Chapter 1

# Introduction

### Problem Definition

Medical coding is a major facet in maintaining patient records and in obtaining healthcare reimbursements. Medical coding takes the descriptions of diseases, injuries, and healthcare procedures from the health care provider and transforms it into a numeric or alphanumeric code to accurately describe the diagnosis of the procedures performed. Proper medical coding is very important from billing and reimbursement perspective.

In United States the medical codes used, include the International Classification of Diseases and Related Health Problems (ICD) codes. The 2016 edition of the ICD -10 codes is divided into 21 chapters, based about the codes each chapter contains. These codes are often used in not just billing purposes but also for research. In Public Health industry, the use of these standardized codes helps to assimilate healthcare crisis and emergency. The use of such standardized data allows proper and effective sharing of data. However, assigning these codes to clinical documentation is not easy and often requires extensive technical writing and hence involves higher labour costs.

This paper leverages the increasing prominence of electronic health records (EHR), and utilizes natural language processing (NLP) algorithms to support the process of medical coding. This paper aims to provide a solution to the issue, by automatically inferring the respective codes from the data in the EHR.

With the advancement of ICD – 9 to the more complex ICD – 10 codes; a need for effective coding methods becomes especially acute. While, the earlier version of ICD – 9 codes contained only 14,440 codes, the ICD – 10 consists of 68,000 codes with addition of another 368 codes effective 1st October 2011. Hence with increase in codes will lead to unprecedented accumulation of healthcare data and thus, establishing a need for effective technology for ICD coding.

**ICD – 10 Procedure Coding System:** The ICD is a system used by healthcare providers to classify and code all diagnosis, symptoms and procedures recorded in conjugation with hospital care in the United States. It is based on the international classification of diseases by World Health Organization (WHO). The first character of the code is an alpha character excluding “u”, the second to 7th character is alpha numeric. The first three characters categorize the injury and the fourth through sixth characters describe in greater detail the cause, anatomical location and severity of an injury or illness. The seventh character is an extension digit and used to classify an initial, subsequent or sequela (late effect) treatment encounter.

Such complex structure of ICD codes over emphasizes the need for automatic coding to reduce human error. Nevertheless, the attempts to predict the codes as unitary entities are bound to suffer data sparsity problems even with a large training corpus. Due to the complex nature of the code structure, defining approximately 70,000 codes can limit the scope; as to how much can be established by matching the code descriptions and data in the EHR.

### 1.2 Literature Review

Clustering of unstructured text has been a traditional machine learning problem. People from various domains be it healthcare, technology, finance or academic are still trying to solve this puzzle to basically identify patterns in documents to group them and tag them to certain topics based on content of the cluster. Latent Dirichlet Allocation(LDA) and Latent Semantic Analysis(LSA) are two widely used algorithms in this arena. Researchers and academicians have modified these algorithms based on for their need and datasets. One similar approach is being described by Thomas Hofmann in his work where he describes a Probabilistic LSA, which is a statistical technique for analysis of two mode and co-occurrence data. This model has real world implications in information retrieval and filtering, natural language processing, machine learning from text. PLSA is a latent variable model for co-occurrence data which associates unobserved class variable with each observation. The experimental analysis of this approach, performed by the author, focuses on two tasks perplexity minimization and automated indexing of documentation.  The experiments consistently validate the advantages of the PLSA, resulting in performance gains for all the data sets used in the experiment. The experiments demonstrate that the advantages of PLSA are not just limited to perplexity. The PLSA approach is more principled than the standard Latent Semantic Analysis and has a wide range of applications in text learning and information retrieval.

Not only LDA and LSA but many clustering methods are also being used in the industries and by researchers to cluster similar documents. Researchers after performing basic natural language procedures like tokenization, stop word removal, stemming construct document term matrix to have a vector representation of documents. Once achieving the numerical representation of the documents many algorithms like hierarchical clustering, K-means are being used to cluster documents. Lorenzi et.al [LOR] an academician in his works also talks a about a model which uses a surgical complications database to predict post-operative surgical complications. The database consists of CPT codes (Current Procedural Terminology) which groups similar classes of surgeries. In this paper, the team develops a hierarchical clustering algorithm of surgical procedures using the CPT codes, resulting in a binary tree based on the similarity between the data points. The authors of this paper also try to provide a comparison between the method used by them, i.e. Predictive Hierarchical Clustering (PHC) and the traditional Bayesian Hierarchical Clustering. Predictive hierarchical clustering (PHC) is like the traditional agglomerative clustering algorithm with its one-pass bottom-up approach that iteratively merges pairs of clusters. [LOR]. PHC is based on two hypotheses while merging the cluster. First hypothesis states that data of parent cluster is generated with same logistic regression model as of the children whereas second hypothesis states that subtrees of the children are inherited by the parent. The results of this paper are based on computation of two hypotheses which are considered for each potential merge. The implementation of this model is initialized by running regression using the lasso logistic regression using all covariates as main effects with no interactions. At each iteration of the algorithm, the possible models to learn are pushed into separate cores of machine to reduce the turn-around-time. After the hierarchical tree is established, the tree is then cut to retrieve cluster solutions. To validate the algorithm, the authors simulate a data resulting in formation of dendrogram, confirming that this algorithm allows the user to find the underlying structures in the data. Furthermore, the authors successfully test that these new clusters improve the prediction of the logistic regression.

The outcome of this model is a software to assist doctors; which displays the predicted complications and suggest interventions based on these predictions. The resulting clusters have learned coefficients which describes the relationship between the predictors and the outcome. The software provides a personalized result for each patient based on the surgical procedure the patient will undergo using mapping between the clusters and CPT codes. The uniqueness of this PHC that they provide learned clusters of nested subgroups in the data which improves prediction. The paper also talks about the limitations of PHC; which is the inability to work with sparse outcomes.

Another team of researchers lead by David M. Blei et.al [DAV] in their work explains the major drawback of the topic modeling, i.e. scalability, and aims to provide a solution for large scale hierarchical topic modeling(HTC). The paper uses an existing parallel implementation of Latent Dirichlet Allocation (LDA) to provide a scalable mechanism for learning hierarchies from large and complex data sets. The approach used in this paper learns a top – down hierarchy. First, they learn a topic model of k topics using the entire corpus using the LDA. Next, they create k new datasets and allocate each document in the corpus to zero, one, or more of the k datasets using method SPLIT-CORPUS and the θ values learned by the topic model. Finally, they learn k new topic models using the k datasets constructed in the previous step. Each of these topic models corresponds to subtopics of one of the original k topics. This procedure can be applied iteratively, generating further levels of the hierarchy. Since their method doesn’t add dependencies from subtopics to their parents (or siblings), the process is easily parallelizable by launching each learning task independently. The consequence of these decisions allows us to leverage massive datasets while escaping some of the limitations of current hierarchical topic modeling. Existing models attempt to learn parameters jointly across levels of the hierarchy, while our approach learns at a single level, implicitly condition- ing on ancestors. Instead of learning parameters for each path in the hierarchy, we learn parameters for each node, removing dependencies between nodes and allowing learning to take place in parallel. [DAV] The implementation of the algorithm is performed by using the Mr.LDA package, in addition with certain customization to allow easier parallelization of learning and adding support three models. The support three models are used to provide the functionality of SPLIT COURSE – split single, split multiple, and split select.

This method is applied to two large scale document repositories to check the implementation of the model; resulting in rich output. The team’s approach is to allocate documents to subtasks illustrating different branches of the hierarchy. The results of this approach assess the run time and human evaluations of quality on large data sets. The authors’ HTC modeling approach achieves its aim of scalability which is shown by the results of the two large scale document repositories.

**1.2 Methodology**

Chapter 2

# Data

### 2.1 Data

To get a good training data for ICD-10 we crawled the data from two most reliable source on is the Wikipedia directory and second is <https://icd10.com> which is the place where the ICD-10 codes are described. ICD codes are divided into 22 chapter with each chapter have blocks with children codes. For example, Chapter-1 named contained 200 codes marked from A00-B-99. Table 2.1 shows the complete list of chapters and the subsequent code ranges along with their titles.

Table 2‑1 ICD-10 Chapters List

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Blocks** | **Title** |
| I | A00–B99 | Certain infectious and parasitic diseases |
| II | C00–D48 | Neoplasms |
| III | D50–D89 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| IV | E00–E90 | Endocrine, nutritional and metabolic diseases |
| V | F00–F99 | Mental and behavioural disorders |
| VI | G00–G99 | Diseases of the nervous system |
| VII | H00–H59 | Diseases of the eye and adnexa |
| VIII | H60–H95 | Diseases of the ear and mastoid process |
| IX | I00–I99 | Diseases of the circulatory system |
| X | J00–J99 | Diseases of the respiratory system |
| XI | K00–K93 | Diseases of the digestive system |
| XII | L00–L99 | Diseases of the skin and subcutaneous tissue |
| XIII | M00–M99 | Diseases of the musculoskeletal system and connective tissue |
| XIV | N00–N99 | Diseases of the genitourinary system |
| XV | O00–O99 | Pregnancy, childbirth and the puerperium |
| XVI | P00–P96 | Certain conditions originating in the perinatal period |
| XVII | Q00–Q99 | Congenital malformations, deformations and chromosomal abnormalities |
| XVIII | R00–R99 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| XIX | S00–T98 | Injury, poisoning and certain other consequences of external causes |
| XX | V01–Y98 | External causes of morbidity and mortality |
| XXI | Z00–Z99 | Factors influencing health status and contact with health services |
| XXII | U00–U99 | Codes for special purposes |

##### Since many of the ICD codes are very closely related their descriptions were identical for example, P03.2 (Fetus and newborn affected by forceps delivery ) and P03.3 (Fetus and new-born affected by delivery by vacuum extractor) had identical descriptions.

### 2.2 Word Cloud

### 2.3 Data Pre-processing

Still Writing

### 2.3 Feature Engineering

Still Writing

# BIBLIOGRAPHY

, ,

, ,

, ,

, ,

, ,

, ,

, ,

, ,

, ,