# Predicting Controlled Vocabulary Based on Text and Citations: Case Studies in Medical Subject Headings in MEDLINE and Patents

BY
ADAM KEHOE

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the School of Information Sciences in the Graduate College of the University of Illinois at Urbana-Champaign, 2019
Urbana, Illinois

Doctoral Committee:
Dr. Vetle I. Torvik, Associate Professor (Chair, Research Director)
Dr. J. Stephen Downie, Associate Dean for Research
Dr. David S. Dubin, Research Associate Professor
Dr. Bertram Ludäscher, Professor
Dr. Neil R. Smalheiser, Professor, University of Illinois at Chicago College of Medicine

# Abstract

This dissertation makes three contributions in the area of controlled vocabulary prediction of Medical Subject Headings. The first contribution is a new partial matching measure based on distributional semantics. The second contribution is a probabilistic model based on text similarity and citations. The third contribution is a case study of cross-domain vocabulary prediction in US Patents. Medical subject headings (MeSH) are an important life sciences controlled vocabulary. They are an ideal ground to study controlled vocabulary prediction due to their complexity, hierarchical nature, and practical significance. The dissertation begins with an updated analysis of human indexing consistency in MEDLINE. This study demonstrates the need for partial matching measures to account for indexing variability. Here, I develop a measure combining the MeSH hierarchy and contextual similarity. This measure provides several new tools for evaluating and diagnosing controlled vocabulary models. Next, a generalized predictive model is introduced. This model uses citations and abstract similarity as inputs to a hybrid KNN classifier. Citations and abstracts are found to be complimentary in that they reliably produce unique and relevant candidate terms. Finally, the predictive model is applied to a corpus of approximately 65,000 biomedical US patents. This case study explores differences in the vocabulary of MEDLINE and patents, as well as the prospect for MeSH prediction to open new scholarly opportunities in economics and healthy policy research.

**DRAFT**

I would first like to thank my supervisor, Dr. Vetle Torvik for his patient help and guidance in this project. His support, both practical and moral, have been invaluable. Dr. Torvik's lecture on literature based discovery in my first course on data mining profoundly changed the trajectory of my career. The help of the committee has also made a major impact in my work. Dr. Dubin was incredibly gracious with his time and advice at multiple stages of this project. Dr. Smalheiser's pioneering work in literature based discovery has been a source of inspiration to me from the beginning of my studies in information science. Dr. Downie and Dr. Ludaescher have both provided valuable insights that helped focus my research strategy. I owe a debt of gratitude to Dr. Les Gasser. His incredible insight, creativity, and kindness are sorely missed. Finally, I would like to thank my wife, Kamila Kehoe. This dissertation would not have been possible without her support.

# Contents

# Chapter 1

# Introduction

# Introduction

This dissertation is focused on the automatic prediction of controlled vocabulary, specifically Medical Subject Headings. The following chapters are structured around three contributions:

1. First, this dissertation describes a longstanding problem in the evaluation of controlled vocabulary prediction models, namely an over-reliance on exact matching. The evaluation chapter updates an existing study on the indexing consistency of human annotators. This study also provides a reference dataset for the development of new partial matching measures that quantitatively capture broad relationships between terms based on their distributional semantics. These measures are designed to supplement existing evaluation metrics.

2. Second, this dissertation presents a predictive model that extracts extracts and ranks candidate terms from related records using citations and abstract similarity. The primary finding is that citations and text are complimentary and typically have high recall of the original MeSH. This model is designed to be highly general and applicable to a wide array of bibliographic databases outside of MEDLINE.

3. Third, the predictive model described above is applied in a case study of US patents. This case study demonstrates the application of automated MeSH prediction beyond MEDLINE, and discusses potential uses of MeSH as an information retrieval and policy analysis tool.

The Medical Subject Headings (MeSH) are at the center of my experiments. Arguably, MeSH is the most important controlled vocabulary in the life sciences. Since 1960, the National Library of Medicine (NLM) has applied MeSH to millions of scientific publications[39]. In that time, the NLM has continuously revised MeSH based on scientific developments. The vocabulary spans a wide variety of topics, including many outside of medicine and biology. Beyond scientific papers, MeSH is also used to index several NLM databases, including clinical trials[39]. In short, MeSH is a widely used, practically significant, and complex controlled vocabulary.

One of the practical motivations of this dissertation is that MeSH is not available for many important life science corpora. This dissertation explores potential applications of MeSH to the patent literature in a case study, and describes the significant limitations of patent controlled vocabularies for the life sciences. The wider goal of this work is to develop a flexible MeSH model that can be applied to biomedical documents, generally. This goal informs a modeling strategy that focuses on understanding widely available document features, namely citations and abstract text.

Beyond improving access to these corpora, MeSH indexing could also empower a range of new scholarship. For example, the hierarchical nature of MeSH enables aggregation by higher level topics. This is particularly useful in policy analysis, where control over the level of topical granularity is helpful in tracking the flow of research investments over time[27]. More broadly, controlled vocabulary could be a useful tool for ”science of science” studies, including the patent space[27]. For example, a common MeSH index could help locate connections between academic research and commercial innovation. MeSH is also a helpful tool in many literature based discovery paradigms. In short, extending MeSH beyond MEDLINE would have a number of practical and scientific benefits.

While MeSH is a highly valuable resource, it is also expensive to apply, especially in domains outside of MEDLINE. Manual annotation is costly, and requires wide expertise in both the life sciences and complex indexing practices. The value of MeSH and the high cost of manual indexing has inspired many automated MeSH prediction efforts and competitions. The wide range of MeSH prediction strategies are reviewed in the following chapter.

Past efforts in this area have met with modest success. MeSH prediction involves a combination of special characteristics that make it a challenging machine learning task. MeSH prediction is best described as a hierarchical multilabel classification (HMLC) problem. The formal details of this problem are discussed in greater detail below. In short, unlike binary or multiclass tasks, HMLC predicts a set rather than a single class. A further complication is that MeSH is structured in a hierarchy, where a term may belong to more than one branch. The MeSH vocabulary is large, with many terms that are closely related to each other. The issue of redundancy and conceptual similarity between terms is explored in depth in both the evaluation and modeling chapter.

In terms of modeling, approaches to HMLC problems have historically pursued one of two strategies. In the first, the data is simplified to fit established machine learning algorithms. For example, the multilabel problem can be decomposed into a large set of binary classification tasks. These approaches have the advantage of simplifying the problem, but come at the cost of discarding important information about relationships between the labels. The other strategy involves creating new algorithms that are natively capable of predicting sets. These approaches typically make good use of label relationships, at the cost of tractability and increased risk of overfitting. These approaches are summarized in detail in the following chapter.

The MeSH prediction literature has prominent examples of both strategies. This dissertation seeks to address a common limitation in the existing literature: a narrow focus on performance measures using exact term matching. This focus is understandable, as common, unambiguous performance measures are central to machine learning scholarship. However, this focus has obscured an underlying evaluation problem.

The dissertation begins with experiments comparing human indexing consistency in a set of papers that were inadvertently indexed twice. This study demonstrates, unsurprisingly, that annotators frequently make different vocabulary choices. The important consequence of this rather obvious observation is that performance measures based on exact matching can provide only limited insight into model performance. If we were to treat the duplicate human annotations as the prediction of a model, we would often conclude that the performance of the model is poor – despite the fact that in other circumstances we would happily treat those annotations as a gold standard. This highlights the crucial importance of partial matching. Here, we immediately see that the duplicate annotations are generally similar. However, most existing approaches to partial matching rely heavily on the MeSH hierarchy. This dissertation presents a new approach to partial matching that leverages the shared context of terms in addition to the hierarchy. This permits a flexible, continuous measure of similarity between terms that augments existing graph-based techniques.

The third general area of focus is in cross-domain prediction. MeSH prediction scholarship has almost exclusively focused on MEDLINE. This dissertation focuses on the development of a more general classification strategy that can predict MeSH for a wide range of scientific documents. I begin with the observation that biomedical documents tend to have similar features: titles, an abstract-like summary, and citations to the scientific literature. As such, this dissertation is strongly focused on elucidating the relationship between citations, text similarity and candidate MeSH. The scientific objective here is not to optimize prediction accuracy so much as to understand how these document features function. Such an understanding is a vital basis to develop a principled, empirical basis for optimization in the future. The fourth chapter develops a classification model that applies a hybrid K Nearest Neighbors strategy that avoids classic pitfalls in multilabel classification.

## Research Questions

The following questions summarize my research objectives. A brief explanation of the question is provided, along with pointers to where the question is answered.

## 1.1 RQ1: Given that human inter-rater reliability is modest, how should MeSH prediction systems evaluate accuracy?

The MeSH vocabulary is large and complex, and each paper has a varying number of labels. Previous studies have found that inter-rater reliability is modest, at about 50%[18]. Evaluations based only on strict matching miss valid assignments. They also overly reward high performance on common terms. How can prediction systems quantitatively differentiate between true errors and partial matches? Chapter 3 establishes a framework for the later modeling chapter. This framework develops a new measure of MeSH similarity based on term context within MEDLINE. This approach also leverages the MeSH hierarchy for partial matching.

## 1.2 RQ2A: Are abstracts and citations effective features for predicting medical subject headings in MEDLINE?

Most biomedical documents have abstract-like text, and many have direct citations to MEDLINE. These links can provide candidate MeSH terms, even if the original document is not in MEDLINE. Can these candidate terms be effectively ranked? How well does this approach predict MeSH terms within MEDLINE? Chapter 4 describes the performance of a classification model based on abstracts and citations.

## 1.3 RQ2B: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

The ubiquity of citations and abstracts make them an attractive source of data. But, do they contain complimentary information? Are there significant differences in their candidate term sets? If so, what is the underlying source of this complimentarity? Chapter 4 describes the differences between abstracts and citations in terms of capture rate and temporal span.

## 1.4 RQ3: How do MeSH terms in MEDLINE compare to predicted MeSH in USPTO patents?

MeSH prediction is straightforward to assess in MEDLINE because there are many labeled examples. However, there are many domains outside of the scientific literature that could benefit from MeSH annotation. How do the distributions of MeSH compare between MEDLINE and USPTO patents? Chapter 5 details a case study of MeSH prediction in USPTO patents.

# Chapter 2

# Background and Literature Review

## 2.1 Biomedical Controlled Vocabularies: Medical Subject Headings and Patent Classification Codes

Biomedical controlled vocabularies are at the core of this dissertation. Different scholarly communities and stakeholders have varying views of controlled vocabulary. Some see them as practical tools for information retrieval, and are primarily concerned with their construction and application. The machine learning perspective of controlled vocabulary necessarily involves reducing it to a mathematical structure that is compatible with a classification algorithm. Another perspective views controlled vocabulary as a kind of artificial language. Each of these perspectives contribute important insights that are explored in this dissertation.

### 2.1.1 What is a Controlled Vocabulary?

Each of the three perspectives detailed above describe and define controlled vocabularies differently. Perhaps the simplest perspective is mathematical. From this point of view, a controlled vocabularies is a set consisting of possible lexical units as elements. The annotations of any given document are a subset of the vocabulary itself. These sets can utilize more complex structures, as with medical subject headings, where terms are arranged in a graph. These graph structures commonly take the form of a directed acyclic graph or tree. Other structures, such as semantic triples, can be used to indicate subject-predicate-object relationships. Further distinctions can be made in terms of designating major and minor terms. Many of these complexities are described in detail below as they relate to specific prediction challenges. For the purposes of introduction, it suffices to say that controlled vocabularies can be viewed as a discrete mathematical structure comprised of distinct lexical units.

Most machine learning approaches to vocabulary prediction happily disregard the history, context and purpose of the target vocabulary in favor of a concise mathematical description. Such mathematical descriptions form the foundations of algorithms and other formal methods for manipulating controlled vocabularies. However, reducing controlled vocabulary to the simplest possible discrete structure misses the richness of a view of controlled vocabulary as a living, if stylized, language. It would be as if computational linguistics viewed text as mere lexicology – inert arrays of tokens, devoid of grammar or higher order structure.

A more nuanced view of controlled vocabularies is that they are a form of ”documentary language,” induced over a set of documents[9]. At a basic level, the controlled vocabulary can help reduce the cardinality of potential search terms. More importantly, it also locates a document within a coherent system of knowledge. In other words, the relationship of terms to each other is as important as the relationship of terms to a particular document.

This dissertation is mostly concerned with medical subject headings, a large biomedical vocabulary that has been in continual use over millions of documents for nearly sixty years at the time of writing. Most prediction efforts view the collective body of MeSH assignments as training data, useful for inducing a classification function for new documents. This a productive viewpoint, and many of the experiments below follow in this tradition. But the millions of annotated documents represent more than just assignments of terms to a document. They also represent expert judgments about which terms belong *together*. One of the central ideas of this dissertation is that it is possible to marry the machine learning view (controlled vocabulary as inert discrete structure) and the ”documentary language” view (controlled vocabulary as artificial language). In other words, techniques of computational linguistics and distributional semantics can be useful in recovering patterns in the relationships between terms. Cumulatively, the structure of MeSH as well as the living practice of annotation, form a kind of language of medicine.

To recap, controlled vocabulary at an explicit level is a simple discrete structure, comprised of individual terms. Though the specific arrangement of this structure may vary (simple sets vs. hierarchical graph), they remain at root a list of terms. At an implicit level, the specific terms in the vocabulary and their application to a body of documents arguably represent a particular system of knowledge. One of the major questions of this dissertation is if this system of knowledge can be systematically recovered and quantified. Later sections return to this idea by using computational linguistics tools, namely distributional semantics, to develop quantitative means of recovering deeper semantic structures implicit in controlled vocabulary assignments. These models represent terms as a high-dimensional vector space, opening up many useful mathematical tools.

This rather conceptual view of controlled vocabulary might be made clearer with a metaphor. The game of chess can be defined by a board, a set of pieces, and a set of permissible operations (legal moves) over those pieces. But, chess is more than its rules – we distinguish the higher order strategy of chess from the power set of legal moves. Similarly, a controlled vocabulary can be defined atomically as a set of terms, their configuration (flat list, hierarchy, etc), and rules for their assignment. But the complexity of controlled vocabulary is only appreciable over time, in their large scale assignment. The higher order structure of controlled vocabulary only emerges through application, or intelligent ”play” in the chess metaphor. This dissertation necessarily engages with the basic definition and mechanics of MeSH, but it is also concerned with a statistical exploration of the ”moves” made by countless annotators over its 60 year history. The cornerstone of this approach is appreciating that the context and relationships between terms to each other is as much valuable as data as the mapping of particular terms to particular documents.

### 2.1.2 What Problem Does MeSH Solve?

To review: a controlled vocabulary is a set of authoritative vocabulary terms reflecting a larger system of knowledge. What problem does this solve? The initial goal of the Medical Subject Headings was to create a unified index for searching the biomedical literature that would be both simple for users and efficient for the National Library of Medicine to maintain and use[39]. The desirability of such a unified index was juxtaposed with the difficulty of reconciling conceptual ambiguity from the outset. The authors of the Medical Subject Headings quoted the following from Swanson (1959) in their organizing principles:

> [Medical information] has been drawn from such a wide span of time and such a diversity of specialized fields that its doctrines belong to several different systems and its language problem is almost as bad as that of India. There is at least one major language for each major department, and each of these has several dialects. The situation is made even worse because in each language we teach a mixture of doctrines which range from Newtonian absolutism to Einsteinian relativism, including additive, reciprocal, exponential, and circular structures. It is tragic to contemplate the amount of effort we now waste because of our conflicting doctrines, and intriguing to wonder to what heights we might soar, each in his own way, once we manage to resolve the internal contradictions in the system by which we live and work.[39]

In response to this problem, the organizing principles go on to state:

> There will be less frustration on the part of librarians and other users of catalogs, indexes, and bibliographies if it is realized that the complexities of the field are such that simple, unequivocal solutions to the problem of the form and substance of medical subject headings are not easy to find...[f]rom one point of view, subject headings may be looked at as an *artificial language* which bears only superficial resemblances to the natural language. Subject headings are more stilted, more stereotyped. From another point of view, if subject headings conceived of as pointers, rather than as labels, a certain amount of ambivalence is tolerable.[39]

Here, we see the recognition that controlled vocabulary is more than a set of terms, but rather an ”artificial language,” in of itself, even if ”stereotyped.” Rather than attempting to untangle the deep complexities and contradictions of scientific knowledge, the creators of MeSH sought to create useful ”pointers” that could be widely understood by users. Ambiguity would be unavoidable, but the practical usefulness of the vocabulary would overcome the inevitable faults in its internal consistency or descriptiveness.

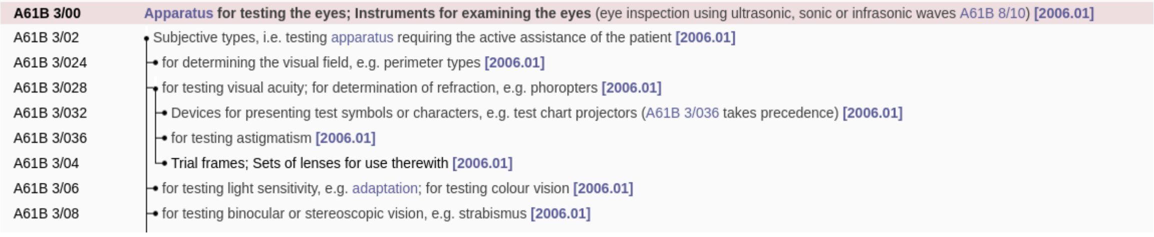
In other words, MeSH was not designed to solve the fragmentation of scientific language, though it was a tempting prospect, so much as to manage it. Just as natural language is necessarily an incomplete description of reality, the artificial language of controlled vocabulary is an incomplete description of a conceptual landscape. Also as in natural language, arguably what is unsaid is often as important as what is said.

These conceptual underpinnings are important and under-appreciated in the literature of machine learning prediction of controlled vocabulary. The formal tradition of machine learning has emphasized controlled vocabulary as a static set of terms and rules. It has been less appreciative that a controlled vocabulary can be a kind of language. This distinction is important both in terms of modeling and particularly in evaluation. For example, the current paradigm asks the question of which model most accurately recovers the exact individual terms applied to a given paper. Another way to see the problem is: which model best expresses the aggregate meaning of the assigned terms? For example, a model that accurately predicts species terms but never predicts substances may have high precision in some parts of the literature. Such a model is intuitively less desirable than a model that returns all branches with approximately correct terms. The second model, while perhaps less faithful to the individual "words," better recovers the "sentence."

The challenge for formal methods in this area is precisely the problem of measuring "meaning." The problem of meaning is formidable, ancient, and deeply philosophical. No solution is offered here. However, in the simplified realm of controlled vocabulary, the problem is perhaps better reframed as the problem of partial matching. Most quantitative evaluation methodologies are variations on the theme of counting how many predicted terms match "true terms." Such an approach necessarily reduces the measure of a language to its vocabulary. As stated above, one way to resolve this problem is to map the relationships of terms to each other based on their historical application. This approach creates a second representation of a controlled vocabulary: its space of "meaning" in the sense of which terms share a context, based on the view of distributional semantics that the meaning of a word depends on its relationship with other words.

### 2.1.3 Patent Classification and MeSH

As in the scientific literature, the complexity and difficulty of keyword search motivated the development of several large scale controlled vocabulary systems in patents. An additional complexity in the historical development of patent controlled vocabulary is the role of national patent offices. The United States Patent Classification (USPC), developed in 1889, has been a mainstay in patent controlled vocabulary,



| | |
|---|---|
| **A61B 3/00** | **Apparatus for testing the eyes; Instruments for examining the eyes** (eye inspection using ultrasonic, sonic or infrasonic waves A61B 8/10) **[2006.01]** |
| A61B 3/02 | • Subjective types, i.e. testing apparatus requiring the active assistance of the patient **[2006.01]** |
| A61B 3/024 | • for determining the visual field, e.g. perimeter types **[2006.01]** |
| A61B 3/028 | • for testing visual acuity; for determination of refraction, e.g. phoropters **[2006.01]** |
| A61B 3/032 | • Devices for presenting test symbols or characters, e.g. test chart projectors (A61B 3/036 takes precedence) **[2006.01]** |
| A61B 3/036 | • for testing astigmatism **[2006.01]** |
| A61B 3/04 | • Trial frames; Sets of lenses for use therewith **[2006.01]** |
| A61B 3/06 | • for testing light sensitivity, e.g. adaptation; for testing colour vision **[2006.01]** |
| A61B 3/08 | • for testing binocular or stereoscopic vision, e.g. strabismus **[2006.01]** |

Figure 2.1: Examples of IPC Codes and Subcodes

along with the widely used International Patent Classification (IPC) vocabulary [16]. The IPC was developed subject to the Strasbourg Agreement in 1971 via the World Intellectual Property Organization (WIPO). Since 2013, the United States and the European Patent Office have jointly developed the Cooperative Patent Classification system. The CPC has also been adopted by the Chinese, Korean and Mexican patent offices, as well as the Russian Federation. The CPC is largely modeled on the IPC [16].

The relatively widespread adoption of CPC make it an ideal target for comparison with MeSH in the biomedical space. Its predecessor, the IPC, has been widely studied and applied. Prior to the CPC, the IPC was used by over 100 patent-issuing bodies worldwide. It is comprised of about 70,000 hierarchically organized classes. There are many national variants of the IPC – the European Classification (ECLA) has 132,000 classes, and the Japanese variant contains 170,000. Various tools exist to help reconcile the versions with each other, or in a limited fashion, with the USPC.[16]

Given the historical importance of IPC, it is the natural vocabulary to juxtapose with MeSH. The IPC is organized into eight top level categories corresponding to very high level domains such as "Electricity" or "Textiles;paper." The following table gives an illustrative example of a biomedical classification:

In this example, the starting code of A indicates the "Human Necessities" category. Code 61 references medical, veterinary or hygiene. B indicates inventions related to diagnostics.[16]

One of the more complex aspects of interpreting IPC codes involves the subcodes. In the above example 3/00 indicates devices for testing the eyes. The code 3/04 is a child of "for testing visual acuity." Though the code 3/06 appears to be at the same level of the hierarchy, it is in fact at a higher level. The numerical codes do not directly correspond to the structure of the controlled vocabulary. Additionally, many classification codes require combining their superordinate classes in order to fully understand the description of the invention.[16]

The general complexity of this system has led to a system that is difficult to use, even for experts. As a result, there has been a longstanding interest in annotating

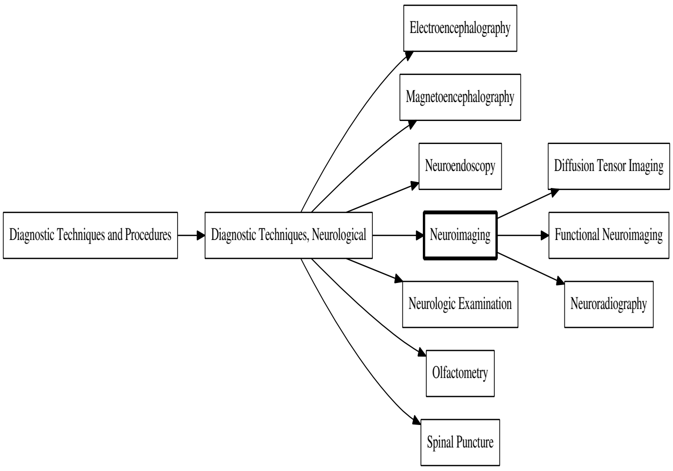


Figure 2.2: MeSH term with parent and child terms

patents with more scientifically familiar classification schemes, notably MeSH.[16]

### 2.1.4 Medical Subject Headings in Practice

To review, the Medical Subject Heading (MeSH) vocabulary was created as an indexing tool and thesaurus in 1960 by the National Library of Medicine[39]. The vocabulary was created in the context of early computerization, as well as continued growth of the biomedical literature. The modern version of MeSH contains approximately twenty eight thousand descriptors, organized in an eleven level hierarchy across sixteen categories [37]. Examples of these categories include anatomic terms, drugs and diseases and organisms. MeSH terms are currently primarily assigned manually by expert indexers at the NLM. The annotation process utilizes a computer recommender system that provides indexers with suggestions that are then manually filtered. Most MEDLINE records have an average of 13 annotations per document, although this can vary depending on the domain [23]. Cost estimates vary, but a common figure is that one MEDLINE article costs approximately 9 USD to annotate as of 2013[38]. The MeSH vocabulary has grown far beyond its original application as an indexing tool and is currently used for a variety of tasks, including query expansion, document summarization and other text mining tasks[33, 27, 54].

There has been growing interest in MeSH prediction as the scale of the literature continues to grow, and as MeSH continues to take on useful applications. In the following section, I will briefly review the MeSH prediction problem more formally, multilabel classification models generally and an examination of several prominent MeSH prediction strategies.

## 2.2 Defining the MeSH Prediction Problem

MeSH prediction is a hierarchical, multilabel classification problem[42, 54]. In a traditional classification problem, the goal is to associate an instance $x_i \in X$ with one class $c_j \in C$. In MeSH prediction specifically and multilabel classification generally, the task is more complex: an instance is simultaneously classified with a set $C_j \in C$[65]. In MeSH prediction, the classes are also organized in a hierarchical structure. In hierarchical classification these structures are either Directed Acyclic Graphs (DAGS) or trees. In such hierarchies, superclass relationships are represented with a partial order such that $c_1, c_2 \in C, c_1 \prec_h c_2 \iff c_1$ is a superclass of $c_2$. Hierarchical classification tasks are further divided by whether they predict classes strictly from the leaf nodes of their hierarchy, or whether intermediate nodes are also permitted. MeSH prediction falls into the latter category of ”optional leaf-node prediction” problems, as terms can be predicted at any level.

MeSH prediction is a particularly complex example of hierarchical multilabel classification because it involves a very large set of classes. The MeSH hierarchy is a DAG, where terms can have multiple parents and exist in more than one branch of the hierarchy. Terms at every level of the hierarchy are used, with highly varying frequencies. The definition of the vocabulary itself has changed substantively over time. This is highly significant, as training samples will putatively contain annotations from different versions of the hierarchy.

## 2.3 Hierarchical and Multilabel Classification

Multilabel classification involves a number of special challenges: necessity of special evaluation methods, adapting either the data or the algorithm to accommodate the use of multiple labels, and the exponential relationship between the number of labels and the outcome space[65].

A common strategy for addressing these challenges is to leverage information about the relationships between labels[65]. There are three common approaches: the simplest ”first order” strategy simply ignores interrelationships between labels. The ”second order” strategy uses pairwise relationships between labels to enhance classification. This might involve simple co-occurence frequency, or ranking between a relevant and irrelevant pair. Finally, the ”higher order” strategy uses multiple relationships between labels, potentially including relationships across the entire set of labels. All of these strategies have been used with varying degrees of success in MeSH classification.

## Defining Multilabel Classification

Suppose a d-dimensional instance space $X = R^d$, and a label space $Y = \{y_1, y_2, y_3, ..., y_q\}$ denoting q possible classes. Multilabel classification learns the function $h : X \rightarrow 2^y$ from a training set $D = \{\mathbf{x}_i, Y_i\}$. $\mathbf{x}_i$ is a d-dimensional feature vector $(x_1, x_2, ..., x_d$ representing an instance, and $Y_i \subseteq y$ is the set of labels for that instance.[65]

### Multilabel Evaluation Metrics

As stated above, multilabel models predict more than one class, necessitating modified forms of common machine learning performance measures. The primary ”flat” evaluation measures used in HMLC are as follows[65]:

$$oneError = \frac{1}{p}\sum_{i=1}^{p}[[argmax_y \in \gamma f(\mathbf{x}_i, y)] \notin \Gamma_i]]$$

The ”one error” metric measures the proportion of incorrect top ranked terms.

$$subsetAcc(h) = \frac{1}{p}\sum_{i=1}^{p}[[h(\mathbf{x}_i) = Y_i]]$$

The subset accuracy metric measures how many of the predicted terms are correct.

$$Jaccard_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|Y_i \cup h(\mathbf{x}_i)|}$$

The Jaccard index measures the ”intersection over the union” between the predicted and true classes.

$$Precision_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|h(\mathbf{x}_i)|}$$

The precision metric measures the proportion of predicted classes that are correct.

$$Recall_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|Y_i|}$$

The recall metric measures the proportion of the true classes that are recovered.

#### Hierarchical Multilabel Measures: Lowest Common Ancestor

The complexity of many multilabel hierarchies requires partial matching evaluation measures. The BioASQ MeSH prediction challenge has traditionally included two such measures: ”hierarchical” precision, recall and f-score and the ”lowest common ancestor (LCA)” precision, recall and f-score[55]. The ”hierarchical” version simply adds all of the ancestors of the predicted and true classes to a set of augmented classes ($Y_{aug}$ and $\hat{\text{Y}}_{aug}$, respectively)[30]. However, this approach penalizes nodes with many ancestors.

The LCA measures attempt to overcome this problem by augmenting the predicted and true classes more selectively by using their shared lowest common ancestor and all of the terms connecting them[30]. Here, the lowest common ancestor is the graph theoretic concept of the lowest node in a tree T that is an ancestor of a pair of terms $n_1$ and $n_2$. The LCA procedure collects all of the lowest common ancestors for each predicted class against each true class, and prunes redundant LCAs. This in turn yields a more targeted augmented set. The precision, recall and f-score metrics are calculated as above, replacing the predicted and true classes with the augmented versions. This dissertation uses an implementation of the LCA algorithm by its authors, called HEMkit.[30]. The precision, recall and F-score LCA variants are defined below:

$$P_{LCA}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|\hat{\text{Y}}_{aug} \cap Y_{aug}|}{|\hat{\text{Y}}_{aug}|}$$

$$R_{LCA}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|\hat{\text{Y}}_{aug} \cap Y_{aug}|}{|Y_{aug}|}$$

$$F_{LCA}(h) = \frac{2P_{LCA}R_{LCA}}{P_{LCA} + R_{LCA}}$$

### 2.3.1 Problem Transformation Strategy: The Label Powerset

The simplest example of the problem transformation approach is to convert multilabel classification into a large multi-class (but single label) problem. This is done by transforming the label space into a powerset of possible labels.[65]

Suppose that $\sigma_y : 2^y \rightarrow$ maps the power set of y to the natural numbers. The label power set transforms the multilabel classes of $D$ to a set of distinct labels as *individual* classes:

$$D^{\dagger}_{\gamma} = \{x_i, \sigma_y(Y_i))|1 <= i <= m\}$$

The advantage of this approach is that it transforms the problem in a way that can be approached using traditional classification methodologies. Perhaps more importantly, this method takes into account the relationships between labels explicitly. The obvious disadvantage is that it creates an exponentially large space of possible classes. Further, this approach cannot capture label combinations outside of the training data. Finally, this approach may result in a small number of examples for some label combinations.

A modification of this approach uses smaller, random subsets of labels in applying the label powerset technique[56]. The label space $Y$ is partitioned into $Y^k$, comprising all possible distinct label sets of k-labelsets. The total size of $Y^k$ is $|L^k| = \binom{|L|}{k}$.

From this set, k-labelsets are iteratively sampled and then used to induct a label powerset classifier. Each iteration contributes to an ensemble of classifiers. A new instance document $\mathbf{x}$ is processed by each classifier, accumulating binary $h_i(\mathbf{x}, y_j)$ for each $y_j$ in the corresponding labelset. The final classification is produced by calculating an average decision of the ensemble classifiers, with a threshold $t$, often set to 0.5. [56]

### 2.3.2 Problem Transformation Strategy: Binary Classifiers

An alternative approach to problem transformation involves training a binary classifier for each label in the label space[65, 46]. In most formulations, any instance $\mathbf{x}_i$ containing a class $y_i$ is considered a positive instance, and any instance $\mathbf{x}_j$ not containing the class is considered a negative instance. The strongest advantage of this approach is its simplicity: any binary classification method can be induced on the decomposed labels. The greatest disadvantage of this approach is that it disregards any information about relationships between the labels. Additionally, class imbalances can be a significant problem when the label space is large and sparse, as it is in MeSH[65].

As with binary classification methods broadly, the classifier chain trains a single classifier per label, but incorporates the predictions of previous classifiers. This approach is highly sensitive to the ordering of the classifiers, as errors propagate from the top of the chain.

### 2.3.3 Directly Optimizing the F-Score: Reverse Multilabel Learning

Many researchers have observed that the techniques described above optimize for proxies of the desired performance measures. An alternative approach originated by Petterson and Caetano reverses the prediction problem in order to permit optimization on a convex relaxation of the F-score [43]. That is, they predict a set of instances given a label. Further, they use a constraint generation strategy (most violated constraint) to make the optimization problem tractable. A subsequent paper extends this methodology by accounting for relationships between labels.[44].

### 2.3.4 Algorithm Adaptation Strategy: KNN methods

The previously described approaches all transform multilabel problems into new problems that are easier to solve using existing technique. In the case of random label sets, the problem is transformed from a multilabel prediction problem into a multiclass problem. More sophisticated versions of this class of technique use ensemble methods to make predictions more robust or tractable. An even simpler approach, binary classification, decomposes the problem into independent binary predictions.

An entirely different approach to multilabel classification modifies classic machine learning algorithms to multilabel data, rather than transforming multilabel data to standard supervised learning problems. One such example of this approach is a modification of the K-Nearest Neighbors algorithm to the multilabel setting [64].

Given an instance $\mathbf{x}$, this approach retrieves a set of KNNs. The class membership of each neighbor is collected in a vector:

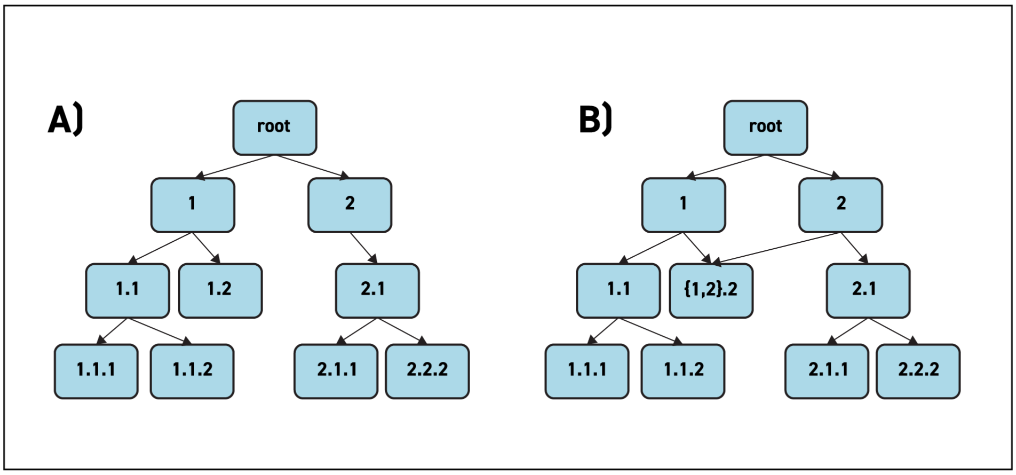
$$\mathbf{y}_t = argmax_{b \in \{0,1\}} P(H_b^l | E_{C_t(l)}^l), l \in Y$$

Using Bayes rule, this can be written:

$$= argmax(P(H_b^l)P(E_{C_t(l)}^l | H_b^l)$$

### 2.3.5 Hierarchical Multilabel Classification

In multilabel classification, an object is associated with more than one class. Another category of tasks involves classification problems where the classes are additionally structured within a tree or a directed acyclic graph (DAG)[8]. Due to this structure, this type of classification task is often described as hierarchical classification. When the classification task involves assigning multiple paths in the class hierarchy it is labeled hierarchical multilabel classification (HMLC). HMLC problems can be found in a wide array of bioinformatics such as protein function prediction, in text classification and image annotation[13, 57, 3, 10, 4]. MeSH prediction can be best described as an HMLC problem.



In many HMLC problems, the classification task is to assign leaf nodes in the hierarchy to each object. In the case of MeSH classification, "shallower" intermediate categories are the classification target. This raises additional complexities. As the number of classes grows and the depth in the hierarchy increases, the classification task becomes more complex due to the smaller number of training examples.

As in multilabel classification, most approaches to HMLC focus on either problem transformation or algorithm adaptation. The simplest approach is to simply flatten the class hierarchy and predict each class separately.

There are broadly two approaches to HMLC: local and global prediction. In local prediction, an ensemble of classifiers predict labels in different parts of the class hierarchy and then combine predictions together. One common approach is to train a cascade of classifiers that predict particular nodes or hierarchical levels using a top-down strategy[29, 13, 26]. This strategy tends to be computationally expensive, since individual classifiers are required for every class or every level of the hierarchy. Additionally, errors may propagate from classifier to classifier.

In global prediction, a single classifier predicts the classes for the hierarchy as a whole. Typically global classifiers are computationally cheaper and do not suffer from the error propagation problems. However, they often discard information from the hierarchy. Examples include the reverse multilabel learning paradigm discussed above.[43]

## 2.4 Notable MeSH Prediction Systems

MeSH classification is a particularly challenging multilabel problem, both due to the complexity and the size of the MeSH vocabulary. A multilabel problem with only twenty class labels has over a million possible label sets. The MeSH vocabulary has a very large number of class labels (approximately 27,000, one for each MeSH term) and a strongly uneven underlying frequency distribution. Inducing a classification function from existing annotations is difficult for a number of reasons. As will be explored in some detail later, annotators frequently disagree on appropriate annotations. Further, indexing practices and the vocabulary itself change over time. In short, due to the size and complexity of the MeSH vocabulary, there are many plausibly valid MeSH assignments for any given record, making relevance judgments difficult.

The high cost of annotation and ever increasing volume of newly published biomedical research attract many researchers to MeSH prediction. A wide range of algorithms and approaches have been applied. Broadly, most approaches use one of two strategies. The first is thesaurus-oriented, focusing exclusively on the information available in the MeSH thesaurus itself[54]. These systems match free text to MeSH terms, and their synonyms and short text descriptions. The second strategy is concept-oriented and typically entails building statistical models for each MeSH term[60, 58, 50, 42]. In the second strategy, features may be extracted directly from the available text or from related documents. Nearest neighbor methods is a predominate approach, particularly via related citations [23, 54, 48, 28].

More recent approaches have used hybrid methods that draw upon ensembles employing both strategies. Two state of the art systems, DeepMeSH and MeSHLabeler, use a combined thesaurus-oriented text matching classifier and a large series of independent MeSH classifiers. These systems also use second-order relationships between labels, namely pairwise correlation, to enhance their final predictions. Finally, both systems also use a model to predict the appropriate number of labels to further reduce the possible output space.

The following sections will present a high level overview of two prominent MeSH prediction systems. The first is MTI, the official MeSH recommender system developed by the National Library of Medicine. The second system is DeepMeSH, the current state of the art system in MeSH prediction.

### 2.4.1 MTI: MetaMap and PRC

Perhaps the most well known MeSH prediction tool is the Medical Text Indexer (MTI) system. The MTI system assists NLM indexers in providing MeSH terms[38]. The MTI system takes inputs of an identifier, title and abstract but is also capable of processing arbitrary biomedical text [38]. Recommendations are computed using two methods: MetaMap indexing and PubMed Related Citations (PRC), a K-nearest neighbors algorithm that identifies similar citations [37]. MetaMap processes the title and abstract to identify UMLS Metathesaurus concepts that can then be mapped to MeSH. Precision and recall performance for the MTI system is typically around .60 [37].

Natural language techniques for matching variants of free text terms to the MeSH vocabulary is a common approach to MeSH prediction. These approaches have demonstrated some promise, but continue to be limited by the inherent ambiguity of biomedical text. The MetaMap component of MTI is capable of generating a large number of variants from text and has some capabilities for word sense disambiguation. However, these capabilities come at considerable computational cost at processing time.

### 2.4.2 DeepMeSH

Recent research efforts have focused on ensemble approaches that blend information drawn from text pattern matching algorithms with machine learning approaches that train models for each term. DeepMeSH builds upon existing hybrid models by incorporating distributional semantics based features[42]. DeepMeSH is itself based on a hybrid/ensemble system developed by Liu, et. al called MeSHLabeler[32]. This system uses MTI predictions as inputs in a more complex ranking algorithm.

The DeepMeSH system consists of two components systems, a MeSH ranking algorithm and a MeSH number prediction system. The MeSH number algorithm is based on prior work by Liu, et. al and predicts the number of MeSH terms given the number of annotated MeSH (MH) terms of citations from the same journal, the number of annotated MH of nearest neighbor documents, the number of terms recommended by MTI and cut offs based on scoring from binary classifiers and the MeSH ranking algorithm[42].

The MeSH ranking algorithm is complex and incorporates several sources of evidence: scoring from a per-MH binary classifier trained on MEDLINE, scores from similar citations obtained by KNN, parwise MeSH correlations between candidate terms, and pattern matching using the MetaMap portion of MTI. DeepMeSH develops upon this foundation by using distribution features, ie doc2vec and doc2vec-TFIDF. Final predictions are made by selecting the top terms, with the cutoff determined by the MeSH number model[42].

The DeepMeSH system is of interest in that it leverages both a second-order strategy by incorporating pairwise MeSH correlations, and dense semantic representations (ie doc2vec) in the input text[34]. This approach has demonstrated significant success; in the BioASQ3 challenge DeepMeSH outperformed MTI in micro F-measure by 12%, and the previous state of the art MeSHLabeler by 2%[42].

However, in the context of MeSH prediction beyond MEDLINE, both of these strategies are untenable. The putative second-order relationships between vocabulary terms are essentially unknown outside of MEDLINE because there is no labeled corpus to draw upon. Indeed, it seems likely that the pairwise correlations would be significantly different in other bibliographic databases. Similarly, the word embeddings trained in MEDLINE are likely to be based on different distributions of term occurrences than those found in other literatures where style, context and vocabulary differ markedly.

## 2.5 MeSH Prediction Beyond MEDLINE

Relatively little research has been developed with respect to MeSH prediction outside of MEDLINE, though there have been several attempts to apply the MeSH vocabulary to patents. In “Annotating Patents with MEDLINE MeSH Codes via Citation Mapping” Thomas Griffin et. al presented a system which matched patent references to MEDLINE records and extracted MeSH terms [20]. This system retrieves the MeSH terms and arranges them alphabetically or by frequency of the term. A patent held by IBM titled “System and Method for Annotating Patents with MeSH Data” proposes a similar procedure that extracts non-patent references directly using the MeSH vocabulary of the cited documents [12].

These approaches were both inspired by the clear information retrieval value of the MeSH vocabulary. Thomas Grin et. al [16] developed an analysis comparing the IPC classification system and MeSH, finding that the MeSH vocabulary is better suited to describing biomedical research. However, neither classification system discussed above attempts to rank or filter MeSH terms beyond frequency measures.

# Chapter 3

# Explaining Prediction Accuracy: Beyond Exact Matching

## 3.1 Revisiting the Consistency of MeSH Indexing: An Argument For Partial Matching

This chapter addresses the following research question:

### RQ1: Given that human inter-rater reliability is modest, how should MeSH prediction systems evaluate accuracy?

Before posing the question of how consistently an algorithm can assign MeSH, it is worth considering how consistent human annotators are. If human indexers are highly consistent, we may conclude that the vocabulary evaluation should be treated in a fairly strict, exact manner. Likewise, if there is limited consistency by human indexers, we may conclude that the vocabulary permits multiple valid annotations for a given document, or at least that indexing behavior is heterogeneous.

Either of these situations have important consequences for automatic annotation. If there are multiple valid annotations, evaluating automatically assigned labels must take into account that measures based on exact match will fail to recognize plausible terms from implausible terms. Additionally, if human annotators apply widely varying standards to their annotations, then any training data derived from human annotated papers will also contain a complex mixture of judgments about term relevance. Likewise, evaluation measures of such a prediction model will inevitably capture a limited representation of the system's ability to replicate a mosaic of annotation styles.

Determining how consistent indexers are is a difficult problem. The high cost of manual indexing means that the National Library of Medicine actively avoids duplication of effort. Indeed, the high cost of annotation is the primary motivation for many studies of automatic MeSH assignment. Studies of indexing consistency are also hampered by the labor required to generate a significant sample. These difficulties were addressed in ”Indexing consistency in MEDLINE”, Funk et. al by identifying 760 papers that were accidentally indexed twice[18]. The bulk of these papers were annotated twice due to the same article being published in more than one venue.

Funk et. al used Hooper's consistency metric to measure agreement between annotators. Hooper's consistency metric is given as:

$$CP(\%) = \frac{100A}{A + |M| + |N|} \tag{3.1}$$

A is the number of terms in agreement

M is the number of terms used by M but not N

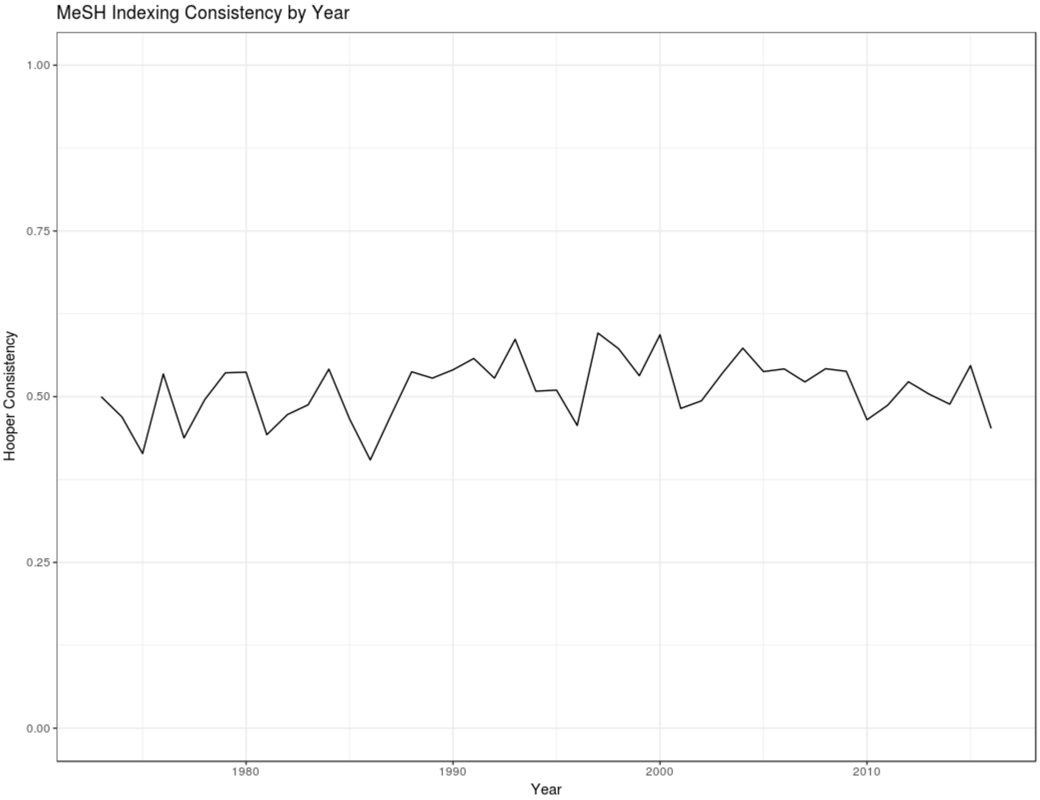


Figure 3.1: Average Consistency: 1973-2016

N is the number of terms used by N but not M

Funk et. al found a mean consistency of 48.2% among main headings. Consistency varied significantly between branches of the MeSH hierarchy, with the lowest level of consistency in branches E,F,H and N. Branches A, B and D had higher levels of consistency. Additionally, they found that the language, indexing priority and length of the article had no significant effect on overall indexing consistency.[18]

### 3.1.1 MeSH Consistency: 1973 - 2016

The analysis by Funk et al. assessed twice annotated papers up to 1983. I have performed a similar analysis by querying all papers with the same title, authors and year from 1973 to 2016. This yielded 3689 pairs of papers. Following the methodology of Funk et. al, I calculated the Hooper consistency metric between the MeSH of each pair. The following plot details the mean consistency over the period:

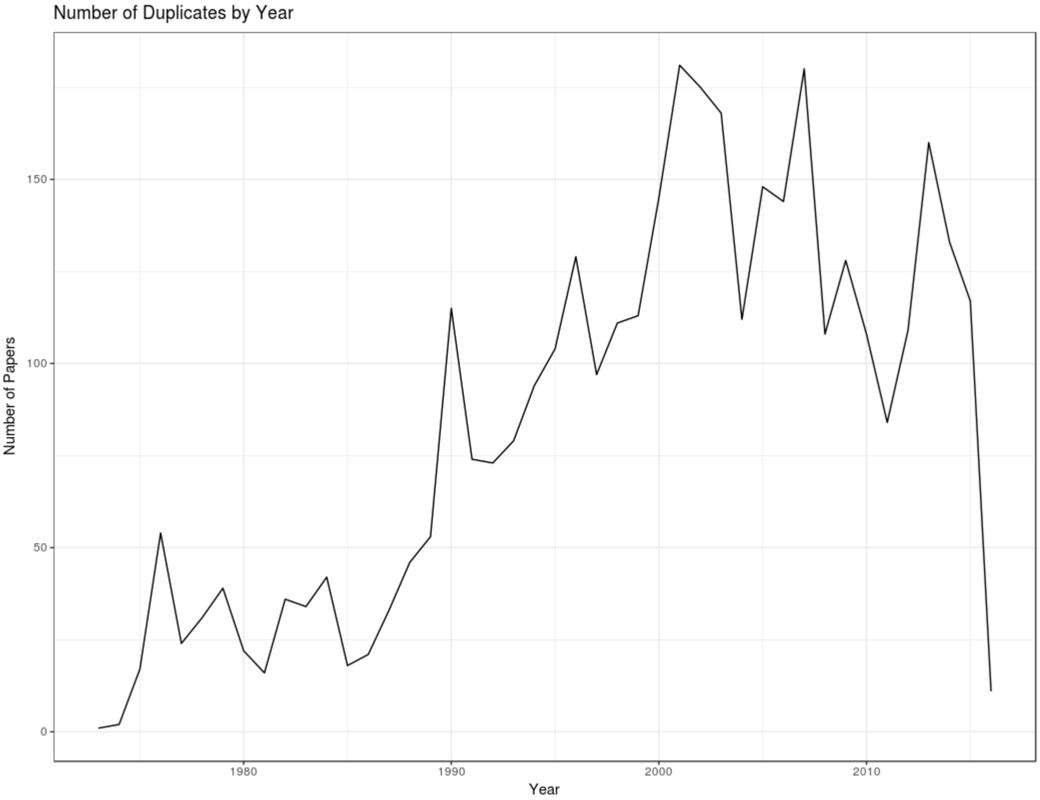


Figure 3.2: Total Number of Duplicate Pairs Per Year: 1973-2016

The mean consistency of papers in this larger set is 52.3%, a slight increase from the 1983 study. Indexing consistency has remained largely flat over time. The relative stability of indexing consistency is notable given the growth in publications and the complexity of the MeSH vocabulary. The following plot details the number of duplicate papers each year:

Replicating the original Funk study shows that consistency remains uneven between MeSH branches:

Taken together, this data indicates that consistency is "high" relative to the overall complexity of the MeSH vocabulary. However, from the standpoint of evaluation, consistency is limited. Treating human annotators as a model would reflect modest accuracy in an exact matching evaluation.

New questions arise in light of this data. How can the differences between vocabulary be characterized? For instance, are most discrepancies due to an omission on the part of one annotator? Or are they due to the choice of a more specific term? Are there terms that are largely synonymous? In short, to fully assess the consistency

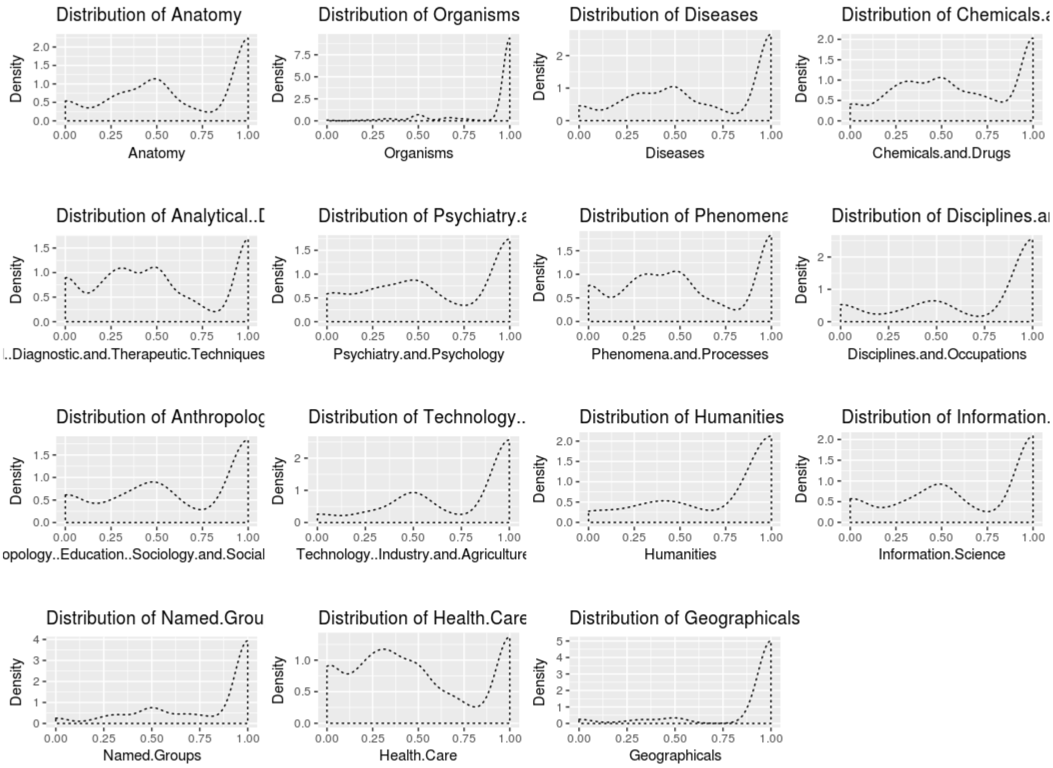


Figure 3.3: Consistency by MeSH Branch

between terms, a robust partial matching metric is necessary.

## 3.2 Partial Matching in the MeSH Hierarchy

The key problem of measuring both indexing consistency and prediction accuracy in MeSH is differentiating the degrees to which two terms match. As shown above, using only exact matching shows that human indexers are about 50% consistent. This is impressive given the size and complexity of the MeSH vocabulary, but also an indication of the difficulty of the prediction problem. While exact matches provide important, straightforward evidence to the accuracy of a model, they have limited ability to describe the overall plausibility of a set of predicted terms.

One approach to this problem is to use a relaxed form of matching by leveraging the MeSH hierarchy itself. In this scheme, two terms can be considered to be a partial match with a score determined by the type of relationship. For example, an exact match is scored at 1, a parent-child relationship at .5 and a sibling relationship at

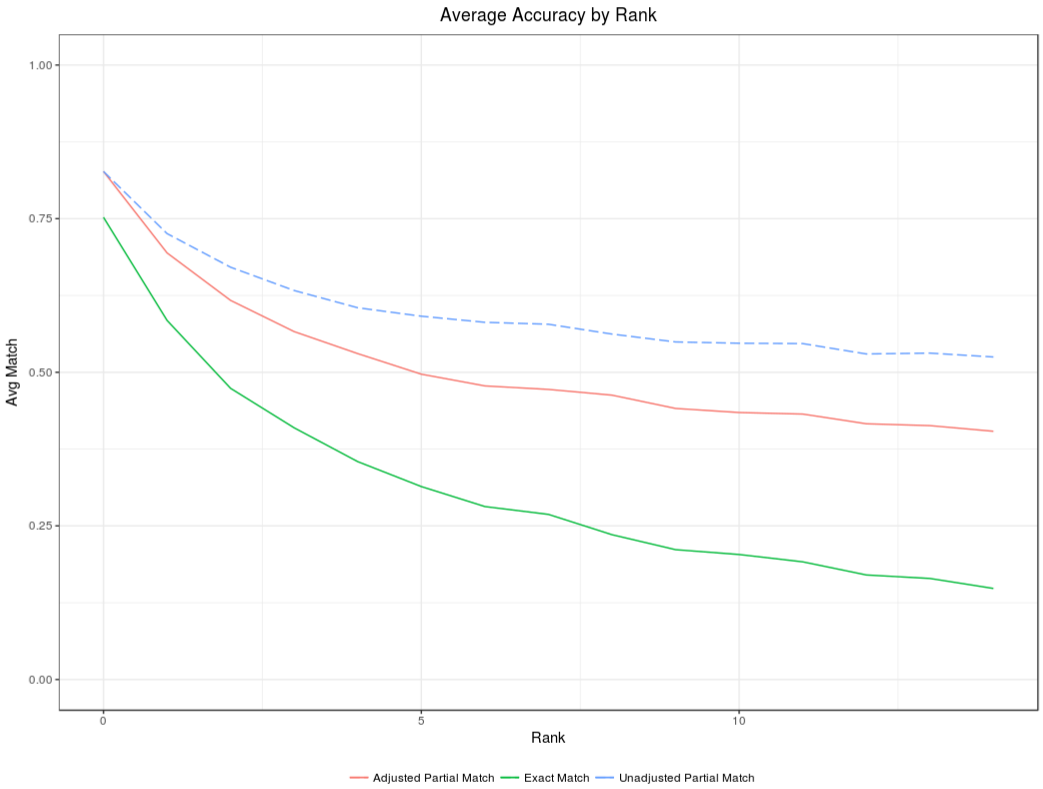


Figure 3.4: Accuracy by Rank in Partial Matching and Exact Schemes

.25. More sophisticated versions of this scheme are possible by taking into account co-occurrences within each MeSH branch.

The shortcoming of this approach is that the MeSH hierarchy is highly inconsistent in terms of the grouping of terms. For example ”Death” (C23.550.260) and ”Dehydration” (C23.550.274) are sibling terms under the parent of ”Pathological Processes” (C23.550). Another example: ”Epidemiologic Methods” (E05.318) and ”Protein Folding” (E05.790) are siblings, despite covering entirely different biological areas. Partial matching via the hierarchy also cannot take into account semantic relationships between different categories, for example between disease and symptom terms.

Other approaches, described above in the literature review section, use vocabulary hierarchies to augment the predicted and true classes. However, these approaches are vulnerable to the same limitations of the structure of the vocabulary. For instance, the Lowest Common Ancestor algorithm would find the same shortest path between the example sibling terms described previously.

### 3.2.1 Partial Matching via Distributional Semantics

One of the key ideas of this dissertation is that MeSH term similarity can also be measured through the lens of shared context. Most approaches to controlled vocabulary prediction focus on learning a function that maps a set of inputs to a set of labels. The training data derives from human judgments about which documents go with which labels. However, there is an additional set of information provided by human annotations: that is, a large set of judgments about which terms belong with other terms.

In natural language analysis, robust measures of word similarity have been developed using neural network techniques. The most notable of these has been word2vec and successor approaches. This approach attempts to learn a vector representation of a vocabulary based on the distributional hypothesis. The distributional hypothesis states that linguistic items with similar distributions have similar meanings[34]. Predicting a term by its linguistic context yields a vector of weights that also serve as a vector representation of the term.

The following approach proposes a quantitative method for measuring the similarity of two MeSH terms based on their shared context. The method encodes each MeSH term as an arbitrarily high dimensional vector. Such an approach permits a continuous-scale measurement of similarity between two terms, irrespective of their location in the MeSH hierarchy, using cosine similarity. This resolves both problems of dissimilar but related terms (distinguishing ”death” and ”dehydration”, and highly similar terms that are not colocated in the hierarchy.

This idea also takes inspiration from Swanson in describing controlled vocabularies as an artificial language. Though word2vec was originally intended for natural language, the idea of learning a vector representation based on shared context is applicable to artificial languages, as well. To train the MeSH word embeddings, the following procedure was used:

1. Each MeSH term was tokenized. In other words, the complete term phrase is considered an independent lexical unit.

2. The order of terms was randomized to prevent artifacts due to alphabetical arrangement of terms.

3. The context window of the word2vec model was set to 15 (a typical number of MeSH assignments) to reflect the unimportance of word order.

For training data, a set of 14.3 million papers’ annotations were used, representing every available MeSH assignment through 2017.

#### 3.2.1.1 Cosine Similarity vs Hooper's Consistency as Partial Match Measure

The evaluation of the measure described above relies on the concept that highly related documents should have high match scores, and unrelated documents should have very low scores. I have compiled a set of papers with the following characteristics:

1. Twice-annotated ”duplicate” papers that are expected to have the same or highly similar annotations.
2. The same set of papers with one randomly selected citation with MeSH, representing a related but non-identical paper.
3. The same set of papers with a randomly selected pair paper. Compared to the first two datasets, the similarity should be somewhat evenly distributed.

The first plot details both Hooper's Consistency and the cosine similarity of MeSH word embeddings in the duplicate set. Hooper's Consistency reflects a density of papers with exact matches, and another large density slightly below 0.5. The figure on the bottom represents the cosine similarity, with a much higher density around 1.

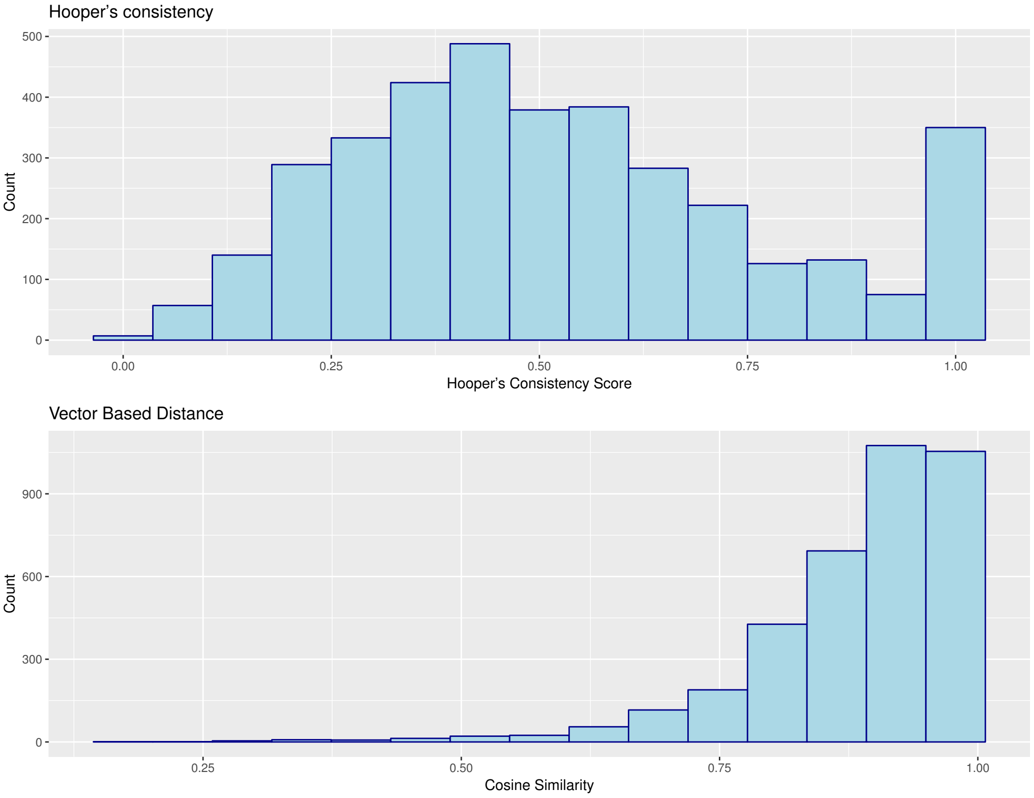


Figure 3.5: Distribution of Hooper's Consistency and Cosine Similarity in Duplicate Paper Set

The following plot compares the distribution of cosine similarity in the duplicate papers, random pairs and random citations with Hooper's Consistency. Here, the random pairs and random citations reflect a much lower degree of relatedness. Taken together, the cosine similarity measure arguably better captures the consistency of MeSH. In the duplicate set, the cosine similarity measure is close to 1. Citation papers show a modest degree of relatedness, and random pairs have a very wide variance.

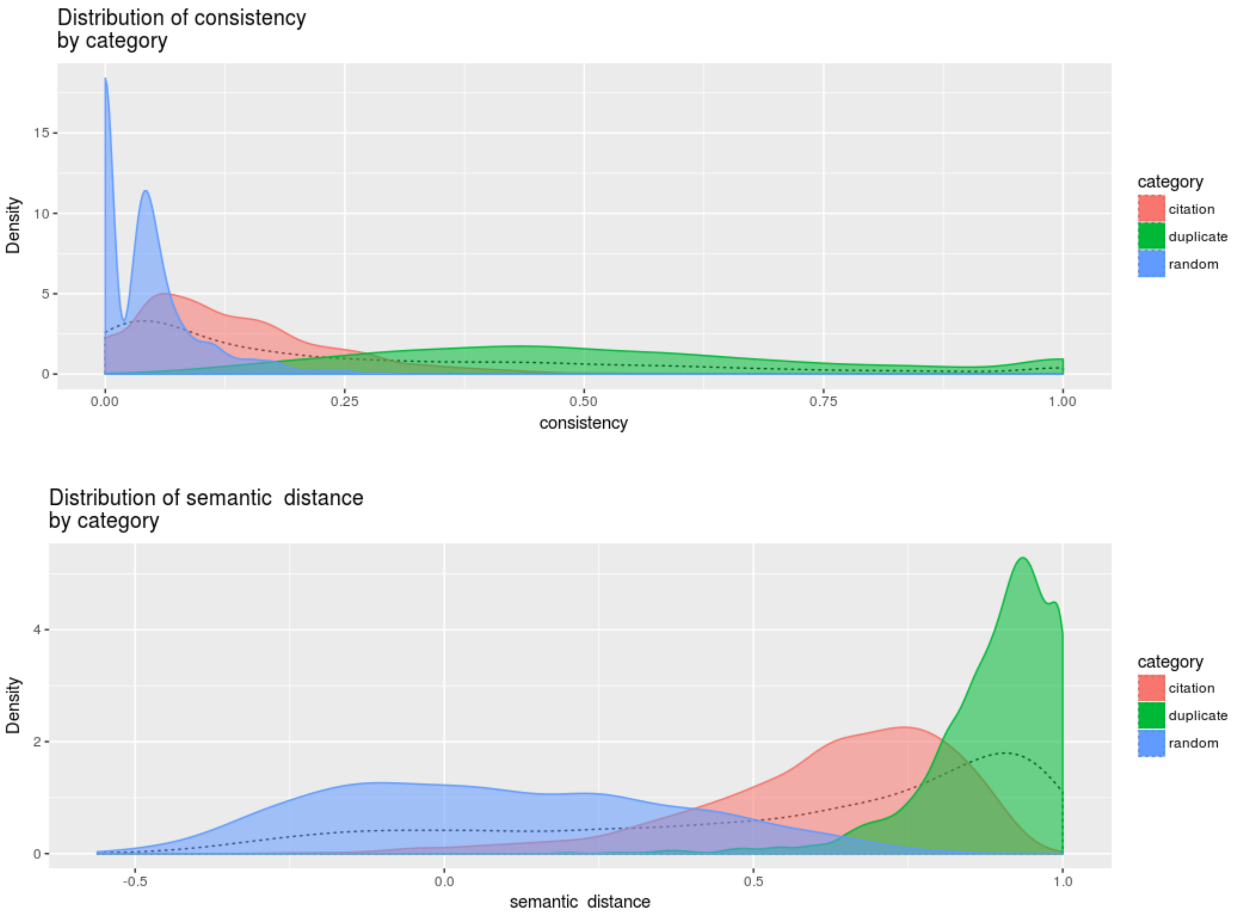


Figure 3.6: Comparison of Hooper's Consistency and Cosine Distance Across Duplicates, Random pairs, and Citations

## 3.3 Combining Word2Vec and MeSH Hierarchy

A continuous measure of similarity between two terms in the MeSH hierarchy based on their context opens many analytical opportunities. For instance, a key problem in measuring partial matches is the inconsistency of semantic similarity in different parts of the MeSH hierarchy. Above, I introduced the example of ”Dehydration” and ”Death” as siblings – clearly, terms with very different meanings. Combining both the information provided by the MeSH hierarchy and information about their context permits a quantitative description of how semantically consistent relationships are throughout the vocabulary.

In the following figure, I plot the average pairwise cosine similarity of sibling terms, aggregated by depth. As may be expected, siblings at the top of the hierarchy are less similar. Similarity generally increases at progressively lower (and therefore more granular) levels of the hierarchy. An exception to this is at the 11th level of the hierarchy, where similarity levels dip.

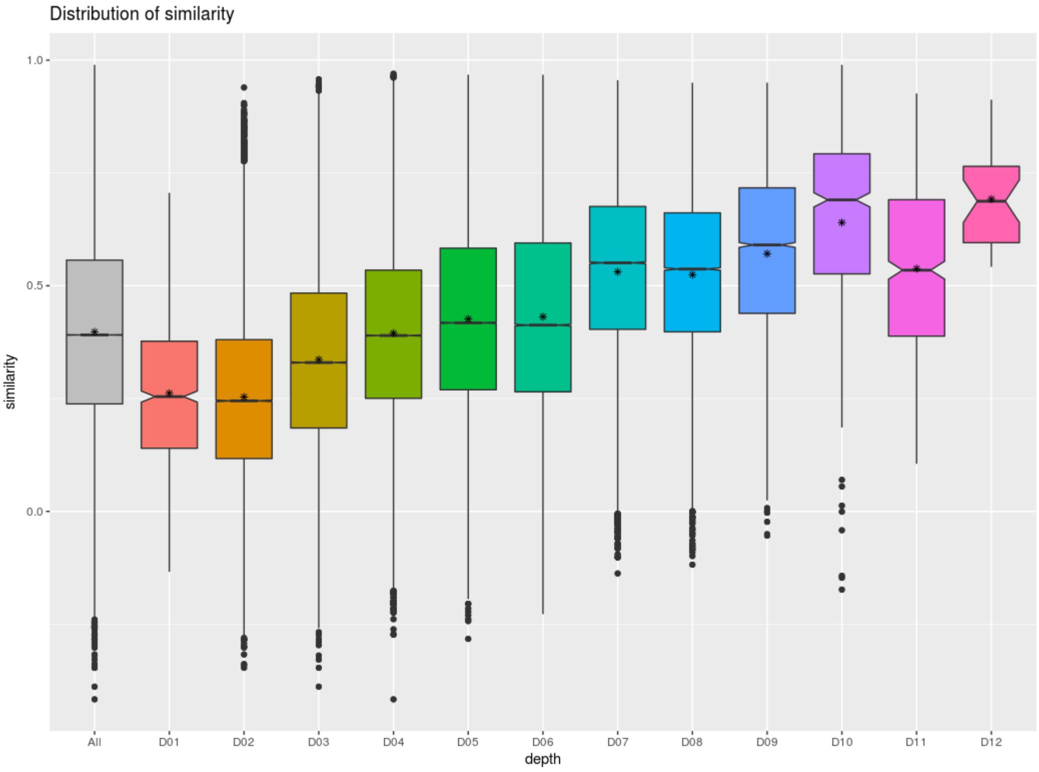


Figure 3.7: Pairwise Similarity Between Siblings by Depth

A similar analysis can be done for similarity within each MeSH branch. Here, we see that the 'Geographicals' branch is among the most consistent. This view of the MeSH hierarchy maps out semantically diverse and homogeneous branches of the hierarchy. For the purpose of calculating partial matching based on relationships, the branch similarity can be used to further weight the match quality.

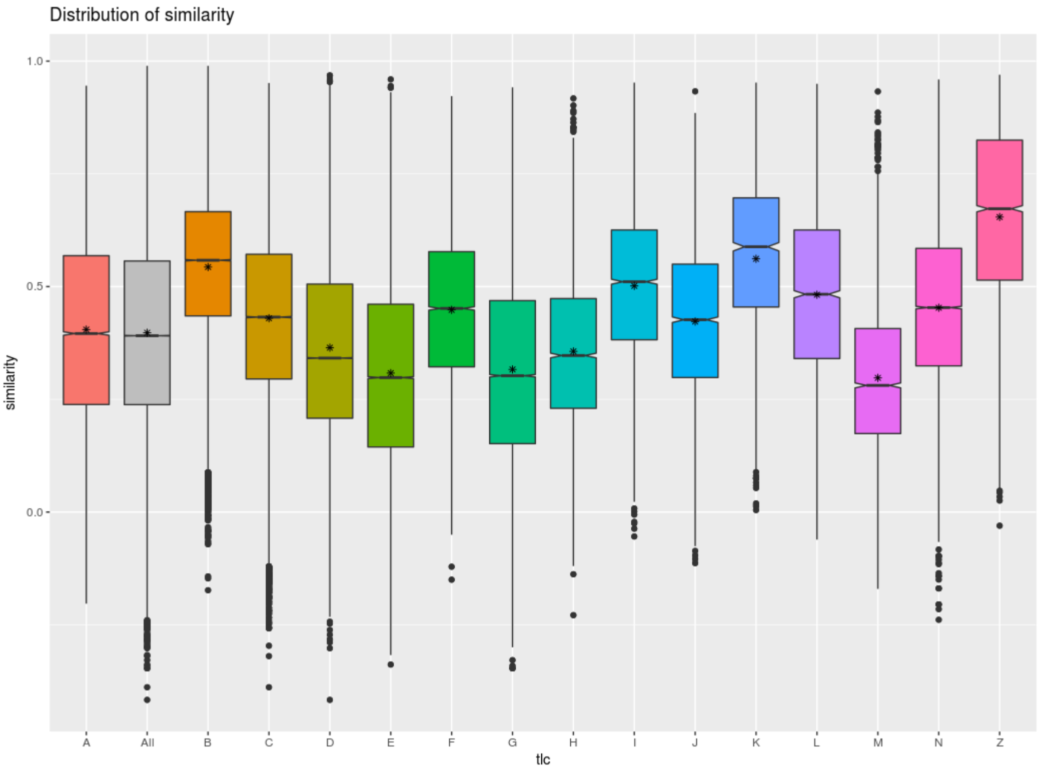


Figure 3.8: Pairwise Similarity Between Siblings by MeSH Branch

The analysis can also be extended to other kinds of relationships. For example, the following plot demonstrates pairwise similarity between parent/child pairs in MeSH, aggregated by MeSH branch. This approximately measures how steep of a conceptual change occurs across and within the branches.

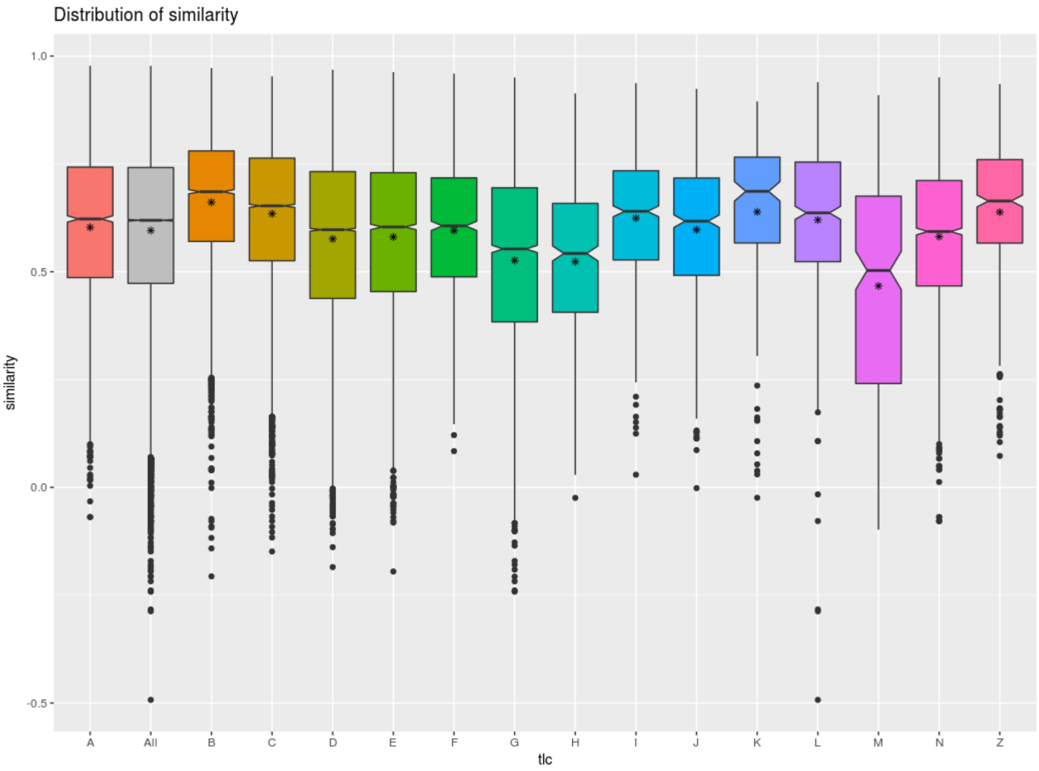


Figure 3.9: Pairwise Similarity Between Parent/Child by MeSH Branch

The next plot shows the distribution of similarity between parent/child pairs, aggregated by depth. Again, the pattern shows a general trend towards increasing similarity at greater levels of depth, with a small reduction at the extremes.

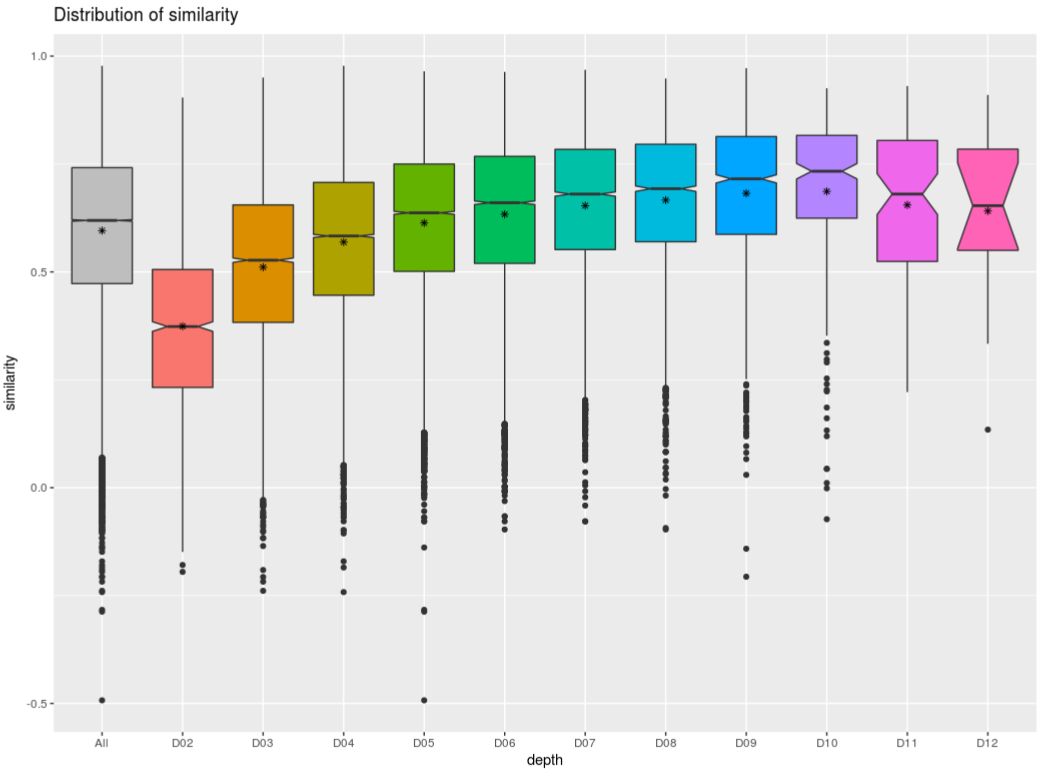


Figure 3.10: Pairwise Similarity Between Parent/Child by Depth

For the purposes of MeSH prediction, the combination of the MeSH hierarchy and cosine similarity permit more finely tuned approaches to evaluation than were previously possible. For instance, partial matching can be calculated with arbitrary weights based on relationship, or by an adjusted weighting scheme based on the relative consistency of each branch. The cosine similarity itself can also be used as a partial match measure. Both of these possibilities are explored below.

Beyond prediction, the vocabulary-wide mapping can be used by controlled vocabulary designers to identify areas of potential inconsistency. Outlier analysis may also be particularly helpful in identifying anomalies in the vocabulary. The modeling chapter provides further examples of applications of the semantic similarity measure as a diagnostic and evaluation tool. The discussion section also explores applications outside of controlled vocabulary prediction that are beyond the scope of this dissertation.

## Evaluation Measures

In constructing evaluation measures based on cosine similarity, it is important to introduce two additional concepts. The first is hierarchical versus non-hierarchical restrictions. The cosine similarity measure intentionally is not based on the hierarchy. Therefore, categorically different terms may have a very high similarity if they appear together often – for example, in disease-drug combinations. Likewise, the hierarchical measures take advantage of relationships, but weight them equally. Therefore, a measure could use both a hierarchy-restricted accuracy measure that uses the cosine similarity of related terms as a weight, and an unrestricted version that simply takes the highest cosine similarity. The interpretive difference is that in the hierarchy restricted versions, terms are guaranteed to belong to the same branch. Therefore, for example, a drug cannot be rated as a high similarity match to a disease. The non-hierarchical measure is intended to be a looser reflection of the plausibility of the assignment.

The second concept is between "free" and "restricted" matching. This primarily concerns the issue of redundant predictions. For example, a set of predictions may contain many synonyms for a few true classes, poorly representing the original term assignments. To address this, one metric uses a "restricted" scheme where only one prediction can match to any true term. The "free" version removes this restriction and allows matching multiple predicted terms against a true term. Here again, the interpretive difference is between a more strict view that seeks to measure how well the original classes are represented as a group, versus a more relaxed view that seeks to address how plausible the predicted terms are individually.

The following expresses the basic measure definitions mathematically:

$$HierarchyFree(H) = \frac{1}{p}\sum_{i=1}^{p} \frac{\sum_{j=1}^{k} Cos_{best}(h_j, y) \forall y \in Y, \forall h \in H : HasRelationship(h_j, y)}{|H|}$$

$$NonHierarchyFree(H) = \frac{1}{p}\sum_{i=1}^{p} \frac{\sum_{j=1}^{k} Cos_{best}(h_j, y) \forall y \in Y, \forall h \in H}{|H|}$$

In summary:

1. "HierarchyFree": Calculates partial matching between terms with a hierarchy relationship (child, parent, sibling, etc), weighted by their pairwise cosine similarity. The "free" component means that any predicted term is free to match against any true term; in other words, multiple predicted terms may match against the same true term. This metric is meant to capture the overall "sensibility" of the predictions balanced against a close match to the true terms position in the hierarchy.

2. ”NonHierarchyFree”: Measure reflects the best cosine similarity match between predicted and true terms. Unlike above, this is not restricted to terms with hierarchy relationships. Because this is less restricted, the value is typically higher. However, due to the lack of constraint in hierarchy relationship, this may provide partial match credit to term pairs that have a common context but are categorically different; e.g. a disease-drug pair. As above, this version allows multiple predicted terms to match to one true term.

3. ”HierarchyRestricted”: This metric is the same as HierarchyFree, but only permits each true term to be matched once. This prevents prediction sets from having a high score if they only approximately match one or two true classes.

4. ”NonHierarchyRestricted”: This is the same as NonHierarchyFree metric, but only permits each true term to be matched once.

# Chapter 4

# Predicting MeSH in MEDLINE: Leveraging Citations and Abstracts

# Introduction

This chapter addresses the following research questions:

### RQ1: Are abstracts and citations effective features for predicting MeSH terms in MEDLINE?

### RQ2: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

RQ1 is addressed by training a model to predict MeSH terms in a large set of MEDLINE papers. This model uses features derived from the frequency of candidate terms in both abstract and citation sets. Details are given below about the procedure for extracting MeSH terms using abstracts and citations. An exploratory analysis of candidate terms demonstrates the high rate at which abstracts and citation sources capture the original MeSH of a target document.

RQ2 is addressed by an analysis of the complimentarity of candidate terms captured using abstracts and citations. This analysis demonstrates how often unique terms are captured. Additionally, the analysis includes a temporal aspect, analyzing how often papers retrieved by abstract similarity are published after an initial paper.

Four models are assessed: one with citations alone, one with text similarity alone, one with citations and text, and an extended model utilizing citations, text, and ”citations of citations” and ”citations of text.” Each model is assessed in three separate test datasets. The first is sampled with ”ideal” conditions where citations and abstracts are both plentiful. The second and third are sampled with low citations and short abstracts, respectively. Additionally, the models are evaluated in terms of their performance in each branch and at each level in the MeSH hierarchy.

The chapter concludes with an examination of partial matching methods introduced in the previous chapter. The word embedding approach described previously is used to explore highly similar terms and the role of redundancy. Finally, the best performing model is evaluated using a systematic partial matching scheme.

# Candidate Identification Procedure

Candidate terms are identified using two procedures. The first uses citations from the document to extract MeSH vocabulary directly. This pool of terms is then optionally expanded by retrieving the citations of those papers, and extracting their vocabulary.

The second approach collects the abstract of the target paper and runs a BM25 text similarity query against a database of MEDLINE abstracts[53]. These candidate sets are referred to throughout the following experiments as the ”Absim” (abstract similarity) set. The top 10-15 absim papers are retrieved, and their MeSH terms extracted. The citations of the absim papers are also optionally retrieved, with their vocabulary extracted as well. Importantly, all absim papers were filtered to ensure that the target paper is not included, since the most ”similar” paper will always be itself.

Using these two procedures, the result is four set of candidate terms: the citation set, the citation of citation set, the absim set, and the citations of the absim set. In the first three models described, only the citation and the absim set are used. In the final model, all four are used.

## Training Data

The training data for the model were collected from 25,000 MEDLINE papers. The following selection criteria were used:

1. The paper must have assigned MeSH.
2. The paper must have an abstract.
3. The paper must have at least one citation.

A population of papers meeting these requirements were collected, and the 25,000 training papers were randomly sampled from that population. The papers range from the years of 1971 to 2015. The training data was intentionally sampled to contain a wide range of papers in terms of the total number of citations and the length of the abstract, and to make minimal assumptions about the documents.

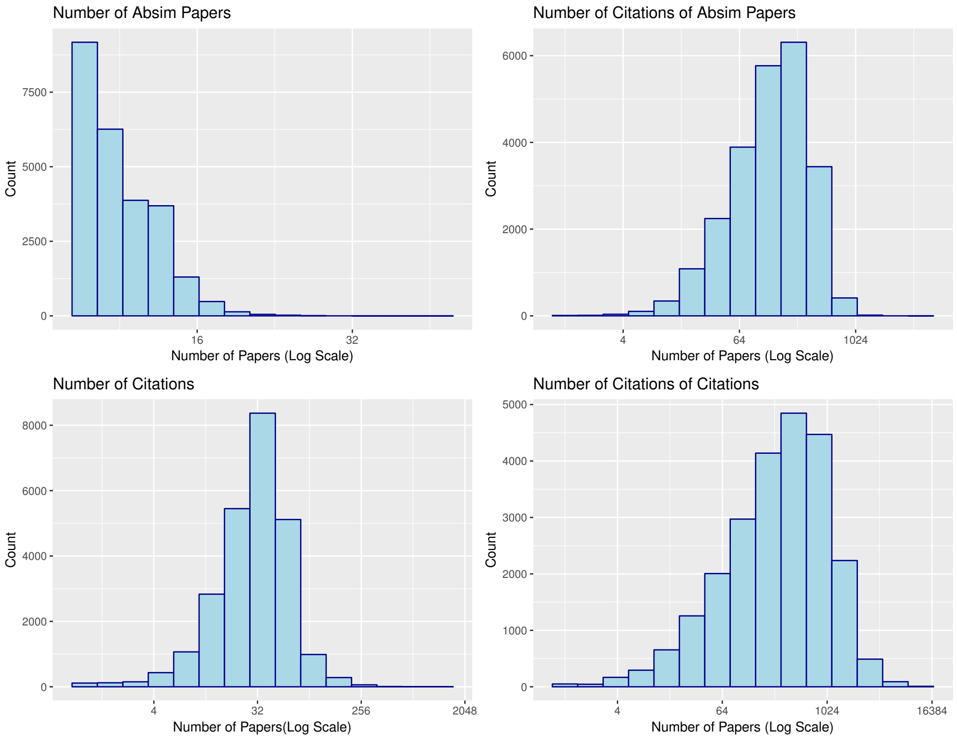


Figure 4.1: Key Training Data Characteristics

The training data contained a total of 27,014,307 candidate terms (taking the union of each of the four sets), with 26,640 unique terms. These unique terms represent approximately 99% of the MeSH vocabulary. Figure 4.1 plots the number of papers from each candidate set at log scale.

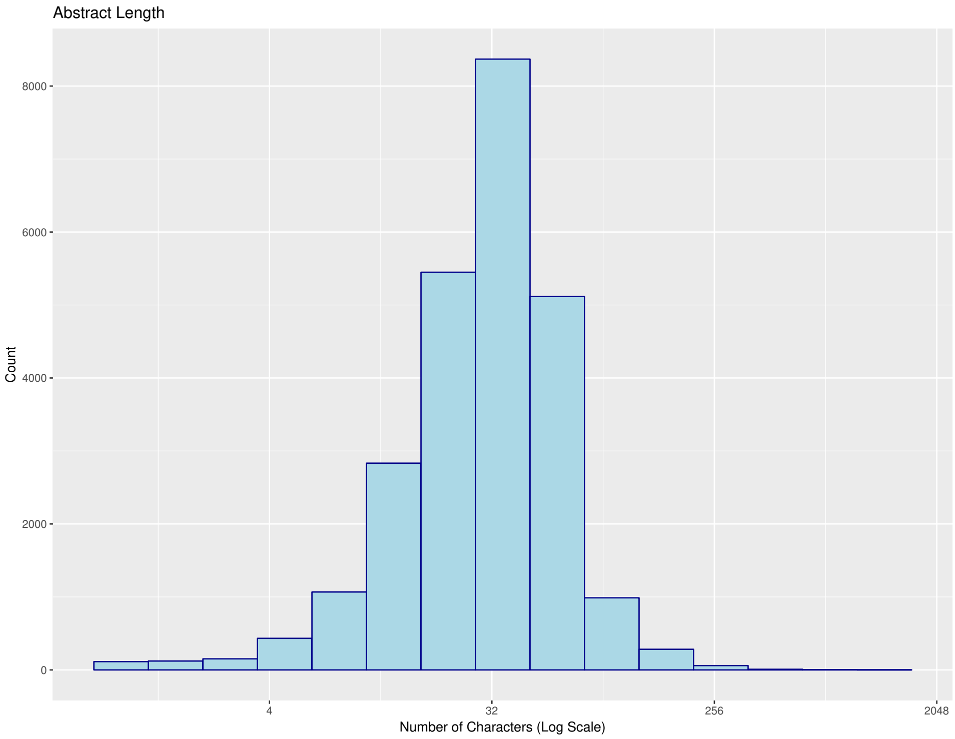


Figure 4.2: Character count of training data abstracts, log scale

The following table summarizes the total number of related records, both in terms of citations and absim records, as well as the length of the abstract in characters.

Table 4.1: Training Set Characteristics

| | **Minimum** | **Median** | **Mean** | **Max** |
|---|---|---|---|---|
| **Year** | 1971 | 2010 | 2006 | 2016 |
| **Citations** | 1 | 32 | 36.85 | 1255 |
| **Citations of Citations** | 0 | 291 | 517 | 12169 |
| **Similar Abstract Citations** | 10 | 11 | 11.58 | 49 |
| **Abstract Length (Characters)** | 95 | 1425 | 1407 | 6093 |

## How Well Do Candidate Sets Capture the Target Terms?

Because the candidate sets determine what terms can be predicted, the degree to which they capture the target MeSH is extremely important. In the following sections, I use ”capture rate” to describe the percentage of label terms that are covered by the candidate sets. For example, if all of the paper’s MeSH are captured by the candidates, this would be considered a capture rate of 100%. The capture rate of both citations and absim records is high with a mean of 86% and 74%, respectively. Further, the capture rate of citations and absim are weakly correlated ($R^2$=.32). The combination of the citation and absim candidates captures 91% of the target MeSH on average with a median of 93%. The following figure shows the distribution of capture rates in the citation and absim sets, as well as the citation of citation and citation of absim sets.

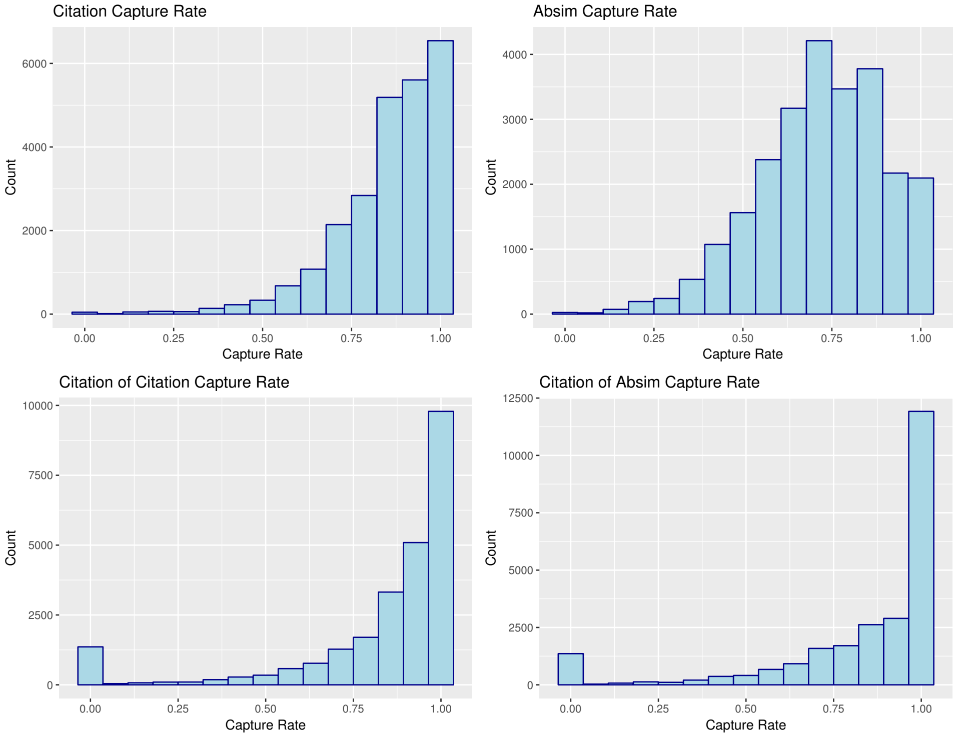


Figure 4.3: Capture Rates in Citations, Absim, Citations of Citations and Citations of Absim

The following figure displays the increase in average cumulative recall by each added citation. On average, 50% of terms are captured within 5 citations. A slightly greater number of absim records are required to achieve the same capture rate.

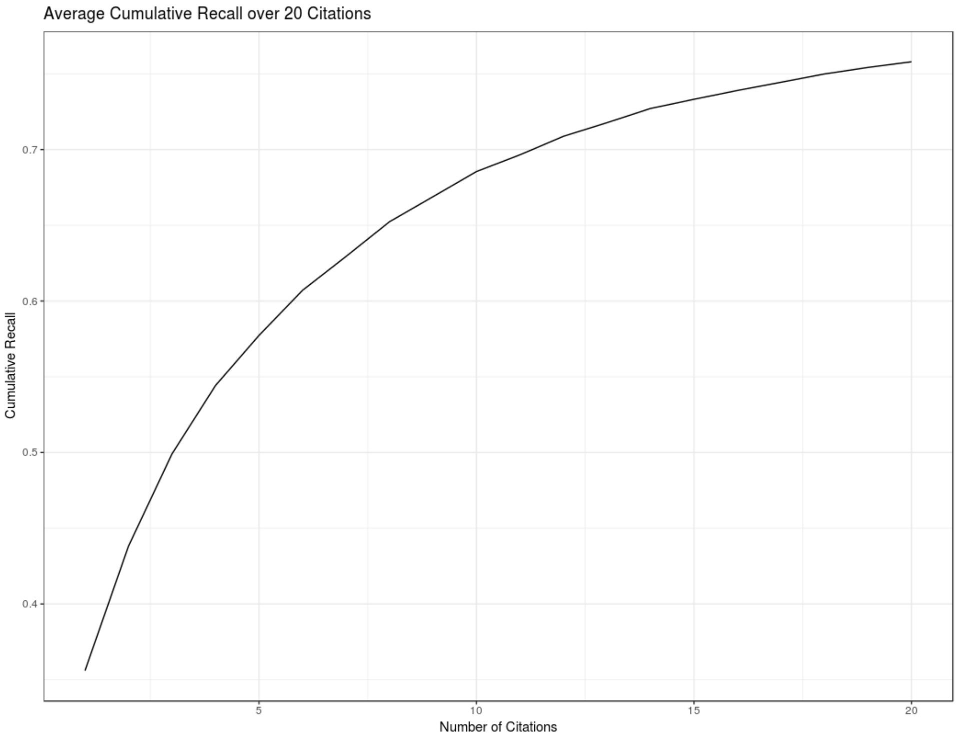


Figure 4.4: Average Cumulative Recall over 20 Citations

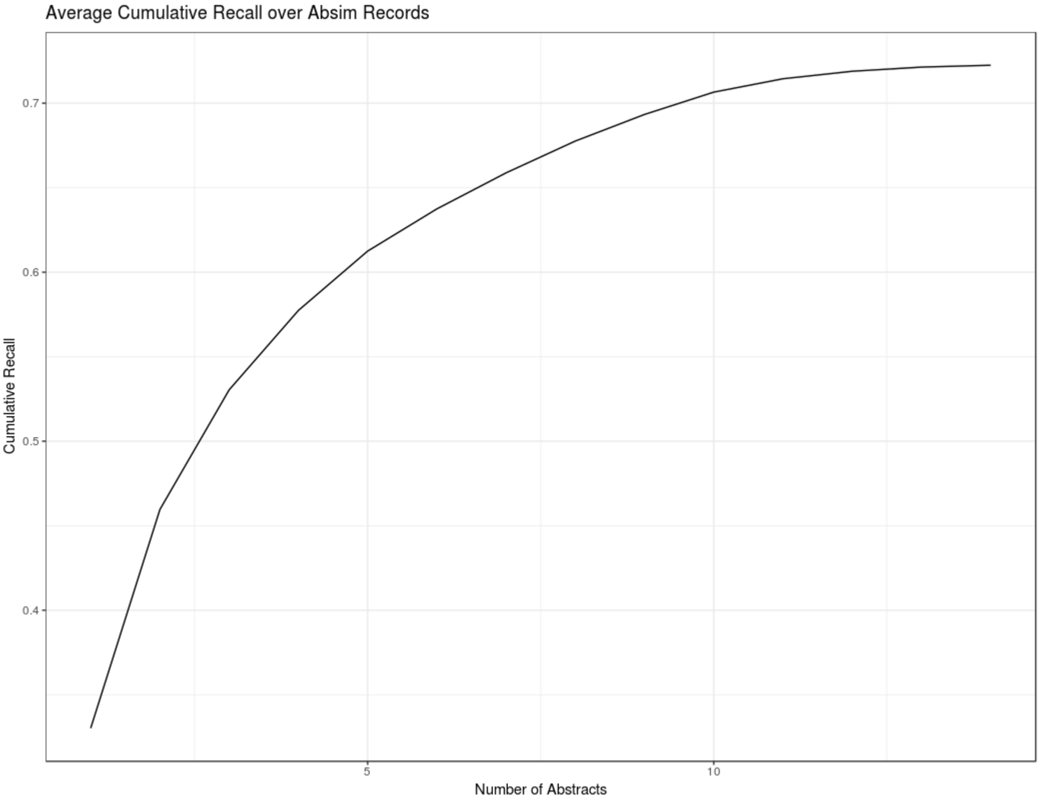


Figure 4.5: Average Cumulative Recall over 20 Absim Records

## 4.1 Complimentarity of Citations and Text

To summarize, the candidate collection procedure described above produces four sets of MeSH terms, each of which have high recall of the target paper's MeSH. A natural question is the degree to which these sets are complimentary, that is, the degree to which they produce unique and useful terms. Although "complimentarity" may have many interpretations, here complimentarity is measured by the symmetric difference between the "true" MeSH terms of any two candidate sets. In other words, complimentarity is meant to describe the unique and correct terms captured by a pair of candidate sets. Note that complimentarity is primarily useful for comparing two fundamentally dissimilar candidate sets – for example, citations and absim. The distinction tends to be less relevant when using derived sets, such as "citations of citations" or "citations of absim", as these reflect underlying differences in the original candidate sets.

Within the training data, in 89% of papers there is at least one unique, correct term within the absim or citation sets. On average, a paper has 3.1 unique and correct terms provided by either the absim or citation sets, constituting 23% of the overall MeSH. There are two situations which account for the remaining 12.51% of papers without complimentarity. In 93.43% of these cases there is no complimentarity because both the absim and the cit sets each captured all of the relevant terms – making it impossible for either set to contribute a unique term. In the remaining minority of cases there is no complimentarity because neither the absim or citation sets had any true terms at all. Of those, 46 had no recall in the citation set, and 25 had no recall from the absim set. 3 papers out of 25,000 had no recall from either citations or absim.

The above establishes that there is complimentarity between the two sets. A deeper question is why there is complimentarity. One obvious mechanism of complimentarity is temporal. Since citations are inherently retrospective (i.e. it is only possible to cite what has already been published), they are limited by whatever the MeSH vocabulary and indexing practices were at the time of publication. Abstract similarity can retrieve related records published after the target record, and thus capture more relevant terms. On average, the absim set has 5 records published after the target date, or 45% newer. The impact of temporality varies significantly; the older the paper, the greater the number of newer absim papers ($R^2 = -.58$).

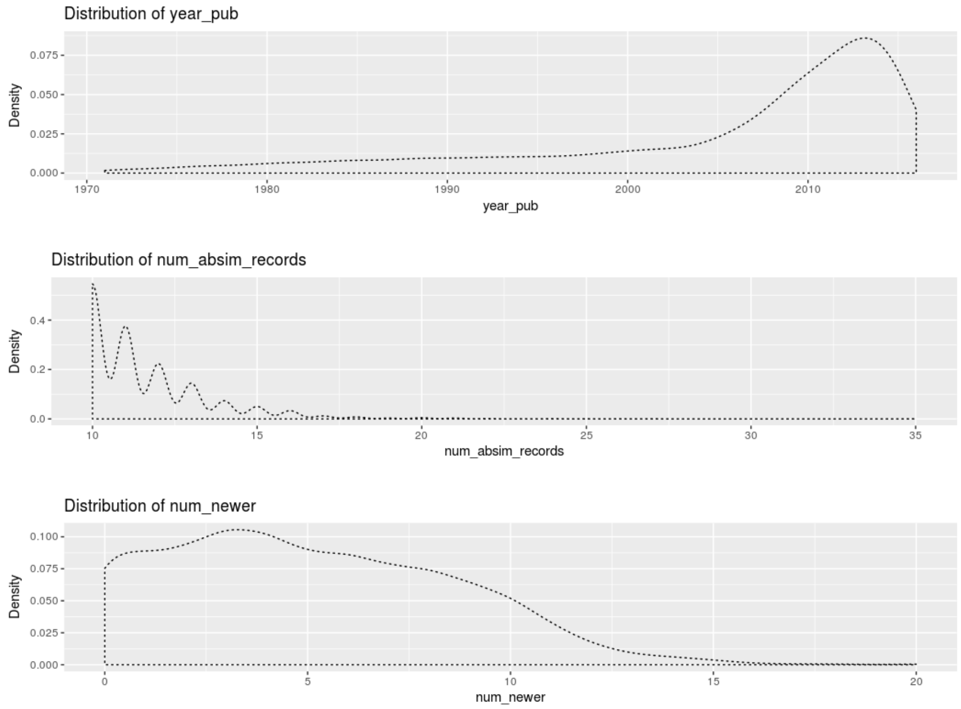


Figure 4.6: Caption

Another mechanism of complimentarity is citing behavior. It is unrealistic for researchers to comprehensively cite the literature. The absim set may reflect highly related papers that were not cited because the researcher was unaware of the work, or because the work had indirect bearing on the work at hand. Overlap is limited, with an average of 1 shared paper between the citations and absim sets. Again, temporal factors are significant. The older the paper, the more likely there is to be newer absim papers, and by definition there can be no overlap between citations and newer absim papers. The greater the number of newer absim papers, the total number of overlapping papers falls ($R^2 = -.32$).

## 4.2 Identifying Relevant Candidate Terms: A Hybrid K-Nearest Neighbors and Binary Classification Approach

Above, I describe how citations and text-based term sets are complimentary – that is, that they uniquely contribute ”true” candidate terms. The underlying mechanisms of this complimentarity are temporal factors and citing behaviors. The next important research question is how well citations and abstracts serve as features in a predictive model.

The model collects a large set of individual candidate terms from training set papers. Each candidate term is indicated as either relevant or irrelevant based on whether it was selected by a human annotator. Final predictions are obtained by taking the top k (throughout, k is set to 15) terms, ranked by probability of relevance.

The inherent class imbalance between relevant and irrelevant candidate terms leads to relatively modest precision and recall in the binary model. That is, the exact probability of any given candidate term tends to be very low due to the overwhelming number of irrelevant terms. As a result, the evaluation here focuses on the final multilabel setting rather than the intermediate binary model.

## Test Data

Three evaluation datasets were constructed to test each of the models above, each comprised of 5,000 records:

1. Few citations: fewer than 5 citations, and abstract of at least 250 characters or greater
2. Balanced: at least 15 citations and an abstract of at least 250 characters or greater (average of 1300 characters)
3. Short abstracts: at least 15 citations and abstract of fewer than 750 characters

In the following sections, each model is briefly described along with the multilabel evaluation measures for each of the three test datasets.

## Citation Only Model

In the first model, term relevance is predicted based on the log of the number of times the candidate term appeared in the citation set, as well as the log of the total number of citation candidate terms. Additionally, the log frequency of the term in MEDLINE is included to adjust for the overall rarity of the term. The model was trained using 10 fold cross-validation using the dataset of 25,000 papers described above.

This version of the model is highly sensitive to the number of available citations. In the "low citation" test set, mean accuracy drops by 29%. Notably, the median one-error is the worst of all models at 0.

Table 4.2: Citation Only Model Coefficients

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | -1.13 | .019 | -56.99 | <2e-16 |
| Citation Count | 2.49 | .003 | 757.38 | <2e-16 |
| Total Citations | -1.36 | .004 | -338.73 | <2e-16 |
| Log(MEDLINE Frequency) | -.03 | .001 | -19.85 | <2e-16 |

### Balanced Citations and Abstracts

Table 4.3: Citation Only Model: Full Citations and Abstracts

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 |
| 1st. Qu. | .33 | .40 | 1.00 | .23 |
| Median | .47 | .50 | 1.00 | .30 |
| Mean | .45 | .51 | .90 | .31 |
| 3rd Qu. | .53 | .62 | 1.00 | .39 |
| Max | .93 | 1.00 | 1.00 | .82 |

### Limited Citations

Table 4.4: Citation Only Model: Low Citation

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 |
| 1st. Qu. | 0 | 0 | 0 | 0 |
| Median | .13 | .17 | 0 | .08 |
| Mean | .16 | .24 | .37 | .11 |
| 3rd Qu. | .27 | .40 | 1.00 | .18 |
| Max | .89 | 1.00 | 1.00 | .89 |

### Short Abstract

Table 4.5: Citation Only Model: Short Abstracts

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 |
| 1st. Qu. | .27 | .42 | 1.00 | .20 |
| Median | .33 | .53 | 1.00 | .27 |
| Mean | .37 | .55 | .88 | .28 |
| 3rd Qu. | .47 | .67 | 1.00 | .35 |
| Max | .87 | 1.00 | 1.00 | .77 |

## Absim Only Model

In the next version of the model, only the absim features are used. Performance overall is lower than in the citation-only version in the normal and short abstract versions. However, unlike the citation only model, performance remains very stable, even in the short abstract test set. There are several possible mechanisms for this difference. Whereas having fewer citations dramatically reduces the amount of information available to the citation model, a short abstract still provides a great deal of information. The overall average number of related records is not reduced. Secondly, the cumulative recall curve is steeper for abstract records than for citations. In other words, more terms are captured with fewer additional papers.

Though this model is more robust in all three settings, it is still has lower performance than the citation only model, likely due to the additional noise introduced by the text similarity algorithm.

Table 4.6: Absim Only Model Coefficients

| | Estimate | Std. Error | z value | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -3.72 | .041 | -90.84 | 2e-16 |
| Citation Count | 3.09 | .003 | 1004.80 | 2e-16 |
| Total Citations | -1.19 | .016 | -75.24 | 2e-16 |
| Log(MEDLINE Frequency) | -.11 | .001 | 108.98 | 2e-16 |

### Balanced Citations and Abstracts

Table 4.7: Absim Only Model: Normal

| | Subset Accuracy | Recall | One Error | microF | Jaccard Index |
|---|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 | 0 |
| 1st. Qu. | .33 | .36 | 1.00 | 0.33 | .20 |
| Median | .40 | .47 | 1.00 | .44 | .28 |
| Mean | .42 | .48 | .91 | .44 | .29 |
| 3rd Qu. | .53 | .60 | 1.00 | .54 | .37 |
| Max | .93 | 1.00 | 1.00 | .89 | .81 |

### Limited Citations

Table 4.8: Absim Only Model: Low Citation

| | Subset Accuracy | Recall | One Error | microF | Jaccard Index |
|---|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 | 0 |
| 1st. Qu. | .20 | .38 | 1.00 | .29 | .17 |
| Median | .33 | .50 | 1.00 | .40 | .25 |
| Mean | .35 | .52 | .88 | .40 | .27 |
| 3rd Qu. | .47 | .67 | 1.00 | .52 | .35 |
| Max | 1.00 | 1.00 | 1.00 | .94 | .89 |

### Short Abstract

Table 4.9: Absim Only Model: Short Abstract

| | Subset Accuracy | Recall | One Error | microF | Jaccard Index |
|---|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 | 0 |
| 1st. Qu. | .27 | .38 | 1.00 | .29 | .17 |
| Median | .33 | .50 | 1.00 | .40 | .25 |
| Mean | .35 | .52 | .88 | .40 | .26 |
| 3rd Qu. | .47 | .67 | 1.00 | .50 | .33 |
| Max | .93 | 1.00 | 1.00 | .93 | .88 |

## Citation and Absim

In the next iteration of this approach, I include the log counts of the candidate term in the absim set, as well as the overall number of absim candidate terms. The inclusion of the absim set produces a significant increase in the accuracy of the model. Further, the model remains robust across all three datasets.

Table 4.10: Citation and Absim Model Coefficients

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | -0.63 | .045 | -14.20 | <2e-16 |
| Citation Count | 1.84 | .003 | 571.92 | <2e-16 |
| Total Citations | -1.03 | .004 | -295.58 | <2e-16 |
| Absim Count | 1.36 | .004 | 329.50 | <2e-16 |
| Total Absim | -0.51 | .016 | -30.90 | <2e-16 |
| Log(MEDLINE Frequency) | -0.06 | .001 | -51.23 | <2e-16 |

### Balanced Citations and Abstracts

Table 4.11: Simple Model: Full Citations and Abstracts

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | .07 | .07 | 0 | .03 |
| 1st. Qu. | .33 | .43 | 1.00 | .25 |
| Median | .47 | .54 | 1.00 | .33 |
| Mean | .48 | .54 | .92 | .34 |
| 3rd Qu. | .60 | .64 | 1.00 | .42 |
| Max | 1.00 | 1.00 | 1.00 | .83 |

### Limited Citations

Table 4.12: Simple Model: Low Citation

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0.00 |
| 1st. Qu. | .27 | .40 | 1.00 | .18 |
| Median | .33 | .55 | 1.00 | .27 |
| Mean | .37 | .54 | .86 | .28 |
| 3rd Qu. | .47 | .67 | 1.00 | .38 |
| Max | 1.00 | 1.00 | 1.00 | .89 |

### Short Abstract

Table 4.13: Simple Model: Short Abstract

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0.00 |
| 1st. Qu. | .27 | .45 | 1.00 | .22 |
| Median | .40 | .57 | 1.00 | .29 |
| Mean | .40 | .58 | .90 | .31 |
| 3rd Qu. | .47 | .71 | 1.00 | .38 |
| Max | .93 | 1.00 | 1.00 | .88 |

## Citations, Abstracts, and Related Records

In order to further enrich the underlying candidate set, I included ”citations of citations” and ”citations of absim” records. Expanding the number of related records increases the overall candidate size and provides source of evidence about the relevance of the term.

Table 4.14: Citation, Absim and Related Record Coefficients

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | -0.40 | .045 | -8.31 | 2e-16 |
| Citation Count | 1.69 | .005 | 363.39 | <2e-16 |
| Total Citations | -0.87 | .005 | -168.35 | <2e-16 |
| Citation of Citation Count | -0.01 | .003 | -4.97 | 6.83e-07 |
| Total Citations of Citations | -0.05 | .002 | -18.46 | <2e-16 |
| Absim Count | 1.17 | .005 | 255.69 | <2e-16 |
| Total Absim | -0.40 | .017 | -24.01 | <2e-16 |
| Citations of Absim Count | 0.23 | .003 | 80.02 | <2e-16 |
| Total Citations of Absim | -0.16 | .002 | -70.39 | <2e-16 |
| Log(MEDLINE Frequency) | -.07 | .001 | -65.54 | <2e-16 |

### Balanced Citations and Abstracts

Table 4.15: Full Model: Full Citations and Abstracts

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | .07 | .06 | 0 | .03 |
| 1st. Qu. | .33 | .42 | 1.00 | .24 |
| Median | .47 | .53 | 1.00 | .32 |
| Mean | .47 | .54 | .91 | .33 |
| 3rd Qu. | .60 | .64 | 1.00 | .41 |
| Max | 1.00 | 1.00 | 1.00 | .83 |

### Limited Citations

Table 4.16: Full Model: Low Citations

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 |
| 1st. Qu. | .27 | .40 | 1.00 | .18 |
| Median | .33 | .56 | 1.00 | .27 |
| Mean | .37 | .55 | .87 | .28 |
| 3rd Qu. | .47 | .66 | 1.00 | .37 |
| Max | 1.00 | 1.00 | 1.00 | .81 |

### Short Abstract

Table 4.17: Full Model: Short Abstract

| | Subset Accuracy | Recall | One Error | Jaccard Index |
|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 |
| 1st. Qu. | .27 | .45 | 1.00 | .22 |
| Median | .40 | .57 | 1.00 | .29 |
| Mean | .39 | .58 | .89 | .30 |
| 3rd Qu. | .47 | .70 | 1.00 | .38 |
| Max | .93 | 1.00 | 1.00 | .88 |

## 4.3 Partial Matching Evaluation

This section has three components. The first is an evaluation of the models described above using existing hierarchical partial matching measures, namely the "hierarchical" and Least Common Ancestor precision, recall and f-score. The second section applies the novel word embedding based metrics introduced in Chapter 3. These metrics reflect how well the predicted classes represent the true classes in terms of their cosine similarity and hierarchy position. The final section demonstrates how the word embedding based evaluation measures can be used for intensive model debugging. For example, the similarity based measure allows identification of the best and least well represented terms, as well as commonly "confused" terms. Further, the similarity measure is applied to an analysis of highly similar and redundant prediction terms.

## Hierarchical Measures

The "hierarchical" measures naively augment the true and predicted classes with all of the ancestor terms. The Lowest Common Ancestor metric, as described in Chapter 2, is a more sophisticated approach that selectively augments the true and predicted classes by using the shortest paths between the lowest common ancestor terms. These measures reflect a high degree of similarity with the flat hierarchical measures. For each model, the evaluation was run against the "normal" conditions dataset (i.e. the dataset with average citations and abstracts). As in the flat evaluation, the performance is relatively stable across models but is highest in the simple model.

Table 4.18: Hierarchical Evaluation Measures: "Normal" Datasets

| | **hP** | **hR** | **hF** | **LCA-P** | **LCA-R** | **LCA-F** |
|---|---|---|---|---|---|---|
| Simple | .67 | .68 | .66 | .43 | .45 | .43 |
| Full | .67 | .67 | .65 | .43 | .45 | .42 |
| Cit Only | .65 | .67 | .64 | .42 | .44 | .41 |
| Absim Only | .62 | .64 | .61 | .40 | .42 | .40 |

## Combined Word Embedding and Hierarchical Measures

The following table collects the performance of the models in the normal conditions dataset using the word embedding based metrics described above. To review from Chapter 3, the four metrics here reflect different views of accuracy:

1. HierarchyFree: This calculates partial matching between terms with a hierarchy relationship (child, parent, sibling, etc) weighted by their pairwise cosine similarity. The "free" component means that any predicted term is free to match against any true term; in other words, multiple predicted terms may match against the same true term. This metric is meant to capture the overall "sensibility" of the predictions balanced against a close match to the true terms place in the hierarchy.

2. NonHierarchyFree: This measure reflects the best cosine similarity match between predicted and true terms. Unlike above, this is not restricted to terms with hierarchy relationships. Because this is less restricted, the value is typically higher. However, due to the lack of constraint in hierarchy relationship, this may provide partial match credit to term pairs that have a common context but are categorically different; e.g. a disease-drug pair. As above, this version allows multiple predicted terms to match to one true term.

3. HierarchyRestricted: This metric is the same as HierarchyFree, but only permits each true term to be matched once. This prevents prediction sets from having a high score if they only approximately match one or two true classes.

4. NonHierarchyRestricted: This is the same as NonHierarchyFree metric, but only permits each true term to be matched once.

There is a somewhat more clear difference in evaluation measures here than in the traditional hierarchical measures. The difference between the HierarchyFree and HierarchyRestricted performances indicate some degree of redundancy in the prediction sets, with a drop of .07 in each model.

Table 4.19: Embedding Based Partial Match Measures: "Normal" Datasets

| | **HierarchyFree** | **NonHierarchyFree** | **HierarchyRestricted** | **NonHierarchyRestricted** |
|---|---|---|---|---|
| Simple | .52 | .72 | .45 | .52 |
| Full | .51 | .71 | .45 | .51 |
| CitOnly | .50 | .71 | .43 | .51 |
| AbsimOnly | .47 | .68 | .40 | .49 |

The overall semantic similarity of predictions varies between and within branches. Figure 4.7 plots the distribution of the best match semantic similarity of predictions to labels in each branch. Some suggestive patterns are clear: branches that are relatively conceptually narrow (Technology and Industry and Health Care) have a tighter range and slightly higher average matching. Broader categories like Diseases and Phenomena and Processes have some of the widest ranges.

Intriguingly, Information Science has the highest median matching, with the widest difference between median and mean. Further work is required to make stronger claims about the significance, but one possible cause is the combination of the unusual conceptual broadness of Information Science and a significant frequency imbalance in terms. The Information Science branch contains an extremely wide range of terms, from "Algorithm" to "Phylogeny" to "Interlibrary Loans". However, as will be discussed later, the most frequent Information Science terms tend to be squarely biomedical, such as "Molecular Sequence Data."

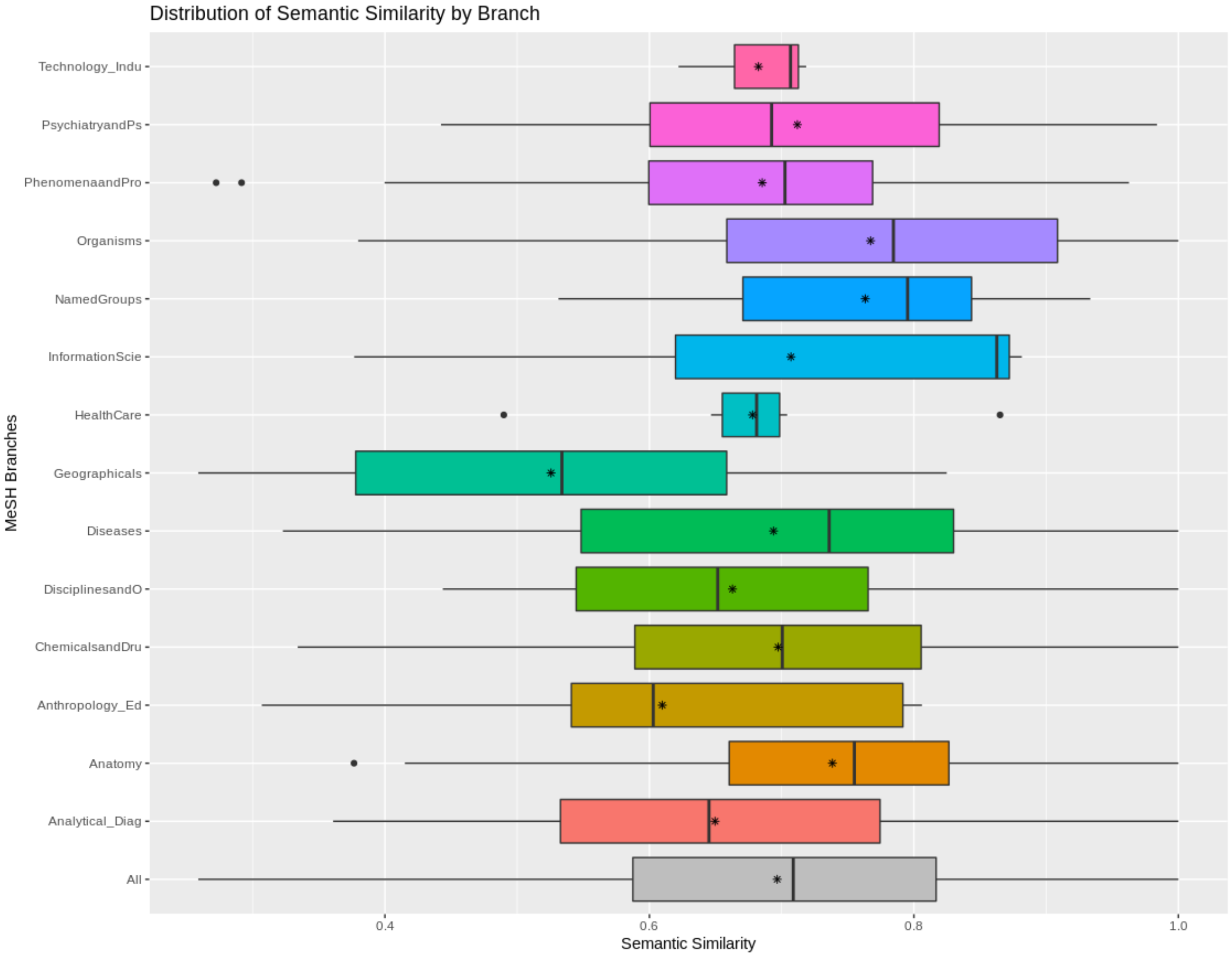


Figure 4.7: Semantic Similarity of Predictions by Branch

### Using Word Embeddings for Model Performance Diagnostics

The semantic similarity approach can be useful for diagnosing model performance with respect to individual terms, as well. Table 4.26 represents the 5 terms with lowest average semantic matching in the predictions. The ”Frequent Matches” column describes the 5 terms that were the most frequent best match to the target. In several cases, the best matches are close: for example, ”Gene Expression” has an overall low matching score but is most frequently confused with ”Gene Expression Regulation”, its sibling term. In two cases, the term itself is the most frequent match but with a high degree of confusion with other terms. Table 4.27 provides further detail on the best matches with ”Young Adult.” Here, we find that 20% of the time Young Adult is successfully matched with itself, but 34% of the time is matched with other

Table 4.21: Disambiguation of Young Adult Terms

| **Term** | **Branch** | **Semantic Similarity** | **Frequency** |
|---|---|---|---|
| Young Adult | Named Group | 1.0 | 20% |
| Middle Aged | Named Group | .27 | 18% |
| Child, Preschool | Named Group | .33 | 10% |
| Child | Named Group | .31 | 6% |
| Cross-Sectional Study | Analytical, Health Care | .39 | 6% |

age related named groups like ”Middle Aged” and ”Child, Preschool.” While these terms are clearly related to age demographics, they have low semantic similarity because they have different contexts – the topics connected to children are different from those of young adults. The model difficulty in differentiating or picking the appropriate term in this case. This suggests that further work is required to address this difficulty, either by using a dedicated model or applying a policy like consolidating age groups to the summary term ”Age Groups” if there are multiple ages predicted or if confidence is low.

Table 4.20: Lowest Average Semantic Similarity Terms

| **Term** | **Branch** | **Avg Similarity** | **Frequent Matches** |
|---|---|---|---|
| Time Factors | Phenomena | .40 | Female, Time Factors, Humans, Male, Age Factors |
| Gene Expression | Phenomena | .41 | Gene Expression Regulation, ”Transcription, Genetic”, Cell Differentiation |
| Young Adult* | Named Group | .45 | Young Adult, Middle Aged, ”Child, Preschool”, Child, Cross-Sectional Studies |
| Recombinant Proteins | Chemicals | .46 | Recombinant Proteins, Amino Acid Sequence, ”Cloning, Molecular ”,Transfection, Protein Binding |
| Models, Biological | Analytical... | .48 | ”Models, Biological”, Signal Transduction, Protein Binding, Computer Simulation, Biological Transport |

Table 4.28 collects the best matching terms. Here, we see that the frequency basis of the models leads to high performance in highly common terms. Terms like ”Animals” and ”Humans” are overwhelming correctly matched to themselves. However, there is an interesting continuity with the worst results – the term ”Middle Aged” is correctly matched 90% of the time, but when it is confused, it is frequently confused with ”Child” and ”Young Adult.”

Table 4.22: Highest Average Semantic Similarity Terms

| Term | Branch | Avg. Similarity | Frequent Matches |
|---|---|---|---|
| Animals | Organisms | .99 | Animals (99%), Middle Aged, Sequence Alignment, Structure-Activity Relationship, Phosphorylation |
| Humans | Organisms | .99 | Humans (99%), Female, Calcium, DNA-Binding Proteins, Mycobacterium tuberculosis |
| Male | NA | .96 | Male (96%), Cell Differentiation, Humans, Transcription Factors, Cell Cycle Proteins |
| Female | NA | .95 | Female (95%), Humans, Bacterial Proteins, Immunohistochemistry, nfluenza, Human" |
| Middle Aged | Named Group | .95 | Middle Aged (90%), Animals, Humans, Child, Young Adult |

### Analyzing Highly Similar MeSH Terms

The modeling methodology described above essentially makes independent judgments about the relevance of a MeSH term. As previously shown in studies of duplicate annotations, there are often many plausible, highly related MeSH terms that might be appropriate for a given paper. It follows that there may be a possibility of ranking several similar terms highly. These terms may be so similar as to be redundant. A method for measuring the degree of redundancy is an essential first step to consolidating highly similar MeSH terms.

The evaluation results above show a marked difference (-.07) between the "free" and "restricted" metrics, indicating that multiple predictions tend to be associated with one true term. A natural question arises: how often are "redundant" terms used in real annotations? How often do they appear in predictions? Are there patterns in the types of redundant terms that appear?

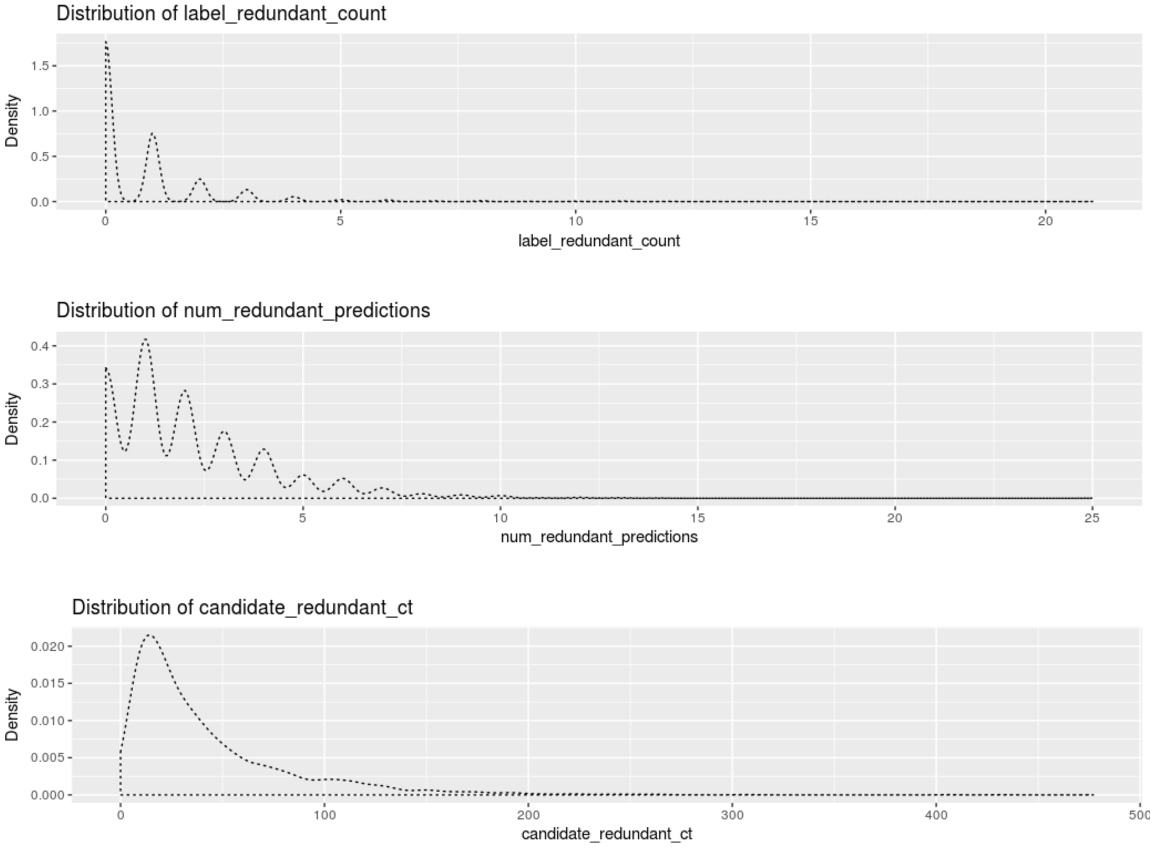
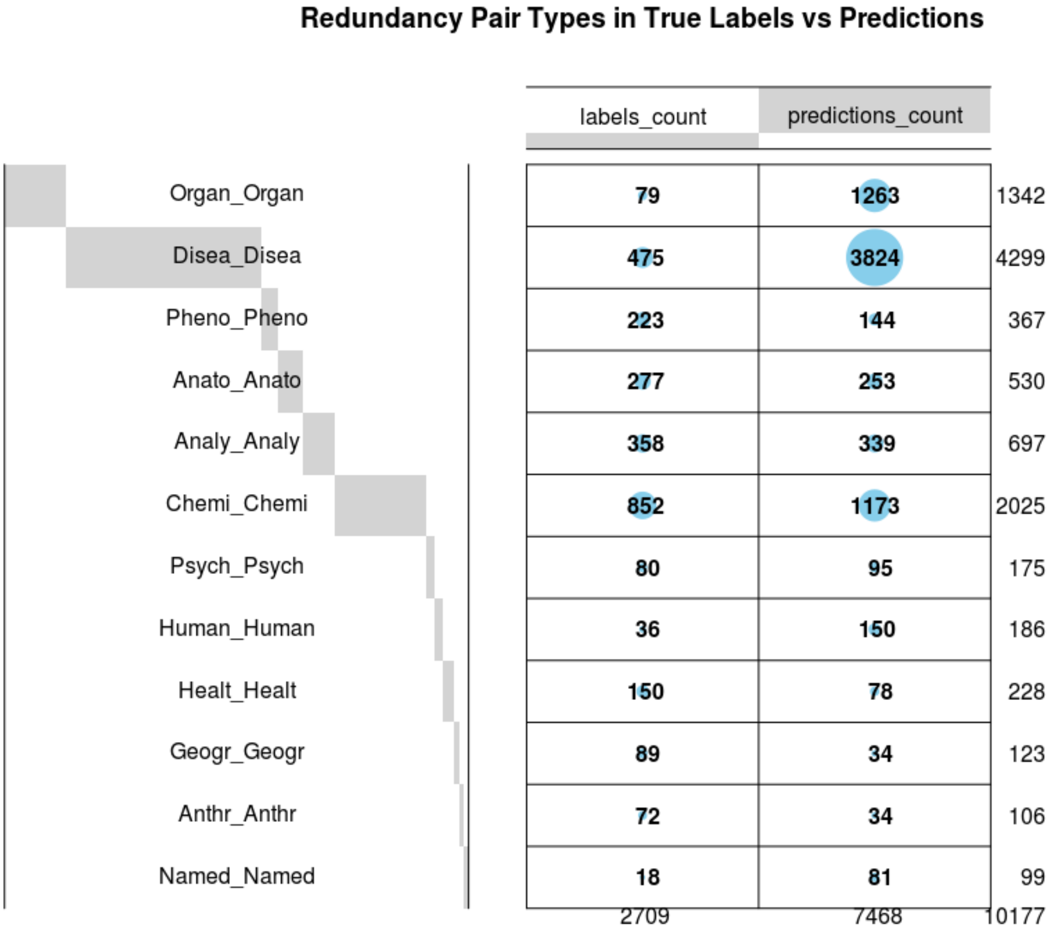


Figure 4.8: Redundancy of Labels, Predictions and Candidates

To address this question, the predicted terms, actual terms and candidate terms were partitioned into possibly redundant and non-redundant terms using a cosine similarity threshold of 0.8. Naturally, many term pairs will have highly similar contexts in true labels. The challenge is to differentiate between naturally co-occurring terms and potentially redundant predictions. Figure 4.7 shows the distribution of redundancy in each set. Most papers contain an expected degree of redundancy in labels, and a higher degree of redundancy among predictions.



Redundancy Pair Types in True Labels vs Predictions

| | labels_count | predictions_count | |
|---|---|---|---|
| Organ_Organ | 79 | 1263 | 1342 |
| Disea_Disea | 475 | 3824 | 4299 |
| Pheno_Pheno | 223 | 144 | 367 |
| Anato_Anato | 277 | 253 | 530 |
| Analy_Analy | 358 | 339 | 697 |
| Chemi_Chemi | 852 | 1173 | 2025 |
| Psych_Psych | 80 | 95 | 175 |
| Human_Human | 36 | 150 | 186 |
| Healt_Healt | 150 | 78 | 228 |
| Geogr_Geogr | 89 | 34 | 123 |
| Anthr_Anthr | 72 | 34 | 106 |
| Named_Named | 18 | 81 | 99 |
| | 2709 | 7468 | 10177 |

Figure 4.9: Redundancy of Labels and Predictions by MeSH Branch

The distribution of this redundancy differs between branches in true terms versus predictions. Notably, the degree of redundancy is higher among disease, organism and chemical terms, as visualized in Figure 4.8. The pattern is less clear when analyzed by the type of relationship between similar terms (sibling, parent-child, etc), other than there is greater overall redundancy among predictions.

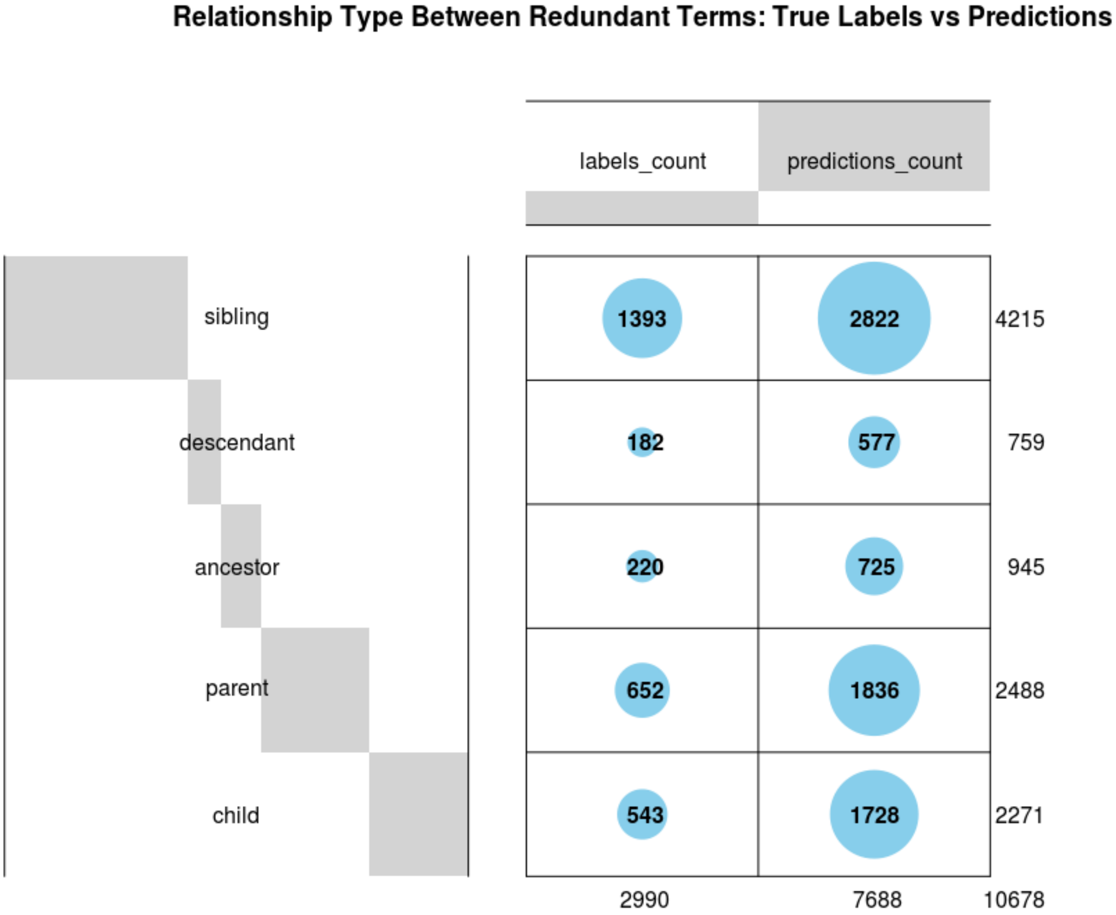


Figure 4.10: Redundancy of Labels and Predictions by MeSH Branch

Table 4.24 provides examples of the top 5 redundant term pairs among predictions. In many cases, terms are so similar as to be considered synonymous (ex: Anxiety Disorders vs Phobic Disorders". In other cases, the relationship seems to be one of subcategory ("Diabetes Mellitus, Type 2" versus "Diabetes Mellitus"). In a few cases, the terms are clearly not synonyms – for example, "Men" and "Women", though putatively their context in MeSH is highly similar.

Table 4.23: Top Redundant Terms in Predictions

| **Relationship Type** | **Redundant Pair** |
|---|---|
| Organism | Rats, Sprague-Dawley \| Rats, Wistar |
| Organism | Rats, Sprague-Dawley \| Rats, Inbred Strains |
| Organism | Rats, Wistar \| Rats, Inbred Strains |
| Organism | Rats, Wistar \| Rats, Sprague-Dawley |
| Organism | Mice, Inbred C57BL \| Mice, Inbred BALB C |
| Disease | HIV Infections \| Acquired Immunodeficiency Syndrome |
| Disease | Cardiovascular Diseases \| Coronary Disease |
| Disease | Diabetes Mellitus, Type 2 \| Diabetes Mellitus |
| Disease | Cardiovascular Diseases \| Hypertension |
| Disease | Coronary Disease \| Myocardial Infarction |
| Phenomenon | Genetic Variation \| Genotype |
| Phenomenon | Protein Conformation \| Protein Structure, Secondary |
| Phenomenon | Algorithms \| Artificial Intelligence |
| Phenomenon | Ion Channel Gating \| Membrane Potentials |
| Phenomenon | Stress, Mechanical \| Biomechanical Phenomena |
| Anatomy | Neurons \| Axons |
| Anatomy | Cerebral Cortex \| Brain |
| Anatomy | Epidermis \| Skin |
| Anatomy | Hematopoietic Stem Cells \| Bone Marrow Cells |
| Anatomy | Neurons \| Synapses |
| AnalyticTechnique | Palliative Care \| Terminal Care |
| AnalyticTechnique | Models, Biological \| Models, Theoretical |
| AnalyticTechnique | Randomized Controlled Trials as Topic \| Clinical Trials as Topic |
| AnalyticTechnique | Equipment Design \| Evaluation Studies as Topic |
| AnalyticTechnique | Retrospective Studies \| Prospective Studies |
| Chemicals | Anti-Inflammatory Agents, Non-Steroidal \| Cyclooxygenase Inhibitors |
| Chemicals | Membrane Proteins \| Carrier Proteins |
| Chemicals | Luteinizing Hormone \| Follicle Stimulating Hormone |
| Chemicals | Anti-Inflammatory Agents, Non-Steroidal \| Cyclooxygenase 2 Inhibitors |
| Chemicals | Follicle Stimulating Hormone \| Luteinizing Hormone |
| Psychology | Risk-Taking \| Adolescent Behavior |
| Psychology | Risk-Taking \| Sexual Behavior |
| Psychology | Anxiety Disorders \| Phobic Disorders |
| Psychology | Visual Perception \| Pattern Recognition, Visual |
| Psychology | Reward \| Reinforcement (Psychology) |
| Healthcare | Safety Management \| Patient Safety |
| Healthcare | Medicaid \| Medicare |
| Healthcare | Medicare \| Medicaid |
| Healthcare | Population \| Population Characteristics |
| Healthcare | Health Care Costs \| Cost-Benefit Analysis |
| NamedGroups | African Americans \| Hispanic Americans |
| NamedGroups | Infant, Premature \| Infant, Low Birth Weight |
| NamedGroups | Men \| Women |
| NamedGroups | European Continental Ancestry Group \| African Continental Ancestry Group |
| NamedGroups | European Continental Ancestry Group \| Continental Population Groups |

Calculation of highly similar terms permits identifying papers with highly similar terms. Table 4.25 collects five papers with ”redundant” pairs. The top example has MeSH assignments for every subtype of Hyperlioproteinemia.

Table 4.24: Top Papers with Redundant Labels

| **PMID** | **Redundant Pairs** | **Labels** |
|---|---|---|
| 6346238 | 21 | Humans;Hyperlipidemia, Familial Combined;Hyperlipoproteinemia Type I; Hyperlipoproteinemia Type II;Hyperlipoproteinemia Type III; Hyperlipoproteinemia Type IV;Hyperlipoproteinemia Type V; Hyperlipoproteinemias;Lipoproteins;Lipoproteins, HDL |
| 12816106 | 15 | Adolescent;Adult;Animals;Animals, Domestic;Child; Child, Preschool;Contact Tracing;Disease Outbreaks; Female;Humans;Illinois;Indiana;Infant;Kansas;Male; Middle Aged;Missouri;Monkeypox virus;Muridae; Ohio;Poxviridae Infections;Sciuridae;Wisconsin |
| 6177960 | 14 | Adrenergic alpha-Antagonists;Adrenergic beta-Antagonists; Adult;Antihypertensive Agents;Cholesterol;Cholesterol, LDL; Cholesterol, VLDL;Drug Therapy, Combination;Humans; Hypertension;Lipids;Lipoproteins;Lipoproteins, LDL; Lipoproteins, VLDL;Male;Middle Aged;Triglycerides |
| 12341391 | 14 | Africa;Africa South of the Sahara;Africa, Southern;Agriculture; Demography;Developing Countries;Economics; Emigration and Immigration;Employment; Health Manpower; Income;Lesotho;Population;Population Dynamics;Social Planning; Socioeconomic Factors;South Africa;Technology |
| 6815875 | 13 | Animals;Candidiasis;Coccidioidomycosis;Cryptococcosis; Digestive System Diseases; Entomophthora;Geotrichosis; Haplorhini;Monkey Diseases;Mucormycosis;Mycoses; Pan troglodytes;Papio;Paracoccidioidomycosis;Primates |

Perhaps more importantly, this approach also allows identification of concentration of highly similar terms in the predictions. The following table provides examples of the top 5 papers with highly redundant predictions:

Table 4.25: Top Papers with Redundant Predictions

| PMID | Predictions |
|---|---|
| 6293146 | Animals;Macaca mulatta;Primates;Macaca fascicularis;<br>Haplorhini;Callitrichinae;Galago;Saguinus;Callithrix;<br>Cebidae;Papio;Aotus trivirgatus;Saimiri;Cercopithecus aethiops;<br>Gastrointestinal Neoplasms |
| 16009392 | Humans;Depth Perception;Vision, Binocular;<br>Vision Disparity;Photic Stimulation;Vision, Monocular;<br>Visual Perception;Animals;Contrast Sensitivity;<br>Pattern Recognition, Visual;Form Perception;<br>Psychophysics;Adult;Optical Illusions;Visual Cortex |
| 1553698 | Middle Aged;Adult;Wounds and Injuries;Adolescent;<br>Aged;Child;Child, Preschool;Aged, 80 and over;<br>Wounds, Nonpenetrating;Young Adult;<br>Infant;Multiple Trauma;Abdominal Injuries;<br>Thoracic Injuries;Wounds, Penetrating |
| 12507406 | Stroke;Animals;Cardiovascular Diseases;<br>Hydroxymethylglutaryl-CoA Reductase Inhibitors;<br>Atherosclerosis;Hypertension;Arteriosclerosis;<br>Postoperative Complications;Vascular Diseases;<br>Inflammation;Coronary Artery Disease;Heart Failure;<br>Carotid Artery Diseases;Thrombosis;Carotid Stenosis |
| 1349908 | Ovarian Neoplasms;Breast Neoplasms;Receptor, ErbB-2;<br>Mice;Animals;Neoplasm Invasiveness;Laryngeal Neoplasms;<br>Genital Neoplasms, Female;Adenocarcinoma;Endometrial Neoplasms;<br>Uterine Neoplasms;Carcinoma;Carcinoma, Squamous Cell;<br>Fallopian Tube Neoplasms;Uterine Cervical Neoplasms |

The first example paper with PMID 6293146 is titled ”Gastrointestinal neoplasms in nonhuman primates: a review and report of eleven new cases,” and include the terms ”Callitrichinae” and ”Macaca mulatta” as MeSH assignments. The abstract further references ”Saguinus” and ”Galago crassicaudatus”. Here, the related records return an abundance of other primate species. This example provides a difficult case for annotation heuristics. MeSH indexers are instructed to combine sibling terms using a parent term. This rule could be applied here to simplify the annotation, at the cost of granularity.

## Summary

In summary, the word embedding based similarity measure has several uses. First, it can help characterize several aspects of the degree to which predictions match true terms. In the hierarchically constrained version, it can adjust the partial matching values to better represent context. The non-constrained version gives some reflection of the overall closeness of predictions, and helps identify non-sequitur terms or highly related terms from other branches. As shown in the term-based analysis, the word embedding approach can be useful for spotting difficult terms, such as "Young Adult." Comparing the "free" and the "restricted" metrics provides some insight into the role of redundancy by controlling how many predicted terms can be matched to each true term. Redundancy can be broken down further using the hierarchy. Here, too, the cosine similarity measure is a useful tool to identify commonly confused predicted terms. At the level of the paper, the redundancy and cosine similarity analysis can help locate extreme examples of "true" redundancy and prediction redundancy.

Further work is required to develop and assess the uses and interpretation of the word embedding metric. It is not intended as a replacement for hierarchy based partial matching measures. As shown in this chapter, both hierarchy and context are perhaps best used together. By itself, hierarchy measures are inherently limited by the structure of the controlled vocabulary. While this can be an effective strategy for partitioning terms based on semantic type, it is more difficult to compare similarity consistently across the hierarchy. The word embedding similarity measures effectively represent how often terms appear together, but it cannot independently control for the semantic type of the term. Combining the two together allows one to judge that a sibling pair like "Heterocyclic Compounds, 3-Ring" and "Heterocyclic Compounds, 4-Ring" are quite similar, whereas another sibling pair like "Dehydration" and "Death" are not. In terms of specific findings, the partial matching analysis suggests the following:

1. There are more highly similar pairs in predictions than labels, suggesting a degree of redundancy. The distribution of redundancy varies by branch, and is highest in disease and chemical terms.

2. The degree of redundancy influences the overall quality of annotations. This is especially important due to the use of a constant threshold for determining terms. Future work is required to address methods for consolidating potentially redundant terms.

# Chapter 5

# Beyond MEDLINE: A Case Study in Patents

# Overview

This chapter addresses the final research question:

**RQ3: How do MeSH terms in MEDLINE compare to predicted MeSH in USPTO patents?**

MeSH prediction has been widely studied, but almost entirely in the context of MEDLINE. This chapter provides a case study in applying MeSH prediction to patents. First, I examine key differences between patents and scientific papers, particularly as they relate to citations and text in terms of the proposed model. Second, the probabilistic model described in the previous chapter is applied to a sample of approximately 65,000 biomedical patents. The resulting predicted terms are then compared to MeSH terms from a comparison sample of papers drawn from MEDLINE as well as the model predictions for those papers. Finally, I conclude by examining potential applications of MeSH prediction in patents, namely in information retrieval and policy analysis.

## 5.1 Purpose, Styles of Attribution and Language: Patents vs the Scientific Literature

The primary objective of this case study is to apply a probabilistic model that relies on citations and abstract similarity to predict MeSH. As such, it is important to outline what both citations and abstracts *are* in the context of patents versus the scientific literature in order to establish that a model trained in the scientific literature has face validity in the patent domain. I argue that while citations and abstracts function very differently in each domain, they share common features that satisfy assumptions of the model. This viability stems from distinct economic, legal and normative motivations.

At root, both patents and scientific papers are formal documents designed to convey precise information. Both are intended to clearly elucidate an idea, and provide the reader with insight into how the idea was developed and tested. Further, there is significant overlap in both the institutions and people involved: many researchers are also patent holders, and many industry scientists have rich publication histories. However, patents are a distinct "genre" of document from scientific papers in a number of respects.

First, the strategic intent of a patent is to define an invention for the purpose of protecting intellectual property. Economic interests are at the heart of patent writing; they are necessarily both technical and legal documents. As such, the key task of a patent is to clearly delineate an idea and to demonstrate its originality. The strategic disclosure of information is key: enough to establish the claims of the inventor, but not so much as to empower competitors.

This is in distinction to the scientific literature. Claims of novelty are common in science, but secondary to claims of truth. Although scientific papers often discuss matters of considerable financial and social value, they are principally concerned with the communication of knowledge rather than the assertion of property. This is not to imply that the scientific community is entirely free of economic or quasi-economic motivation. Citing behaviors in particular are influenced and sometimes corrupted by social incentives[17].

Second, issues of quality and originality are adjudicated differently between patents and scientific papers. Patents are issued by the state after a complex review process. The issuance of a patent bears a strong legal assumption of validity [14]. Patent disputants engage in legal proceedings, with strongly defined precedents and procedures. Title 35 of the U.S. Patent Code governs strict criteria of patentability. A patent application can be rejected for the following reasons[14]:

1. **35 U.S.C. §101: Subject Matter Eligibility or Utility:** rejecting the claimed invention because it either is directed to ineligible subject matter, such as a law of nature, physical phenomena, or abstract idea, or is not useful.

2. **35 U.S.C. §102: Novelty:** rejecting the claimed subject matter because it is not novel at the time of invention and is described in a printed publication or publicly used or sold in the United States more than one year prior to the filing of the patent application.

3. **35 U.S.C. §103: Non-Obviousness:** rejecting the claimed invention because it is an obvious advance over what was known at the time of invention.

4. **35 U.S.C. §112: Disclosure:** rejecting the claimed invention because the patent fails to adequately describe and enable others to practice the invention or fails to clearly define what is claimed.

5. **35 U.S.C. §121: Restriction requirements:** restricting the patent application to a single invention because the application includes two or more independent and distinct inventions

Peer review, by contrast, is administered by the scientific community: its specific practices and norms vary by field and by journal. There is arguably significant overlap between scientific values, particularly in terms novelty and non-obviousness (35 U.S.C. §102 and 35 U.S.C. §103). Even 35 U.S.C. §121 is reflected in the common emphasis on parsimony and the practice of the "least publishable unit". However, the risk calculus is markedly different in each setting, and has significant implications.

The complex set of factors – strategic, normative, and legal – influence the way in which patent writers and scientists cite and write. The following sections unpack how these factors influence citing behavior, as well as linguistic choices.

The difference of strategic aims is perhaps most clearly reflected in styles of citation. For patent writers, the originality and distinctiveness of the invention are at issue. As such, the purpose of a citation in patents is to position the invention with respect to previous work. The most likely challenge to a patent will come from an argument that the idea was either not new, or that it was so widely known as to be trivial at the time of publication. In other words, citations are materially important in potential patent litigation.

In the scientific realm, citations broadly act to record the intellectual history of a work. Philosophies and citation styles differ between individuals and between fields, but citations are commonly used to properly attribute ideas and to provide readers with further information. Unlike in patents, there is no constraint to protect or legally define an idea. In fact, in distinction to patents, scientists are often judged in terms of various impact factors that reward prolific publishing. As such, scientists are incentivized to self-cite, and to cooperatively cite others.

Further, the language of patent and scientific papers are also different. This is largely due to their formal construction. Patents are arguably both scientific and engineering documents, but they are also legal documents. As such, distinctive and sometimes formulaic language is used in patent text. The definition of the invention is also subject to strategic considerations. The patent holder wants the broadest possible definition possible, but also does not want to over extend themselves in opposition proceedings.

## Differing Goals With Common Mechanisms: Linkages Between Related Work

In sum:

1. Patents focus on defining and protecting intellectual property. Scientific papers aim to communicate new ideas and evidence.

2. Patents use attribution as part of an argument of originality. Scientific papers use attribution to create a record of intellectual labor.

3. Patents use formulaic, legal language narrowly focused on a claim and designed to withstand legal scrutiny. Scientific papers emphasize communication of claims and are comparatively less formal.

The differing styles of attribution and abstract writing have implications for the MeSH prediction strategy. These are mostly positive. While patents may not cite as exhaustively as scientific papers, their writers are strategically motivated to provide key citations to avoid litigation challenges. The narrow focus of patents also means that cited papers are directly related to the invention. Likewise, patent abstracts tend to be narrowly focused on the key claims and invention. However, unusual and formulaic language can be a potential barrier to text similarity methods like BM25 which are designed to emphasize distinctive terms. The technological language of patents (”invention” and ”inventor”) are less common in MEDLINE and may pose problems for text similarity methods. Future work is required to systematically examine how the underlying language differs between scientific papers and patents, and whether this has any significant impact on the absim approach.

# Prediction Methodology

Table 5.1: Top Patent vs Pubmed Frequency Differences (abstract length in characters)

| | **N** | **Mean Cit.** | **Mean Abstract** | **Median Cit.** | **Median Abstract** |
|---|---|---|---|---|---|
| Patent Dataset | 62671 | 47 | 640 | 30 | 568 |
| Pubmed Dataset | 62671 | 38 | 1411 | 33 | 1425 |

This case study applies the models described in the previous chapter, specifically the model using citations, absim, as well as citations of citations and abstract citations. A corpus of patents was selected based on citations to the biomedical literature, and on having citation and abstract characteristics similar to the baseline evaluation dataset. Citations to the literature were retrieved using PATCI, a tool for mapping patent citations to MEDLINE[2]. Specifically, the patents had to have the following characteristics:

1. At least 15 biomedical citations.

2. An abstract of at least 250 words.

This resulted in a set of 67,621 biomedical patents. For comparative purposes, a matched dataset of MEDLINE papers with the same characteristics was sampled. No criteria was placed on the year of either patents or MEDLINE papers. As in the previous chapter, the top ranked 15 predicted terms were taken for each patent. The MEDLINE sample was also processed using the model to compare predicted terms.

## 5.2 Patent MeSH: Human-Oriented and Drug Focused

The following table collects the relative percentage of each MeSH branch, compared against the MEDLINE sample. The dominant category is Organisms, followed by Chemicals and Drugs. As a whole, patents are less diverse than the MEDLINE sample, with no term terms reported in the ”Psychiatry and Psychology”, ”Healthcare”, ”Geographicals” or ”Anthropology, Education, Sociology and Social Phenomena” branches. Neither the MEDLINE sample or the patent sample had an appreciable number of papers with Humanities terms. Notably, Humanities and Geographical terms were predicted in similar proportions (¡.05%) in the MEDLINE prediction set. This is suggestive that the lack of these terms in the patent set is not an artifact of the modeling process.

Information science terms were dramatically more common (.36) in the patent set, and significantly (+.12) more common in the comparison predictions. This is likely due to the high prevalence of the ”Molecular Sequence Data” term, amounting to nearly half of the papers in the patent set. Likewise, ”Amino Acid Sequence” and ”Base Sequence” were also highly prevalent, and both have dual locations in the Phenomena and Process branch as well as the Information Science branch. The Chemicals and Drugs branch is also more prevalent (+.19) in the patent sample than in the MEDLINE set. This branch is only moderately more common in the MEDLINE predictions. Both patents and the Pubmed sample had a relatively large percentage of special terms that do not belong within the regular MeSH hierarchy, primarily the ”Female” and ”Male” term. These terms are denoted with NA in the table below.

Broadly considered, the MEDLINE predictions are close to the true term proportions. Except for the Information Science and the Named Groups branches, all other branches are within ¡= .05 of the true proportions.

Table 5.2: MeSH Branch Proportion of Total Predicted Vocabulary in Patents vs Pubmed Samples

| **MeSH Branch** | **Patent Proportion** | **Pubmed Proportion** | **Difference** |
|---|---|---|---|
| Anatomy | .58 | .50 | +.08 |
| Organisms | .99 | .95 | +.04 |
| Diseases | .35 | .50 | -.15 |
| Chemicals... | .89 | .70 | +.19 |
| Analytical... | .69 | .73 | +.04 |
| Psychiatry... | .02 | .15 | -.13 |
| Phenomena... | .86 | .69 | +.17 |
| Disciplines... | .06 | .10 | -.04 |
| Anthropology... | .00 | .08 | -.08 |
| Technology... | .02 | .03 | -.01 |
| Humanities... | .00 | .01 | -.01 |
| InformationSci... | .50 | .14 | +.36 |
| Named Groups | .26 | .30 | -.04 |
| Healthcare | .01 | .13 | -.12 |
| Geographicals | .00 | .14 | -.14 |

Table 5.3: MeSH Branch Proportion of Total Predicted Vocabulary in Pubmed vs Actual Terms

| **MeSH Branch** | **Predictions Proportion** | **Pubmed Proportion** | **Difference** |
|---|---|---|---|
| Anatomy | .54 | .50 | +.04 |
| Organisms | .99 | .95 | +.04 |
| Diseases | .48 | .50 | -.02 |
| Chemicals... | .74 | .70 | +.04 |
| Analytical... | .68 | .73 | +.05 |
| Psychiatry... | .13 | .15 | -.02 |
| Phenomena... | .72 | .69 | +.03 |
| Disciplines... | .08 | .10 | -.02 |
| Anthropology... | .06 | .08 | -.02 |
| Technology... | .03 | .03 | .00 |
| Humanities... | .01 | .01 | .00 |
| InformationSci... | .26 | .14 | +.12 |
| Named Groups | .42 | .30 | +.12 |
| Healthcare | .09 | .13 | -.04 |
| Geographicals | .09 | .14 | -.05 |

An examination of the most frequent MeSH terms shows broad similarities in the top terms in both patents and Pubmed papers by ranking, but a striking difference in frequency. ”Humans” is by far the most dominant patent MeSH term, with 92% of patents containing the term, compared to 62% in the MEDLINE comparison. Animals is similarly dominant, at 88%. Though it is difficult to assess the significance without further evaluation of the underlying predictions, the dominance of ”Humans” and ”Animals” suggests a stronger applied medical focus in patents versus the broader biomedical literature. This is consistent with 35 U.S.C. §101 prohibitions against patents on physical laws or basic phenomena.

Table 5.4: Top Pubmed and Patent Terms in Sample

| **Rank** | **Patent Term** | **Patent Frequency** | **Pubmed Term** | **Pubmed Frequency** |
|---|---|---|---|---|
| 1 | Humans | 62196 | Humans | 42134 |
| 2 | Animals | 59752 | Animals | 24262 |
| 3 | Female | 36985 | Female | 21878 |
| 4 | Male | 35662 | Male | 21446 |
| 5 | Molecular Sequence Data | 31695 | Adult | 11641 |
| 6 | Mice | 31499 | Middle Aged | 10365 |
| 7 | Amino Acid Sequence | 24184 | Mice | 9190 |
| 8 | Base Sequence | 22926 | Aged | 7301 |
| 9 | Rats | 17104 | Molecular Sequence Data | 5072 |
| 10 | Adults | 14900 | Adolescent | 4721 |

Table 5.5: Top Patent vs Pubmed Frequency Differences

| **Term** | **MeSH Branch** | **Difference** | **Patent Frequency** | **Pubmed Frequency** |
|---|---|---|---|---|
| Animals | Organisms | 35490 | 59752 | 24262 |
| Molecular Sequence Data | InfoSci | 26623 | 31695 | 5072 |
| Mice | Organisms | 22309 | 31499 | 9190 |
| Amino Acid Sequence | Phenomena/InfoSci | 20876 | 24184 | 3308 |
| Humans | Organisms | 20062 | 62196 | 42134 |
| Base Sequence | Phenomena/InfoSci | 19403 | 22926 | 3523 |
| Female | NA | 15107 | 36985 | 21878 |
| Male | NA | 14216 | 35662 | 21446 |
| Rats | Organisms | 12891 | 17104 | 4213 |
| Cell Line | Anatomy | 9493 | 12309 | 2816 |

Table 5.6: Top Chemicals and Drugs Patent Terms

| Term | Patent Frequency | Pubmed Frequency |
|---|---|---|
| DNA | 7701 | 1287 |
| RNA, Messenger | 6702 | 2102 |
| Antibodies, Monoclonal | 5396 | 731 |
| Recombinant Proteins | 5271 | 1131 |
| Peptides | 2962 | 733 |
| Proteins | 2764 | 850 |
| Bacterial Proteins | 2522 | 1716 |
| Antineoplastic Agents | 2038 | 805 |
| Recombinant Fusion Proteins | 1949 | 685 |
| DNA, Viral | 1831 | 583 |

There are several similarly suggestive differences in the frequency of individual terms. Examining the top Chemicals and Drugs terms shows higher frequencies for molecular biology terms: DNA is nearly 6 times as prevalent, and RNA is  3.3 as prevalent. Two terms with wide applications in pharmaceuticals and diagnostics, "Antibodies, Monoclonal" and "Recombinant Proteins" were 7.3x and 4.6x as prevalent in patents. Given the putative connection of patents to pharmaceuticals, the top pharmaceutical terms are provided in table 5.8. "Antineoplastic Agents" is by far the most frequent, and nearly 2.5 more common than in the MEDLINE comparison.

Table 5.7: Top Pharmaceutical Categories in Patents

| Term | Patent Frequency | Pubmed Frequency |
|---|---|---|
| Antineoplastic Agents | 2038 | 805 |
| Enzyme Inhibitors | 1147 | 514 |
| Anti-Bacterial Agents | 1083 | 950 |
| Antiviral Agents | 918 | 286 |
| Oligonucleotide Probes | 585 | 102 |
| Adjuvants, Immunologic | 493 | 91 |
| DNA Probes | 432 | 100 |
| Antioxidants | 418 | 303 |
| Contrast Media | 403 | 171 |
| Protease Inhibitors | 352 | 102 |

Diseases were relatively underrepresented in the patent predictions (-.15) and nearly the same between the MEDLINE predictions and terms (-.02). However, there are some suggestive patterns in the top terms. For example, "Neoplasms" is 1.7x as common in patents. More general conditions like "Obesity" and "Inflammation" were relatively more common in MEDLINE. Further work is required to examine if these reflect systematic differences, or are artifacts of the modeling process.

Table 5.8: Top Disease Subjects in Patents

| **Term** | **Patent Frequency** | **Pubmed Frequency** |
|---|---|---|
| Neoplasms | 1823 | 1058 |
| Breast Neoplasms | 1088 | 1014 |
| Alzheimer Disease | 928 | 348 |
| Disease Models, Animal | 808 | 1770 |
| Neovascularization, Pathologic | 638 | 233 |
| Prostatic Neoplasms | 577 | 445 |
| Coronary Disease | 555 | 166 |
| Obesity | 472 | 709 |
| Postoperative Complications | 469 | 426 |
| Inflammation | 458 | 713 |

The most over-represented category in patents was ”Information Science”. Examining terms only from the Information Science branch (i.e, removing terms with multiple locations like ”Amino Acid Sequence”) confirms that ”Molecular Sequence Data” is the main driver of this difference. The striking frequency of this term may be due to indexing practices. From 1988-2016, this term was automatically applied to any paper containing databank accession numbers, or any papers with molecular sequences in the text. Given that the underlying modeling method is based on the frequency of term assignments, this term may be artificially high.

Table 5.9: Top Information Science Branch Terms

| **Term** | **Patent Frequency** | **Pubmed Frequency** |
|---|---|---|
| Molecular Sequence Data | 31695 | 5072 |
| Image Processing, Computer-Assisted | 943 | 496 |
| Computer Simulation | 634 | 936 |
| Software | 508 | 625 |
| Signal Processing, Computer-Assisted | 358 | 108 |
| Databases, Factual | 173 | 310 |
| User-Computer Interface | 162 | 187 |
| Computers | 162 | 62 |
| Internet | 119 | 407 |
| Computer-Aided Design | 98 | 32 |

# Key Findings

In sum:

1. Patents and scientific papers use citations and abstracts very differently. Patents use citations as part of a carefully considered legal strategy to define the precise claim of their invention. Scientists use citations more variably, but generally as a way to provide credit and to document their intellectual path.

2. Although the motivations behind particular citations and writing choices differ, they both provide rich links to papers with representative MeSH. In the case of patents, the careful selection of citations provides a reasonable expectation that the cited work will have a close relationship to the subject of the patent. Likewise, both scientists and patent writers are motivated to write crisp abstracts that focus on the key contribution or claim they are making.

3. In the examine sample, the total number of citations were roughly equivalent between patents and MEDLINE. Abstracts tended to be much longer in patents, perhaps due to legal writing style.

4. The patent vocabulary reflected significant differences from the MEDLINE vocabulary. Terms related to chemicals and drugs and information science are much more prevalent(+19% and +36%), due to a much larger number of terms related to pharmaceuticals and molecular biology. These categories are also higher in the predicted terms of the MEDLINE sample, but to a lesser extent (.04% and +.12%, respectively).

5. Predicted terms also differ in terms of diseases studied. Disease terms were lower in patents (-.15%) and essentially the same in the predicted termset (-.02%). However, the most prevalent disease subjects were related to neoplasms and Alzheimer disease and appear much more frequently than the MEDLINE comparison (1.72x and 2.66x, respectively)

6. Many of the less applied branches relating to the humanities, geographic locations, and healthcare nearly disappear in the patent set. They remain relatively stable in the predicted terms of the PUBMED set.

7. While the results are suggestive, further work is required to assess their significance. On the whole, the differences described suggest that patents have a more applied, human-centric and pharmaceutical focus than MEDLINE. These findings are consistent with the nature of the patent literature.

8. It is difficult to evaluate the accuracy of individual MeSH assignments in patents. However, policy analysts have proposed summarizing MeSH terms to relatively high levels of the hierarchy to examine broad patterns of investment in biomedicine. [27] The same approach could be employed in patents, making highly granular precision less important.

# Chapter 6

# Discussion

This dissertation centered around four research questions. The following summarizes each research question, as well as the major findings and contributions.

## RQ1: Given that human inter-rater reliability is modest, how should MeSH prediction systems evaluate accuracy?

In the evaluation chapter, an early study on MeSH indexing consistency was updated to the present day. This study found that inter-rater reliability (as measured by Hooper's consistency) has remained relatively modest though stable at 50%. This measure relies on exact matching, which disregards valid, related assignments. In order to develop a more comprehensive picture, several partial matching systems were introduced. The first were graph based measures which give a partial score for matching a term closely in the MeSH hierarchy. Secondly, a partial matching mechanism using distributional semantics was introduced. This measure treats MeSH assignments as an artificial language and trains word embeddings using the Word2Vec algorithm. Term similarity was measured using cosine similarity between term vectors in the word embedding space. This approach provides a direct similarity metric, as well as a way to filter relationships in the hierarchy. Using these two approaches, inter-rater reliability was found to be much more robust. This approach also highlighted the importance of evaluation metrics in MeSH prediction broadly, and the challenge of interpreting model predictions.

## RQ2A: Are abstracts and citations effective features for predicting medical subject headings in MEDLINE?

In the third chapter a probabilistic model was introduced for assigning MeSH terms. This model uses candidate terms drawn from citations and MEDLINE records with similar abstracts. Several variants based on this concept were introduced: a model using only citations, only abstract similarity, both together, and a final model using related records (citations of citations) from both the citation and abstract similarity sets. Each model was tested in three different settings, including a sample with few citations, a sample with short abstracts, and a "normal" sample with typical citations and abstracts. The primary finding is that the citation only model performs relatively well compared to more complex models, but is highly sensitive to citation sparsity. Likewise, the abstract only model performs more modestly but also more robustly, even when abstracts are short. The best performance is achieved by combining citation and abstract similarity features. Including citations of citations does not improve the model significantly in the test datasets.

A partial matching analysis using the findings above reflected that redundancy may be a significant challenge. Here, I found that there are more highly similar term pairs in predictions than in labels, particularly in the disease and drug related branches.

## RQ2B: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

The model evaluation described above indicates that performance improves by combining the abstract and citation features. There are several factors at play. The abstract and citation sets were found to contain unique, relevant terms 88% of the time. The majority of cases where no unique terms were found was due to both sets perfectly covering the target paper. Prior to modeling, including both these sources of information improves the underlying candidate sets. There are at least two mechanisms underlying this complimentarity. The first is temporality. Citations are inherently retrospective, and are limited by the state of the vocabulary at the time they were indexed. Abstract similarity allows the inclusion of papers published after the target, and thus a potentially broader selection of terms. Secondly, citation behavior is always constrained by the awareness and motivation of researchers. Abstract similarity may include papers that were overlooked or otherwise uncited. As described above, the abstract similarity approach is also less sensitive to sparse data. A short abstract has less impact on the overall model performance than few citations – largely because even a short abstract can still yield a significant number of related papers. However, the abstract similarity sets are also ”noisier”, returning larger termsets and more spurious terms. This is due in large part to the inexact nature of text similarity matching. Unlike citations, there is no assurance of a close relationship between the target and the related paper. In summary, abstracts and citations are complimentary both in terms of their contributions to the candidate sets, and in terms of mitigating data sparsity.

## RQ3: How do MeSH terms in MEDLINE compare to predicted MeSH in USPTO patents?

The best model from the earlier chapters was applied to a sample of approximately 65,000 biomedical US patents. The biomedical nature of the patent was established by the patent having citations to MEDLINE. An equivalent sample of MEDLINE papers was constructed, along with the model's predicted terms. Analysis of the resulting vocabulary found significant differences between the distribution of terms. Many of the less applied branches relating to the humanities, geographic locations, and healthcare nearly disappear in the patent set. Terms related to chemicals and drugs and information science are much more prevalent(+19% and +36%), due to a much larger number of terms related to pharmaceuticals and molecular biology. These categories are also higher in the predicted terms of the MEDLINE sample, but to a lesser extent (.04% and +.12%, respectively). Disease terms were lower in patents (-.15%) and essentially the same in the predicted termset (-.02%). However, the most prevalent disease subjects were related to neoplasms and Alzheimer disease and appear much more frequently than the MEDLINE comparison (1.72x and 2.66x, respectively). Further work is required to assess the significance of these differences. These early results are suggestive that patents reflect a more applied, pharmaceutical focus than MEDLINE as a whole.

## Future Directions

In the course of this research, several promising directions for future work became apparent. Namely:

1. Revisiting the patent case study using a comparative framework between patent classification codes (CPC) and MeSH.

2. Comprehensively studying the impact of different text similarity algorithms. For example, using a word2vec or doc2vec approach instead of the BM25 text similarity tool used here.

3. Optimizing the modeling approach through the use of ensemble models and models for determining the term cutoff threshold.

4. Expanding on the MeSH word embedding concept as a basis for postprocessing prediction rankings

5. New scholarly opportunities in economic and policy analysis via MeSH summarization.

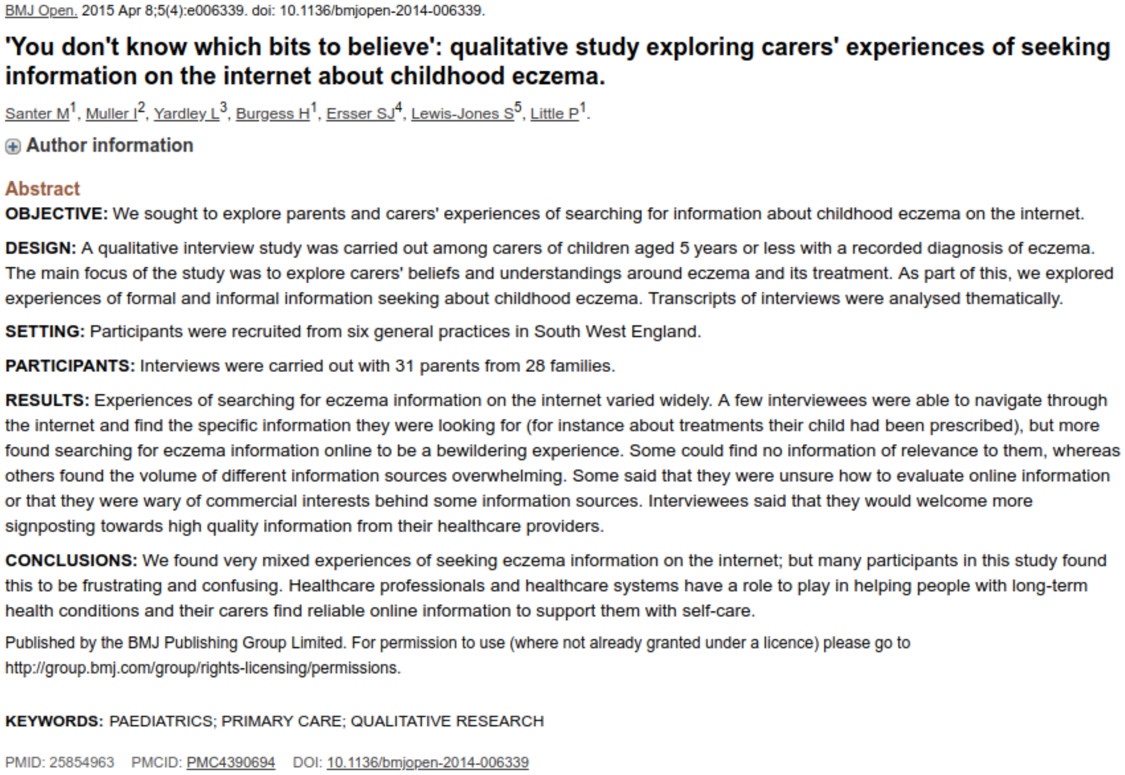
## Comparing Patent Classifications with MeSH

The case study presented in Chapter 5 describes empirical differences between a sample of patents and PUBMED papers. However, this study is only suggestive as to the accuracy of the terms. A gold standard dataset is difficult and expensive to obtain, as patents are not currently indexed using MeSH. A large body of qualified annotators would be necessary to properly index a sample, but also prohibitively difficult to obtain. The entire endeavor of MeSH prediction in different domains is limited by a lack of validation data. To that end, future work might attempt to reconcile aspects of the patent classification system (CPC) and MeSH in order to provide an estimate of accuracy. By linking similar terms between the two systems, it may be possible to more rigorously investigate the underlying model performance without expensive human annotation. However, significant work would need to be undertaken to make this approach possible – especially given the complexity of patent classification codes.

## 6.1 Limitations of Absim: Rare, Distinct and Misleading Terms

One of the early goals of this research was to establish a flexible, modular approach to MeSH prediction. As such, the text similarity algorithm used is intended to be replaceable with other methods. One of the known limitations of the Absim method is that it can be misled by statistically unusual language. For example, regional idiosyncrasies in terms can create an artificially ”rare” term that skews Absim towards inappropriate results. A case of this can be found in the paper ”’You don't know which bits to believe': qualitative study exploring carers' experiences of seeking information on the internet about childhood eczema.” The term ”carer” appears several times throughout the abstract. In the UK, a family member or paid helper is referred to as a ”carer”. In most other countries, this role is referred to either as a caregiver or home health aide. As a result, the top Absim result for this paper is a paper titled ”The role of district nursing: perspectives of cancer patients and their carers before and after hospital discharge.” Though the majority of the first paper is concerned with information seeking behavior and technology, the rare, distinct but marginally relevant term ”carer” dominates the Absim results.

The distinctiveness of ”carer” is largely an artifact of discrepancies in language usage. Geographic differences in language provide particularly stark examples of artificially distinctive terms, but other issues abound – namely the classic problem of word sense disambiguation. These limitations all stem from a reliance on the precise words used in the abstract.



BMJ Open. 2015 Apr 8;5(4):e006339. doi: 10.1136/bmjopen-2014-006339.

**'You don't know which bits to believe': qualitative study exploring carers' experiences of seeking information on the internet about childhood eczema.**

Santer M[1], Muller I[2], Yardley L[3], Burgess H[1], Ersser SJ[4], Lewis-Jones S[5], Little P[1].

⊕ **Author information**

**Abstract**

**OBJECTIVE:** We sought to explore parents and carers' experiences of searching for information about childhood eczema on the internet.

**DESIGN:** A qualitative interview study was carried out among carers of children aged 5 years or less with a recorded diagnosis of eczema. The main focus of the study was to explore carers' beliefs and understandings around eczema and its treatment. As part of this, we explored experiences of formal and informal information seeking about childhood eczema. Transcripts of interviews were analysed thematically.

**SETTING:** Participants were recruited from six general practices in South West England.

**PARTICIPANTS:** Interviews were carried out with 31 parents from 28 families.

**RESULTS:** Experiences of searching for eczema information on the internet varied widely. A few interviewees were able to navigate through the internet and find the specific information they were looking for (for instance about treatments their child had been prescribed), but more found searching for eczema information online to be a bewildering experience. Some could find no information of relevance to them, whereas others found the volume of different information sources overwhelming. Some said that they were unsure how to evaluate online information or that they were wary of commercial interests behind some information sources. Interviewees said that they would welcome more signposting towards high quality information from their healthcare providers.

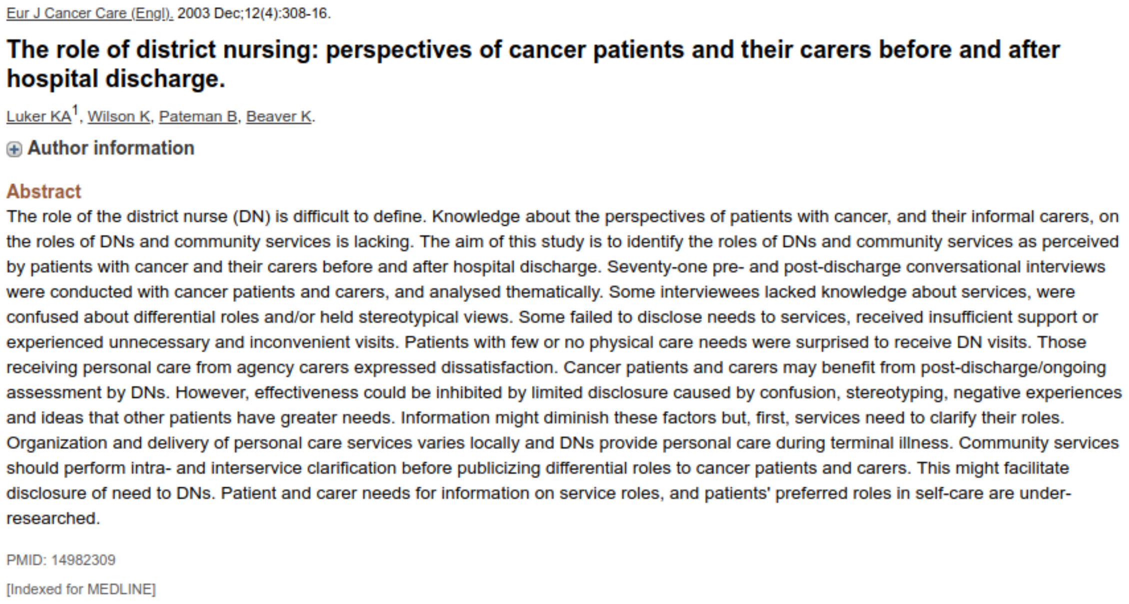
**CONCLUSIONS:** We found very mixed experiences of seeking eczema information on the internet; but many participants in this study found this to be frustrating and confusing. Healthcare professionals and healthcare systems have a role to play in helping people with long-term health conditions and their carers find reliable online information to support them with self-care.

Published by the BMJ Publishing Group Limited. For permission to use (where not already granted under a licence) please go to http://group.bmj.com/group/rights-licensing/permissions.

**KEYWORDS:** PAEDIATRICS; PRIMARY CARE; QUALITATIVE RESEARCH

PMID: 25854963 PMCID: PMC4390694 DOI: 10.1136/bmjopen-2014-006339

Figure 6.1: Reference Paper



Eur J Cancer Care (Engl). 2003 Dec;12(4):308-16.

## The role of district nursing: perspectives of cancer patients and their carers before and after hospital discharge.

Luker KA[1], Wilson K, Pateman B, Beaver K.

⊕ Author information

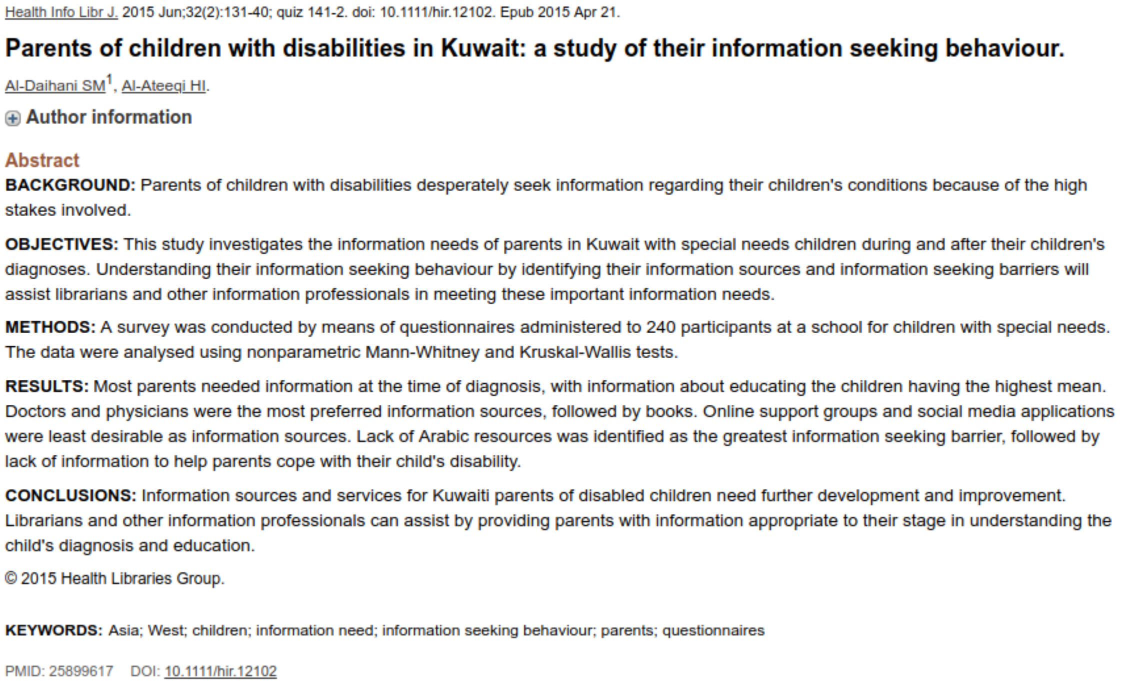
**Abstract**

The role of the district nurse (DN) is difficult to define. Knowledge about the perspectives of patients with cancer, and their informal carers, on the roles of DNs and community services is lacking. The aim of this study is to identify the roles of DNs and community services as perceived by patients with cancer and their carers before and after hospital discharge. Seventy-one pre- and post-discharge conversational interviews were conducted with cancer patients and carers, and analysed thematically. Some interviewees lacked knowledge about services, were confused about differential roles and/or held stereotypical views. Some failed to disclose needs to services, received insufficient support or experienced unnecessary and inconvenient visits. Patients with few or no physical care needs were surprised to receive DN visits. Those receiving personal care from agency carers expressed dissatisfaction. Cancer patients and carers may benefit from post-discharge/ongoing assessment by DNs. However, effectiveness could be inhibited by limited disclosure caused by confusion, stereotyping, negative experiences and ideas that other patients have greater needs. Information might diminish these factors but, first, services need to clarify their roles. Organization and delivery of personal care services varies locally and DNs provide personal care during terminal illness. Community services should perform intra- and interservice clarification before publicizing differential roles to cancer patients and carers. This might facilitate disclosure of need to DNs. Patient and carer needs for information on service roles, and patients' preferred roles in self-care are under-researched.

PMID: 14982309

[Indexed for MEDLINE]

Figure 6.2: Top Absim Result: The Nursing Connection



Health Info Libr J. 2015 Jun;32(2):131-40; quiz 141-2. doi: 10.1111/hir.12102. Epub 2015 Apr 21.

## Parents of children with disabilities in Kuwait: a study of their information seeking behaviour.

Al-Daihani SM[1], Al-Ateeqi HI.

⊕ Author information

**Abstract**

**BACKGROUND:** Parents of children with disabilities desperately seek information regarding their children's conditions because of the high stakes involved.

**OBJECTIVES:** This study investigates the information needs of parents in Kuwait with special needs children during and after their children's diagnoses. Understanding their information seeking behaviour by identifying their information sources and information seeking barriers will assist librarians and other information professionals in meeting these important information needs.

**METHODS:** A survey was conducted by means of questionnaires administered to 240 participants at a school for children with special needs. The data were analysed using nonparametric Mann-Whitney and Kruskal-Wallis tests.

**RESULTS:** Most parents needed information at the time of diagnosis, with information about educating the children having the highest mean. Doctors and physicians were the most preferred information sources, followed by books. Online support groups and social media applications were least desirable as information sources. Lack of Arabic resources was identified as the greatest information seeking barrier, followed by lack of information to help parents cope with their child's disability.

**CONCLUSIONS:** Information sources and services for Kuwaiti parents of disabled children need further development and improvement. Librarians and other information professionals can assist by providing parents with information appropriate to their stage in understanding the child's diagnosis and education.

© 2015 Health Libraries Group.

**KEYWORDS:** Asia; West; children; information need; information seeking behaviour; parents; questionnaires

PMID: 25899617 DOI: 10.1111/hir.12102

Figure 6.3: Top Kamaji Result: The Information Seeking Connection

As described in some length above, recent work in distributional semantics has sought to address these shortcomings by calculating continuous space vector representations of words. This approach has been widely and successfully used in document retrieval and summarization [34]. To address rare but misleading terms encountered in absim, I conducted a small number of experiments using a prototype Word2Vec-based text matching system[41]. This system, called Kamaji, matches arbitrary input text to the average of Word2vec vectors in PUBMED abstracts using a simple approximate nearest neighbors index.[7]

A full comparison of kamaji and absim will need to be undertaken in future work, but the paper referenced above (figure 6.1) offers an instructive example. The top Absim result for this paper is a match for the unusual term ”carer”, but is otherwise unrelated to the health information themes that are central in the target paper. The top Kamaji result matches the information seeking behavior aspect, but is otherwise unrelated to the target. Because Kamaji does not rely on exact word matching, it successfully overcomes the ”carer” pitfall. However, the representation of the input text to a single point in the vector space is highly limited.

Experimentally, Kamaji has a higher degree of complimentarity (unique and relevant terms) and a larger overall vocabulary size than Absim. However, this initial version did not perform as well as a replacement to absim in the overall relevance model described in Chapter 4. The larger vocabulary size suggests that more irrelevant terms are introduced. The averaging scheme used in Kamaji can be revisited and improved using more sophisticated approaches. The use of distributional semantics for MeSH prediction and biomedical text processing is a highly promising direction for future research.

## 6.2 Improving Prediction Results with MeSH Distributional Semantics

As described in Chapter 3 and Chapter 4, the MeSH word embedding approach is a useful evaluation tool. In future work, it may also have a more direct application in a probabilistic model. One approach might be to use the similarity measure to prune highly redundant terms – pairwise terms with very high similarity. Other more sophisticated approaches could also attempt to find non-sequitur terms. For example, terms that are extremely dissimilar from other terms.

There are several promising directions for improving the underlying model. Two clear opportunities are in using ensemble models, as in DeepMeSH. Secondly, adding an additional model to predict the number of MeSH terms could help improve accuracy by tuning the cut-off threshold k.

## 6.3 Beyond Patents: Empowering New Scholarly Opportunities in Health Policy and Information Retrieval

Although patents were the focus of the non-MEDLINE case study, other types of biomedical documents are compatible with the predictive model described in this dissertation. For example, conference proceedings and NIH grant awards both typically contain abstracts and citations. MeSH prediction in these corpora could yield a number of benefits.

The most obvious application would be in practical information retrieval tools. A unified annotation system could help enhance retrieval applications by providing users with a familiar and highly nuanced vocabulary. Applying MeSH across a range of life sciences bibliographic databases could be a first step towards more comprehensive search systems. Such a system would have the advantage of covering not only published scientific research, but also applications (in patents) and prospective work (in NIH grants).

This work also has implications for more theoretical areas of information science. The distributional semantics approach to MeSH could be adapted for a variety of information retrieval tasks. For example, cosine similarity could be used to construct term ”neighborhoods” for query expansion. In such an approach, a user could select a similarity threshold for a query consisting of a set of MeSH. Either automatically or manually, similar terms could be added to their query. The MeSH hierarchy could also be applied to enhance this approach by filtering terms by categories like ”Chemicals and Drugs” or ”Diseases.” Simple similarity searching could also be used in conjunction with the hierarchy to construct a naive concept graph. For instance, a user could formulate a query identifying the most similar pharmaceutical terms to a given disease term. More advanced applications might leverage the hierarchy and contextual similarity together to find additional terms via connected components.

The vector properties of word embeddings have been widely studied in information retrieval, particularly in terms of search by analogy [34]. As described above, vector arithmetic can be used to find analogous concepts, as in the ”King - Man + Woman = Queen” example. The MeSH word embeddings can also support these kinds of queries. A simple prototype system built on this concept is available online at: http://meshexplorer.adamkehoe.com. Using the analogy vector addition framework, one can juxtapose biologically related terms. For example, a user can input a pair of terms related to a molecule and a genetic disease (”alpha 1-Antitrypsin” and ”alpha 1-Antitrypsin Deficiency”). A third target term is introduced (”Phenylketonurias”). The vector offset between the first pair can be applied to the third term to locate a similar concept. In the example query, ”Phenylalanine hydroxylase” is returned.

The term results could be further ranked based on their rarity, their closeness to the target vector, or other properties such as their location in the MeSH hierarchy.

MeSH embeddings could also be used to support topic identification. For example, the MeSH of a target paper could be clustered based on their location in the embedding space. The term nearest to the centroid of each cluster could be used to represent the core areas of a paper. Another prototype system along these lines is available at http://meshexplorer.adamkehoe.com/clustering.

Medical subject heading prediction also has applications beyond information retrieval. Economic and policy analyses of the impact of the National Institute of Health have focused on the impact of single grants, or portfolios of grants related to disease areas. [**?**] Colleagues have suggested that MeSH could be a natural choice to expand and extend such analyses with its ”comprehensive and rigorous classification” [27] The breadth of the MeSH vocabulary is particularly important due to the unexpected way in which many scientific discoveries develop. NIH Associate Director Carrie D. Wolinetz, Ph.D. noted in 2016 that ”The pathways from research to practice to changes in public health are typically non-linear and unpredictable. For a scientific discovery to make that journey may take decades or more and involves a complex ecosystem”[61]. Empirically tracing the flow of federal funding through translational research requires a broad methodology. For example, neurostimulation technologies provides an example of how NIH funding can generate ideas that cross disciplinary boundaries [15, 27]. Beginning in the late 1960s, researchers supported by NIH began experimenting with using electrodes for the purpose of restoring hearing loss which evolved into more advanced cochlear implants by the mid-1990s[15, 27]. Motivated by the initial research auditory rehabilitation, researchers in 1973 began examining the relationship between electrical function and Parkinson’s disease eventually leading to the development of treatments that successfully reduced the intensity of tremors[15, 27]. Recently, this research has served as a foundation for new methods for treating spinal cord injuries and vision loss[15, 27]. Narrower methods based on a single disease would be unable to trace this development. MeSH is better suited to describing the complex evolution of ideas that are commonplace in the life sciences.

Connecting NIH funding with the MeSH taxonomy could create a foundation for a systematic examination of the impact of federal funding on the research and innovation ecosystem in terms of the generation and flow of both scholars and ideas. [27] MeSH could serve both as a means of categorizing research efforts, and as a kind of connective tissue linking together different domains. For example, with both patents and NIH grants annotated, it may be possible to track not only federal investment, but also private research. MeSH annotation could be a powerful first step in opening several new realms of scholarly inquiry in policy and economics.

# Bibliography

[1] The research, condition, and disease categorization process. https://report.nih.gov/rcdc/index.aspx, 2016. Accessed: 2016-01-26.

[2] S. Agarwal, M. Lincoln, H. Cai, and V. I. Torvik. Patci: A probabilistic citation matcher. http://abel.lis.illinois.edu/cgi-bin/patci/search.pl. Accessed: 2016-01-26.

[3] D. Aleksovski, D. Kocev, and S. Dzeroski. Evaluation of distance measures for hierarchical multilabel classification in functional genomics. In *1st Workshop on Learning from Multi-Label Data (MLD) held in conjunction with ECML/PKDD*, pages 5–16, 2009.

[4] R. T. Alves, M. R. Delgado, and A. A. Freitas. Multi-label hierarchical classification of protein functions with artificial immune systems. In *Brazilian Symposium on Bioinformatics*, pages 1–12. Springer, 2008.

[5] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[6] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers. The nlm indexing initiative's medical text indexer. *Medinfo*, 11(Pt 1):268–72, 2004.

[7] E. Bernhardson. Annoy: Approximate nearest neighbors.

[8] W. Bi and J. T. Kwok. Multi-label classification on tree-and dag-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 17–24, 2011.

[9] M. K. Buckland et al. Vocabulary as a central concept in library and information science. In *CoLIS*, 1999.

[10] R. Cerri, R. C. Barros, and A. C. de Carvalho. Hierarchical classification of gene ontology-based protein functions with neural networks. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.

[11] R. Cerri, R. C. Barros, A. C. de Carvalho, and Y. Jin. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC bioinformatics*, 17(1):373, 2016.

[12] I. B. M. Corporation. System and method for annotating patents with mesh data, 2007. US Patent Application: US 20070112833 A1.

[13] E. P. Costa, A. C. Lorena, A. C. P. L. F. Carvalho, A. A. Freitas, and N. Holden. Comparing several approaches for hierarchical classification of proteins with decision trees. In *Proceedings of the 2Nd Brazilian Conference on Advances in Bioinformatics and Computational Biology*, BSB'07, pages 126–137, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] C. A. Cotropia, M. A. Lemley, and B. Sampat. Do applicant patent citations matter? *Research Policy*, 42(4):844–854, 2013.

[15] S. Dodson. Nih case studies of research impact. *Science of Science Policy*, 2016.

[16] D. Eisinger, G. Tsatsaronis, M. Bundschus, U. Wieneke, and M. Schroeder. Automated patent categorization and guided patent search using ipc as inspired by mesh and pubmed. *Journal of Biomedical Semantics*, 4(1):S3, 2013.

[17] I. Fister, I. Fister, and M. Perc. Toward the discovery of citation cartels in citation networks. *Frontiers in Physics*, 4:49, 2016.

[18] M. E. Funk and C. A. Reid. Indexing consistency in medline. *Bulletin of the Medical Library Association*, 71(2):176, 1983.

[19] Y. P. Gerasimenko, D. C. Lu, M. Modaber, S. Zdunowski, P. Gad, D. G. Sayenko, E. Morikawa, P. Haakana, A. R. Ferguson, R. R. Roy, et al. Noninvasive reactivation of motor descending control after paralysis. *Journal of neurotrauma*, 32(24):1968–1980, 2015.

[20] T. D. Griffin, S. K. Boyer, and I. G. Councill. *Annotating Patents with Medline MeSH Codes via Citation Mapping*, pages 737–744. Springer New York, New York, NY, 2010.

[21] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu. Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints. *IEEE transactions on cybernetics*, 43(4):1265–1276, 2013.

[22] U. J. House WF. Long term results of electrode implantation and electronic stimulation of the cochlea in man. *The Annals of otology, rhinology, and laryngology*, 82:504–517, 1973.

[23] M. Huang, A. Névéol, and Z. Lu. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667, 2011.

[24] A. Jimeno Yepes, J. G. Mork, B. Wilkowski, D. Demner Fushman, and A. R. Aronson. Medline mesh indexing: lessons learned from machine learning and future directions. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 737–742. ACM, 2012.

[25] A. J. Jimeno-Yepes, L. Plaza, J. G. Mork, A. R. Aronson, and A. Díaz. Mesh indexing based on automatically generated summaries. *BMC bioinformatics*, 14(1):208, 2013.

[26] A. K. Kehoe and V. I. Torvik. Predicting medical subject headings based on abstract similarity and citations to medline records. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*, pages 167–170. IEEE, 2016.

[27] A. K. Kehoe, V. I. Torvik, N. R. Smalheiser, and M. B. Ross. Predicting mesh beyond medline. In *Workshop on Scholarly Web Mining (SWM), 2017 IEEE/ACM Joint Conference on*. IEEE, 2017.

[28] W. Kim, A. R. Aronson, and W. J. Wilbur. Automatic mesh term assignment and quality assessment. In *Proceedings of the AMIA Symposium*, page 319. American Medical Informatics Association, 2001.

[29] S. Kiritchenko, S. Matwin, and F. Famili. Hierarchical text categorization as a tool of associating genes with gene ontology codes. 2004.

[30] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

[31] J. Lin and W. Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8, 2007.

[32] K. Liu, S. Peng, J. Wu, C. Zhai, H. Mamitsuka, and S. Zhu. Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347, 2015.

[33] Z. Lu, W. Kim, and W. J. Wilbur. Evaluation of query expansion using mesh in pubmed. *Information retrieval*, 12(1):69–80, 2009.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[35] S. Milojević. How are academic age, productivity and collaboration related to citing behavior of researchers? *PloS one*, 7(11):e49176, 2012.

[36] S. P. F. G. H. Moen and T. S. S. Ananiadou. Distributional semantics resources for biomedical text processing.

[37] J. G. Mork, D. Demner-Fushman, S. Schmidt, and A. R. Aronson. Recent enhancements to the nlm medical text indexer. In *Working Notes for CLEF 2014 Conference, Sheffield, UK*, pages 1328–1336, 2014.

[38] J. G. Mork, A. Jimeno-Yepes, and A. R. Aronson. The nlm medical text indexer system for indexing biomedical literature. In *BioASQ@ CLEF*, 2013.

[39] T. N. L. of Medicine. History of mesh. https://www.nlm.nih.gov/mesh/intro_preface.html#pref_rem. Accessed: 2017-04-27.

[40] F. E. Otero, A. A. Freitas, and C. G. Johnson. A hierarchical multi-label classification ant colony algorithm for protein function prediction. *Memetic Computing*, 2(3):165–181, 2010.

[41] I. Pavlopoulos, A. Kosmopoulos, and I. Androutsopoulos. Bioasq releases continuous space word vectors obtained by applying word2vec to pubmed abstracts.

[42] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, and S. Zhu. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics*, 32(12):i70, 2016.

[43] J. Petterson and T. S. Caetano. Reverse multi-label learning. In *Advances in Neural Information Processing Systems*, pages 1912–1920, 2010.

[44] J. Petterson and T. S. Caetano. Submodular multi-label learning. In *Advances in Neural Information Processing Systems*, pages 1512–1520, 2011.

[45] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 995–1000. IEEE, 2008.

[46] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

[47] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7(Jul):1601–1626, 2006.

[48] P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664, 2006.

[49] F. B. Simmons, C. J. Mongeon, W. R. Lewis, and D. A. Huntington. Electrical stimulation of acoustical nerve and inferior colliculus: Results in man. *Archives of OtolaryngologyHead & Neck Surgery*, 79(6):559, 1964.

[50] S. Sohn, W. Kim, D. C. Comeau, and W. J. Wilbur. Optimal training sets for bayesian prediction of mesh® assignment. *Journal of the American Medical Informatics Association*, 15(4):546–553, 2008.

[51] B. B. Sougata Mukherjea. Biopatentminer: An information retrieval system for biomedical patents. *Proceedings of the 30th VLDB Conference*, 2004.

[52] A. Sun, E.-P. Lim, and W.-K. Ng. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology*, 54(11):1014–1028, 2003.

[53] V. I. Torvik. Absim: A tool for calculating bm25 similarity among pairs of abstracts in pubmed. http://abel.lis.illinois.edu/cgi-bin/absim/search.py. Accessed: 2016-01-26.

[54] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[55] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

[56] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pages 406–417. Springer, 2007.

[57] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73(2):185, 2008.

[58] M. Wahle, D. Widdows, J. R. Herskovic, E. V. Bernstam, and T. Cohen. Deterministic binary vectors for efficient automated indexing of medline/pubmed abstracts. In *AMIA annual symposium proceedings*, volume 2012, page 940. American Medical Informatics Association, 2012.

[59] J. Wehrmann, R. Cerri, and R. Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5225–5234, 2018.

[60] W. J. Wilbur and W. Kim. Stochastic gradient descent and the prediction of mesh for pubmed records. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1198. American Medical Informatics Association, 2014.

[61] C. Wolinetz. Capturing impact: A method for measuring progress, 2016.

[62] J. D. Wren, A. Valencia, and J. Kelso. Reviewer-coerced citation: case report, update on journal policy and suggestions for future prevention. 01 2019.

[63] J. H. Zaragoza, L. E. Sucar, E. F. Morales, C. Bielza, and P. Larranaga. Bayesian chain classifiers for multidimensional classification. In *IJCAI*, volume 11, pages 2192–2197, 2011.

[64] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[65] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.