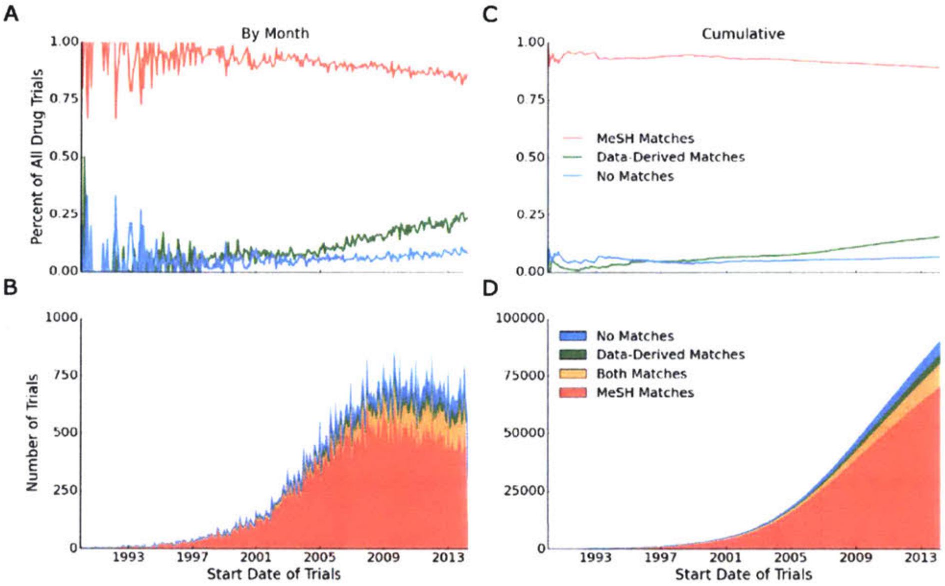


Figure D-4: Categorization of drug strings over time show similar patterns to disease string categorizations. (A) The percentage of drug trials each month that have a drug string that matches a MeSH term decreases over time, while the percentage of those that match a data-derived term increases over time. (B) The number of trials in each category by month shows that MeSH matches peak around 2008, as in the case of condition strings. (C) The cumulative sum of the data in (A). (D) The cumulative sum of the data in (B). Approximately 80% of the trials have a MeSH match.

D.4 Sampling and Validation of Drug Strings

We sampled 100 trials and examined the text of each trial relative to the drug names automatically extracted using our thesaurus. Of the 100 trials, 0 trials had an incorrect drug, 4 trials had a missed drug and 6 trials had an approximate match to a drug with some loss of detail (see Table D.6). Of the missing, 3 were missing because they were not listed in the drug field of the trial or because the drug was not in MeSH and did not occur often enough to create a new category. In the 100 trials there were 216 drugs total, so the error rate by drug string is actually lower.

We also sampled 100 trials that did not have any drugs extracted. Of those, 92 were correctly ignored given our criteria (true negatives). The 92 trials were not matched because there were either too few of the drugs listed to form a new term, too few diseases associated with the drug to form a new term, the drugs listed were too general or were not actually drugs (see Table D.7). The 8 trials with drugs that were not in the thesaurus were automatically filtered because they only showed up once (5 trials), were in the English dictionary (2 trials) or were manually filtered (1 trial). The 5 trials that only showed up once could have been mapped to other terms in the thesaurus, but were not found using Levenshtein distance because of unique word ordering or the use of acronyms.

Error	Number	Notes
Missing a drug	4	3 of the 4 were missing because they were not listed in the drug field or because the drug did not occur often enough to become its own term in the vocabulary.
Incorrect	0	
Approximate match	6	1. Lipid A vs. Glucopyranosly Lipid A 2. Meningococcal Vaccines vs. Meningococcal B Vaccine 3. Peginterferon vs. Peginterferon a2 4. CD133 vs. CD133+ cells 5. Insulin vs. Basal Insulin 6. t-cells vs. Xcellerated T Cells

Table D.6: Evaluation of 216 randomly selected mappings from drug strings. There were three possible types of errors: "Missing a drug" indicates that the extracted term was partially correct, but did not include all drugs listed in the original string. "Incorrect" indicates that the drug in the original string and the drug in our data set are not the same. "Approximate match" indicates the original drug is somewhat different from the drug extracted to our data set. The notes section under "Approximate match" shows the extracted term compared to the original term. We randomly selected 100 trials and those trials contained 216 drugs total.

Reason no match	**Number**	**Notes**
Too few trials	45	The drug only showed up in 2 or fewer trials.
Too few disease	18	Though the drug show up in 3-10 trials, it only showed up with 2 or fewer diseases.
Too general	17	These terms could be considered drugs, but were too general or too vague to be its own term (e.g. "hormonal contraceptive" or "Alcon investigational agent" where Alcon was the company).
Not a drug	12	Examples: "blood collection," "phenotypical approach," "1mg," "5mg," "10mg"
Missed mappings	8	1. "AC-170" [5 trials] manually filtered 2. "donor natural killer cell infusion" [various trials] filtered because of English words 3. "psilocybin" [14 trials] In English dictionary because a recreation drug 4. "Multivitamins (including B, C and E)" [1 trial] word order 5. "EIA chemotherapy" [1 trial] too few, but EIA is an acronym for 3 drugs 6. "IC Green" [1 trial] too few, acronym for indocyanine green 7. "acid Zoledronic" [1 trial] too few, word order 8. "EBV-Specific CTLs and CD45 Mab" [1 trial] too few, uses acronyms

Table D.7: Evaluation of 100 random drug strings that were not mapped to drug terms. The left column gives the reasons why a string was not mapped to a term in our data set and the middle column shows how many and the right column gives an explanation of the reasons. The missed mappings represent drugs that could have been included in our data set, representing a false negative rate of 8%. The notes for missed mappings includes the drug substring that was missed, how many trials it showed up in and why it was not mapped to a drug term in our data set.

D.5 Disease Representation

The inspiration for the drug-based disease representation came from an examination of specific blockbuster drugs. We noted that even though a blockbuster drug was approved for a only one or a few indications, it was often tested on a wide variety of diseases in clinical trials. It appears that once a blockbuster drug is approved, the number of trials involving goes up exponentially. With sufficiently many trials creating relationships between drugs and diseases, it is possible to analyze them as a connected web. Such an analysis is reminiscent how social networks are analyzed. By now there are sufficient trials and sufficient diversity to create a representation based of disease based on drugs.

We further noted in our explanation that there appeared to be two types of trials performed on a blockbuster drug. One type revolves around the approved indication where similar diseases are tested or the original indication is tested along with a secondary indication. The other type of exploration appears unrelated to the original indication with trials including a wide variety of diseases. This insight allowed us to filter the data to preserve the strongest relationships between disease noted by the first type of exploration.

D.6 Network Connectivity

For space constraints we do not visualize the Disease-Trial-Drug Network (DTDN) nor the complete Disease-Disease Network (DDN), but we plot the distribution of node connectivity for each in D-5.

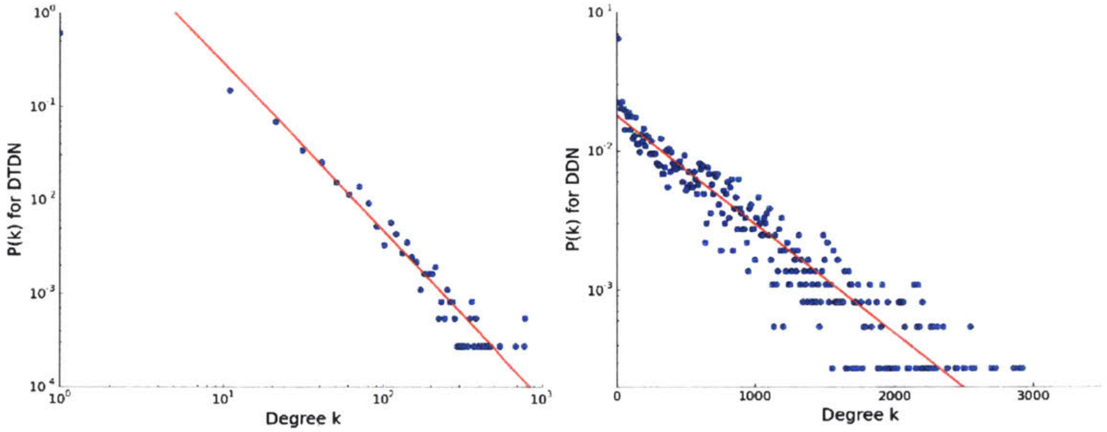


Figure D-5: Degree distribution of nodes in the Disease-Trial-Drug Network (DTDN) and the Disease-Disease Network (DDN). For each plot, the y-axis shows proportion of nodes P(k) while the x-axis shows the degree. Note that (a) log-log plot while (b) is a log-linear plot. (a) The DTDN Disease-Drug Network degree distribution follows a power law with coefficient $\gamma = 1.8$ as shown with the red fit line. (b) The DDN degree distribution follows an exponential distribution with coefficient $\gamma = .0018$ as shown with the red fit line.

D.7 Network Layout

Our purpose in displaying the network is to show similarity between nodes approximated by proximity, so a natural approach for a layout is a force-direct algorithm.[280] The basic idea is set up a system of forces drawing nodes together and pushing them apart, then solve for the minimum energy of the system. The "spring force" attracts nodes proportional to some function of distance and weighting of the edge.[281, 282, 283] A repulsive force can be used to space nodes apart, which is often proportional to 1/distance between the nodes. It could also be proportional to degree of the two nodes[283] which can be useful for scale-free networks[225] that have some highly connected nodes as in our case. We chose to experiment with the Fruchterman and Reingold,[281] Kamada and Kawai[282] and ForceAtlas2[284] methods.

All of the three algorithms use simulated annealing to minimize the energy of the system because it is more likely to find a global optimum.[247] Each algorithm starts with nodes at random positions and the final position may depend on those starting positions. Running an algorithm different times will yield different absolute positions, but we are interested in the relative position of each node to others.

To ensure that relative distance between nodes did not depend on starting position and did not vary strongly by algorithm, we measured the variance of distance between every node for 10 independent layouts generated by each of our three algorithms. We plot a histogram of the more than 125,000 variances and normalize the histogram so the integral equals 1 as seen in Figure D-6. The histogram indicates what proportion of nodes stay the same distance away from each other over multiple runs of the layout algorithm. If our layout is consistent, we would expect the distance between most pairs of nodes to be relatively constant while distance between some pairs of nodes would be relatively random because they are unrelated or distant on the graph. For a reference distribution we simulated a distribution of randomly placed nodes on the same size plot. We see our layouts have variance distributions shifted strongly toward zero indicating the relative distances in our network plot can be reproduced consistently.

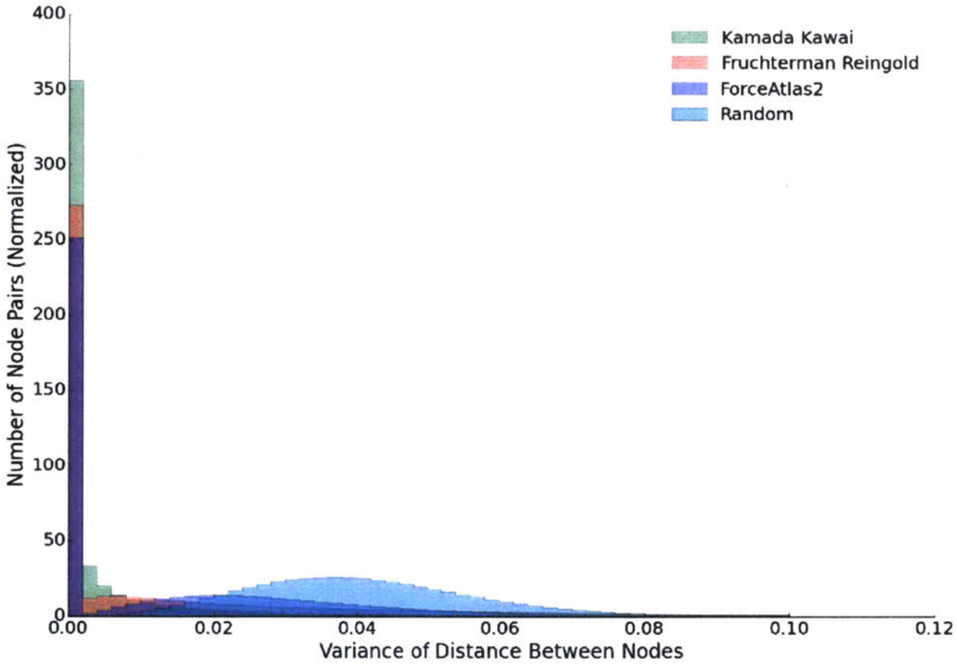


Figure D-6: Small variations in distance between node pairs indicates the plotting algorithms are consistent. Three different force-direct network layout algorithms were used to plot the disease-disease network 10 times each on the same scale. The variance in Euclidean distance between the same pairs of nodes was calculated and a histogram of all such variances is plotted. There were more than 125,000 node pair distances for each plot. The variance for a random layout is also shown. The vast majority of node pair distance change very little independent of the algorithm or if the algorithm is run several times. There are some node pairs that do appear to be more random, which may be the result of weakly connected nodes in the network that therefore end up in more random locations.

D.8 MeSH Categories

We use the subtrees "Diseases[C]" and "Pyschiatry and Psychology[F]" to categorize the diseases and to compare the DDN with the HDN (see Table D.8). We do not include "Bacterial Infections and Mycoses [C01]," "Virus Diseases [C02]," "Parasitic Diseases [C03]," "Male Urogenital Diseases [C12]," "Female Urogenital Diseases and Pregnancy Complications [C13]" because they were not included in the HDN. We also do not include "Disorders of Environmental Origin [C21]," "Animal Diseases [C22]," "Pathological Conditions, Signs and Symptoms [C23]," "Occupational Diseases [C24]," "Chemically-Induced Disorders [C25]," and "Wounds and Injuries [C26]" because they do not partition the human disease space.

MeSH Categories	**MeSH Tree Number**	**HDN Categories**
Bone and Joint	C05.116/C05.550	Bone
Neoplasms	C04	Cancer
Cardiovascular Diseases	C14	Cardiovascular
Connective Tissue Diseases	C17.300	Connective tissue
Skin Diseases	C17.800	Dermatological
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16	Developmental
Otorhinolaryngologic Diseases and Stomatognathic Diseases	C09/C07	Ear, Nose, Throat
Endocrine System Diseases	C19	Endocrine
Digestive System Diseases	C06	Gastrointestinal
Hemic and Lymphatic Diseases	C15	Hematological
Immune System Diseases	C20	Immunological
Metabolic Diseases	C18.452	Metabolic
Muscular Diseases	C05.651	Muscular
Nervous System Disease	C10	Neurological
Nutrition Disorders	C18.654	Nutritional
Eye Diseases	C11	Ophthamological
Psychiatry and Psychology	F	Psychiatric
-	-	Renal
Respiratory Tract Diseases	C08	Respiratory
-	-	Skeletal
Multiple	-	Multiple
None	-	Unclassiffied

Table D.8: MeSH categories and Human Disease Network categories. MeSH categories for disease nodes were selected based on the subtrees of the "Diseases [C]" and "Psychiatry and Psychology [F]." We modified these categories slightly to be able to make direct comparisons with the Human Disease Network.

Each disease that is a MeSH descriptor has one or more tree numbers. Those that are a MeSH concept point to one or more descriptors. For each disease we can therefore trace it to one or more tree numbers and move up the tree to the categories listed in the table. If it fits under more than one heading (e.g. - 'Colorectal Neoplasms' shows up under 'Digestive System Diseases' and 'Neoplasms'), we assign it to one of the categories if more than 66% of the categories are the same, otherwise it is classified as "multiple."

D.9 Binomial Testing

We use a one-tailed binomial test to compare overlap of the disease-disease network (DDN) with other data sources. The test is based on the binomial distribution . For comparing the DDN to MeSH, we test how many edges are between nodes of the same MeSH category compared to randomly distributed edges. In this case, n is the number of edges in our graph, k is the number of edges between nodes of the same MeSH category and p is the probability of randomly placing an edge between nodes of the same category. We calculate p as the total possible edges between nodes of the same category divided by the total number of possible edges in the graph.

For comparing the DDN to the HDN, we look at the subset of nodes that are exactly the same in each network and test how many edges are the same in each graph compared to randomly distributed edges. In this case, n is the number of edges in the DDN subset network, k is the number of edges exactly the same in both subset networks, and p in the probability of randomly placing an edge correctly in the HDN subset. We calculate p as the number of edges in the HDN subset divided by the total possible edges in the HDN subset.

For comparing the DDN with MeSH and the HDN for each category, we ran 1000 Monte Carlo simulations to randomly assign edges for the nodes in the network. To calculate the NNI we also ran the layout algorithm for each of the 1000 randomly connected networks. Table D.9 shows the p-values calculated for comparing the DDN and the MeSH taxonomy by categories. Table D.10 show the p-values calculated for comparing the DDN and the HDN by categories.

MeSH Categories	**NNI p-value**	**Shortest Path p-value**
Bone and Joint	<.001	<.001
Cardiovascular	<.001	<.001
Connective Tissue	.160	.722
Developmental	.877	<.001
Digestive	<.001	<.001
Ear, Nose and Throat	.034	<.001
Endocrine	.016	<.001
Eye	<.001	<.001
Hemic and Lymphatic	<.001	<.001
Immunological	.245	<.001
Metabolic	<.001	<.001
Muscular	.333	.065
Neoplasms	<.001	<.001
Nervous System	.002	<.001
Nutrition	.003	<.001
Pyschiatric	<.001	<.001
Respiratory	<.001	<.001
Skin	.339	.981

Table D.9: Statistical significance of clustering by category relative to MeSH. One-tailed p-values were calculated using 1000 Monte Carlo simulations to randomly assign edges between the nodes in the network and layout each network with the same algorithm used to display the DDN. Of the 15 categories with more than 10 nodes, 13 have a significantly smaller NNI (p<.05) and 14 have a significantly smaller average shortest path within the category compared to without (p<.05). The NNI represents a clustering of the physical layout in 2 dimensions and the shortest path represents clustering strictly in terms of network connectivity.

MeSH Categories	**Degree Fraction p-value**	**Shortest Path p-value**
Bone and Joint	<.001	<.001
Cardiovascular	<.001	<.001
Connective Tissue	<.001	.950
Developmental	<.001	..002
Digestive	<.001	<.001
Ear, Nose and Throat	<.001	<.001
Endocrine	<.001	<.001
Eye	<.001	<.001
Hemic and Lymphatic	<.001	<.001
Immunological	<.001	<.001
Metabolic	<.001	<.001
Muscular	-	.253
Neoplasms	<.001	<.001
Nervous System	<.001	.001
Nutrition	.003	<.001
Pyschiatric	<.001	<.001
Respiratory	<.001	<.001
Skin	-	.941

Table D.10: Statistical significance of clustering by category relative to the Human Disease Network (HDN). The one-tailed p-values were calculated using 1000 Monte Carlo simulations to randomly assign edges between the same collection nodes of nodes as in the DDN. The categories that had at least one pair of nodes directly connected all showed significant differences except for "Connective Tissue." The degree fraction represents connectivity within the same category and shortest path represents connectivity but takes into account indirect connections.

D.10 Drug Predictions Using MeSH

We used distance in the MeSH hierarchy to predict drugs in the test set for a disease based on proximity to drugs in the training set. We define distance as the number of links in the tree structure between two drugs in MeSH. If the drug shows up multiple times in MeSH it is the minimum of all such distances. We then rank potential test drugs by the minimum distance between the test drug and all training drugs. The highest ranked predictions are then those that are closest to one or more of the training drugs. The histogram of all AUC's in our data is plotted in Figure D-7 compared to the histogram for random predictions.

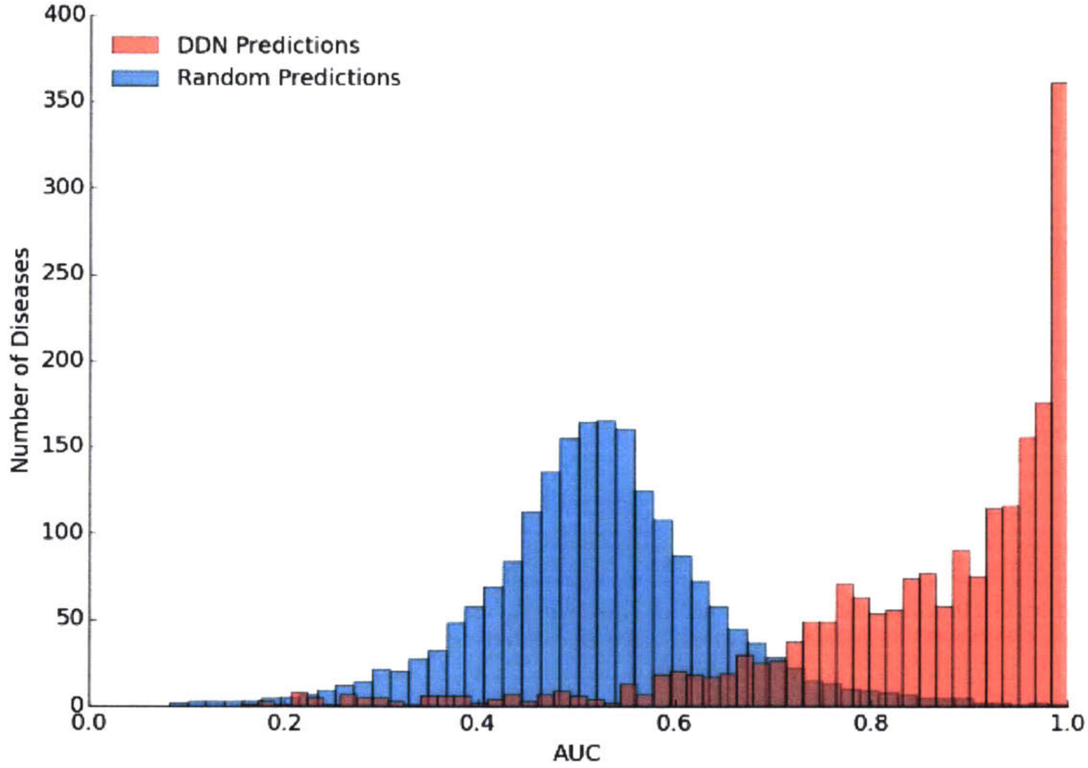


Figure D-7: Histogram of AUC's for drug predictions for each disease compared to random. Histogram of all disease AUC's (red) along with the distribution of AUC's for random predictions (cyan). **For each disease, a score was obtained by collaborative filtering which was used to prioritize** drugs. An AUC was then calculated by comparing the drug scores and the drugs tested in trials for that disease. The collection of all AUC's was used to create the histogram in red. The training data set includes all trials before January, 2011 given and the test data set includes trials after January, 2011. The blue histogram was calculated by generating random scores and comparing to the test data.

Specific examples of predictions are shown in Table D.11. A large MeSH prediction often meant that the predicted drug was close to one of the test drugs, such as "Testosterone" and "testosterone undecanoate" or "Botulinum Toxins" and "Botulinum Toxins, Type A." Canakinumab is close to Interleukin-1 in MeSH because it is an Interleukin-1 beta blocker. To provide intuition for the meaning of AUCs, there are 7,671 drugs in the data set, so an AUC of 0.99 would indicate drugs in the test set would be included in approximately the top 77 predicted drugs.

D.11 Neighborhood Connectivity

One of the advantage of a network model is to capture indirect relationships among elements. Direct relationships are modeled as the edges between nodes in a network, but two nodes can have a relationship indirectly through other nodes. For example, if node A and B have no edge, but A and B are both connected to C, D and E then we might conclude that A and B are indeed related. Figure D-8 plots the number of trials and drugs for a given disease along with the degree of neighbors. There are many cases where the neighborhood size does not scale with the number of trials or drugs, indicating information in the network connectivity.

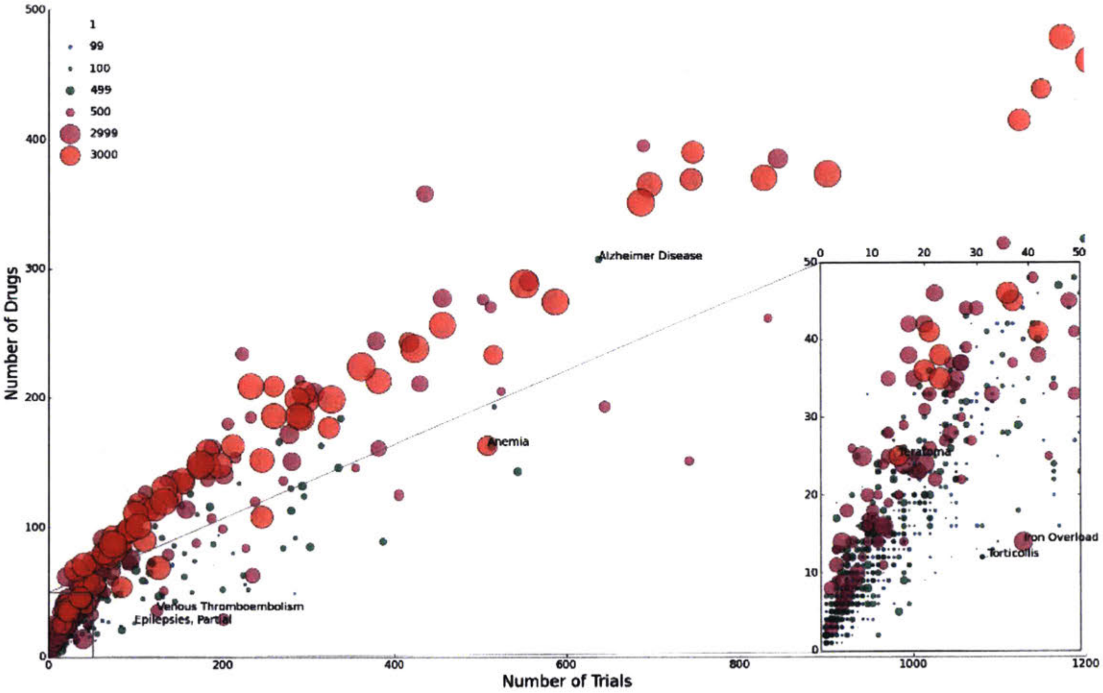


Figure D-8: Neighborhood connectivity of diseases in the DDN. Each circle is a disease in the DDN with the size of the circle proportional to the sum total degree of all neighbors of the disease. A larger number indicates a highly connected neighborhood while a smaller number indicates a sparsely connected neighborhood where additional links will more likely change the location of that node in the network. Identifying such sparsely connected diseases may help predict where a clinical trial will provide the most information in defining the disease landscape. The number of trials and number of drugs are on the axes to show that local topology in the DDN does not always match direct measures of how much learning has taken place.

Disease	Drugs in Training Set	DDN AUC	Nave AUC	Unique Drugs in Test Set
Brain Concussion	· Magnesium Sulfate · Metoclopramide	0.98	0.58	· Melatonin · Fatty Acids, Omega-3[285] · Progesterone · docosahexanoic acid[285] · resveratrol · Ondansetron
Polypoidal Choroidal Vasculopathy	· bevacizumab · combretastatin · fosbretabulin · ranibizumab · verteporfin	0.99	0.49	· 1-phenyl-3,3-dimethyltriazene · Aflibercept[286] · salicylhydroxamic acid
Heart Block	· Adrenal Cortex Hormones · Dexamethasone · Immunoglobulins, Intravenous	0.97	0.08	· Hydroxychloroquine[287]
Bulimia	· duloxetine · orlistat · sibutramine · topiramate	0.97	0.81	· Liraglutide[288]
Adenomatous Polyps	· Aspirin · beta-Glucans · Calcitriol · Calcium · Curcuminoids · DFMO · Eflornithine · Eicosapentaenoic Acid · erlotinib · Polyethylene Glycols	0.99	0.17	· Metformin[289]
Adenomatous Polyposis Coli	· celecoxib · Curcumin · Eicosapentaenoic Acid · erlotinib · Inulin · MK0966 · Probiotics · rofecoxib · Starch · Ursodeoxy · DFMO · Eflornithine · Sulindac	0.98	0.38	· Metformin[289]
Eisenmenger Complex	· bosentan · sildenafil · tadalafil	0.96	0.42	· Citrulline · Epoprostenol · Iloprost[290] · malic acid
Schnitzler Syndrome	· givinostat hydrochloride · rilonacept · Interleukin-1	1.0	0.99	· Canakinumab[291, 292]
Hypogonadotrophic Males	· testosterone undecanoate · vardenafil	1.0	0.99	· Testosterone
Crow's Feet	· onabotulinumtoxinA · Botulinum Toxins, Type A	1.0	0.99	· Botulinum Toxins

Table D.11: Drugs in training and test sets for select diseases with large AUCs showing prediction potential of DDN data. Diseases are shown on the left with the drugs in the training set (trials before 2011), the AUC for that disease using our data, the AUC for that disease using a closest drugs in MeSH approach (see Appendix:Methods) and the unique predicted drugs in the test set. References are given for some test drugs predicted showing recent research that suggests a confirmed link between that drug and the disease. The first seven diseases were selected where the DDN prediction AUCs were much larger than the MeSH prediction AUCs. The final three show cases where the AUCs are close.

Bibliography

[1] Elizabeth Warren. Strengthening Research through Data Sharing. *New England Journal of Medicine*, 375(5):401–403, 2016.

[2] Institute of Medicine. *Best care at lower cost: the path to continuously learning health care in America.* The National Academies Press, Washington, D.C., 2013.

[3] Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, 167(7):748–751, 2010.

[4] Joseph Loscalzo, Isaac Kohane, and Albert-László Barabási. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology*, 3(124):124, 2007.

[5] Eamonn Maguire. Taxonomy in Biology and Visualization. *http: //isa-tab.sourceforge.net/docs/publications/Taxonomy.pdf*, 2012.

[6] Jackie Leach Scully. What is a disease? *EMBO Reports*, 5(7):650–653, 2004.

[7] Bruce McGregor. Constructing a concise medical taxonomy. *Journal of the Medical Library Association*, 93(1):121–123, 2004.

[8] Joachim P. Sturmberg and Carmel M. Martin. Diagnosis - the limiting focus of taxonomy. *Journal of Evaluation in Clinical Practice*, 22(1):103–111, 2016.

[9] Institute of Medicine. *Toward Precision Medicine : Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease Committee on a Framework for Development a New Taxonomy of Disease.* National Academies Press, Washington, D.C., 2011.

[10] Atul J Butte. The ultimate model organism. *Science*, 320(5874):325–7, 2008.

[11] Thomas R Insel. The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry*, 171(4):395–397, 2014.

[12] Institute of Medicine. *Improving Diagnosis in Health Care.* National Academies Press, Washington, D.C., 2015.

[13] Leila Agha and David Molitor. Location Matters: The Adoption of New Medical Technologies. *Stanford Institute of Economic Policy Research*, 2013.

[14] V N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 10(5):988–99, 1999.

[15] Institute of Medicine. *To Err Is Human: Building a safer health system.* Washington, D.C., 2000.

[16] M Saeed, C Lieu, G Raber, and R G Mark. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in cardiology*, 29:641–644, 2002.

[17] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[18] R. Andrew Taylor, Joseph R. Pare, Arjun K. Venkatesh, Hani Mowafi, Edward R. Melnick, William Fleischman, and M. Kennedy Hall. Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3):269–278, 2016.

[19] Gil Alterovitz, Michael Xiang, David P Hill, Jane Lomax, Jonathan Liu, Michael Cherkassky, Jonathan Dreyfuss, Chris Mungall, Midori A Harris, Mary E Dolan, Judith A Blake, and Marco F Ramoni. Ontology engineering. *Nat Biotech*, 28(2):128–130, feb 2010.

[20] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for Precision Medicine. *New England Journal of Medicine*, 366(6):489–491, 2012.

[21] J Larry Jameson and Dan L Longo. Precision Medicine–Personalized, Problematic, and Promising. *New England Journal of Medicine*, 372(23):2229–2234, 2015.

[22] Francis S Fancis S. Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[23] Robert Saunders and Mark D Smith. The Path to Continuously Learning Health Care. *Issues in Science and Technology*, 29(3):27–37, 2013.

[24] Douglas M Berwick. Continuous Improvement as an Ideal in Health Care. *New England Journal of Medicine*, 320(I):53–56, 1989.

[25] K Calman. The profession of medicine. *British Medical Journal*, 309(6962):1140–3, 1994.

[26] Steven H Miles. *The Hippocratic Oath and the ethics of medicine.* Oxford University Press, 2005.

[27] Geoffrey Rose. *The strategy of preventive medicine.* Oxford University Press, 1992.

[28] Lisa Rosenbaum. Bridging the Data-Sharing Divide Seeing the Devil in the Details, Not the Other Camp. *New England Journal of Medicine*, 376(23):2201–2203, 2017.

[29] Hans-Georg Eichler, Frank Petacy, Francesco Pignatti, and Guido Rasi. Access to Patient-Level Trial Data - A Boon to Drug Developers. *New England Journal of Medicine*, 369(17):1577–1579, 2013.

[30] Robert Steinbrook. The European Medicines Agency and the Brave New World of Access to Clinical Trial Data. *JAMA Internal Medicine*, 173(5), 2013.

[31] Perry Nisen and Frank Rockhold. Access to patient-level data from GlaxoSmithKline clinical trials. *New England Journal of Medicine*, 369(5):475–478, 2013.

[32] Peter Groves and David Knott. The 'big data' revolution in healthcare. Technical Report January, McKinsey&Company, 2013.

[33] Mihcelle Mello, Jeffrey Francer, Marc Wilenzick, Patricia Teden, Barbara E Bierer, and Mark Barnes. Preparing for Responsible Sharing of Clinical Trial Data. *New England Journal of Medicine*, 369(17):1651–1658, 2013.

[34] Travis B. Murdoch and Allan S. Detsky. The Inevitable Application of Big Data to Health Care. *Journal of the American Medical Association*, 309(13), 2013.

[35] Joseph S Ross and Harlan M Krumholz. Ushering in a new era of open science through data sharing: the wall must come down. *Journal of the American Medical Association*, 309(13):1355–1356, 2013.

[36] Laura Merson, Oumar Gaye, and Philippe J. Guerin. Avoiding Data Dumpsters Toward Equitable and Useful Data Sharing. *New England Journal of Medicine*, 374(25):2414–2415, 2016.

[37] In-Hyun Park, Natasha Arora, Hongguang Huo, Nimet Maherali, Tim Ahfeldt, Akiko Shimamura, M William Lensch, Chad Cowan, Konrad Hochedlinger, and George Q Daley. Disease-specific induced pluripotent stem cells. *Cell*, 134(5):877–86, 2008.

[38] Graham J Lieschke and Peter D Currie. Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics*, 8(5):353–367, 2007.

[39] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33, 2016.

[40] David L Sackett, William MC Rosenberg, and JA Muir Gray. Evidence based medicine: what it is and what it isn't. *Brithish Medical Journal*, 312:71–72, 1996.

[41] P Michael Ho, Pamela N Peterson, and Frederick a Masoudi. Evaluating the evidence: is there a rigid hierarchy? *Circulation*, 118(16):1675–1684, 2008.

[42] Edzard Ernst. How much of general practice is based on evidence? *The British Journal of General Practice*, 54(501):313–328, 2004.

[43] John Garrow. How Much of Orthodox Medicine Is Evidence Based? *British Medical Journal*, 335(7627):951, 2007.

[44] Thomas R Frieden. Evidence for Health Decision Making Beyond Randomized, Controlled Trials. *New England Journal of Medicine*, 377(5):465–475, 2017.

[45] Paul a James, Suzanne Oparil, Barry L Carter, William C Cushman, Cheryl Dennison-Himmelfarb, Joel Handler, Daniel T Lackland, Michael L LeFevre, Thomas D MacKenzie, Olugbenga Ogedegbe, Sidney C Smith, Laura P Svetkey, Sandra J Taler, Raymond R Townsend, Jackson T Wright, Andrew S Narva, and Eduardo Ortiz. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). *Journal of the American Medical Association*, 311(5):507–20, 2014.

[46] Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[47] Uli Niemann, Myra Spiliopoulou, Henry Volzke, and Jens-Peter Kuhn. Interactive Medical Miner: Interactively Exploring Subpopulations in Epidemiological Datasets. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8726 of *Lecture Notes in Computer Science*, pages 460–463. Springer Berlin Heidelberg, 2014.

[48] Marc Ereshefsky. Defining 'health' and 'disease'. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 40(3):221–227, 2009.

[49] Ray Moynihan. A new deal on disease definition. *British Medical Journal*, 342:d2548, 2011.

[50] Alon Halevy, Peter Norvig, and Fernando Pereira. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[51] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, 2007.

[52] Anthony H V Schapira. Mitochondrial disease. *Lancet*, 368(9529):70–82, 2006.

[53] Saurabh Sahar and Paolo Sassone-Corsi. Metabolism and cancer: the circadian clock connection. *Nature Reviews Cancer*, 9(12):886–896, 2009.

[54] Horacio Fabrega. The need for an ethnomedical science. *Science*, 189(4207):969–975, 1975.

[55] Francesca Mangialasche, Alina Solomon, Bengt Winblad, Patrizia Mecocci, and Miia Kivipelto. Alzheimer's disease: clinical trials and drug development. *The Lancet Neurology*, 9(7):702–716, 2010.

[56] Melvin D. Cheitlin, Adolph M. Hutter, Ralph G. Brindis, Peter Ganz, Sanjay Kaul, Richard O. Russell, and Randall M. Zusman. Use of Sildenafil (Viagra) in Patients With Cardiovascular Disease. *Circulation*, 99(1):168–177, 1999.

[57] Stephen R Braun. Promoting Low T: A medical writer's perspective. *Journal of American Medical Association Internal Medicine*, 173(15):1458–1460, 2013.

[58] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A Margolin, Laura J van't Veer, Nuria Lopez-Bigas, Peter W Laird, Benjamin J Raphael, Li Ding, A Gordon Robertson, Lauren A Byers, Gordon B Mills, John N Weinstein, Carter Van Waes, Zhong Chen, Eric A Collisson, Christopher C Benz, Charles M Perou, and Joshua M Stuart. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944, 2014.

[59] Nancy E. Lane, Kenneth Brandt, Gillian Hawker, Elena Peeva, Edward Schreyer, Wayne Tsuji, and Marc Hochberg. OARSI-FDA initiative: Defining the disease state of osteoarthritis. *Osteoarthritis and Cartilage*, 19(5):478–482, 2011.

[60] Elliott Antman, Jean-Pierre Bassand, Werner Klein, Magnus Ohman, Jose Luis Lopez Sendon, Lars Rydén, Maarten Simoons, and Michael Tendera. Myocardial infarction redefined–a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *Journal of the American College of Cardiology*, 36(3):959–69, 2000.

[61] Daniela Berg, Ronald B. Postuma, Bastiaan Bloem, Piu Chan, Bruno Dubois, Thomas Gasser, Christopher G. Goetz, Glenda M. Halliday, John Hardy, Anthony E. Lang, Irene Litvan, Kenneth Marek, José Obeso, Wolfgang Oertel, C. Warren Olanow, Werner Poewe, Matthew Stern, and Günther Deuschl. Time to redefine PD? Introductory statement of the MDS Task Force on the definition of Parkinson's disease. *Movement Disorders*, 29(4):454–462, 2014.

[62] Carol A. Bean. *Relationships in the Organization of Knowledge.* Springer, 2001.

[63] Heather Hedden. Chapter 1 What Are Taxonomies. In *The Accidental Taxonomist.* Information Today, Inc, 2 edition, 2010.

[64] Shu-hsien Liao. Knowledge management technologies and applicationsliterature review from 1995 to 2002. *Expert Systems with Applications*, 25(2):155–164, 2003.

[65] Olivier Bodenreider. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearbook of Medical Informatics*, 3841:67–79, 2008.

[66] Iwao M. Moriyama, Ruth M. Loy, and Alastair H.T. Robb-Smith. *History of the statistical classification of diseases and causes of death.* National Center for Health Statistics, 2011.

[67] Margaret H. Coletti and Howard L. Bleich. Medical Subject Headings Used to Search the Biomedical Literature. *Journal of the American Medical Informatics Association*, 8(4):317–323, 2001.

[68] Kent A Spackman, Keith E Campbell, and Roger A Côté. SNOMED RT: a reference terminology for health care. In *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 640–644, 1997.

[69] Warren A. Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J. Mungall, Janos X. Binder, James Malone, Drashtti Vasant, Helen Parkinson, and Lynn M. Schriml. Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, 43(D1):D1071–D1078, 2015.

[70] Lynn M. Schriml and Elvira Mitraka. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. *Mammalian Genome*, 26(9-10):584–589, 2015.

[71] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):940–946, 2012.

[72] George Weisz. *Divide and Conquer: A Comparative History of Medical Specialization.* Oxford University Press, 2006.

[73] David C Hsia, Cathaleen A Ahern, Brian P Ritchie, Linda M Moscoe, and W. Mark Krushat. Medicare reimbursement accuracy under the prospective payment system, 1985 to 1988. *Journal of the American Medical Association*, 268(7):896–899, 1992.

[74] George L. Engel. The Need for a New Medical Model: A Challenge for Biomedicine. *Science*, 196(4286):129–136, 1977.

[75] Sandro Santagata, Ankita Thakkar, Ayse Ergonul, Bin Wang, Terri Woo, Rong Hu, J Chuck Harrell, George McNamara, Matthew Schwede, Aedin C Culhane, David Kindelberger, Scott Rodig, Andrea Richardson, Stuart J Schnitt, Rulla M Tamimi, and Tan A Ince. Taxonomy of breast cancer based on normal cell phenotype predicts outcome. *The Journal of Clinical Investigation*, 124(2):859–870, 2013.

[76] Brian F Gage, Amy D Waterman, William Shannon, Michael Boechler, Michael W Rich, and Martha J Radford. Validation of Clinical Classification Schemes Results From the National Registry of Atrial Fibrillation. *Journal of the American Medical Association*, 285(22):2864–2870, 2001.

[77] Jorg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-Laszlo Barabasi. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 2015.

[78] Annemarie Jutel. Sociology of diagnosis: a preliminary review. *Sociology of Health & Illness*, 31(2):278–299, 2009.

[79] Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century.* National Academies Press, Washington, D.C., 2001.

[80] C. William Hanson and Bryan E Marshall. Artificial intelligence applications in the intensive care unit. *Critical Care Medicine*, 29(2):427–435, 2001.

[81] Sancy A. Leachman and Glenn Merlino. Medicine: The final frontier in cancer diagnosis. *Nature*, 542(7639):36–38, 2017.

[82] Catherine De Angelis, Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin, Christine Laine, Ana Marusic, A. John P.M. Overbeke, Torben V. Schroeder, Hal C. Sox, and Martin B. Van Der Weyden. Clinical Trial Registration : A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, 351(12):1250–1251, 2004.

[83] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug Target Identification Using Side-Effect Similarity. *Science*, 321(5886):263–266, 2008.

[84] Dong-Sheng Cao, Liu-Xia Zhang, Gui-Shan Tan, Zheng Xiang, Wen-Bin Zeng, Qing-Song Xu, and Alex F Chen. Computational Prediction of Drug-Target Interactions Using Chemical, Biological, and Network Features. *Molecular Informatics*, 33(10):669–681, 2014.

[85] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5), 2012.

[86] Hui Huang, Ping Zhang, Xiaoyan A Qu, Philippe Sanseau, and Lun Yang. Systematic prediction of drug combinations based on clinical side-effects. *Scientific reports*, 4:7160, 2014.

[87] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, Ryan Whaley, Richard a Glennon, Jérôme Hert, Kelan L H Thomas, Douglas D Edwards, Brian K Shoichet, and Bryan L Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

[88] Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125–31, 2012.

[89] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, and Nicholas P Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, 9(9):2147–2163, 2014.

[90] Wenhui Wang, Sen Yang, Xiang Zhang, and Jing Li. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30(20):2923–2930, 2014.

[91] Guangxu Jin and Stephen T C Wong. Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines. *Drug Discovery Today*, 19(5):637–644, 2014.

[92] Steven E Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, 356(24):2457–2471, 2007.

[93] Emma Marris. Diabetes drugs under scrutiny in a post-Vioxx world. *Nature Reviews Drug Discovery*, 6(7):505–6, 2007.

[94] Advisory Committee Meeting for NDA 21071 Avandia (rosiglitazone maleate) tablet. 2010.

[95] Summary Minutes of the Joint Meeting of the Endocrinologic and Metabolic Drugs Advisory Committee and the Drug Safety and Risk Management Advisory Committee. *Toxicologic Pathology*, 35(2):1–14, 2010.

[96] Janet Woodcock, Joshua M Sharfstein, and Margaret Hamburg. Regulatory action on rosiglitazone by the US Food and Drug Administration. *New England Journal of Medicine*, 363(16):1489–1491, 2010.

[97] Janet Woodcock. Decision on continued marketing of rosiglitazone. Technical report, Center for Drug Evaluation and Research, 2010.

[98] Clifford J Rosen. The rosiglitazone story - lessons from an FDA Advisory Committee meeting. *New England Journal of Medicine*, 357(9):844–846, 2007.

[99] Clifford J Rosen. Revisiting the rosiglitazone story–lessons learned. *The New England Journal of Medicine*, 363(9):803–6, 2010.

[100] Summary Minutes of the Joint Meeting of the Endocrinologic and Metabolic Drugs Advisory Committee and the Drug Safety and Risk Management Advisory Committee. *Toxicologic Pathology*, 35(2):1–14, 2013.

[101] Mike Mitka. Panel recommends easing restrictions on rosiglitazone despite concerns about cardiovascular safety. *Journal of the American Medical Association*, 310(3):246–7, 2013.

[102] Jeffrey M Gimble, Claudius E Robinson, Xiyang Wu, Katherine A Kelly, Brenda R Rodriguez, Steven A Kliewer, Jurgen M Lehmann, and David C Morris. Peroxisome proliferator-activated receptor-gamma activation by thiazolidinediones induces adipogenesis in bone marrow stromal cells. *Molecular Pharmacology*, 50(5):1087–1094, 1996.

[103] Eric Wooltorton. Rosiglitazone (Avandia) and pioglitazone (Actos) and heart failure. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 166(2):219, 2002.

[104] Hannele Yki-Järvinen. Thiazolidinediones. *New England Journal of Medicine*, 351(11):1106–1118, 2004.

[105] Harlan M Krumholz. A perspective on the American Heart Association/American College of Cardiology science advisory on thiazolidinedione drugs and cardiovascular risks. *Circulation: Cardiovascular quality and outcomes*, 3(3):221–222, 2010.

[106] Sanjay Kaul, Ann F Bolger, David Herrington, Robert P Giugliano, and Robert H Eckel. Thiazolidinedione drugs and cardiovascular risks: a science advisory from the American Heart Association and American College of Cardiology Foundation. *Circulation*, 121(16):1868–1877, 2010.

[107] Ralf Gold, Ludwig Kappos, Douglas L Arnold, Amit Bar-Or, Gavin Giovannoni, Krzysztof Selmaj, Carlo Tornatore, Marianne T Sweetser, Minhua Yang, Sarah I Sheikh, and Katherine T Dawson. Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis. *New England Journal of Medicine*, 367(12):1098–1107, 2012.

[108] Robert J. Fox, David H. Miller, J. Theodore Phillips, Michael Hutchinson, Eva Havrdova, Mariko Kita, Minhua Yang, Kartik Raghupathi, Mark Novas, Marianne T. Sweetser, Vissia Viglietta, and Katherine T. Dawson. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *New England Journal of Medicineurnal of medicine*, 367(12):1087–1097, 2012.

[109] W Schweckendiek. Heilung von Psoriasis vulgaris [Treatment of psoriasis vulgaris]. *Medizinische Monatsschrift*, 13(2):103–104, 1959.

[110] Mir N Islam. Inhibition of Mold in Bread By Dimethyl Fumarate. *Journal of Food Science*, 47(5):1710–1712, 1982.

[111] Ana Gimenez-Arnau, Juan Francisco Silvestre, Pedro Mercader, Jesus la Cuadra, Isabel Ballester, Fernando Gallardo, Ramón M Pujol, Erik Zimerson, and Magnus Bruze. Shoe contact dermatitis from dimethyl fumarate: clinical manifestations, patch test results, chemical analysis, and source of exposure. *Contact Dermatitis*, 61(5):249–260, 2009.

[112] Dinanda N Kolbach and Cornells Nieboer. Fumaric acid therapy in psoriasis: Results and side effects of 2 years of treatment. *Journal of the American Academy of Dermatology*, 27(5, Part 1):769–771, 1992.

[113] Hok Bing Thio, Jan G van der Schroeff, Wieke M Nugteren-Huying, and B J Vermeer. Long-term systemic therapy with dimethylfumarate and monoethylfumarate (Fumaderm®) in psoriasis. *Journal of the European Academy of Dermatology and Venereology*, 4(1):35–40, 1995.

[114] Ulrich Mrowietz, Peter Altmeyer, Thomas Bieber, Martin Röcken, Rudolf E Schopf, and Wolfram Sterry. Treatment of psoriasis with fumaric acid esters (Fumaderm®). *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 5(8):716–717, 2007.

[115] Robert J Fox, Mariko Kita, Stanley L Cohan, Lily Jung Henson, Javier Zambrano, Robert H Scannevin, John O'Gorman, Mark Novas, Katherine T Dawson, and J Theodore Phillips. BG-12 (dimethyl fumarate): a review of mechanism of action, efficacy, and safety. *Current Medical Research and Opinion*, 30(2):251–262, 2013.

[116] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[117] Euan A Ashley. The precision medicine initiative: A new national effort. *Journal of the American Medical Association*, 313(21):2119–2120, 2015.

[118] Rui Chen and Michael Snyder. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):73–82, 2013.

[119] Clinical Trials Transformation Initiative. AACT Database.

[120] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

[121] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, 2005.

[122] Erin M Ramos, Douglas Hoffman, Heather a Junkins, Donna Maglott, Lon Phan, Stephen T Sherry, Mike Feolo, and Lucia a Hindorff. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics*, 22(1):144–7, 2014.

[123] Richard J Roberts. PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*, 98(2):381–382, 2001.

[124] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms - disease network. *Nature Communications*, 5:4212, 2014.

[125] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Lászlό Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, 2007.

[126] Wei-Qi Wei, Robert M Cronin, Hua Xu, Thomas a Lasko, Lisa Bastarache, and Joshua C Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–61, 2013.

[127] Jacques Bertillon. *International classification of causes of sickness and death / Revised by the International commission at the session of Paris , July 1 to 3 , 1909, for use begining January 1, 1910, and until December 31, 1919.* Government Printing office, 1910.

[128] National Center for Health Statistics (US) and Others. *The International Classification of Diseases: 9th Revision, Clinical Modification: ICD-9-CM.* 1991.

[129] Kevin Donnelly. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279–90, 2006.

[130] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2014). *National Cancer Institute, DCCPS, Surveillance Research Program*, 2017.

[131] N Howlader, AM Noone, M Krapcho, D Miller, K Bishop, CL Kosary, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Cronin, editors. *Cancer Statistics Review, 1975-2013 - SEER Statistics.* National Cancer Institute, Bethesda, MD, 2017.

[132] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(343):343, 2010.

[133] Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Research*, 36(Suppl_1):D684–D688, 2008.

[134] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Suppl_1):D901–D906, 2008.

[135] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Suppl_1):D1035–D1041, 2011.

[136] Helena Chmura Kraemer. The Reliability of Clinical Diagnoses: State of the Art. *Annual Review of Clinical Psychology*, 10(1):111–130, 2014.

[137] Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[138] Godfrey S. Getz and Catherine A. Reardon. Diet and murine atherosclerosis. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 26(2):242–249, 2006.

[139] Stephen H. Schneider, Angelica Vitug, and Neil Ruderman. Atherosclerosis and physical activity. *Diabetes/Metabolism Reviews*, 1(4):513–553, 1986.

[140] Aldons J. Lusis. Genetics of atherosclerosis. *Trends in Genetics*, 28(6):267–275, 2012.

[141] Amparo C. Villablanca, Muthuvel Jayachandran, and Carole Banka. Atherosclerosis and sex hormones: current concepts. *Clinical Science*, 119(12):493–513, 2010.

[142] Aldons J Lusis. Atherosclerosis. *Nature*, 407(6801):233–241, 2000.

[143] Guanghui Hu and Pankaj Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS ONE*, 4(8), 2009.

[144] Bryan Haslam and Luis Perez-breva. Learning disease relationships from clinical drug trials. *Journal of the American Medical Informatics Association*, 289(10):1–11, 2016.

[145] Daniel S. Chen and Ira Mellman. Oncology meets immunology: The cancer-immunity cycle. *Immunity*, 39(1):1–10, 2013.

[146] Adrian R Hill. Making decisions in ophthalmology. *Progress in Retinal Research*, 6:207–244, 1987.

[147] Mike A Nalls, Cory Y McLean, Jacqueline Rick, Shirley Eberly, Samantha J Hutten, Katrina Gwinn, Margaret Sutherland, Maria Martinez, Peter Heutink, Nigel M Williams, John Hardy, Thomas Gasser, Alexis Brice, T Ryan Price, Aude Nicolas, Margaux F Keller, Cliona Molony, J Raphael Gibbs, Alice Chen-Plotkin, Eunran Suh, Christopher Letson, Massimo S Fiandaca, Mark Mapstone, Howard J Federoff, Alastair J Noyce, Huw Morris, Vivianna M Van Deerlin, Daniel Weintraub, Cyrus Zabetian, Dena G Hernandez, Suzanne Lesage, Meghan Mullins, Emily Drabant Conley, Carrie A M Northover, Mark Frasier, Ken Marek, Aaron G Day-Williams, David J Stone, John P A Ioannidis, and Andrew B Singleton. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *The Lancet Neurology*, 14(10):1002–9, 2015.

[148] Elio Tonutti and Nicola Bizzaro. Diagnosis and classification of celiac disease and gluten sensitivity. *Autoimmunity Reviews*, 13(45):472–476, 2014.

[149] Nataliya Razumilava and Gregory J Gores. Classification, Diagnosis, and Management of Cholangiocarcinoma. *Clinical Gastroenterology and Hepatology*, 11(1):13–21, 2013.

[150] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423, 1948.

[151] Peter N. Robinson. Deep phenotyping for precision medicine. *Human Mutation*, 33(5):777–780, 2012.

[152] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics*, 133(1):e54–e63, 2014.

[153] Andrew R Joyce and Bernhard O Palsson. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.

[154] Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(Database issue):D136–43, 2012.

[155] GezaP. Balint, W.Watson Buchanan, and Jan Dequeker. A brief history of medical taxonomy and diagnosis. *Clinical Rheumatology*, 25(2):132–135, 2006.

[156] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

[157] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems.* World Health Organization, 2004.

[158] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, Gavin Sherlock, and The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[159] Georg Schett, Dirk Elewaut, Iain B Mcinnes, Jean-michel Dayer, and Markus F Neurath. How cytokine networks fuel inflammation. *Nature Medicine*, 19(7):822–826, 2013.

[160] Heather Christenson. Hathi Trust: A Research Library at Web Scale. *Library Resources & Technical Services*, 55(2):93–102, 2010.

[161] Jaime Huerta-Cepas, Joaquín Dopazo, and Toni Gabaldón. ETE: a python Environment for Tree Exploration. *BMC bioinformatics*, 11(1):24, 2010.

[162] Joan L Warren, Carrie N Klabunde, Deborah Schrag, Peter B Bach, and Gerald F Riley. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical Care*, 40(8 Suppl):IV–3–18, 2002.

[163] James B Yu, Cary P Gross, Lynn D Wilson, and Benjamin D Smith. NCI SEER Public-Use Data: Applications and Limitations in Oncology Research. *Oncology*, 23:288–95, 2009.

[164] National Library of Medicine. Rare Books and Journals (https://www.nlm.nih.gov/hmd/collections/books/), 2017.

[165] Hans Lilja, David Ulmert, and Andrew J. Vickers. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring (Corrigendum). *Nature Reviews Cancer*, 8(5):403–403, 2008.

[166] Olivera J Finn. Cancer Immunology. *New England Journal of Medicine*, 358(25):2704–2715, 2008.

[167] Charles E Rosenberg. Disease in history: frames and framers. *The Milbank Quarterly*, 67(Suppl 1):1–15, 1989.

[168] Emma Wicks. Cystic Fibrosis. *British medical journal*, 334(7606):1270–1271, 2007.

[169] Alexa T. McCray and Nicholas C. Ide. Design and Implementation of a National Clinical Trials Registry. *Journal of the American Medical Informatics Association*, 7(3):313–323, 2000.

[170] Michael S. Lauer and Ralph B. D'Agostino. The Randomized Registry Trial - The Next Disruptive Technology in Clinical Research. *New England Journal of Medicine*, 369(17):1579–1581, 2013.

[171] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang Zhong Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[172] Mihail Popescu and Dong Xu. *Data Mining in Biomedicine Using Ontologies.* Artech House, 2009.

[173] Steffen Staab and Rudi Studer. *Handbook on Ontologies.* Springer, 2004.

[174] Anima Singh, Girish Nadkarni, John Guttag, and Erwin Bottinger. Leveraging hierarchy in medical codes for predictive modeling. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 96–103, 2014.

[175] Atul J Butte and Isaac S Kohane. Creation and implications of a phenome-genome network. *Nature Biotechnology*, 24(1):55–62, 2006.

[176] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, Melissa A Basford, David S Carrell, Peggy L Peissig, Abel N Kho, Jennifer A Pacheco, Luke V Rasmussen, David R Crosslin, Paul K Crane, Jyotishman Pathak, Suzette J Bielinski, Sarah A Pendergrass, Hua Xu, Lucia A Hindorff, Rongling Li, Teri A Manolio, Christopher G Chute, Rex L Chisholm, Eric B Larson, Gail P Jarvik, Murray H Brilliant, Catherine A McCarty, Iftikhar J Kullo, Jonathan L Haines, Dana C Crawford, Daniel R Masys, and Dan M Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1110, 2013.

[177] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629, 2016.

[178] Marc Vidal, Michael E Cusick, and Albert-Laszlo Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[179] Sir George Knibbs. History of the development of the ICD (http://www.who.int/classifications/icd/en/HistoryOfICD.pdf). *World Health Organization.*

[180] James R. Egner. AJCC Cancer Staging Manual. *Journal of the American Medical Association*, 304(15):1726, 2010.

[181] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue):D115–9, 2004.

[182] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808, 2006.

[183] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd Annual meeting of the association for computational linguistics*, pages 133–138, Las Cruces (New Mexico), 1994.

[184] Linton C Freeman. Centrality in Social Networks. *Social Networks*, 1(1968):215–239, 1978.

[185] Philip Resnik. Using Information Content to Evalutate Semantic Similarity in a Taxonomy. *14th International Joint Conference on Artificial Intelligence, IJCAI 1995*, 1:448–453, 1995.

[186] Philip Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artiicial Intelligence Research*, 11(3398):95–130, 1999.

[187] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the International Conference on Machine Learning*, pages 296–304, 1998.

[188] Liu Yang and R Jin. Distance metric learning: A comprehensive survey. *http://www.cse.msu.edu/~yangliu1/frame_survey_v2.pdf*, 2006.

[189] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[190] Alexander Kraskov, Harald Stögbauer, Ralph G Andrzejak, and Peter Grassberger. Hierarchical clustering using mutual information. *Europhysics Letters (EPL)*, 70(2):278–284, 2005.

[191] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 76–83, 1989.

[192] Paul Viola and William M Wells. Alignment by maximization of mutual information. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 16–23. IEEE, 1995.

[193] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[194] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[195] George Hripcsak and Adam S. Rothschild. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005.

[196] Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of German Society for Computational Linguistics (GSCL 2009)*, pages 31–40, 2009.

[197] Panagiotis Peitsidis, T. Datta, I. Pafilis, O. Otomewo, Edward G Tuddenham, and Rezan A Kadir. Bernard Soulier syndrome in pregnancy: A systematic review. *Haemophilia*, 16(4):584–591, 2010.

[198] Rebecca F. Rosen, Aaron S. Farberg, Marla Gearing, Jeromy Dooyema, Patrick M. Long, Daniel C. Anderson, Jeremy Davis-Turak, Giovanni Coppola, Daniel H. Geschwind, Jean Francois Paré, Timothy Q. Duong, William D. Hopkins, Todd M. Preuss, and Lary C. Walker. Tauopathy with paired helical filaments in an aged chimpanzee. *Journal of Comparative Neurology*, 509(3):259–270, 2008.

[199] Renee L. Cohen, Robert E. Tepper, Carlos Urmacher, and Seymour Katz. Kaposi's sarcoma and cytomegaloviral ileocolitis complicating long-standing Crohn's disease in an HIV-negative patient. *American Journal of Gastroenterology*, 96(10):3028–3031, 2001.

[200] Friedman Scott L, Teresa L Wright, and David F Altman. Gastrointestinal Kaposi's sarcoma in patients with acquired immunodeficiency syndrome. Endoscopic and autopsy findings. *Gastroenterology*, 89(1):102–108, 1985.

[201] Angela M D'Alessandro and Gregory J Mulford. An unusual case of electromyrographic recorded myokymic potentials: A case report. *Archives of Physical Medicine and Rehabilitation*, 83(5):727–729, 2002.

[202] Nancy S Sung, Crowley William F, Myron Genel, Patricia Salber, Lewis Sandy, Louis M Sherwood, Stephen B Johnson, Veronica Catanese, Hugh Tilson, Kenneth Getz, Elaine L Larson, David Scheinberg, E. Albert Reece, Harold Slavkin, Adrian Dobs, Jack Grebb, Rick A Martinez, Allan Korn, and David Rimoin.

Central challenges facing the national clinical research enterprise. *Journal of the American Medical Association*, 289(10):1278–1287, 2003.

[203] Christopher Paul Adams and Van Vu Brantner. Spending on new drug development. *Health Economics*, 19(2):130–141, 2010.

[204] Jeffrey M. Drazen. Sharing Individual Patient Data from Clinical Trials. *New England Journal of Medicine*, 372(3):201, 2015.

[205] Deborah A. Zarin. Participant-Level Data and the New Frontier in Trial Transparency. *New England Journal of Medicine*, 369(5):468–469, 2013.

[206] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[207] Cesar A Hidalgo, Nicholas Blumm, Albert-Laszlo Barabasi, and Nicholas A Christakis. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology*, 5(4):e1000353, 2009.

[208] Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. Drug-target network. *Nature Biotechnology*, 25(10):1119–1126, 2007.

[209] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Jensen, and Peer Bork. Drug Target Identification Using Side-Effect Similarity. *Science*, 321(2008):263–266, 2008.

[210] Albert-Laszlo Barabasi, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2011.

[211] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

[212] Gerhard Klebe, Ute Abraham, and Thomas Mietzner. Molecular Similarity Indexes in A Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. *Journal of Medicinal Chemistry*, 37(24):4130–4146, 1994.

[213] Camille G. Wermuth. Similarity in drugs: reflections on analogue design. *Drug Discovery Today*, 11(7-8):348–354, 2006.

[214] ClinicalTrials.gov.

[215] Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J. McCourt, and Ricardo Pietrobon. The database for aggregate analysis of Clinicaltrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS ONE*, 7(3), 2012.

[216] Robert M. Califf. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010. *Journal of the American Medical Association*, 307(17):1838, 2012.

[217] Bradford R Hirsch, Robert M Califf, Steven K Cheng, Asba Tasneem, John Horton, Karen Chiswell, Kevin A Schulman, David M Dilts, and Amy P Abernethy. Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. *Journal of the American Medical Association Internal Medicine*, 173(11):972–9, 2013.

[218] Jula K. Inrig, Robert M. Califf, Asba Tasneem, Radha K. Vegunta, Christopher Molina, John W. Stanifer, Karen Chiswell, and Uptal D. Patel. The landscape of clinical trials in nephrology: A systematic review of clinicaltrials.gov. *American Journal of Kidney Diseases*, 63(5):771–780, 2014.

[219] Menghua Wu, Marina Sirota, Atul J Butte, and Bin Chen. Characteristics of drug combination therapy in oncology by analyzing clinical trial data on ClinicalTrials.gov. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 68–79, 2015.

[220] Zhe He, Simona Carini, Tianyong Hao, Ida Sim, and Chunhua Weng. A Method for Analyzing Commonalities in Clinical Trial Target Populations. In *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 1777–1786, 2014.

[221] Tianyong Hao, Alexander Rusanov, Mary Regina Boland, and Chunhua Weng. Clustering clinical trials with similar eligibility criteria features. *Journal of Biomedical Informatics*, 52:112–120, 2014.

[222] Carolyn E Lipscomb. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266, 2000.

[223] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[224] Deborah A. Zarin and A Keselman. Registering a Clinical Trial in ClinicalTrials.gov. *Chest*, 131(3):909–912, 2007.

[225] Albert-Laszlo Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[226] Philip J Clark and Francis C Evans. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecological Society of America*, 35(4):445–453, 2012.

[227] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative Filtering Recommender Systems. In *Lecture Notes in Computer Science*, pages 291–324. Springer, Berlin, 2007.

[228] AP Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[229] Bernard V North, David Curtis, and Pak C Sham. A Note on the Calculation of Empirical P-Values from Monte Carlo Procedures. *American Journal of Genetics*, 71(2):439–441, 2002.

[230] Paul Brassard, Maria Vutcovici, Pierre Ernst, Valerie Patenaude, Maida Sewitch, Samy Suissa, and Alan Bitton. Increased incidence of inflammatory bowel disease in Québec residents with airway diseases. *European Respiratory Journal*, 45(4):962–968, 2015.

[231] KP Murphy, Y Weiss, and Michael I. Jordan. Loopy Belief Propagation for Approximate Inference: An Empiricial Study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 467–475, San Francisco, CA, 1999.

[232] Igor Kononenko. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[233] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, 2008.

[234] Hector O. Ventura and Mandeep R. Mehra. Bloodletting as a cure for dropsy: Heart failure down the ages. *Journal of Cardiac Failure*, 11(4):247–252, 2005.

[235] Brian P O'Sullivan and Steven D Freedman. Cystic fibrosis. *Lancet*, 373(9678):1891–904, 2009.

[236] Steven M Rowe, Stacey Miller, and Eric J Sorscher. Cystic fibrosis. *New England Journal of Medicine*, 352:1992–2001, 2005.

[237] Tiinamaija Tuomi, Nicola Santoro, Sonia Caprio, Mengyin Cai, Jianping Weng, and Leif Groop. The many faces of diabetes: A disease with increasing heterogeneity. *The Lancet*, 383(9922):1084–1094, 2014.

[238] Lauren P. Wadsworth, Natacha Lorius, Nancy J. Donovan, Joseph J. Locascio, Dorene M. Rentz, Keith A. Johnson, Reisa A. Sperling, and Gad A. Marshall. Neuropsychiatric symptoms and global functional impairment along the Alzheimer's continuum. *Dementia and Geriatric Cognitive Disorders*, 34(2):96–111, 2012.

[239] Elie Dolgin. Big pharma moves from 'blockbusters' to 'niche busters'. *Nature medicine*, 16(8):837, aug 2010.

[240] M. Miles Braun, Sheiren Farag-El-Massah, Kui Xu, and Timothy R. Coté. Emergence of orphan drugs in the United States: a quantitative assessment of the first 25 years. *Nature Reviews Drug Discovery*, 2010.

[241] Irena Melnikova. Rare diseases and orphan drugs. *Nature review drug discovery*, 11:267–268, 2012.

[242] Dan L. Longo and Jeffrey M. Drazen. Data Sharing. *New England Journal of Medicine*, 374(3):276–277, 2016.

[243] Jeffrey M. Drazen. Data Sharing and the Journal. *New England Journal of Medicine*, 374(19):e24, 2016.

[244] Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, 2011.

[245] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[246] Dimitri P. Bertsekas. *Convex Analysis and Optimization.* Athena Scientific, 2003.

[247] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[248] James C Spall. *Introduction to Stochastic Search and Optimization.* Wiley, 2003.

[249] Paul E. Utgoff, Neil C Berkman, and Jeffery A Clouse. Decision Tree Induction Based on Efficient Tree Restructuring. *Machine Learning*, 29(1):5–44, 1997.

[250] Lior Rokach and Oded Maimon. Top-Down Induction of Decision Trees ClassifiersA Survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.

[251] Rachna Shah and Peter T. Ward. Lean manufacturing: Context, practice bundles, and performance. *Journal of Operations Management*, 21(2):129–149, 2003.

[252] Merriam Webster Dictionary Online, 2017.

[253] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[254] Vladimir N. Vapnik. *The Nature of Statistical Learning.* Springer, 1995.

[255] Richard Kempter, Wulfram Gerstner, and J. Leo Van Hemmen. Hebbian learning and spiking neurons. *Physical Review E*, 59(4):4498–4514, 1999.

[256] Yuriy V. Pershin and Massimiliano Di Ventra. Experimental demonstration of associative memory with memristive neural networks. *Neural Networks*, 23(7):881–886, 2010.

[257] Xiaojin Zhu and Andrew B Goldberg. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.

[258] Raja Giryes and Michael Elad. Reinforcement Learning: A Survey. *European Signal Processing Conference*, pages 1475 – 1479, 2011.

[259] Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics : A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2015.

[260] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[261] Vladimir N Vapnik. *Statistical Learning Theory*, volume 2. John Wiley & Sons, Inc., 1998.

[262] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Li. *Applied Linear Statistical Models.* Chicago, IL, 1996.

[263] John A. Rice. *Mathematical Statistics and Data Analysis.* 2001.

[264] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[265] Andrea Bonarini. An Introduction to Learning Fuzzy Classifier Systems. In *Learning Classifier Systems. From Foundations to Applications*, pages 83–106, 2000.

[266] Trevor Hastie. Regularization Paths. *Statistics*, 33(1):1–22, 2006.

[267] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer, 2001.

[268] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[269] Jonathan Baxter. Theoretical Models of Learning to Learn. In *Learning to Learn*, chapter 4. Kluwer Academic Publishers, 1998.

[270] Jonathan Baxter. Learning model bias. *Advances in neural information processing systems*, 8:169–175, 1996.

[271] Rodney A. Brooks. Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1-2):3–15, 1990.

[272] Philip Bille. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3):217–239, 2005.

[273] Kaizhong Zhang. A constrained edit distance between unordered labeled trees. *Algorithmica*, 15(3):205–222, 1996.

[274] Tatsuya Akutsu, Daiji Fukagawa, Atsuhiro Takasu, and Takeyuki Tamura. Exact algorithms for computing the tree edit distance between unordered trees. *Theoretical Computer Science*, 412(4-5):352–364, 2011.

[275] Mateusz Pawlik and Nikolaus Augsten. RTED : A Robust Algorithm for the Tree Edit Distance. *The 38th International Conference onVery Large Data Bases*, pages 334–345, 2011.

[276] Ian Roberts and Emma Sydenham. Barbiturates for acute traumatic brain injury. *The Cochrane Library*, 1999.

[277] Vojtech Huser and James J Cimino. Evaluating adherence to the International Committee of Medical Journal Editors' policy of mandatory, timely clinical trial registration. *Journal of the American Medical Informatics Association*, 20(e1):e169 e174, 2013.

[278] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Journal of Soviet Physics Doklady*, 10(8):707–709, 1966.

[279] JW Ratclif and D Metzener. Pattern Matching: the Gestalt Approach. *Dr. Dobb's Journal*, 46, 1988.

[280] Stephen G Kobourov. Force-directed drawing algorithms. *Handbook of Graph Drawing and Visualization (Discrete Mathematics and Its Applications)*, pages 383–408, 2013.

[281] Thomas M J Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[282] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[283] Peter A Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.

[284] Mathieu Jacomy, Tommaso Venturini, Sebastien Hermann, and Mathieu Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6):e98679, 2014.

[285] James D Mills, Kevin Hadley, and Julian E Bailes. Dietary supplementation with the omega-3 fatty acid docosahexaenoic acid in traumatic brain injury. *Neurosurgery*, 68(2):474–481, 2011.

[286] Maiko Inoue, Akira Arakawa, Shin Yamane, and Kazuaki Kadonosono. Short-Term Efficacy of Intravitreal Aflibercept in Treatment-Naive Patients With Polypoidal Choroidal Vasculopathy. *Retina*, 34(11):2178–2184, 2014.

[287] Robert D. Tunks, Megan E B Clowse, Stephen G. Miller, Leo R. Brancazio, and Piers C A Barker. Maternal autoantibody levels in congenital heart block and potential prophylaxis with antiinflammatory agents. *American Journal of Obstetrics and Gynecology*, 208(1):64.e1–7, 2013.

[288] Susan L. McElroy, Anna I. Guerdjikova, Nicole Mori, and Anne M. O'Melia. Pharmacological management of binge eating disorder: Current and emerging treatment options. *Therapeutics and Clinical Risk Management*, 8:219–241, 2012.

[289] Kunihiro Hosono, Hiroki Endo, Hirokazu Takahashi, Michiko Sugiyama, Eiji Sakai, Takashi Uchiyama, Kaori Suzuki, Hiroshi Iida, Yasunari Sakamoto, Kyoko Yoneda, Tomoko Koide, Chikako Tokoro, Yasunobu Abe, Masahiko Inamori, Hitoshi Nakagama, and Atsushi Nakajima. Metformin suppresses colorectal aberrant crypt foci in a short-term clinical trial. *Cancer Prevention Research*, 3(9):1077–1083, 2010.

[290] Song I. Yang, Wook Jin Chung, Sung Hwan Jung, and Deok Young Choi. Effects of inhaled iloprost on congenital heart disease with eisenmenger syndrome. *Pediatric Cardiology*, 33(5):744–748, 2012.

[291] Robbie Pesek and Roger Fox. Successful Treatment of Scnitzler Syndrome With Canakinumab. *Cutis*, 94(3):E11–E12, 2014.

[292] Steven Vanderschueren and Daniël Knockaert. Canakinumab in Schnitzler Syndrome. *Seminars in Arthritis and Rheumatism*, 42(4):413–416, 2013.