Prevalence and Evaluation of Potential Abbreviations in Intensive Care Documentation

A Clinical Language Exploration, Annotation Research: Utilizing Open Source and Commercial Applications

By

David M. Brundage

A Dissertation Submitted
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics School of Health Professions

Rutgers, the State University of New Jersey

November 2019

RUTGERS
School of Health Professions

Final Dissertation Defense Approval Form

Prevalence and Evaluation of Potential Abbreviations in Intensive Care Documentation

BY

David Brundage

Dissertation Committee:

Shankar Srinivasan PhD

Frederick Coffman PhD

Suril Gohel PhD

Approved by the Dissertation Committee:

_____ Date: _____

_____ Date: _____

_____ Date: _____

_____ Date: _____

_____ Date: _____

TABLE OF CONTENTS

ABSTRACT

Introduction: Abbreviations are often used in clinical documentation to reduce time spent documenting in electronic health records and to save space during documentation. Abbreviations represent a specific challenge in healthcare as they can often contain multiple means. This ambiguous use of abbreviations is a patient safety issue for clinicians who do not properly understand the intended use of the abbreviation and presents a health literacy issue to patients as they try and understand what a provider's note says about the care provided. Plenty of research has been done on a clinician's ability to disambiguate abbreviations, but little work has been done to assess how clinicians are using abbreviations or creating tools to assist administrators and clinicians to explore the documentation of their providers.

Methods: A semi-supervised approach was taken to identify potential abbreviations within

the MIMIC-III database. Over 400 million-word tokens were compared to a list approved abbreviation for Beth Israel Deaconess Hospital. The results of this semi-supervised identification were used to analyze the use of abbreviations and prevalence of abbreviations within the dataset.

Results: 463,175,566 raw word tokens were compared to a list of 1,742 approved abbreviations. On average, every document within MIMIC contained almost 14 abbreviation tokens, or roughly 9% of an average note is comprised of potential abbreviations. Some notes contained almost 26% of potential abbreviation tokens. The average count of potential abbreviations for a note created by an RN is 21.87, and the average count of potential abbreviations in a note created by an MD is 11.39. There is a substantial difference in the number of abbreviations used in a note by an RN and MD.

MIMIC note events contain a substantial amount of Using the Medrec2vec word embedding model we extracted the ten most similar terms for each approved abbreviation at Beth-Israel Deaconess (BID) and assessed if vector space contained the semantic meaning for the abbreviated term. Of the 1,743 abbreviations approved by BID the word embedding model was able to accurately extract the semantic relationship for 963 terms.620 abbreviation terms were not able to extract the appropriate semantic term, and 160 terms were not found within the vector space of the model. Our model achieved a precision of .60, a recall of .85, and an F1 of .71, while our model performed decently only using term similarity, it struggled when abbreviations had multiple meanings.

Conclusion: Using the MIMIC data set we have shown that clinical abbreviations and complex clinical jargon make up a specific amount of provider documentation. 8.22% of total words within the MIMIC note events table is a term found within the Beth Israel Deaconess approved abbreviation list. We have also shown that there is the capability to replace abbreviations in medical text to provide additional context to patients.

abbreviations>=5

Acknowledgment

For Sara, who stood by me through everything and constantly reassured me that I was good enough, never once doubted that I would succeed, and who maybe one day will call me Doctor; I love you more than I could ever express and I hope you know that your dedication and unwavering resolve in the face of adversity has made me the proudest man I could ever hope to be. To my loving Mother who cared for me when I was sick and stood up for me when I had no voice, your strength only taught me to never give up when faced with a challenge. To my Sister and Eli, for always being in my corner and proving that family can accomplish anything.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background of Problem

Of the different types of data collected during a patient's care the most common types are structured and unstructured data. Unstructured data makes up a large portion of data collected within the Electronic Medical Record (EMR) and presents a specific challenge for data collection and analysis. Beyond just the EMR, unstructured data is also utilized amongst varying applications and databases. Unstructured data can be entered into external systems through either abstraction and integration directly with the electronic medical record, or through manual entry transcribed from different forms and reports. This unstructured data can present problems within third party databases when it comes to the information stored.

Unstructured data is continuing to grow rapidly within the healthcare field. In order to effectively use this data, we must overcome potential hurdles. Data quality is an important aspect of data governance and can have a profound effect on unstructured data. Unstructured data also represents a challenge in usability. For the unstructured data to be most efficient for analysis, the data must be easily located, extracted, and organized into an easily accessible structure. These challenges can be remedied with a lengthy process utilizing manual review of each unstructured data field and inputting the appropriate meta-data. With large data sets this would require a huge amount of time spent to go back

through years of data that have already been collected.

As advancements in technology have been made, so has our ability to analyze large, complex, and unstructured data. Utilizing advanced machine learning algorithms, it is possible to analyze unstructured data sets without the manual intervention. The subfield of machine learning that utilizes computer language and algorithms to bring meaningful information and knowledge from unstructured data is called Natural Language Processing (NLP). Natural Language Processing focuses on the interactions between human language and computers and sits at the intersection of computer science, artificial intelligence, and computational linguistics.

What can be done with NLP, and how does it bring value to unstructured data? NLP can summarize blocks of text using and extracting the most important and central ideas while ignoring irrelevant information. NLP can be used to automatically generate keyword tags from content to create additional meta-data. Generating keywords and meta-data also allows for technique that discovers topics contained within a body of text. Sentiment analysis can be performed to identify the sentiment of a string of text, from very negative to neutral to very positive. Words can also be reduced to their root, or stem, or you can even break up text into tokens. By using all these different methodologies in tandem, additional context, and semantic layers can be added to unstructured data. Using NLP, we can perform content enrichment on data to increase the usability/searchability of the data.

1.2 Statement of the problem

As providers care for patients, documentation of the care provided is captured within the electronic medical record. As mentioned in chapter 1.1 the two most common ways of capturing data are through structured fields, and unstructured notes. Structured data, is entered and coded by a registered health data professional that can input the data in a way that makes it easy to view, share and access. Structured data is programmed to allow clinicians to easily spot trends in vital health statistics. Examples of structured data includes lab results, category lists, and even simple check boxes. Unstructured data includes all the provider notes, interpretations, narratives, and any other free text field. While structured data provides an easy way to graph and trend results, unstructured data requires an additional layer of analysis for meaningful information.

The general consensus is that unstructured data accounts for 80 percent of the data in business organizations, and the same percentage applies to healthcare organizations a well [1]. It is estimated that over one billion clinical documents are created in the united states every year. It is also estimated that nearly 60% of these documents contain information is clinically valuable for patient care. That means there are over five hundred thousand documents that contain information trapped in an unstructured format waiting for information to be extracted. The expected growth of data is driven by unstructured data. Unstructured data is growing by 42.5% per year, compared to structured data at 22.4%[2].

Utilizing unstructured data in order to facilitate clinical decision making is an important part of the clinical process. To use this information to the fullest extent clinicians must take time to manually review each note in order to assess the information that it provides. This process is lengthy, time consuming, and often providers may not know which note contains valuable information for clinical decision making. However, by utilizing natural language processing, and text mining, it may be possible to make the unstructured data better at providing information. Without working to make unstructured data within the electronic medical record better at providing information the expanse of available data will cause meaningful data to be lost.

Using the data, information, knowledge, wisdom pyramid it is easy to identify a hierarchy of how information can be consumed. The first step to allow the unstructured data to be consumed appropriately for decision making, is to draw out actionable data. Content Enrichment can be applied to the unstructured data to accomplish this feat. Content enrichment is the utilization of modern content processing techniques like machine learning, AI and natural language processing to add structure, context and metadata to content to make it more useful to humans and computers. By mining the data to increase metadata, semantic layers/fingerprints, ontologies, and taxonomies, it may be possible to provide better information and knowledge within the patient's chart. The tools and resources to complete this task are becoming increasingly accessible.

Completing this research will provide a much-needed benefit to the patient's treatment and care. Additionally, the information and data captured from unstructured

provider notes may be used in unique and novel ways. By analyzing the way clinicians document and use abbreviations in the record it may be possible to provide better patient care and increase patient safety. One example of a unique and novel use of natural language processing is adding NLP extracted data and word sense disambiguation to increase health literacy and reduce both patient and clinician frustration in understanding what was documented.

1.3 Research Objectives

Specifically, the objectives were:

- Perform a comprehensive analysis of the MIMIC-III database and analyze the unstructured note data.

- Develop additional semantic layers, taxonomies, and metadata to increase the information derived from the MIMIC-III database using machine learning and natural language processing.

- Use the data collected from the natural language processing and machine learning to increase clinical understanding, and abbreviation awareness.

- Replace clinical jargon with plain language using clinical named entity recognition extraction and replacement utilizing unsupervised machine learning.

1.4 Need and Rationale

Unstructured data is rapidly growing within the electronic health record, and among different business uses within healthcare. As the data continues to grow pertinent information that could be used for clinical decision making will become trapped and lost.

A survey of the literature has shown that natural language processing of unstructured data, coupled with structured data fields, can increase the effectiveness of clinical decision-making tools. Failure to explore factors associated natural language processing and content enrichment, such as, semantic layers, metadata, and new taxonomies or ontologies could result in delay in patient care and decreased outcomes. Utilizing these new data points could play a pivotal role in clinicians understanding and patient safety.

1.5 Research Hypotheses

The aim of this study is to explore the prevalence and use of abbreviations found within clinical unstructured data such as nursing notes, ancillary reports, and discharge notes combined with advanced analytics and machine learning to create an interactive tool to explore unstructured clinical data. This study will be determined by the following research hypotheses.

- **Hypothesis 1**: Documentation found in the MIMIC Note Events uses a significant amount (>5%) of medical abbreviations.
- **Hypothesis 2**: There is a significant difference in the amount of abbreviations used in documentation between Registered Nurses and Physicians.
- **Hypothesis 3**: Abbreviations and potential synonyms can be replaced with more descriptive terms using unsupervised machine learning applications

CHAPTER II

LITERATURE REVIEW

2.1 Literature Search and Search Strings

The literature search consisted of a review of various articles published on the topic of Natural Language Processing of biomedical data, as well as, articles specific to analyzing the MIMIC database. Articles were searched in the PubMed and Ovid database which includes Medline. Google searches were also utilized to identify potential search strings. The total number of articles and/or abstracts reviewed were *387*. Of the *387*, approximately *109* were reviewed in detail. The following are several of the search strings used to locate articles from the databases: Search terms used were:

Relating to MIMC and Natural Language Processing

- "MIMIC Database"
- "MIMIC Database Natural Language Processing"
- "Abbreviations in Healthcare"
- "MIMIC Text Analysis"
- "Python Natural Language Processing"
- "NLP and Content Enrichment"
- "Natural Language Processing Health Data"
- "Word Sense Disambiguation MIMIC"
- "Word Embeddings MIMIC – III"

2.2 Current State of Research

MIMIC has been used in multidisciplinary research since the early 2000's. Research has been performed using the data provided in MIMIC to answer questions throughout medicine. MIMIC has been used as a reference source, primary data source, and to compare the state of clinical charting. MIMIC has been presented in conferences, international journals, and throughout academia. One of the benefits to MIMICs current state is that a large breath of research and data are available to support new and novel research topics.

MIMIC-III is an extension of MIMIC-II: it incorporates the data contained in MIMIC-II (collected between 2001 - 2008) and updates it with newly collected data between 2008 - 2012. Many of the data elements in MIMIC-III have been regenerated from the raw data in a way that improves the quality of the underlying data without sacrificing the original structure[3].

One of the biggest challenges of adding new data to MIMIC-III was due to a change in a data management platform at Beth Israel Deaconess Medical Center. The hospital replaced their original data collection platform, Philips CareVue system which provided data from 2001 to 2008. The new system implemented in 2008 Metavision data management system is currently used to date[3].

2.3 MIMIC Database

MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are III) is a large, freely-accessible database that consists of deidentified health-related data pertaining to over forty thousand patients who stayed in the intensive care units of the Beth Israel Deaconess Medical Center, for 11 years, between 2001 and 2012.[3]

The data stored within the database includes demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). MIMIC has been used in a diverse range of analytic studies within multiple fields of research including epidemiology, clinical decision-rule improvement, and electronic tool development. MIMIC has three unique features that make it specifically useful for research:

- MIMIC is a freely available dataset for research worldwide.
- MIMIC is comprised of a very large and diverse patient population with the critical care departments.
- MIMIC contains precise, time sensitive data including lab results, electronic documentation, and bedside monitoring trends and waveforms.

Table 1

Critical care unit	CCU	CSRU	MICU	SICU	TSICU	Total
Distinct patients, no. (% of total admissions)	5,674 (14.7%)	8,091 (20.9%)	13,649 (35.4%)	6,372 (16.5%)	4,811 (12.5%)	38,597 (100%)
Hospital admissions, no. (% of total admissions)	7,258 (14.6%)	9,156 (18.4%)	19,770 (39.7%)	8,110 (16.3%)	5,491 (11.0%)	49,785 (100%)
Distinct ICU stays, no. (% of total admissions)	7,726 (14.5%)	9,854 (18.4%)	21,087 (39.5%)	8,891 (16.6%)	5,865 (11.0%)	53,423 (100%)
Age, years, median (Q1-Q3)	70.1 (58.4–80.5)	67.6 (57.6–76.7)	64.9 (51.7–78.2)	63.6 (51.4–76.5)	59.9 (42.9–75.7)	65.8 (52.8–77.8)
Gender, male, % of unit stays	4,203 (57.9%)	6,000 (65.5%)	10,193 (51.6%)	4,251 (52.4%)	3,336 (60.7%)	27,983 (55.9%)
ICU length of stay, median days (Q1-Q3)	2.2 (1.2–4.1)	2.2 (1.2–4.0)	2.1 (1.2–4.1)	2.3 (1.3–4.9)	2.1 (1.2–4.6)	2.1 (1.2–4.6)
Hospital length of stay, median days (Q1-Q3)	5.8 (3.1–10.0)	7.4 (5.2–11.4)	6.4 (3.7–11.7)	7.9 (4.4–14.2)	7.4 (4.1–13.6)	6.9 (4.1–11.9)
ICU mortality, percent of unit stays	685 (8.9%)	353 (3.6%)	2,222 (10.5%)	813 (9.1%)	492 (8.4%)	4,565 (8.5%)
Hospital mortality, percent of unit stays	817 (11.3%)	424 (4.6%)	2,859 (14.5%)	1,020 (12.6%)	628 (11.4%)	5,748 (11.5%)

CCU is Coronary Care Unit; CSRU is Cardiac Surgery Recovery Unit; MICU is Medical Intensive Care Unit; SICU is Surgical Intensive Care Unit; TSICU is Trauma Surgical Intensive Care Unit

MIMIC-III contains data associated with over 53,000 unique hospital admissions for adult patients (aged 16 years or above) admitted to the critical care units between 2001 and 2012. The MIMIC database also contains data for 7870 neonatal admissions between 2001 and 2008. This data covers 38,597 distinct adult patients and 49,785 hospital admissions. The median age of adult patients is 65.8 years (Q1–Q3: 52.8–77.8), 55.9% patients are male, and in-hospital mortality is 11.5%. The median length of an ICU stay is 2.1 days (Q1–Q3: 1.2–4.6) and the median length of a hospital stay is 6.9 days (Q1-Q3: 4.1–11.9). A mean of 4,579 charted observations and 380 laboratory measurements are available for each hospital admission. Table 1 provides a breakdown of the adult population by care unit[4]

2.3.1 Current Applications of the MIMIC Data Set

One identified barrier into the generation of high quality, and robust clinical data is the lack of reproducibility in the study. The MIMIC database benefits from having a centralized code repository. This centralized code repository creates a code base that allows for researchers to create reproducible studies on the critical care dataset. The repository provides code that assists clinicians to load the data into a relational database schema, allows for the creation of data extracts, and even provides the capability to reproduce entire research studies and analysis plans. Utilizing the code repository and scripts researchers can extract comorbidity statuses, severity of illness scores,

administrative definitions, medication and treatment administration and more[3].

The code repository benefits from executable documents that provide user tutorials and templates that allow researchers to replicate procedures and processes. Additional benefits of having a centralized code repository is the ability for the community to discuss the data and concepts and collaborate to improve the tool. Issue tracking is built into the repository and provides the community the ability to track and maintain known issues and resolutions. By providing open source code alongside the freely accessible MIMIC-III database, researchers have enabled end-to-end reproducibility of electronic health record analysis[5].

Code within the repository is available as standardized scripts in multiple programming languages including R, Python, and Structured Query Language (SQL). The scripts are modified to allow an individual who has been granted access to the MIMIC-III database to generate several different "views" of the data, with each view being an extraction from the raw data. The repository has each script associated with an automatically generated unique commit hash that acts as an identifier for the code. The commit hash has the benefits of allowing publications that use the repository to cite the commit hash, allowing other researchers to download a copy of the code used regardless of any modifications since. This provides an additional layer of reproducibility to the MIMIC data set and helps to relieve one of the identified barriers in clinical research, a lack of reproduction in research

MIMIC has been used in multiple applications and research studies. While MIMIC has shown that it has a variety of uses, it is not without its own hinderances. MIMIC is a public

dataset, but there has been an identified barrier in access, the technological requirements to allow medical researchers to become proficient in MIMIC research. Currently, MIMIC requires in depth knowledge of the SQL programming language and an understanding of the database structure and schema of MIMIC. These are challenging requirements especially for health researchers and clinicians who may have limited computer proficiency.

In order to overcome this challenge, interactive, web-based visualization platforms have been developed that allow, for the first-time, MIMIC users to easily explore and navigate the database. The interactive tool offers two features an Explore feature, and a Compare feature. Explore allows a user to select a specific patient cohort from MIMIC and visualize the relationships, and distributions of data among clinical, and administrative variables.

The Compare feature enables users to select two patient cohorts and visually compare them with respect to a variety of variables. The tool is also helpful to experienced MIMIC researchers who can use this tool to increase the speed at which they develop their SQL queries to manually extract and visualize the data. This tool has provided a benefit to MIMIC research by allowing a quick and convenient way for researchers to perform an initial analysis. This research also provides a new way for MIMIC researchers to learn the characteristics of the MIMIC data and the relationships within the dataset. This tool will hopefully create a more informed MIMC research base. The MIMIC visualization project does not include any analytics tools or capabilities and leaves this as a future endeavor[6].

Clinicians in intensive care units are required to make rapid decisions based on

physiological observations. These observations are used to assess the clinical deterioration of patients and are a major interest to researchers in the field of biomedical engineering and informatics. By investigating these biological parameters researchers have been able to assess the parameters for use in risk assessment models.

In a study of 127 adult ICU patients selected from the MIMIC II (predecessor to MIMIC III) database researchers used continuous temporal monitoring of physiological data points such as, heart rate, blood pressure, and oxygen saturation. The number of random variables under consideration by the model were reduced and feature selection and feature extraction were performed. This dimensionality reduction utilized a deep learning autoencoder and were used to train a support vector machine model. Utilizing multiple statistical methods such as random forest, and fuzzy c-means clustering (FCM) the researchers were able to determine patient risk stratification.

Researchers were able to stratify patients in groups of stable or deteriorating patients. Performance assessment of these groups was done using the receiver operating characteristic (ROC). The area under the ROC (AUROC) was 93.2 (95% CI (92.9–93.4)) with sensitivity and specificity values of 0.80 and 0.89, respectively. The suggested fuzzy risk levels using the combined method of the FCM clustering and RF achieved an accuracy of 1 (0.9999, 1), with both sensitivity and specificity values.

The research performed by Dervishi has shown that the MIMIC-III database can be used to infer clinically relevant information. The risk assessment models have been shown to be effective in the estimation of the patient's stability. One constraint of this research application is that it was a retrospective analysis and further studies will be needed to assess

the clinical impact of the model[7].

2.4 Predicting Clinical Outcomes with MIMIC

One of the most beneficial aspects of the MIMIC-III database is the ability to test clinically relevant predictive algorithms. In subsection 2.3.1 we reviewed some of the current applications and analysis of the MIMIC-III database. In this section we will review some of the predictive models and research that have been employed using MIMIC. These research articles will provide a firm background on applying advanced analytics and machine learning using MIMIC data.

In reviewing the research MIMIC has been shown to be used to assist in the prediction of critically ill patients. Using a prognostic model, researchers have been able to predict the 60-day case fatality rate in patients requiring renal replacement therapy. The study was validated through an independent cohort due to the lack of prognostic models in clinical practice. The study followed 1,053 critically ill patients requiring RRT from the MIMIC-III database for analysis. The models' discrimination was evaluated using c-statistics. Calibration was evaluated by Hosmer-Lemeshow (H-L) test and GiViTi calibration belt.

The results from the study show that in a case-mix population, including patients with normal or altered serum creatinine (sCr) at ICU admission, discrimination was moderate, with a c-statistic of 0.71 in the non-integerized risk model. In patients with altered baseline sCr, better discrimination was achieved with the integer risk model (0.7695%CI 0.71–0.81). As for the calibration, although the H-L test was good only in patients with normal/slightly altered sCr at admission, the calibration belt disclosed no significant

deviations from the bisector line for any of the models in patients, regardless of admission sCr.

This study showed that the prognostic model can be useful in a larger group of critically ill patients and could provide benefit to patients beyond the cohort. One issue with the model is that it had some slight discrimination capacity for patients that presented with an elevated sCr at admission. With the addition of a refitted model the results did show improvement, and highlighted the need for external validation and continual reiteration of the prognostics model over time before being implement in clinical practice[8].

In reviewing this piece of literature, we can understand the value that the MIMIC dataset provided. This study use of the MIMIC data shows that clinically relevant data for decision making can be extracted from the data set. Coupling the data from MIMIC with known clinical indicators can increase the value derived from the database.

As electronic health records continue to amass large amounts of health-related data, the use of predictive analytics could help transform medicine with Predictive, Preventive, and Personalist (PPPM) medicine. The use of predictive analytics can benefit both quality and the costs associate with healthcare. Due to the complexity of the data involved, data driven decision making methods are not easily translated into clinical care models. As we have discussed, applying cutting edge predictive methods, and the process required to extract, transform, and load (ETL) the data requires in-depth programming skills and limits its ability to be easily accessible to clinicians. This leaves a disparity between the potential of the data and how the data is used.

By focusing on an open framework utilizing visual environments these issues are easily

accessible by the medical community. Research completed by Pouke Et Al showed how a such a framework could be developed. By integrating data from critical care patients from the MIMIC-II database into a visual data mining environment (RapidMiner) a framework was created to support scalable predictive analytics. The ETL process, as recommended by the Cross-Industry Standard Process for Data Mining (CRISP-DM) began by retrieving data from the MIMIC-II tables. Using visual tools for ETL on Hadoop and predictive modeling in RapidMiner, a robust process for automatic building, parameter optimization and evaluation of various predictive models, under different feature selection schemes. Because these processes can be easily adopted in other projects, this environment is attractive for scalable predictive analytics in health research[9].

While the research presented the use of MIMIC with RapidMiner in order to create a scalable predictive analysis platform, where our methodology will focus on the use of the Python programming language and framework. The use of RapidMiner has the benefit of providing a system that has the capability to manipulate, extract, process, and analyze large complex data sets without using any coding. While this platform provides a benefit for clinicians who may have little experience with coding it lacks in the level of configurability and control that comes from an object-oriented programming language.

2.5 MIMIC and Natural Language Processing

MIMIC has been used in multiple research projects in coordination with natural language processing. Research has been performed to classify illnesses in chronically ill patient to assist in decision support for clinicians. This was achieved by multi-label

classification of multivariate time series from the medical records of chronically ill patients found in MIMIC, using methods such as bag of words, and other classification algorithms. Additionally, Zuffrey et al, compared supervised dimensionality reduction techniques to multi-label classification algorithms[10].

The results from the study showed that a non-linear dimensionality reduction approach is applicable to clinical time series data using a bag of words algorithm. The bag of words algorithm is comparable to other multi-label classification algorithms. By chaining the projected features, the performance of the algorithm could be increased for a binary approach. The evaluation shows the feasibility of representing medical health records using the bag of words algorithm for multi-label classification tasks. The research completed by Zuffrey et al shows that MIMIC is a suitable candidate from natural language processing.

Natural language processing has been used to mine the MIMIC database to beyond just novel methods. By combining learned structure of clinical concepts derived from the unstructured free text within nursing notes, along with physiological data, ICU patients could be stratified by risk. This risk stratification can assist in the prediction of patient mortality. By using Hierarchical Dirichlet Processes (HDP), a non-parametric topic modeling technique, researchers were able to automatically discover groups of co-occurring clinical concepts. The success and utility of the topic structure for predicting mortality was evaluated against 14,739 unstructured nursing notes from the MIMIC-II database. The results showed the by using the learned topic structure from nursing notes acquired in the first 24 hours of an ICU admission, a clinical scoring system, SAPS-I,

could be improved. The combination of physiologic data from the first 24 hours, coupled with nursing note text, was able to increase the area under the curve (AUC) for predicting mortality was 0.82. In comparison, the AUC for mortality prediction using physiologic data alone in the SAPS-I algorithm only had an AUC of 0.72. This shows that the extracted clinical topics used to modify the SAPS-I algorithm can greatly improve the baseline impact[8].

The research completed by Lehman et al has shown that there is value in the combination of unstructured nursing data with clinical scoring systems. [11]

MIMIC has also been used to assist in treatment, and prevention of severe sepsis and septic shock. Sepsis and septic shock affect millions of patients and has a mortality rate of nearly 50%. Due to the sever nature of this disease the Center for Medicare and Medicaid Services has laid a series of guidelines that clinicians should follow to improve clinical quality outcomes. One way to improve these outcomes is through the early identification of at-risk patients. With the advent of Electronic Health Records surveillance tools have been developed that can assist in automatically recognize early sepsis symptoms. One of the largest constraints to finding accurate, and timely data for sepsis identification is that this data is mainly captures in unstructured clinical notes. Research has been developed to assist in the automatic monitoring of nursing notes for clinical indicators of sepsis. This method created an annotated dataset through text analysis and could then be combined with a machine learning model to achieve a predictive value [12].

A critical task in analysis of electronic health records consists of correctly identifying the concepts and diagnosis within the record. In many cases, the most valuable and relevant information for an accurate classification of medical conditions exist only in the unstructured clinical narratives. The most commonly used approach to this problem relies on extracting multiple clinician-defined medical concepts from text and using machine learning techniques to identify whether a patient has a certain condition. However, recent advances in deep learning and NLP enable models to learn a rich representation of (medical) language.

A study by Gehrmann et al used Convolutional neural networks (CNN) for text classification. This approach allows for the augmentation of existing techniques and leverages the representation of language to learn which phrases in a text are relevant. In this work, Gehrmann et al compare concept extraction-based methods with convolutional neural networks and other commonly used models in NLP in ten phenotyping tasks using 1,610 discharge summaries from the MIMIC-III database. Their study showed that the convolutional neural network routinely outperformed concept extraction methods in many tasks. The convolutional neural network has an improvement by 26 for the F1-score and a 7 point increased in the ROCAUC [13]. Gehrmann et al's research shows that a deep learning approach and model can be built from the MIMIC dataset.

Other deep learning applications using the MIMIC dataset have also been completed. Jauregi et al developed a recurrent neural network with specialized word embeddings for health-domain named-entity recognition. Previous state-of-the-art systems on Drug Name Recognition (DNR) and Clinical Concept Extraction (CCE) have focused on a

combination of text “feature engineering” and conventional machine learning algorithms. Recurrent neural networks (RNNs) have proved capable of automatically learning effective features from either random assignments or automated word “embeddings”. Jauregi et al created a domain specific word embedding by using health domain datasets such as MIMIC-III. Two deep learning methods, namely the Bidirectional LSTM and the Bidirectional LSTM-CRF, are evaluated. The domain specific embeddings helped to cover unusual words in the data. Domain specific word embeddings has allowed Jauregi to avoid costly feature engineering and achieve higher accuracy [14]. Our approach utilizes a domain specific word embedding model also trained on MIMIC to be used downstream in health literacy applications.

Research has been completed on replacement of text in medical records through machine learning applications. Our research has a focus on extract texting and replacing ambiguous terms to increase health literacy. Medical researchers are legally required to protect patients' privacy by removing personally identifiable information from medical records before sharing the data with other researchers. Douglass et al propose a method for computer-assisted removal and replacement of protected health information (PHI) from free-text nursing notes collected in the intensive care unit as part of the MIMIC II project. The sensitivity of human experts working alone to perform PHI deidentification ranged from 0.63 to 0.93, with an average of 0.81. An algorithm generated few false negatives but many false positives. Its sensitivity was 0.85, but its positive predictive value was only 0.37 [15].

2.6 Abbreviations in Healthcare

The use of abbreviations in healthcare are routinely used to save space and time. Clinicians work in a high paced environment and are required to document how care was delivered. The combination of this stressful environment and documentation requirement has led to the creation and use of multiple abbreviations. Research has been completed on the use of abbreviations by healthcare providers. In Sinha et al's "Use of abbreviations by healthcare professionals: what is the way forward?" the authors compiled a list of abbreviations from clinical notes and presented a questionnaire to healthcare professionals and asked to evaluate the abbreviation. An abbreviation was defined as a shortened word or phrase, an acronym, contracture, or an initialism. A curated list of 30 abbreviations were selected from a total of 100 extracted abbreviations from 50 clinical notes. The questionnaires were distributed to 225 participants and 216 were completed with a correct response of only 43% [16]. This research shows that abbreviations within multiple specialties can be confusing outside of the original authors intended audience. The research also demonstrates that healthcare professionals may have poor knowledge of common abbreviations. The definition of a healthcare abbreviation will be used for our research.

An additional cross-sectional study was reviewed as part of this literature. Tsina et al's "Use of Abbreviations and Acronyms among Healthcare Workers in a Resource Limited Setting" presented in the Journal of Healthcare Communications further illustrates the disparity between clinicians in their ability to identify abbreviations. Tsina et al collected 1,693 abbreviations from 57 inpatient charts and presented a self-administered survey to randomly selected clinicians. In the study, healthcare workers could only correctly identify 73% of the abbreviations used. This research also demonstrated that specific healthcare

provider types demonstrated significantly different results in identifying the appropriate abbreviations. Allied Health workers were shown to score the lowest on the questionnaire with physicians and nurses performing similarly. Tsina et al's research also showed that the abbreviations had alternative meanings. Participants reported that 58.1% of the selected abbreviations had an alternate meaning[17]. This research shows that clinicians from different healthcare settings are familiar with different abbreviations. This furthers our hypothesis that clinicians from different paths use abbreviations differently and that there is a difference in the amount of abbreviations each clinician type uses when documenting patient care.

What we have been able to identify from the literature review on abbreviations in healthcare is that there have been robust studies on the clinician's ability to identify abbreviations. What has been missing from the research has been HOW clinicians are using abbreviations. We need to understand who the most prominent users of abbreviations are in medical documentation to determine if documentation standards need to be revised, or if there is area for improvement in the electronic health records.

2.7 Health Literacy

Health.gov defines Health Literacy as "…The degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions". Only 12 percent of adults have Proficient health literacy, according to the National Assessment of Adult Literacy. In other words, nearly

nine out of ten adults may lack the skills needed to manage their health and prevent disease. Low literacy has been linked to poor health outcomes such as higher rates of hospitalization and less frequent use of preventive services [18].

Health literacy is dependent on individual and systemic factors such as: level of communication skills, the patient and provider culture, demands of healthcare and public health systems (resourcing, financial, technological) and demands of the situation all applied to the context of the individual. Health literacy directly affects people's ability to navigate healthcare systems, share health history, engage or participate in self-care, and understand how to manage their health choices.

As technology has continued to expand in the healthcare sector, computer applications have become involved in our care delivery. Computer-based health literacy interventions for older adults have been developed in previous studies. From September 2007 to June 2009 Xie et al conducted a study on a total of 218 adults between the ages of 60–89. The four week-long curricula covered two National Institutes of Health (NIH) websites: NIHSeniorHealth.gov and MedlinePlus.gov. Computer and Web knowledge significantly improved from pre- to post-intervention ($p<.01$ in both cases). Most participants found both sites easy to use and were able to find needed information on both. Most participants (78%) reported that what they learned had affected their participation in their own health care. Participants had positive feedback on the intervention [19]. The Xie et al study shows that the findings support the effectiveness and popularity of computer-based interventions to health illiteracy.

2.8 Abbreviation Disambiguation

Abbreviations are commonplace in medical documentation. Many abbreviations can even have multiple meanings or different abbreviations can mean the same thing. In computational linguistics, word-sense disambiguation (WSD) is a problem that attempts to identify the context in which a word is used in a sentence. Multiple computational linguistic use cases, such as improving relevance of search engines, cataphora/anaphora resolution, coherence, inference rely on word sense disambiguation as a part of their solution. Abbreviation Disambiguation is a subset of word-sense disambiguation.

Wu et al examined the use of neural word embeddings for clinical abbreviation disambiguation. Three different methods for deriving word embeddings from a large unlabeled clinical corpus: one existing method called Surrounding based embedding feature (SBE), and two newly developed methods: Left-Right surrounding based embedding feature (LR_SBE) and MAX surrounding based embedding feature (MAX_SBE) were trained on MIMIC-II[20].

Biomedical abbreviations and acronyms are widely used in biomedical literature, thus making biomedical literature adequate training sources. Since many of them represent important content in biomedical literature, information retrieval and extraction benefits from identifying the meanings of those terms. Yu et al, presents a semi-supervised method that applies MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. After training the machine learning model Yu et al predicted the full forms of abbreviations in full-text journal articles by

applying supervised machine-learning algorithms in a semi-supervised fashion. This study reported up to 92% prediction precision and up to 91% coverage [21]. Both studies by Yu et al and Wu et al utilize supervised machine learning to make predictions and decisions. Our approach is to utilize unsupervised learning to assist in abbreviation disambiguation and allow patients to make their own determinations of what abbreviations should be disambiguated.

2.9 Text Analysis in Python

Python is a free, open-source, cross-platform programming environment. While Python excels in data science, and statistical applications it has the added benefit of shallow learning curve for novice programmers. Python's large community and libraries allows for users to quickly pickup on syntax and easily be able to find documentation. While this climb to becoming an expert in Python may be challenging the Python community, and open-source nature, means that there is a vast amount of knowledge and access to tools.

One of the keys to Python capabilities has been its densely populated collection of extension software libraries, known in Python terminology as packages, supplied and maintained by Python's extensive user community. Each package extends the functionality of the base Python language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package.

The best-known package repository for analytical purposes is, Anaconda, currently Anaconda has over 1,500 packages that are published, and which have gone through an

extensive screening for procedural conformity and cross-platform compatibility before being accepted by the archive. Python thus features a wide range of inter-compatible packages, maintained and continuously updated by scholars, practitioners, and projects such as Spyder and Jupyter. Furthermore, these packages may be installed easily and safely from within the Python environment using a single command. Python thus provides a solid bridge for developers and users of new analysis tools to meet, making it a very suitable programming environment for scientific collaboration. The tools available in Python for carrying out text analysis allow for cutting-edge text analysis using relatively few commands. The powerful and complex packages are ideal to our project and will be easily applied to the MIMIC data set.

Text analysis has become particularly well established in Python. The number of packages dedicated to text processing and text analysis have become increasingly more available. These text analysis packages focus on techniques that range from low-level string operations all the way to advanced text modeling such as Latent Dirichlet Allocation models. There has also been an increased effort among Python developers to coordinate and create additional complex packages to solve text analysis and natural language processing problems. One of the major benefits of text analysis in Python is the ability to quickly switch between packages or combine them. Natural Language Processing and Machine Learning applications have plenty of packages available including Scikit-Learn, NLTK, SpaCy, and Gensim.

Chapter III

Methodology

3.1 Introduction

This research is comprised of three major components:

1. Creation of a text extraction, interpretation, annotation, pipeline.

2. Abbreviation disambiguation through Natural Language Processing and Word Embeddings

3. Analysis of the unstructured text to evaluate the prevalence and use of abbreviations in clinical text.

Our first phase of the research was the development of a text extraction, interpretation, and annotation pipeline to standardize the clinical language found within the note events. Using both commercial, and open source applications, we developed a novel approach to extract textual references to valuable medical information such as medical condition, treatment, tests and test results, medication (including dosage, frequency, method of administration) and standardize the language back a codified ontology SNOMED.

For the identification of potential abbreviations component of the research, we implemented a semi-supervised method. Each of the 461,501,598 tokens were compared to the approved abbreviations list supplied by Beth Israel Deaconess. The approved abbreviation list contained 1,743 abbreviations, both our total corpus and the potential abbreviation list were both cleaned before being manipulated. Each set of tokens had

punctuation and stop words removed. Our goal was to categorize each token in the note events table as a potential approved abbreviation.

For the analysis component of this research, we joined multiple tables in the MIMIC dataset that contained additional information on the type of clinician that wrote each note events. We imported all data into an analysis platform and perform targeted queries and analytics around our research questions. The end results included a web application that allows clinicians to visualize the relationship of terms extracted in our vector space. This web application can act as a powerful tool to allow clinicians to better understand clinical language or be utilized as a tool to monitor how clinical language is used in a facility.

3.2 Data Aggregation and Tools

This research uses a variety of data sources and tools that are both open sources and commercially available. We also developed a domain specific word2vec model to assist in the identification of potential emerging synonyms not identified by our data.

3.2.1 Data - MIMIC

Data for this research was obtained from the MIMIC database on the World Wide Web as described in Section 2.3. For the text extraction and machine learning experiments, we extracted the archive (zip) files containing the patient records for each of the study years (2001 – 2008) into a text file. Our experiments utilized the contents in TEXT, CATEGORY and SUBJECT_ID fields in NOTEEVENTS records. The TEXT field was used for text extraction, annotation, and classification and SUBJECT_ID field was used to map the record to information of interest in other files. CATEGORY was used to identify only note

types of interest and exclude notes that would not contain information that was pertinent to the experiment. We ignored the other fields. Table 2 lists the file names, table name, description of tables purpose, and the number data records contained:

File Name	Description	Table Name	Row Count
ADMISSIONS.csv.gz	Define a patient's hospital admission,	admissions	65,116
CALLOUT.csv.gz	Provides information when a patient was READY for discharge from the ICU, and when the patient was discharged from the ICU.	callout	38,235
CAREGIVERS.csv.gz	Defines the role of caregivers.	caregivers	7,696
CHARTEVENTS.csv.gz	Contains all charted data for all patients.	chartevents	390,726,794
CPTEVENTS.csv.gz	Contains current procedural terminology (CPT) codes, which facilitate billing for procedures performed on patients.	cptevents	603,244
D_CPT.csv.gz	High-level definitions for current procedural terminology (CPT) codes.	d_cpt	134
D_ICD_DIAGNOSES.csv.gz	Definition table for ICD diagnosis.	d_icd_diagnoses	15,713
D_ICD_PROCEDURES.csv.gz	Definition table for ICD procedures.	d_icd_procedures	3,939
D_ITEMS.csv.gz	Definition table for all items in the ICU databases.	d_items	12,544
D_LABITEMS.csv.gz	Definition table for all laboratory measurements.	d_labitems	753
DATETIMEEVENTS.csv.gz	Contains all date formatted data.	datetimeevents	3,886,274
DIAGNOSES_ICD.csv.gz	Contains ICD diagnoses for patients, most notably ICD-9 diagnoses.	diagnoses_icd	662,052
DRGCODES.csv.gz	Contains diagnosis related groups (DRG) codes for patients.	drgcodes	119,095
ICUSTAYS.csv.gz	Defines each ICUSTAY_ID in the database, i.e. defines a single ICU stay.	icustays	46,595

INPUTEVENTS_CV.csv.gz	Input data for patients.	inputevents_cv	15,035,790
INPUTEVENTS_MV.csv.gz	Input data for patients.	inputevents_mv	3,124,825
LABEVENTS.csv.gz	Contains all laboratory measurements for a given patient, including outpatient data.	labevents	30,322,892
MICROBIOLOGYEVENTS.csv.gz	Contains microbiology information, including tests performed and sensitivities.	microbiology events	658,262
NOTEEVENTS.csv.gz	Contains all notes for patients.	noteevents	1,862,200
OUTPUTEVENTS.csv.gz	Output data for patients.	outputevents	3,902,731
PATIENTS.csv.gz	Contains all charted data for all patients.	patients	39,186
PRESCRIPTIONS.csv.gz	Contains medication related order entries, i.e. prescriptions.	prescriptions	3,965,145
PROCEDUREEVENTS_MV.csv.gz	Contains procedures for patients	procedureevents_mv	205,254
PROCEDURES_ICD.csv.gz	Contains ICD procedures for patients, most notably ICD-9 procedures.	procedures_icd	188,835
SERVICES.csv.gz	Lists services that a patient was admitted/transferred under.	services	77,410
TRANSFERS.csv.gz	Physical locations for patients throughout their hospital stay.	transfers	226,920

Table 2 MIMIC Data source

3.2.2 Tool – Amazon Comprehend Medical

In any Natural Language Processing application often the first stage is preprocessing of the unstructured information. In python we would use a library such as Natural Language Tool Kit (NLTK) to perform common items such as tokenization, lemmatization, removing stop words, part of speech tagging, creating Ngrams. With AWS Comprehend Medical we can leave all the preprocessing and clinical named entity recognition to amazon's state of the art machine learning models. Amazon can quickly gather information related to medical

conditions, procedures, tests, and medications.

The strength of AWS Comprehend Medical includes the output of its clinical named entity recognition. While there are already biomedical ontology annotation applications available, some we even use in our pipeline, many require exact term matches in order to be correctly annotated to an ontology. AWS provides confidence level of all extractions allowing us only accept extractions we are highly confident in and manually reviewing any that do not meet a predefined threshold. In clinical documentation certainty is an extremely important concept, for example: 'Patient has cancer and 'Patient is negative for cancer' have critically different meanings. The AWS tool allows for us to understand the traits of the text and account for annotation.

After setting up the AWS and creating a token we can call the Comprehend Medical API directly from our Python Script:

```
import boto3

client = boto3.client('comprehendmedical')
response = client.detect_entities(
      Text='no signs of CHF')
```

Our response variable will contain a JSON output that provides information regarding our processed text. In our example above "no signs of CHF" returns the data found in table 4 below.

Category	Score	Text	Type	Trait	Trait Score
MEDICAL_CONDITION	0.987581	CHF	DX_NAME	NEGATION	0.917102

Table 3 AWS Comprehend Medical Output

3.2.3 Data – Unified Medical Language System – UMLS

The continued growth of the biomedical field and the rapid integration of technological solutions and products to biomedicine has dramatically increased the amount of available biomedical data. This rapid growth means that researchers are now tasked with the hurdle of having to extract only the data that is needed from the available data. Once our entities have been extracted from the test using AWS these entities still need to be standardized for appropriate use. Biomedical researchers have turned to ontologies and terminologies to structure and annotate their data with ontology concepts for better search and retrieval.

UMLS is a set of terminology and ontology dictionaries an metadata that are available as both a software and file download that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. UMLS has three tools called Knowledge Sources:[22]

- Metathesaurus: Terms and codes from clinical vocabularies, such as Current Procedural Terminology (CPT), International Statistical Classification of Diseases(ICD), and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)

- Semantic Network: Broad categories (semantic types) and their relations

- SPECIALIST Lexicon and Lexical Tools: Natural Language Processing tools

For our research we will be utilizing an API call of the UMLS metathesaurus. The Metathesaurus is a large, multi-purpose, and multilingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships. Their uses incorporate: patient care, health services billing, public health statistics, indexing and cataloging of biomedical literature, basic, clinical, and health services research. Extracting the terms via the UMLS Metathesaurus is not only becoming relatively standard for similar use cases, but it is also regularly updated, leading to a current place in which terms can be drawn from.

3.2.4 Tool – Domain Specific Word Embeddings

One of the benefits of UMLS is the ability to search for concepts based on synonyms. In our current example. AWS extracts the terms CHF; we pass CHF to the UMLS API and we return a SNOMED CT concept for Congestive Heart Failure. This allows us to standardize our terminology for rule building. What this doesn't account for is new and emerging synonyms, or even abbreviations that have not been recognized as synonyms. As a potential solution we have developed a domain specific word embedding model trained on the entire MIMIC NOTEEVENT corpus. Word embeddings can capture the context of a word in a document as well as the semantic and syntactic similarity and relation with other words.

Word2Vec is one of the most popular technique to perform word embeddings using shallow neural network with a low barrier to entry and easy to implement python API wrappers. The Word2Vec algorithm was developed by Tomas Mikolov in 2013 at Google. The objective of a word2vec embedding is for words that contain similar context occupy close spatial positions in a high dimensionality space. Mathematically we can use the cosine of the angle between these vectors within the word embedding to identify words that have similar semantic relationships. If the cosine is close to 1, i.e. the angle close to 0 than these two entities are more similar than entities where a cosine is closer to 0. Here comes the idea of generating distributed representations. In natural language it is understood that some words have a dependence on another words. Word embeddings and word2vec allow the words in context of each other to get greater share of this dependence when represented in a multidimensional space. Figure 2 displays the architecture of a shallow neural network Word2vec model[23].

The input or the context word is a one hot encoded vector of size V. The hidden layer contains N neurons and the output is again a V length vector with the elements being the softmax values.

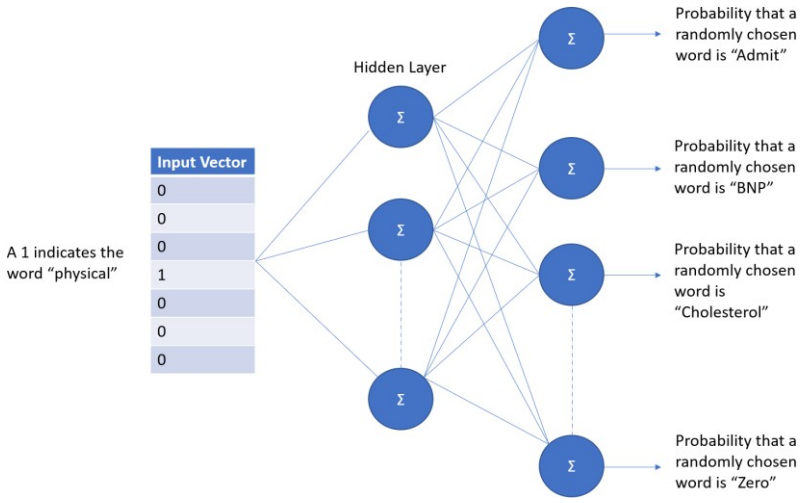


Figure 1: 2-layer shallow neural network architecture

This word embedding allows us to pass a term to the model and return the most similar term in the vector spaces. If AWS had extracted the term 'abx', a common abbreviation for 'antibiotics' and we passed this to UMLS we would not have received an annotation as UMLS does not recognize this concept. With our word embedding model we can pass 'abx' and ask the model to return the most similar terms found in the vector space as seen in figure 2 below:

```
In [130]: print(medrec_w2v.most_similar(["abx"]))
[('antibiotics', 0.6414980292320251), ('antibx', 0.542931854724884),
('antibiotic', 0.5059525370597839), ('anbx', 0.4764251112937927), ('cipro',
0.44416549801826477), ('ceftazimime', 0.407066285610199), ('vanco',
0.39654541015625), ('flagyl', 0.38997602462768555), ('antibiodics',
0.372799813747406), ('zosyn', 0.36688899993896484)]
```

Figure 2: Word2vec Most Similar 'abx'

The top three nearest vectors are ‘antibiotics’, ‘antibx’, and ‘antibiotic’ with the remaining terms being actual antibiotics or even misspellings. Using the nearest vector ‘antibiotics’ as our new UMLS term we successfully identify a SNOMED-CT concept that we can use for standardization.

3.3.0 Clinical Language Entity Extraction Pipeline

Figure 3 displays the diagram of our clinical language entity extraction pipeline. The unstructured free text is extracted from specified note types and sent to the Amazon Comprehend Medical cloud services through a Python API call. The returned entities from Amazon are then sent to the Unified Medical Language System through a second API call and annotated with specific SNOMED-CT concepts. If an extracted entity does not return a SNOMED-CT concept that entity will be passed to the Medrec_Word2ved model which is our domain specific work embedding model. The closest vector in the vector space will be replaced for the original entity and passed back to UMLS for annotation. This will repeat until all terms have been successfully annotated. Entities that are derived from the word embedding model may require human validation.

The extracted and mapped ontologies will then be fed to the rules engine to asses each document for clinical indicators and quality measures. After all rules have completed the results will then be output to an interactive web application that will allow users to view documents that met or failed our clinical indications or measures. The web application will also allow for users to reassess any selections made by the word embedding model and redetermine scoring. Future state may include a SOLR search engine to allow specific

clinicians to search across all their documentation, on any patient and view specific indicators.

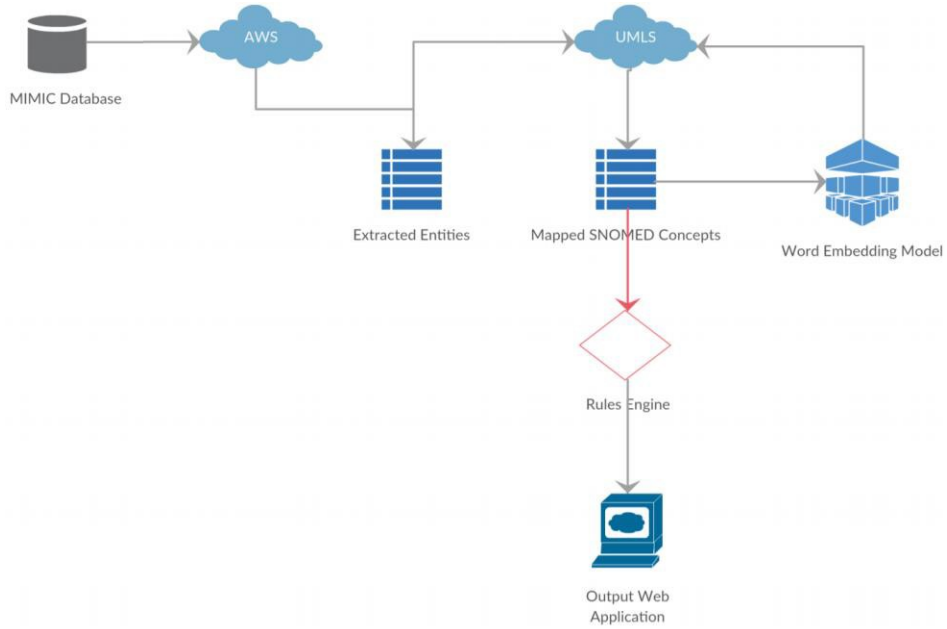


Figure 3 Process and Data Map

3.6.0 Procedures for Data Access

Both the MIMIC and UMLS databases are available to the public for use in research though there are specific procedures required for access. Both required data user agreements, compliance training access to specific web servers. The procedures to obtain these databases are outlined below

3.6.1 MIMIC

To obtain access to MIMIC a user must first create a PhysioNet account. Once the user has been granted access to PhysioNet the user will need to Complete the CITI (Collaborative Institutional Training Initiative at the University of Miami) “Data or Specimens Only Research” course as an MIT affiliate. After completing the CITI training the user will need to sign the PhysioNet Clinical Database Restricted Data Use Agreement.

3.6.2 UMLS

UMLS licenses are only issues to individuals and are not granted to groups or organizations. To create a UMLS Terminology Service account to access the UMLS services and terminology browsers users must accept the terms of the UMLS Metathesaurus License. Users will first sign up for an account on the UTS homepage. After reading and accepting the license and its appendices users will need to complete and submit a license request form. After receiving an email from the NLM users will be able to authenticate their license. Once the license has been reviewed and approved access to the UTS services will be granted to the user.

Chapter IV

Analysis and Results

4.1 Introduction

The word embedding model we created was processed over 1.8 million records to extract word embeddings for over 115 thousand individual word vectors. Now that we have our domain specific word embeddings Medrec2Vec, we can use this word embedding matrix to assist in additional Natural Language Processing applications to increased health literacy. Our research focuses on abbreviation disambiguation, concept extraction, and replacement of medical concepts.

4.2 Evaluating Potential Abbreviations in MIMIC

MIMIC-III note events corpus contains 463,175,566 raw word tokens, each token was compared to the list of approved abbreviations for Beth Israel Deaconess Medical Center. The list of approved abbreviations contains 1,742 abbreviations and their associated meaning. After filtering the potential abbreviations out of each document, we can evaluate the count, frequency, distribution of abbreviations within the corpus. Table 4 below shows the summary statistics of the count of abbreviation tokens, the count of word tokens, and the percentage of abbreviation tokens to word tokens.

	Potential Abbreviations	Word Tokens	Percent Abbreviations
Mean	13.42	318.26	9%
Standard Deviation	12.96	445.06	6%
Minimum	0.00	0.00	0%
1st Quartile	4.00	85.00	4%
Median	9.00	185.00	8%
3rd Quartile	20.00	339.00	13%
Maximum	57.00	10105.00	26%

Table 4 Token Summary Statistics

On average, every document within MIMIC contains almost 14 abbreviation tokens, or roughly 9% of an average note is comprised of potential abbreviations. Some notes contain almost 26% of potential abbreviation tokens. Figure 4 displays the distribution of potential abbreviation token counts among the MIMIC-III dataset.

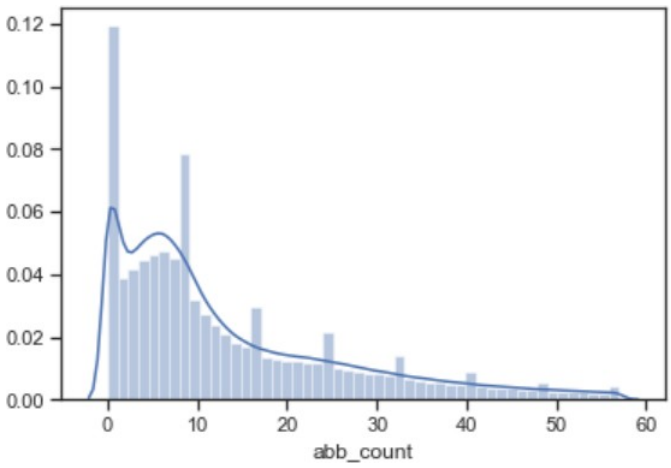


Figure 4: Potential Abbreviation Distribution

We also wanted to visually compare the distribution of our standardized abbreviation token percentage across the different categories within MIMIC. Figure 4 allowed us to visualize

this distribution among all note categories.

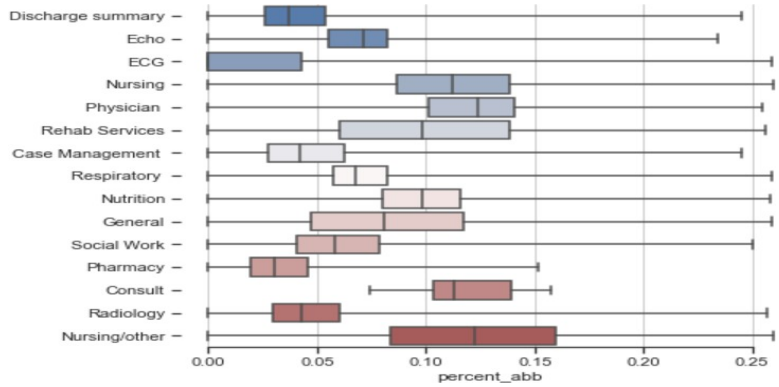


Figure 5: Percent abbreviation box plot

The top 50 most common abbreviations within MIMIC contain 53% of all potential abbreviations used within the corpus. In addition, the most commonly used abbreviation 'PT' makes up almost 10% by itself.

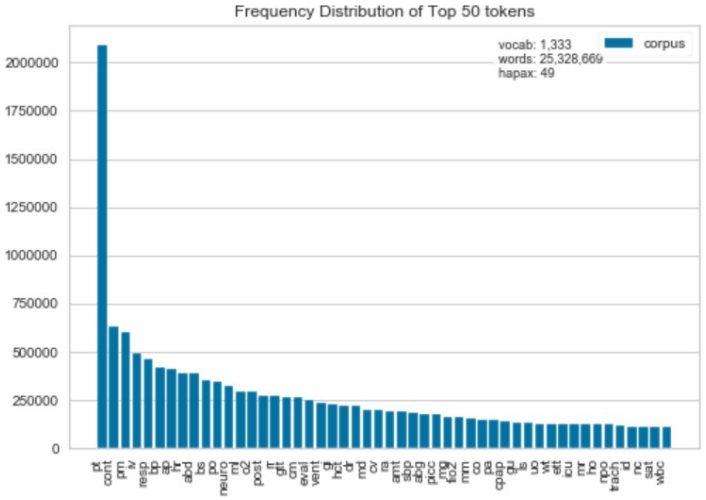


Figure 6: Top 50 Abbreviations

This is not surprising as ‘PT’ has multiple definitions approved by BIDMC. ‘PT’ can reference both ‘Physical Therapy’ and ‘Preterm, and ‘Prothrombin Time’. Our word embedding model shows that the most similar term in vector space to ‘PT’ is the term ‘patient’, this is not an approved abbreviation according the BIDMC policy but is an extremely common abbreviation.

Abbreviation	Definition
P.T.	**physical therapy /Physical Therapist**
PT	**prothrombin time**
PT	**preterm**

Table 5: Approved PT Definitions Example

4.3 Evaluating Potential Abbreviations Between Clinicians

When viewing the box plot for each note category we can see that there is variance between the note categories on the amount of abbreviations used. Our second hypothesis states that there is a statistically significant difference in the number of abbreviations used in documentation between physicians and nurses. MIMIC provides a ‘Caregiver’ table that contains a unique ID for each clinical action taken as well as a ‘Label’ column that provides that users credentials. By joining this table to our notes events tables, we can compare the prevalence of abbreviations within physician and nursing documentation.

By filtering our Note Events data to only caregivers that are either an RN or MD we can begin to compare the abbreviation usage of these two groups. Our comparison data contains

617,642 notes created by an RN and 111,281 notes created by an MD, this is 33% and 6% of our original Note Events data respectively. The average count of potential abbreviations for a note create by an RN is 21.87, and the average count of potential abbreviations in a note created by an MD is 11.39

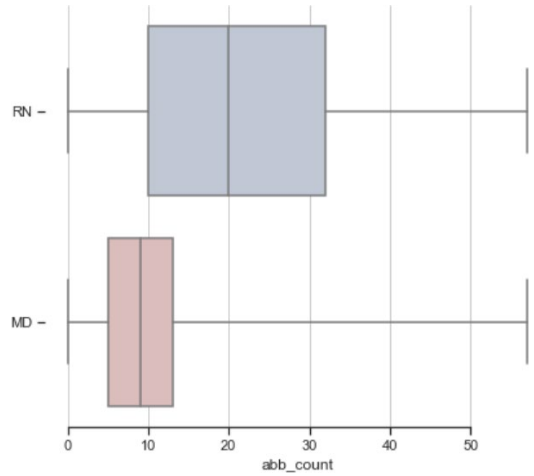


Figure 7: RN vs MD Abbreviation Usage

We can see from the Box plot that there appears to be a difference in the amount of abbreviations used by these two types of clinicians. The first step in determining the appropriate hypothesis testing, is to examine the data's characteristics. The following are true or assumed for this data:

1. The samples are independent of each other and none of the documents could be included in both.

2. The dependent variable, count of abbreviations, is continuous.

3. The independent variable, Clinician Type (‘Label’) has 2 groups, and they are categorical.

4. Outliers have been removed

5. Variances are homogenous, p<.001

```
LeveneResult(statistic=65712.91030482127, pvalue=0.0)
```

It was determined that the distribution does not follow a normal distribution, as illustrated in the Q-Q plot below:

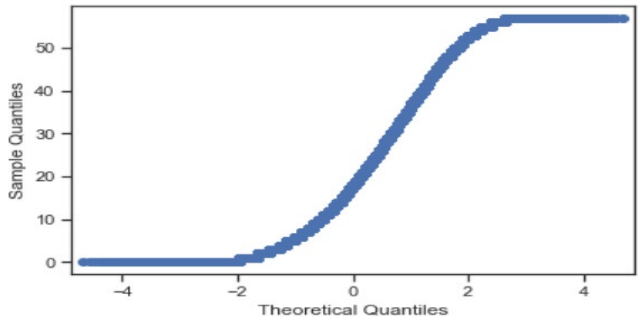


Figure 8: Q-Q plot

The Wilcoxon Rank Sum test was chosen as the non-parametric test of the independent t-test due to the distribution not meeting the assumption of normality. The Wilcoxon Rank only makes the assumptions of independence and equal variance. In a two-sample t test the null hypothesis is equal means, with the Wilcoxon test the null hypothesis is equal medians. The Wilcoxon test tests the assumption that the two populations have the same distribution

with the same median. To reject the null means, we have evidence that one distribution differs and is shifted to either the left or right.

If the Wilcoxon Rank Sum is statistically significant, additional testing must be performed to analyze the effect size. We want to try and quantify the difference between these two distributions, not just prove a statistical difference. As we can see in figure 8 below, the p value of Rank Sum test is P <= 0.001, we would reject the null hypotheses.

```
stat, p = stats.ranksums(RNandMD['abb_count'][RNandMD['LABEL'] == 'RN'], RNandMD['abb_count'][RNandMD['LABEL'] == 'MD'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')
```

```
Statistics=244.823, p=0.000
Different distributions (reject H0)
```

Figure 9 Wilcoxon Rank Sum

To assess the effect size between the two groups we chose to use Cliff's D, which has the benefit of providing a quantifiable metrics, as well as how large of an effect there is. Cliff's D measures of how often the values in one distribution are larger or smaller than the values in another distribution. The benefit of Cliff's D for our experiment is it does not require any assumptions about the shape of the two distributions. Cliff's D shows that there is a medium effect size.

```
In [198]: cliffsDelta(RNandMD['abb_count'][RNandMD['LABEL']=='RN'], RNandMD['abb_count'][RNandMD['LABEL']=='MD'])
Out[198]: (0.4603131833445889, 'medium')
```

4.4 MedRec2Vec – Domain Specific Word Embedding

Our domain specific word embedding is trained on the full MIMIC-III note events corpus. Table 6 below shows the summary statistics of the entire trainable corpus.

	Sentence Tokens	Word Tokens
Mean	19.14	318.26
Standard Deviation	24.43	445.06
Minimum	0.00	0.00
1st quartile	5.00	85.00
Median	12.00	185.00
3rd Quartile	25.00	339.00
Maximum	594.00	10105.00

Table 6: Note Events Summary Statistics

We collected 979,953 unique words from our total corpus of 463,175,566 raw words, and 44,854,934 sentences. We chose to remove all words that were not present a minimum of ten times in our corpus. By utilizing a minimum count, we could ensure that the embedding was not an uncommon, or one-off term. By removing these uncommon words, we were left with 115,727 unique words, only 11% of our original 979,953. While our unique words were dramatically reduced, our training corpus was still at 99% of the original size with 461,501,598 words, dropping 1,673,968 words. Figure 9 visualizes this corpus reduction through the removal of uncommon tokens.

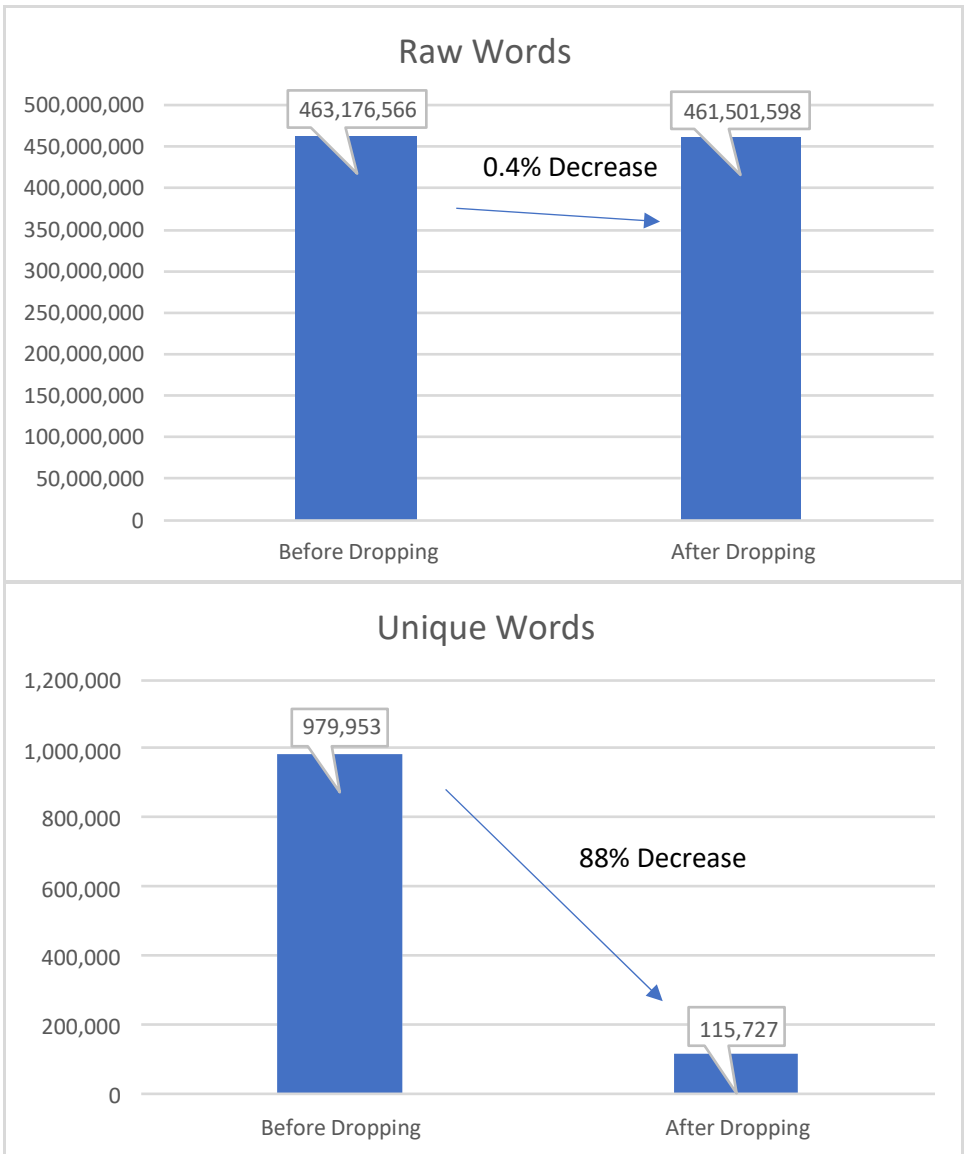


Figure 9: Word2vec Corpus Reduction

We also chose to downsample our corpus for the most common .1% of words, effectively downsampling 36 of our most-common terms. Our model was ran using 96 worker CPU's on 115,727 vocabulary with 200 features using a continuous bag of words approach. We used a window size of 5 allowing 5 words between our current term and predicted word.

4.5 Most Similar Abbreviations – Beth Israel Abbreviations

Using the Medrec2vec word embedding model we extracted the ten most similar terms for each approved abbreviation at Beth-Israel Deaconess (BID) and assessed if vector space contained the semantic meaning for the abbreviated term. Of the 1,743 abbreviations approved by BID the word embedding model was able to accurately extract the semantic relationship for 963 terms.

	True Positive	True Negative
Predicted Positve	963	620
Predicted Negative	160	

Table 7: Performance 10 most similar

620 abbreviation terms were not able to extract the appropriate semantic term, and 160 terms were not found within the vector space of the model. Our model achieved a precision of .60, a recall of .85, and an F1 of .71, while our model performed decently only using term similarity, it struggled when abbreviations had multiple meanings.

While our first approach was to see if the word embedding model contained the relationship for all the approved abbreviations within the first ten most similar vector spaces our result requires a decision to be made on a potential replacement term. We decided to use the most similar vector score as our replacement term for each extracted abbreviation.

4.6 Most Similar Abbreviations – Do Not Use

Using the Medrec2vec word embedding model we extracted the ten most similar terms for each abbreviation on the Joint Commission do not use and assessed if vector space contained the semantic meaning for the abbreviated term. Of the 6 abbreviations listed on

the do not use list the word embedding model was able to accurately extract the semantic relationship for all 6 terms. This result shows promise for replacing potential do not use abbreviations with their appropriate full term.

4.7 Extractions and Replacements

Using our pipeline, we used 20 consult notes from MIMIC-III as our test set. Each document was annotated with the expected medical conditions, tests, procedures, labs, medications, and anatomical terms before being sent through the pipeline. Each document was parsed through Amazon Comprehend Medical and extracted all entities from each document. Tables 8 – 13 display summary statistics for the extractions from amazon comprehend.

Total Extractions	
Mean	192.45
Standard Error	20.57
Median	190.50
Mode	133.00
Standard Deviation	91.99
Sample Variance	8461.63
Kurtosis	2.53
Skewness	0.87
Range	441.00
Minimum	11.00
Maximum	452.00
Sum	3849.00
Count	20.00

Table 8 Total Extractions

Test_Treatment_Procedure	
Mean	56.55
Standard Error	4.866899478
Median	64.00
Mode	66.00
Standard Deviation	21.77
Sample Variance	473.73
Kurtosis	0.12
Skewness	-0.65
Range	85.00
Minimum	4.00
Maximum	89.00
Sum	1131.00
Count	20.00

Table 9 Test_Treatment_Procedure Extraction

Medical_Condition	
Mean	70.6
Standard Error	11.94667979
Median	54.50
Mode	103.00
Standard Deviation	53.43
Sample Variance	2854.46
Kurtosis	1.55
Skewness	1.31
Range	210.00
Minimum	4.00
Maximum	214.00
Sum	1412.00
Count	20.00

Table 10 Medical_Condition Extractions

Anatomy	
Mean	40.6
Standard Error	5.914923143
Median	37.00
Mode	2.00
Standard Deviation	26.45
Sample Variance	699.73
Kurtosis	2.15
Skewness	1.12
Range	113.00
Minimum	2.00
Maximum	115.00
Sum	812.00
Count	20.00

Table 11 Anatomy Extractions

PHI	
Mean	7.25
Standard Error	1.699651667
Median	5.50
Mode	6.00
Standard Deviation	7.60
Sample Variance	57.78
Kurtosis	7.87
Skewness	2.49
Range	34.00
Minimum	0.00
Maximum	34.00
Sum	145.00
Count	20.00

Table 12 PHI Extractions

Medication	
Mean	17.45
Standard Error	3.028352861
Median	16.00
Mode	20.00
Standard Deviation	13.54
Sample Variance	183.42
Kurtosis	0.11
Skewness	0.85
Range	45.00
Minimum	1.00
Maximum	46.00
Sum	349.00
Count	20.00

Table 13 Medication Extractions

The reliance of Amazon Comprehend Medical in our pipeline also meant that we needed to assess the performance of ACM for clinical named entity recognition and ensure that it met that standards needed to extract pertinent clinical information to annotate and disambiguate. Below is our evaluation of the amazon comprehend extraction, and classification performance.

Each extracted entity from Amazon Comprehend Medical was then compared to a list of 1,743 approved abbreviations from Beth Israel Deaconess and if the entity matched an abbreviation it was passed to our word embedding model. The extracted abbreviation was then replaced with the term that was found closest in vector space. We then compared the performance of our method to replace terminology. Of the 3,849 total entity extractions, 705 (18%) entities matched a term on our approved abbreviation list. Using only the nearest term in vector space our word embedding approach properly identified 117 terms for an

accuracy of .17, performance for only the most similar approach is lacking but does show some potential.

Looking at Beth Israel's approved abbreviations and extracting the most similar score we were only able to match the full term 17% of the time. We noticed in our research that the approved abbreviation list contains more than just abbreviations. Our model can only return one token that is most similar in the vector space. By filtering and accounting for acronyms, initialisms, and multi-token responses our test set was reduced from 1,740 terms to 291 terms. After Passing the 291 and evaluating the model performance we achieved a 32% increase in accuracy and matched the appropriate definition with a 49% accuracy. As part of our rules engine we will focus on replacing these 291 terms when extracted from free text notes using Amazon Comprehend Medical.

4.8 Web Application

CLEAR has an interactive web application to assist in increasing understanding of the information contained in medical documentation. This web application helps patients and clinicians increase health literacy in documentation. The web application is comprised of Extract, Explore, and Compare. Each of these tools are designed to assist users in being able to better understand their medical records through interactive visualization and entity extraction.

Extract Entities allows a user to input or import documentation and extract the pertinent clinical terms from the text. This section sends the documentation to AWS Comprehend Medical to extract insights.

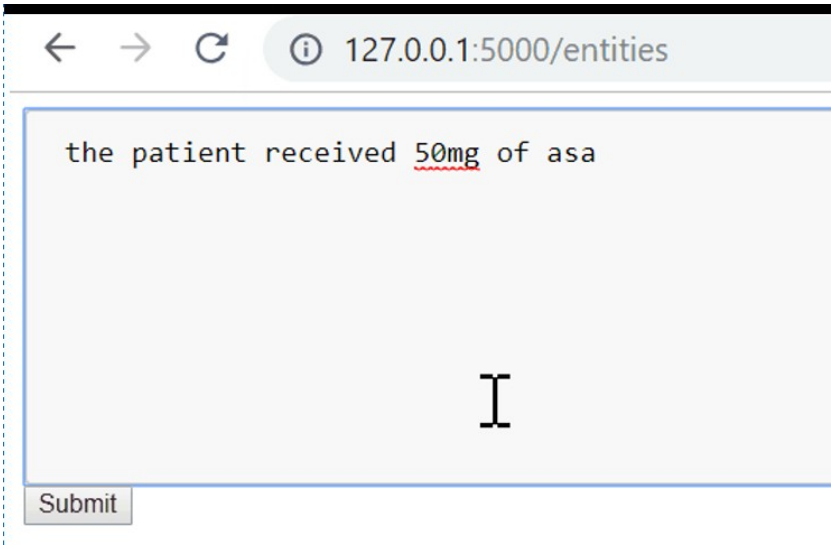
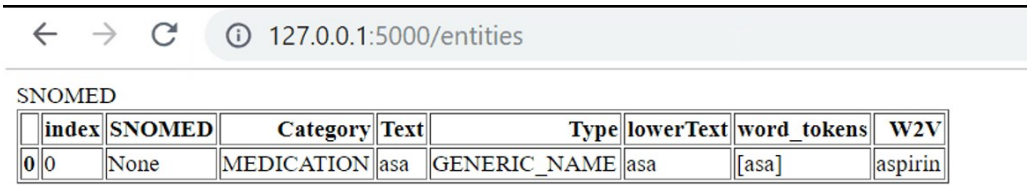


Figure 10: Entities Input

Once the AWS API returns the extracted clinical terms each term is sent to UMLS to extract an annotation. If an annotation is not found for the term the pipeline will then send the term to our word embedding model to try and find a potential replacement term.



	index	SNOMED	Category	Text	Type	lowerText	word_tokens	W2V
0	0	None	MEDICATION	asa	GENERIC_NAME	asa	[asa]	aspirin

Figure 11: Entities extraction/annotation

Explore allows users to interact directly with the word embedding model to help better understand relationships and decisions from Extract. Users can submit lists of terms to understand what relationships exist in the model using a dendrogram. Users can also submit

terms to see the similarity in vector space.

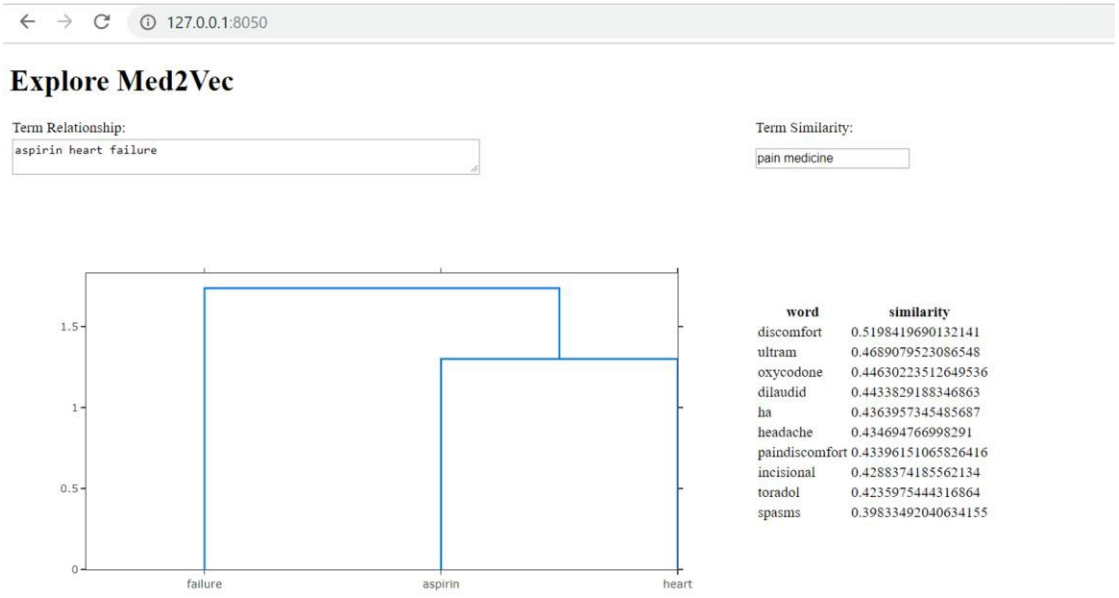


Figure 12: Explore

Compare is an interactive TSNE and Kmeans clustering. TSNE allows dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. Kmeans allows us to cluster our terms and visualize relationships. Compare allows users to submit queries to define relationships and return clusters of similar terms.

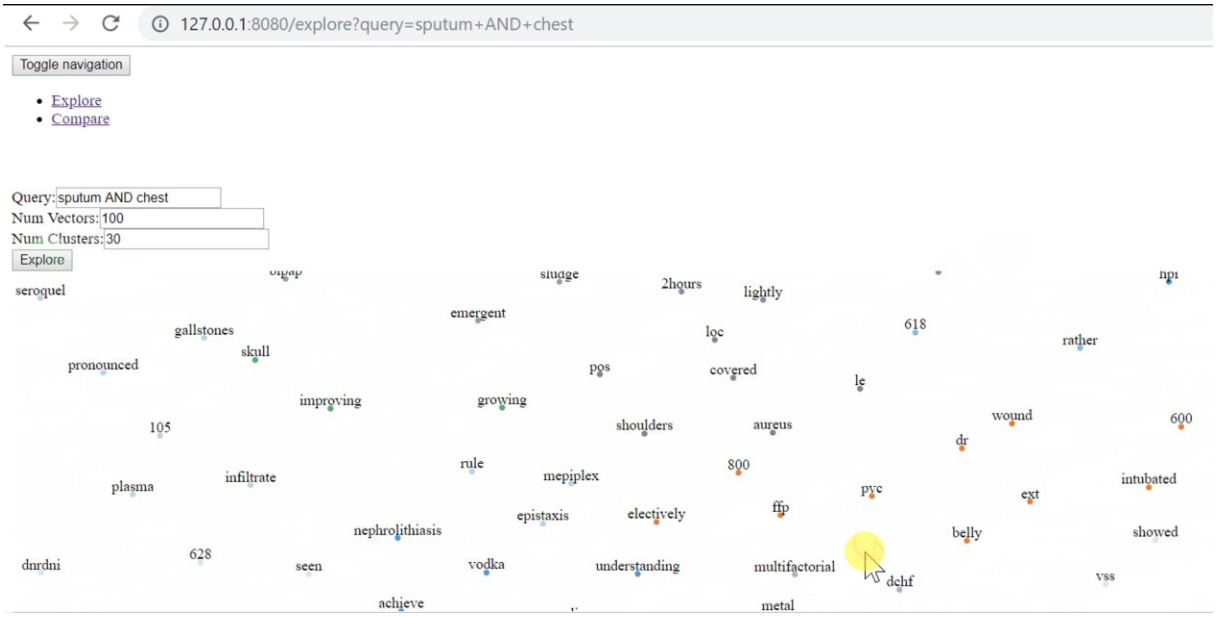


Figure 13: Compare

Chapter V

Summary and Conclusion

5.1 Introduction

With the push for national EHR adoption and the subsequent increase in patients access to their clinical data, including provider notes, health literacy will be a major barrier in patients being actively involved in their care. While patients have this access, we as public health professionals, will need to develop novel application to transform clinical jargon into plain text. This research hopes to add to the current sparseness of data, and research, to demonstrate that novel web applications can be used to better disambiguate medical terminology. Using clinical named entity extraction, domain specific word embedding models, and rules-based replacements health literacy can be increased by developing novel applications to reduce barriers in clinical documentation to the layperson.

Using the MIMIC data set we have shown that clinical abbreviations and complex jargon make up a specific amount of provider documentation. 8.22% of total words within the MIMIC note events table is a term found within the Beth Israel Deaconess approved abbreviation list. We have also shown that there is the capability to replace abbreviations in medical text to provide additional context to patients and providers. Complex natural language processing tools can be used by administrators and clinicians to better understand what and how documentation is being entered into their systems.

5.2 Limitations

The limitations of this study reflect the current lack of data in discovering the overall

increase of health literacy with the use clinical language extraction and replacement. This is the only study known to the researcher specifically examining the use of Amazon Comprehend Medical for clinical named entity extraction on clinical documentation, which weakens any findings until additional research is completed. The increase in health literacy from term extraction and replacement is also outside the purview of this thesis as we set out to only assess the ability of commercial applications associated with clinical named entity recognition and the capability of unsupervised word embeddings to select potential term replacements. Because MIMIC datetime stamps are date shifted in order to protect patient privacy during the MIMIC deidentification process, it is not possible to compare the change in abbreviation use across a span of time. This limitation prevents a longitudinal view of the use of abbreviations within clinical documentation, which would have been helpful in determining the impact of our study in real world applications.

5.3 Application of Research

While this research is exploratory in nature, it is developed with real world applications. The Joint Commission Maintains a "Do Not Use List" of medical abbreviations and ensures facilities maintain abbreviations in their documentation. This standard, Standard MOI.4 states: *" The hospital uses standardized diagnosis and procedure codes and ensures the standardized use of approved symbols and abbreviations across the hospital"* [24]. Using a word embedding model hospitals can monitor and evaluate how abbreviations are being used and proactively correct erroneous documentation.

5.4 Implications for Future Study

The results of this study, while limited, are encouraging to the expansion of knowledge about the use of abbreviations in clinical documentation. This research also has positive impact on the use of health information technology and advanced analytical tools to monitor and evaluate the ever-changing documentation style of clinicians. Additional study should expand on this data by measuring the abbreviation use among additional clinicians and specialties to determine if there are similar effects within these populations. In addition, research should be done on sub-domain (Radiology, Cardiology, Emergency Medicine etc..) specific word embeddings even within the healthcare domain. By creating word embeddings of specialty specific documentation, we can create embeddings that are highly representative of how that specialty documents. A robust clinical natural language processing application can be utilized to monitor documentation within a facility to ensure adherence to internal organization policy and external standards or guidelines. Clinical natural language processing can also help facilities prevent fraud, waste, and abuse by allowing quality and compliance to real time monitor documentation.

REFERENCES

1. Structured and Unstructured Data: What is It? – Sherpa Software. https://sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/. Accessed September 12, 2019.

2. Nadkarni A. Structured Versus Unstructured Data: The Balance of Power Continues to Shift. March 2014.

3. MIMIC. https://mimic.physionet.org/about/mimic/. Accessed February 8, 2018.

4. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3. doi:10.1038/sdata.2016.35

5. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc. 2018;25(1):32-39. doi:10.1093/jamia/ocx084

6. Lee J, Ribey E, Wallace JR. A web-based data visualization tool for the MIMIC-II database. BMC Med Inform Decis Mak. 2016;16. doi:10.1186/s12911-016-0256-9

7. Dervishi A. Fuzzy risk stratification and risk assessment model for clinical monitoring in the ICU. Computers in Biology and Medicine. 2017;87:169-178. doi:10.1016/j.compbiomed.2017.05.034

8. Carvalho GMC de, Leite TT, Libório AB. Prediction of 60-day Case Fatality in Critically-ill Patients Receiving Renal Replacement Therapy: External Validation of a Prediction Model. Shock. 2017;Publish Ahead of Print. doi:10.1097/SHK.0000000000001054

9. Poucke SV, Zhang Z, Schmitz M, et al. Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform. PLoS One. 2016;11(1). doi:10.1371/journal.pone.0145791

10. Bromuri S, Zufferey D, Hennebert J, Schumacher M. Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. Journal of Biomedical Informatics. 2014;51:165-175. doi:10.1016/j.jbi.2014.05.010

11. Lehman L, Saeed M, Long W, Lee J, Mark R. Risk Stratification of ICU Patients Using Topic Models Inferred from Unstructured Progress Notes. AMIA Annu Symp Proc. 2012;2012:505-511.

12. Septic shock; current pathogenetic concepts from a clinical perspective. Medical Science Monitor. https://www.medscimonit.com/download/index/idArt/15400. Accessed March 23, 2018.

13. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLOS ONE. 2018;13(2):e0192360. doi:10.1371/journal.pone.0192360

14. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of Biomedical Informatics. 2017;76:102-109. doi:10.1016/j.jbi.2017.11.007

15. Douglass M, Clifford GD, Reisner A, Moody GB, RG M. Computer-assisted de-identification of free text in the MIMIC II database. In: Computers in Cardiology, 2004. ; 2004:341-344. doi:10.1109/CIC.2004.1442942

16. Sinha S, McDermott F, Srinivas G, Houghton PWJ. Use of abbreviations by healthcare professionals: what is the way forward? Postgrad Med J. 2011;87(1029):450-452. doi:10.1136/pgmj.2010.097394

17. Tsima BM. Use of Abbreviations and Acronyms among Healthcare Workers in a Resource Limited Setting. Journal of Healthcare Communications. 2017;2(3). doi:10.4172/2472-1654.100063

18. Health Literacy - Fact Sheet: Health Literacy Basics. https://health.gov/communication/literacy/quickguide/factsbasic.htm. Accessed March 20, 2019.

19. Xie B. Improving older adults' e-health literacy through computer training using NIH online resources. Library & Information Science Research. 2012;34(1):63-71. doi:10.1016/j.lisr.2011.07.006

20. wu yonghui, Xu J, Zhang Y, Xu H. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. In: Proceedings of BioNLP 15. Beijing, China: Association for Computational Linguistics; 2015:171-176. doi:10.18653/v1/W15-3822

21. Yu H, Kim W, Hatzivassiloglou V, Wilbur WJ. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. Journal of Biomedical Informatics. 2007;40(2):150-159. doi:10.1016/j.jbi.2006.06.001

22. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/index.html. Accessed October 3, 2019.

23. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:13013781 [cs]. January 2013. http://arxiv.org/abs/1301.3781. Accessed March 21, 2019.

24. Use of Codes, Symbols, and Abbreviations. Joint Commission International. https://www.jointcommissioninternational.org/use-of-codes-symbols-and-abbreviations/. Accessed September 12, 2019.

Appendix

Appendix A – Word Embedding Code

```
import pandas as pd
import sklearn
from sklearn import model_selection, preprocessing, linear_model,
naive_bayes, metrics, svm
from sklearn.feature_extraction.text import TfidfVectorizer,
CountVectorizer
from sklearn import decomposition, ensemble
import nltk import gzip import gensim
from gensim.models import word2vec
from gensim.models.word2vec import Word2Vec LabeledSentence =
gensim.models.doc2vec.LabeledSentence import logging
from string import punctuation
import string
import re
from tqdm import tqdm

import pymysql.cursors ###Data Loading###
# Connect to the database.
connection = pymysql.connect(host='localhost',
user=username, password=password, db='mimic', charset='utf8mb4',
cursorclass=pymysql.cursors.DictCursor) print ("connect successful!!")


# SQL Queries
nsql = "Select *        from mimic.noteevents where noteevents.category
= 'Nursing' limit 225000"
#225000
nursing = pd.read_sql(nsql, connection)
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',
level=logging.INFO)

#Load note data and create corpus data = nursing
DF = pd.DataFrame(data) text = DF['TEXT']
```

```
corpus = text.str.cat(sep=' ') #remove deidentification patterns
data['clean_text'] = data.apply(lambda row:
re.sub('\[\*\*(\w*\s*|\(\w*\)|(\d*\-))*\*\*\]', ' ', row['TEXT']),
axis=1)
data['clean_text'] = data['clean_text'].str.replace('[^\w\s]','')

#Stemming
from nltk.stem.porter import PorterStemmer porter_stemmer =
PorterStemmer()
stem = porter_stemmer.stem(corpus)
data['clean_text'] = data['clean_text'].apply(porter_stemmer.stem)

#Toeknize sentences
from nltk.tokenize import sent_tokenize sent_tokenize_list =
sent_tokenize(corpus)
data['sent_tokens'] = data['clean_text'].apply(sent_tokenize)


#Tokenize words
from nltk.tokenize import word_tokenize word_tokenize_list =
word_tokenize(stem)
data['word_tokens'] = data['clean_text'].apply(word_tokenize)

#remove punctuation
cleanword = re.sub(r'[^\w\s]','',stem)
lowercased_sents = [sent.lower() for sent in sent_tokenize_list]
discard_punctuation_and_lowercased_sents = [re.sub(r'[^\w\s]','',sent)
for sent in lowercased_sents]

sent1 = [word_tokenize(sent) for sent in
discard_punctuation_and_lowercased_sents]

from nltk.corpus import stopwords #filter stop words
filtered_words = [word for word in sent1 if word not in
stopwords.words('english')]

#Word2vec Model
medrec_w2v = Word2Vec(size=200, min_count=10, iter=500)
medrec_w2v.build_vocab(filtered_words) medrec_w2v.train(filtered_words,
epochs=medrec_w2v.iter, total_examples=medrec_w2v.corpus_count)

medrec_w2v.save('medrec_w2v')
```

Appendix B – MIMIC Database Loading Code

Loading code provided by MIMIC and modified to create local schema

```
-- csv2mysql with arguments:
--          -o
--          1-define.sql
--          -u
--          -k
--          -p
--          -z
--          ADMISSIONS.csv
--          CALLOUT.csv
--          CAREGIVERS.csv
--          CHARTEVENTS.csv
--          CPTEVENTS.csv
--          DATETIMEEVENTS.csv
--          DIAGNOSES_ICD.csv
--          DRGCODES.csv
--          D_CPT.csv
--          D_ICD_DIAGNOSES.csv
--          D_ICD_PROCEDURES.csv
--          D_ITEMS.csv
--          D_LABITEMS.csv
--          ICUSTAYS.csv
--          INPUTEVENTS_CV.csv
--          INPUTEVENTS_MV.csv
--          LABEVENTS.csv
--          MICROBIOLOGYEVENTS.csv
--          NOTEEVENTS.csv
--          OUTPUTEVENTS.csv
--          PATIENTS.csv
--          PRESCRIPTIONS.csv
--          PROCEDUREEVENTS_MV.csv
--          PROCEDURES_ICD.csv
--          SERVICES.csv
--          TRANSFERS.csv


DROP TABLE IF EXISTS ADMISSIONS;
CREATE TABLE ADMISSIONS (          -- rows=58976 ROW_ID SMALLINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED NOT NULL, ADMITTIME DATETIME NOT NULL,
DISCHTIME DATETIME NOT NULL,
DEATHTIME DATETIME,
ADMISSION_TYPE VARCHAR(255) NOT NULL,              -- max=9
ADMISSION_LOCATION VARCHAR(255) NOT NULL,          -- max=25
```

```
DISCHARGE_LOCATION VARCHAR(255) NOT NULL,              -- max=25
INSURANCE VARCHAR(255) NOT NULL, -- max=10
LANGUAGE VARCHAR(255),             -- max=4 RELIGION VARCHAR(255), --
max=22 MARITAL_STATUS VARCHAR(255), -- max=17
ETHNICITY VARCHAR(255) NOT NULL, -- max=56 EDREGTIME DATETIME,
EDOUTTIME DATETIME,
DIAGNOSIS TEXT,           -- max=189 HOSPITAL_EXPIRE_FLAG TINYINT
UNSIGNED NOT NULL, HAS_CHARTEVENTS_DATA TINYINT UNSIGNED NOT NULL,
UNIQUE KEY ADMISSIONS_ROW_ID (ROW_ID),             -- nvals=58976
UNIQUE KEY ADMISSIONS_HADM_ID (HADM_ID)            -- nvals=58976
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/ADMISSIONS.csv' INTO
TABLE
ADMISSIONS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ADMITTIME,@DISCHTIME,@DEATHTIME,@ADMISSI
ON_TYP
E,@ADMISSION_LOCATION,@DISCHARGE_LOCATION,@INSURANCE,@LANGUAGE,@RELIGIO
N,@MAR
ITAL_STATUS,@ETHNICITY,@EDREGTIME,@EDOUTTIME,@DIAGNOSIS,@HOSPITAL_EXPIR
E_FLAG
,@HAS_CHARTEVENTS_DATA) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ADMITTIME = @ADMITTIME, DISCHTIME = @DISCHTIME,
DEATHTIME = IF(@DEATHTIME='', NULL, @DEATHTIME), ADMISSION_TYPE =
@ADMISSION_TYPE, ADMISSION_LOCATION = @ADMISSION_LOCATION,
DISCHARGE_LOCATION = @DISCHARGE_LOCATION, INSURANCE = @INSURANCE,
LANGUAGE = IF(@LANGUAGE='', NULL, @LANGUAGE), RELIGION =
IF(@RELIGION='', NULL, @RELIGION),
MARITAL_STATUS = IF(@MARITAL_STATUS='', NULL, @MARITAL_STATUS),
ETHNICITY = @ETHNICITY,
EDREGTIME = IF(@EDREGTIME='', NULL, @EDREGTIME), EDOUTTIME =
IF(@EDOUTTIME='', NULL, @EDOUTTIME), DIAGNOSIS = IF(@DIAGNOSIS='',
NULL, @DIAGNOSIS), HOSPITAL_EXPIRE_FLAG = @HOSPITAL_EXPIRE_FLAG,
HAS_CHARTEVENTS_DATA = @HAS_CHARTEVENTS_DATA;

DROP TABLE IF EXISTS CALLOUT;
CREATE TABLE CALLOUT (        -- rows=34499 ROW_ID SMALLINT UNSIGNED
NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT
UNSIGNED NOT NULL, SUBMIT_WARDID TINYINT UNSIGNED,
SUBMIT_CAREUNIT VARCHAR(255),             -- max=5 CURR_WARDID TINYINT
UNSIGNED
CURR_CAREUNIT VARCHAR(255),           -- max=5 CALLOUT_WARDID TINYINT
UNSIGNED NOT NULL, CALLOUT_SERVICE VARCHAR(255) NOT NULL,    -- max=5
REQUEST_TELE TINYINT UNSIGNED NOT NULL, REQUEST_RESP TINYINT UNSIGNED
NOT NULL, REQUEST_CDIFF TINYINT UNSIGNED NOT NULL, REQUEST_MRSA TINYINT
UNSIGNED NOT NULL, REQUEST_VRE TINYINT UNSIGNED NOT NULL,
CALLOUT_STATUS VARCHAR(255) NOT NULL,  -- max=8 CALLOUT_OUTCOME
VARCHAR(255) NOT NULL,                -- max=10 DISCHARGE_WARDID
TINYINT UNSIGNED,
ACKNOWLEDGE_STATUS VARCHAR(255) NOT NULL,             -- max=14
CREATETIME DATETIME NOT NULL,
UPDATETIME DATETIME NOT NULL, ACKNOWLEDGETIME DATETIME, OUTCOMETIME
```

```
DATETIME NOT NULL, FIRSTRESERVATIONTIME DATETIME,
CURRENTRESERVATIONTIME DATETIME,
UNIQUE KEY CALLOUT_ROW_ID (ROW_ID),            -- nvals=34499
UNIQUE KEY CALLOUT_CURRENTRESERVATIONTIME (CURRENTRESERVATIONTIME)

--
nvals=1164
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/CALLOUT.csv' INTO TABLE
CALLOUT
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@SUBMIT_WARDID,@SUBMIT_CAREUNIT,@CURR_WAR
DID,@C
URR_CAREUNIT,@CALLOUT_WARDID,@CALLOUT_SERVICE,@REQUEST_TELE,@REQUEST_RE
SP,@RE
QUEST_CDIFF,@REQUEST_MRSA,@REQUEST_VRE,@CALLOUT_STATUS,@CALLOUT_OUTCOME
,@DISC
HARGE_WARDID,@ACKNOWLEDGE_STATUS,@CREATETIME,@UPDATETIME,@ACKNOWLEDGETI
ME,@OU TCOMETIME,@FIRSTRESERVATIONTIME,@CURRENTRESERVATIONTIME)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
SUBMIT_WARDID = IF(@SUBMIT_WARDID='', NULL, @SUBMIT_WARDID),
SUBMIT_CAREUNIT = IF(@SUBMIT_CAREUNIT='', NULL, @SUBMIT_CAREUNIT),
CURR_WARDID = IF(@CURR_WARDID='', NULL, @CURR_WARDID), CURR_CAREUNIT =
IF(@CURR_CAREUNIT='', NULL, @CURR_CAREUNIT), CALLOUT_WARDID =
@CALLOUT_WARDID,
CALLOUT_SERVICE = @CALLOUT_SERVICE, REQUEST_TELE = @REQUEST_TELE,
REQUEST_RESP = @REQUEST_RESP, REQUEST_CDIFF = @REQUEST_CDIFF,
REQUEST_MRSA = @REQUEST_MRSA, REQUEST_VRE = @REQUEST_VRE,
CALLOUT_STATUS = @CALLOUT_STATUS, CALLOUT_OUTCOME = @CALLOUT_OUTCOME,
DISCHARGE_WARDID = IF(@DISCHARGE_WARDID='', NULL, @DISCHARGE_WARDID),
ACKNOWLEDGE_STATUS = @ACKNOWLEDGE_STATUS,
CREATETIME = @CREATETIME,
```

```
UPDATETIME = @UPDATETIME,
ACKNOWLEDGETIME = IF(@ACKNOWLEDGETIME='', NULL, @ACKNOWLEDGETIME),
OUTCOMETIME = @OUTCOMETIME,
FIRSTRESERVATIONTIME = IF(@FIRSTRESERVATIONTIME='', NULL,
@FIRSTRESERVATIONTIME),
CURRENTRESERVATIONTIME = IF(@CURRENTRESERVATIONTIME='', NULL,
@CURRENTRESERVATIONTIME);

DROP TABLE IF EXISTS CAREGIVERS;
CREATE TABLE CAREGIVERS (      -- rows=7567 ROW_ID SMALLINT UNSIGNED NOT
NULL, CGID SMALLINT UNSIGNED NOT NULL, LABEL VARCHAR(255),  -- max=6
DESCRIPTION VARCHAR(255),              -- max=21
UNIQUE KEY CAREGIVERS_ROW_ID (ROW_ID),             -- nvals=7567
UNIQUE KEY CAREGIVERS_CGID (CGID) -- nvals=7567
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/CAREGIVERS.csv' INTO
TABLE
CAREGIVERS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
  (@ROW_ID,@CGID,@LABEL,@DESCRIPTION) SET
ROW_ID = @ROW_ID, CGID = @CGID,
LABEL = IF(@LABEL='', NULL, @LABEL),
DESCRIPTION = IF(@DESCRIPTION='', NULL, @DESCRIPTION);

DROP TABLE IF EXISTS CHARTEVENTS;
CREATE TABLE CHARTEVENTS (         -- rows=263201375 ROW_ID INT UNSIGNED
NOT NULL,
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL,
HADM_ID MEDIUMINT UNSIGNED, ICUSTAY_ID MEDIUMINT UNSIGNED, ITEMID
MEDIUMINT UNSIGNED NOT NULL, CHARTTIME DATETIME NOT NULL, STORETIME
DATETIME,
CGID SMALLINT UNSIGNED, VALUE TEXT, -- max=91 VALUENUM FLOAT,
VALUEUOM VARCHAR(255),             -- max=17 WARNING TINYINT UNSIGNED,
ERROR TINYINT UNSIGNED,
RESULTSTATUS VARCHAR(255),         -- max=6 STOPPED VARCHAR(255),  --
max=8
UNIQUE KEY CHARTEVENTS_ROW_ID (ROW_ID)             -- nvals=263201375
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/CHARTEVENTS.csv' INTO
TABLE
CHARTEVENTS
```

```
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@ITEMID,@CHARTTIME,@STORETIME
,@CGID
,@VALUE,@VALUENUM,@VALUEUOM,@WARNING,@ERROR,@RESULTSTATUS,@STOPPED) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), ICUSTAY_ID =
IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), ITEMID = @ITEMID,
CHARTTIME = @CHARTTIME,
STORETIME = IF(@STORETIME='', NULL, @STORETIME), CGID = IF(@CGID='',
NULL, @CGID),
VALUE = IF(@VALUE='', NULL, @VALUE),
VALUENUM = IF(@VALUENUM='', NULL, @VALUENUM), VALUEUOM =
IF(@VALUEUOM='', NULL, @VALUEUOM), WARNING = IF(@WARNING='', NULL,
@WARNING), ERROR = IF(@ERROR='', NULL, @ERROR),
RESULTSTATUS = IF(@RESULTSTATUS='', NULL, @RESULTSTATUS), STOPPED =
IF(@STOPPED='', NULL, @STOPPED);

DROP TABLE IF EXISTS CPTEVENTS;
CREATE TABLE CPTEVENTS (          -- rows=573146 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED NOT NULL, COSTCENTER VARCHAR(255) NOT NULL,    --
max=4 CHARTDATE DATETIME,
CPT_CD VARCHAR(255) NOT NULL,             -- max=5 CPT_NUMBER MEDIUMINT
UNSIGNED,
CPT_SUFFIX VARCHAR(255), -- max=1 TICKET_ID_SEQ SMALLINT UNSIGNED,
SECTIONHEADER VARCHAR(255),       -- max=25 SUBSECTIONHEADER TEXT,--
max=169 DESCRIPTION VARCHAR(255), -- max=30
UNIQUE KEY CPTEVENTS_ROW_ID (ROW_ID)          -- nvals=573146
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/CPTEVENTS.csv' INTO
TABLE
CPTEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@COSTCENTER,@CHARTDATE,@CPT_CD,@CPT_NUMBE
R,@CPT
_SUFFIX,@TICKET_ID_SEQ,@SECTIONHEADER,@SUBSECTIONHEADER,@DESCRIPTION)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
COSTCENTER = @COSTCENTER,
CHARTDATE = IF(@CHARTDATE='', NULL, @CHARTDATE),
```

```
CPT_CD = @CPT_CD,
CPT_NUMBER = IF(@CPT_NUMBER='', NULL, @CPT_NUMBER), CPT_SUFFIX =
IF(@CPT_SUFFIX='', NULL, @CPT_SUFFIX), TICKET_ID_SEQ =
IF(@TICKET_ID_SEQ='', NULL, @TICKET_ID_SEQ), SECTIONHEADER =
IF(@SECTIONHEADER='', NULL, @SECTIONHEADER),
SUBSECTIONHEADER = IF(@SUBSECTIONHEADER='', NULL, @SUBSECTIONHEADER),
DESCRIPTION = IF(@DESCRIPTION='', NULL, @DESCRIPTION);

DROP TABLE IF EXISTS DATETIMEEVENTS;
CREATE TABLE DATETIMEEVENTS (          -- rows=4486049 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED,
ICUSTAY_ID MEDIUMINT UNSIGNED, ITEMID MEDIUMINT UNSIGNED NOT NULL,
CHARTTIME DATETIME NOT NULL, STORETIME DATETIME NOT NULL,
CGID SMALLINT UNSIGNED NOT NULL, VALUE DATETIME,
VALUEUOM VARCHAR(255) NOT NULL,            -- max=13 WARNING TINYINT
UNSIGNED,
ERROR TINYINT UNSIGNED,
RESULTSTATUS VARCHAR(255),         -- max=0 STOPPED VARCHAR(255),   --
max=8
UNIQUE KEY DATETIMEEVENTS_ROW_ID (ROW_ID) -- nvals=4486049
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/DATETIMEEVENTS.csv' INTO
TABLE
DATETIMEEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@ITEMID,@CHARTTIME,@STORETIME
,@CGID
,@VALUE,@VALUEUOM,@WARNING,@ERROR,@RESULTSTATUS,@STOPPED) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), ICUSTAY_ID =
IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), ITEMID = @ITEMID,
CHARTTIME = @CHARTTIME, STORETIME = @STORETIME, CGID = @CGID,
VALUE = IF(@VALUE='', NULL, @VALUE), VALUEUOM = @VALUEUOM,
WARNING = IF(@WARNING='', NULL, @WARNING), ERROR = IF(@ERROR='', NULL,
@ERROR),
RESULTSTATUS = IF(@RESULTSTATUS='', NULL, @RESULTSTATUS), STOPPED =
IF(@STOPPED='', NULL, @STOPPED);

DROP TABLE IF EXISTS DIAGNOSES_ICD;
CREATE TABLE DIAGNOSES_ICD (           -- rows=651047 ROW_ID MEDIUMINT
UNSIGNED NOT NULL,
```

```
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT UNSIGNED NOT
NULL, SEQ_NUM TINYINT UNSIGNED,
ICD9_CODE VARCHAR(255),            -- max=5
UNIQUE KEY DIAGNOSES_ICD_ROW_ID (ROW_ID)          -- nvals=651047
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/DIAGNOSES_ICD.csv' INTO
TABLE
DIAGNOSES_ICD
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(@ROW_ID,@SUBJECT_ID,@HADM_ID,@SEQ_NUM,@ICD9_CODE) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
SEQ_NUM = IF(@SEQ_NUM='', NULL, @SEQ_NUM), ICD9_CODE =
IF(@ICD9_CODE='', NULL, @ICD9_CODE);

DROP TABLE IF EXISTS DRGCODES;
CREATE TABLE DRGCODES ( -- rows=125557 ROW_ID MEDIUMINT UNSIGNED NOT
NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT
UNSIGNED NOT NULL, DRG_TYPE VARCHAR(255) NOT NULL,    -- max=4 DRG_CODE
VARCHAR(255) NOT NULL,         -- max=4 DESCRIPTION TEXT,    -- max=193
DRG_SEVERITY TINYINT UNSIGNED, DRG_MORTALITY TINYINT UNSIGNED,
UNIQUE KEY DRGCODES_ROW_ID (ROW_ID)           -- nvals=125557
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/DRGCODES.csv' INTO TABLE
DRGCODES
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@DRG_TYPE,@DRG_CODE,@DESCRIPTION,@DRG_SEV
ERITY,
@DRG_MORTALITY) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
DRG_TYPE = @DRG_TYPE, DRG_CODE = @DRG_CODE,
DESCRIPTION = IF(@DESCRIPTION='', NULL, @DESCRIPTION), DRG_SEVERITY =
IF(@DRG_SEVERITY='', NULL, @DRG_SEVERITY), DRG_MORTALITY =
IF(@DRG_MORTALITY='', NULL, @DRG_MORTALITY);

DROP TABLE IF EXISTS D_CPT;
```

```
CREATE TABLE D_CPT (          -- rows=134 ROW_ID TINYINT UNSIGNED NOT
NULL, CATEGORY TINYINT UNSIGNED NOT NULL,
SECTIONRANGE VARCHAR(255) NOT NULL,       -- max=37 SECTIONHEADER
VARCHAR(255) NOT NULL, -- max=25 SUBSECTIONRANGE VARCHAR(255) NOT NULL,
                                          -- max=11 SUBSECTIONHEADER
TEXT NOT NULL,                            -- max=169 CODESUFFIX
VARCHAR(255), -- max=1 MINCODEINSUBSECTION MEDIUMINT UNSIGNED NOT NULL,
MAXCODEINSUBSECTION MEDIUMINT UNSIGNED NOT NULL,
UNIQUE KEY D_CPT_ROW_ID (ROW_ID), -- nvals=134
UNIQUE KEY D_CPT_SUBSECTIONRANGE (SUBSECTIONRANGE),           --
nvals=134
UNIQUE KEY D_CPT_MAXCODEINSUBSECTION (MAXCODEINSUBSECTION)            -
- nvals=134
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/D_CPT.csv' INTO TABLE
D_CPT FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY
'"' LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@CATEGORY,@SECTIONRANGE,@SECTIONHEADER,@SUBSECTIONRANGE,@SUBSE
CTIONH EADER,@CODESUFFIX,@MINCODEINSUBSECTION,@MAXCODEINSUBSECTION)
SET
ROW_ID = @ROW_ID, CATEGORY = @CATEGORY,
SECTIONRANGE = @SECTIONRANGE, SECTIONHEADER = @SECTIONHEADER,
SUBSECTIONRANGE = @SUBSECTIONRANGE, SUBSECTIONHEADER =
@SUBSECTIONHEADER,
CODESUFFIX = IF(@CODESUFFIX='', NULL, @CODESUFFIX), MINCODEINSUBSECTION
= @MINCODEINSUBSECTION, MAXCODEINSUBSECTION = @MAXCODEINSUBSECTION;

DROP TABLE IF EXISTS D_ICD_DIAGNOSES;
CREATE TABLE D_ICD_DIAGNOSES (        -- rows=14567 ROW_ID SMALLINT
UNSIGNED NOT NULL, ICD9_CODE VARCHAR(255) NOT NULL, -- max=5
SHORT_TITLE VARCHAR(255) NOT NULL,    -- max=24 LONG_TITLE TEXT NOT
NULL,                                 -- max=222
UNIQUE KEY D_ICD_DIAGNOSES_ROW_ID (ROW_ID),           -- nvals=14567
UNIQUE KEY D_ICD_DIAGNOSES_ICD9_CODE (ICD9_CODE)          --
nvals=14567
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/D_ICD_DIAGNOSES.csv'
INTO TABLE D_ICD_DIAGNOSES
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(@ROW_ID,@ICD9_CODE,@SHORT_TITLE,@LONG_TITLE) SET
ROW_ID = @ROW_ID, ICD9_CODE = @ICD9_CODE,
```

```
SHORT_TITLE = @SHORT_TITLE, LONG_TITLE = @LONG_TITLE;

DROP TABLE IF EXISTS D_ICD_PROCEDURES; CREATE TABLE D_ICD_PROCEDURES (
-- rows=3882
ROW_ID SMALLINT UNSIGNED NOT NULL,
ICD9_CODE VARCHAR(255) NOT NULL, -- max=4 SHORT_TITLE VARCHAR(255) NOT
NULL,                                -- max=24 LONG_TITLE TEXT NOT
NULL,                                -- max=163
UNIQUE KEY D_ICD_PROCEDURES_ROW_ID (ROW_ID),          -- nvals=3882
UNIQUE KEY D_ICD_PROCEDURES_ICD9_CODE (ICD9_CODE),            --
nvals=3882
UNIQUE KEY D_ICD_PROCEDURES_SHORT_TITLE (SHORT_TITLE) -- nvals=3882
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/D_ICD_PROCEDURES.csv'
INTO TABLE D_ICD_PROCEDURES
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(@ROW_ID,@ICD9_CODE,@SHORT_TITLE,@LONG_TITLE) SET
ROW_ID = @ROW_ID, ICD9_CODE = @ICD9_CODE,
SHORT_TITLE = @SHORT_TITLE, LONG_TITLE = @LONG_TITLE;

DROP TABLE IF EXISTS D_ITEMS;
CREATE TABLE D_ITEMS (-- rows=12478 ROW_ID SMALLINT UNSIGNED NOT NULL,
ITEMID MEDIUMINT UNSIGNED NOT NULL, LABEL TEXT, -- max=95
ABBREVIATION VARCHAR(255),         -- max=50 DBSOURCE VARCHAR(255) NOT
NULL,                              -- max=10 LINKSTO VARCHAR(255), --
max=18 CATEGORY VARCHAR(255),      -- max=27 UNITNAME VARCHAR(255),
                                   -- max=19 PARAM_TYPE VARCHAR(255), --
max=16 CONCEPTID VARCHAR(255),     -- max=0
UNIQUE KEY D_ITEMS_ROW_ID (ROW_ID),           -- nvals=12478
UNIQUE KEY D_ITEMS_ITEMID (ITEMID)            -- nvals=12478
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/D_ITEMS.csv' INTO TABLE
D_ITEMS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@ITEMID,@LABEL,@ABBREVIATION,@DBSOURCE,@LINKSTO,@CATEGORY,@UNI
TNAME,
@PARAM_TYPE,@CONCEPTID) SET
ROW_ID = @ROW_ID,
```

```
ITEMID = @ITEMID,
LABEL = IF(@LABEL='', NULL, @LABEL),
ABBREVIATION = IF(@ABBREVIATION='', NULL, @ABBREVIATION), DBSOURCE =
@DBSOURCE,
LINKSTO = IF(@LINKSTO='', NULL, @LINKSTO), CATEGORY = IF(@CATEGORY='',
NULL, @CATEGORY), UNITNAME = IF(@UNITNAME='', NULL, @UNITNAME),
PARAM_TYPE = IF(@PARAM_TYPE='', NULL, @PARAM_TYPE), CONCEPTID =
IF(@CONCEPTID='', NULL, @CONCEPTID);

DROP TABLE IF EXISTS D_LABITEMS;  CREATE TABLE D_LABITEMS ( -- rows=755
ROW_ID SMALLINT UNSIGNED NOT NULL, ITEMID SMALLINT UNSIGNED NOT NULL,
LABEL VARCHAR(255) NOT NULL, -- max=36 FLUID VARCHAR(255) NOT NULL, --
max=25
CATEGORY VARCHAR(255) NOT NULL,            -- max=10 LOINC_CODE
VARCHAR(255), -- max=7
UNIQUE KEY D_LABITEMS_ROW_ID (ROW_ID),             -- nvals=755
UNIQUE KEY D_LABITEMS_ITEMID (ITEMID) -- nvals=755
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/D_LABITEMS.csv' INTO
TABLE
D_LABITEMS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(@ROW_ID,@ITEMID,@LABEL,@FLUID,@CATEGORY,@LOINC_CODE) SET
ROW_ID = @ROW_ID, ITEMID = @ITEMID, LABEL = @LABEL, FLUID = @FLUID,
CATEGORY = @CATEGORY,
LOINC_CODE = IF(@LOINC_CODE='', NULL, @LOINC_CODE);

DROP TABLE IF EXISTS ICUSTAYS;  CREATE TABLE ICUSTAYS ( -- rows=61532
ROW_ID SMALLINT UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT
NULL, HADM_ID MEDIUMINT UNSIGNED NOT NULL, ICUSTAY_ID MEDIUMINT
UNSIGNED NOT NULL, DBSOURCE VARCHAR(255) NOT NULL,     -- max=10
FIRST_CAREUNIT VARCHAR(255) NOT NULL,              -- max=5 LAST_CAREUNIT
VARCHAR(255) NOT NULL, -- max=5 FIRST_WARDID TINYINT UNSIGNED NOT NULL,
LAST_WARDID TINYINT UNSIGNED NOT NULL,
INTIME DATETIME NOT NULL,
OUTTIME DATETIME,
LOS FLOAT,
UNIQUE KEY ICUSTAYS_ROW_ID (ROW_ID),           -- nvals=61532
UNIQUE KEY ICUSTAYS_ICUSTAY_ID (ICUSTAY_ID)            -- nvals=61532
)
CHARACTER SET = UTF8;
```

```
LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/ICUSTAYS.csv' INTO TABLE
ICUSTAYS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@DBSOURCE,@FIRST_CAREUNIT,@LA
ST_CAR EUNIT,@FIRST_WARDID,@LAST_WARDID,@INTIME,@OUTTIME,@LOS)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ICUSTAY_ID = @ICUSTAY_ID, DBSOURCE = @DBSOURCE,
FIRST_CAREUNIT = @FIRST_CAREUNIT, LAST_CAREUNIT = @LAST_CAREUNIT,
FIRST_WARDID = @FIRST_WARDID, LAST_WARDID = @LAST_WARDID, INTIME =
@INTIME,
OUTTIME = IF(@OUTTIME='', NULL, @OUTTIME), LOS = IF(@LOS='', NULL,
@LOS);

DROP TABLE IF EXISTS INPUTEVENTS_CV;
CREATE TABLE INPUTEVENTS_CV (          -- rows=17528894 ROW_ID INT
UNSIGNED NOT NULL,
SUBJECT_ID SMALLINT UNSIGNED NOT NULL,
HADM_ID MEDIUMINT UNSIGNED, ICUSTAY_ID MEDIUMINT UNSIGNED, CHARTTIME
DATETIME NOT NULL, ITEMID SMALLINT UNSIGNED NOT NULL, AMOUNT FLOAT,
AMOUNTUOM VARCHAR(255),            -- max=3 RATE FLOAT,
RATEUOM VARCHAR(255),              -- max=8 STORETIME DATETIME NOT  NULL,
CGID SMALLINT UNSIGNED,
ORDERID MEDIUMINT UNSIGNED NOT NULL, LINKORDERID MEDIUMINT UNSIGNED NOT
NULL, STOPPED VARCHAR(255),        -- max=8 NEWBOTTLE TINYINT UNSIGNED,
ORIGINALAMOUNT FLOAT,
ORIGINALAMOUNTUOM VARCHAR(255),        -- max=3 ORIGINALROUTE
VARCHAR(255),                          -- max=20 ORIGINALRATE FLOAT,
ORIGINALRATEUOM VARCHAR(255),          -- max=5 ORIGINALSITE
VARCHAR(255),                          -- max=20
UNIQUE KEY INPUTEVENTS_CV_ROW_ID (ROW_ID) -- nvals=17528894
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/INPUTEVENTS_CV.csv' INTO
TABLE
INPUTEVENTS_CV
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
```

```
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@CHARTTIME,@ITEMID,@AMOUNT,@A
MOUNTU
OM,@RATE,@RATEUOM,@STORETIME,@CGID,@ORDERID,@LINKORDERID,@STOPPED,@NEWB
OTTLE,
@ORIGINALAMOUNT,@ORIGINALAMOUNTUOM,@ORIGINALROUTE,@ORIGINALRATE,@ORIGIN
ALRATE UOM,@ORIGINALSITE)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), ICUSTAY_ID =
IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), CHARTTIME = @CHARTTIME,
ITEMID = @ITEMID,
AMOUNT = IF(@AMOUNT='', NULL, @AMOUNT), AMOUNTUOM = IF(@AMOUNTUOM='',
NULL, @AMOUNTUOM), RATE = IF(@RATE='', NULL, @RATE),
RATEUOM = IF(@RATEUOM='', NULL, @RATEUOM), STORETIME = @STORETIME,
CGID = IF(@CGID='', NULL, @CGID), ORDERID = @ORDERID,
LINKORDERID = @LINKORDERID,
STOPPED = IF(@STOPPED='', NULL, @STOPPED), NEWBOTTLE =
IF(@NEWBOTTLE='', NULL, @NEWBOTTLE),
ORIGINALAMOUNT = IF(@ORIGINALAMOUNT='', NULL, @ORIGINALAMOUNT),
ORIGINALAMOUNTUOM = IF(@ORIGINALAMOUNTUOM='', NULL,
@ORIGINALAMOUNTUOM), ORIGINALROUTE = IF(@ORIGINALROUTE='', NULL,
@ORIGINALROUTE), ORIGINALRATE = IF(@ORIGINALRATE='', NULL,
@ORIGINALRATE), ORIGINALRATEUOM = IF(@ORIGINALRATEUOM='', NULL,
@ORIGINALRATEUOM), ORIGINALSITE = IF(@ORIGINALSITE='', NULL,
@ORIGINALSITE);

DROP TABLE IF EXISTS INPUTEVENTS_MV;
CREATE TABLE INPUTEVENTS_MV (          -- rows=3618991 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED NOT NULL, ICUSTAY_ID MEDIUMINT UNSIGNED,
STARTTIME DATETIME NOT NULL,
ENDTIME DATETIME NOT NULL,
ITEMID MEDIUMINT UNSIGNED NOT NULL,
AMOUNT FLOAT NOT NULL,
AMOUNTUOM VARCHAR(255) NOT NULL, -- max=19 RATE FLOAT,
RATEUOM VARCHAR(255),              -- max=12 STORETIME DATETIME NOT
NULL,
CGID SMALLINT UNSIGNED NOT NULL,
ORDERID MEDIUMINT UNSIGNED NOT NULL,
LINKORDERID MEDIUMINT UNSIGNED NOT NULL, ORDERCATEGORYNAME VARCHAR(255)
NOT NULL, -- max=24 SECONDARYORDERCATEGORYNAME VARCHAR(255), -- max=24
ORDERCOMPONENTTYPEDESCRIPTION VARCHAR(255) NOT NULL, -- max=57
ORDERCATEGORYDESCRIPTION VARCHAR(255) NOT NULL,            -- max=14
PATIENTWEIGHT FLOAT NOT NULL,
TOTALAMOUNT FLOAT,
TOTALAMOUNTUOM VARCHAR(255), -- max=2 ISOPENBAG TINYINT UNSIGNED NOT
NULL,
```

```
CONTINUEINNEXTDEPT TINYINT UNSIGNED NOT NULL, CANCELREASON TINYINT
UNSIGNED NOT NULL, STATUSDESCRIPTION VARCHAR(255) NOT NULL, -- max=15
COMMENTS_EDITEDBY VARCHAR(255),           -- max=15 COMMENTS_CANCELEDBY
VARCHAR(255),                             -- max=15 COMMENTS_DATE
DATETIME,
ORIGINALAMOUNT FLOAT NOT NULL,
ORIGINALRATE FLOAT NOT NULL,
UNIQUE KEY INPUTEVENTS_MV_ROW_ID (ROW_ID) -- nvals=3618991
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/INPUTEVENTS_MV.csv' INTO
TABLE
INPUTEVENTS_MV
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@STARTTIME,@ENDTIME,@ITEMID,@
AMOUNT
,@AMOUNTUOM,@RATE,@RATEUOM,@STORETIME,@CGID,@ORDERID,@LINKORDERID,@ORDE
RCATEG
ORYNAME,@SECONDARYORDERCATEGORYNAME,@ORDERCOMPONENTTYPEDESCRIPTION,@ORD
ERCATE
GORYDESCRIPTION,@PATIENTWEIGHT,@TOTALAMOUNT,@TOTALAMOUNTUOM,@ISOPENBAG,
@CONTI
NUEINNEXTDEPT,@CANCELREASON,@STATUSDESCRIPTION,@COMMENTS_EDITEDBY,@COMM
ENTS_C ANCELEDBY,@COMMENTS_DATE,@ORIGINALAMOUNT,@ORIGINALRATE)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ICUSTAY_ID = IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), STARTTIME =
@STARTTIME,
ENDTIME = @ENDTIME, ITEMID = @ITEMID, AMOUNT = @AMOUNT, AMOUNTUOM =
@AMOUNTUOM,
RATE = IF(@RATE='', NULL, @RATE),
RATEUOM = IF(@RATEUOM='', NULL, @RATEUOM), STORETIME = @STORETIME,
CGID = @CGID, ORDERID = @ORDERID,
LINKORDERID = @LINKORDERID, ORDERCATEGORYNAME = @ORDERCATEGORYNAME,
SECONDARYORDERCATEGORYNAME = IF(@SECONDARYORDERCATEGORYNAME='', NULL,
@SECONDARYORDERCATEGORYNAME),
ORDERCOMPONENTTYPEDESCRIPTION = @ORDERCOMPONENTTYPEDESCRIPTION,
ORDERCATEGORYDESCRIPTION = @ORDERCATEGORYDESCRIPTION, PATIENTWEIGHT =
@PATIENTWEIGHT,
TOTALAMOUNT = IF(@TOTALAMOUNT='', NULL, @TOTALAMOUNT), TOTALAMOUNTUOM =
IF(@TOTALAMOUNTUOM='', NULL, @TOTALAMOUNTUOM), ISOPENBAG = @ISOPENBAG,
CONTINUEINNEXTDEPT = @CONTINUEINNEXTDEPT, CANCELREASON = @CANCELREASON,
STATUSDESCRIPTION = @STATUSDESCRIPTION,
COMMENTS_EDITEDBY = IF(@COMMENTS_EDITEDBY='', NULL,
@COMMENTS_EDITEDBY), COMMENTS_CANCELEDBY = IF(@COMMENTS_CANCELEDBY='',
NULL,
@COMMENTS_CANCELEDBY),
```

```
COMMENTS_DATE = IF(@COMMENTS_DATE='', NULL, @COMMENTS_DATE),
ORIGINALAMOUNT = @ORIGINALAMOUNT,
ORIGINALRATE = @ORIGINALRATE;

DROP TABLE IF EXISTS LABEVENTS;
CREATE TABLE LABEVENTS (          -- rows=27872575 ROW_ID INT UNSIGNED
NOT NULL,
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL,
HADM_ID MEDIUMINT UNSIGNED,
ITEMID SMALLINT UNSIGNED NOT NULL,
CHARTTIME DATETIME NOT NULL, VALUE TEXT,  -- max=100 VALUENUM FLOAT,
VALUEUOM VARCHAR(255),         -- max=10 FLAG VARCHAR(255),  -- max=8
UNIQUE KEY LABEVENTS_ROW_ID (ROW_ID)           -- nvals=27872575
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/LABEVENTS.csv' INTO
TABLE
LABEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ITEMID,@CHARTTIME,@VALUE,@VALUENUM,@VALU
EUOM,@ FLAG)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), ITEMID = @ITEMID,
CHARTTIME = @CHARTTIME,
VALUE = IF(@VALUE='', NULL, @VALUE),
VALUENUM    =    IF(@VALUENUM='',    NULL,    @VALUENUM),    VALUEUOM    =
IF(@VALUEUOM='', NULL, @VALUEUOM), FLAG = IF(@FLAG='', NULL, @FLAG);

DROP TABLE IF EXISTS MICROBIOLOGYEVENTS;
CREATE TABLE MICROBIOLOGYEVENTS (          -- rows=328446 ROW_ID
MEDIUMINT UNSIGNED NOT NULL,
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL,
HADM_ID MEDIUMINT UNSIGNED, CHARTDATE DATETIME NOT NULL, CHARTTIME
DATETIME,
SPEC_ITEMID MEDIUMINT UNSIGNED,
SPEC_TYPE_DESC VARCHAR(255) NOT NULL,              -- max=56 ORG_ITEMID
MEDIUMINT UNSIGNED,
ORG_NAME VARCHAR(255),             -- max=70 ISOLATE_NUM TINYINT
UNSIGNED, AB_ITEMID MEDIUMINT UNSIGNED, AB_NAME VARCHAR(255),     --
max=20
DILUTION_TEXT VARCHAR(255),           -- max=6 DILUTION_COMPARISON
VARCHAR(255),                         -- max=2 DILUTION_VALUE SMALLINT
UNSIGNED,
```

```
INTERPRETATION VARCHAR(255), -- max=1
UNIQUE KEY MICROBIOLOGYEVENTS_ROW_ID (ROW_ID) -- nvals=328446
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/MICROBIOLOGYEVENTS.csv'
INTO TABLE MICROBIOLOGYEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@CHARTDATE,@CHARTTIME,@SPEC_ITEMID,@SPEC_
TYPE_D
ESC,@ORG_ITEMID,@ORG_NAME,@ISOLATE_NUM,@AB_ITEMID,@AB_NAME,@DILUTION_TE
XT,@DI LUTION_COMPARISON,@DILUTION_VALUE,@INTERPRETATION)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), CHARTDATE = @CHARTDATE,
CHARTTIME = IF(@CHARTTIME='', NULL, @CHARTTIME), SPEC_ITEMID =
IF(@SPEC_ITEMID='', NULL, @SPEC_ITEMID), SPEC_TYPE_DESC =
@SPEC_TYPE_DESC,
ORG_ITEMID = IF(@ORG_ITEMID='', NULL, @ORG_ITEMID), ORG_NAME =
IF(@ORG_NAME='', NULL, @ORG_NAME), ISOLATE_NUM = IF(@ISOLATE_NUM='',
NULL, @ISOLATE_NUM), AB_ITEMID = IF(@AB_ITEMID='', NULL, @AB_ITEMID),
AB_NAME = IF(@AB_NAME='', NULL, @AB_NAME),
DILUTION_TEXT = IF(@DILUTION_TEXT='', NULL, @DILUTION_TEXT),
DILUTION_COMPARISON = IF(@DILUTION_COMPARISON='', NULL,
@DILUTION_COMPARISON),
DILUTION_VALUE = IF(@DILUTION_VALUE='', NULL, @DILUTION_VALUE),
INTERPRETATION = IF(@INTERPRETATION='', NULL, @INTERPRETATION);

DROP TABLE IF EXISTS NOTEEVENTS;
CREATE TABLE NOTEEVENTS (          -- rows=2078705 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED,
CHARTDATE DATE NOT NULL, CHARTTIME DATETIME, STORETIME DATETIME,
CATEGORY VARCHAR(255) NOT NULL,            -- max=17 DESCRIPTION
VARCHAR(255) NOT NULL,                     -- max=80 CGID SMALLINT
UNSIGNED,
ISERROR TINYINT UNSIGNED, TEXT MEDIUMTEXT, -- max=55725
UNIQUE KEY NOTEEVENTS_ROW_ID (ROW_ID) -- nvals=2078705
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/NOTEEVENTS.csv' INTO
TABLE
NOTEEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
```

```
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@CHARTDATE,@CHARTTIME,@STORETIME,@CATEGOR
Y,@DES CRIPTION,@CGID,@ISERROR,@TEXT)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), CHARTDATE = @CHARTDATE,
CHARTTIME = IF(@CHARTTIME='', NULL, @CHARTTIME), STORETIME =
IF(@STORETIME='', NULL, @STORETIME), CATEGORY = @CATEGORY,
DESCRIPTION = @DESCRIPTION,
CGID = IF(@CGID='', NULL, @CGID),
ISERROR = IF(@ISERROR='', NULL, @ISERROR), TEXT = IF(@TEXT='', NULL,
@TEXT);

DROP TABLE IF EXISTS OUTPUTEVENTS;
CREATE TABLE OUTPUTEVENTS ( -- rows=4349339 ROW_ID MEDIUMINT UNSIGNED
NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT
UNSIGNED,
ICUSTAY_ID MEDIUMINT UNSIGNED, CHARTTIME DATETIME NOT NULL,
ITEMID MEDIUMINT UNSIGNED NOT NULL, VALUE FLOAT,
VALUEUOM VARCHAR(255),            -- max=2 STORETIME DATETIME NOT  NULL,
CGID SMALLINT UNSIGNED NOT NULL, STOPPED VARCHAR(255),-- max=0 NEWBOTTLE
VARCHAR(255),                     -- max=0 ISERROR VARCHAR(255),  --
max=0
UNIQUE KEY OUTPUTEVENTS_ROW_ID (ROW_ID)           -- nvals=4349339
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/OUTPUTEVENTS.csv' INTO
TABLE
OUTPUTEVENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@CHARTTIME,@ITEMID,@VALUE,@VA
LUEUOM
,@STORETIME,@CGID,@STOPPED,@NEWBOTTLE,@ISERROR) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID,
HADM_ID = IF(@HADM_ID='', NULL, @HADM_ID), ICUSTAY_ID =
IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), CHARTTIME = @CHARTTIME,
ITEMID = @ITEMID,
VALUE = IF(@VALUE='', NULL, @VALUE),
VALUEUOM = IF(@VALUEUOM='', NULL, @VALUEUOM), STORETIME = @STORETIME,
CGID = @CGID,
```

```
STOPPED = IF(@STOPPED='', NULL, @STOPPED), NEWBOTTLE =
IF(@NEWBOTTLE='', NULL, @NEWBOTTLE), ISERROR = IF(@ISERROR='', NULL,
@ISERROR);

DROP TABLE IF EXISTS PATIENTS;  CREATE TABLE PATIENTS ( -- rows=46520
ROW_ID SMALLINT UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT
NULL, GENDER VARCHAR(255) NOT NULL,    --  max=1 DOB VARCHAR(255) NOT
NULL,                                  -- max=19 DOD DATETIME,
DOD_HOSP DATETIME,
DOD_SSN DATETIME,
EXPIRE_FLAG TINYINT UNSIGNED NOT NULL,
UNIQUE KEY PATIENTS_ROW_ID (ROW_ID),           -- nvals=46520
UNIQUE KEY PATIENTS_SUBJECT_ID (SUBJECT_ID)            -- nvals=46520
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/PATIENTS.csv' INTO TABLE
PATIENTS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(@ROW_ID,@SUBJECT_ID,@GENDER,@DOB,@DOD,@DOD_HOSP,@DOD_SSN,@EXPIRE_FLAG)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, GENDER = @GENDER,
DOB = @DOB,
DOD = IF(@DOD='', NULL, @DOD),
DOD_HOSP = IF(@DOD_HOSP='', NULL, @DOD_HOSP), DOD_SSN = IF(@DOD_SSN='',
NULL, @DOD_SSN), EXPIRE_FLAG = @EXPIRE_FLAG;

DROP TABLE IF EXISTS PRESCRIPTIONS;
CREATE TABLE PRESCRIPTIONS (          -- rows=4156848 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED NOT NULL, ICUSTAY_ID MEDIUMINT UNSIGNED,
STARTDATE DATETIME,
ENDDATE DATETIME,
DRUG_TYPE VARCHAR(255) NOT NULL, -- max=8 DRUG VARCHAR(255),-- max=58
DRUG_NAME_POE VARCHAR(255),    -- max=58 DRUG_NAME_GENERIC VARCHAR(255),
                      -- max=49 FORMULARY_DRUG_CD VARCHAR(255),    --
max=17 GSN TEXT,      -- max=125
NDC VARCHAR(255),             -- max=11 PROD_STRENGTH VARCHAR(255),
                              -- max=60 DOSE_VAL_RX VARCHAR(255), --
max=26 DOSE_UNIT_RX VARCHAR(255),      -- max=32 FORM_VAL_DISP
VARCHAR(255),                 -- max=47 FORM_UNIT_DISP VARCHAR(255), --
max=13
```

```
ROUTE VARCHAR(255),            -- max=28
UNIQUE KEY PRESCRIPTIONS_ROW_ID (ROW_ID)          -- nvals=4156848
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/PRESCRIPTIONS.csv' INTO
TABLE
PRESCRIPTIONS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@STARTDATE,@ENDDATE,@DRUG_TYP
E,@DRU
G,@DRUG_NAME_POE,@DRUG_NAME_GENERIC,@FORMULARY_DRUG_CD,@GSN,@NDC,@PROD_
STRENG
TH,@DOSE_VAL_RX,@DOSE_UNIT_RX,@FORM_VAL_DISP,@FORM_UNIT_DISP,@ROUTE)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ICUSTAY_ID = IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), STARTDATE =
IF(@STARTDATE='', NULL, @STARTDATE), ENDDATE = IF(@ENDDATE='', NULL,
@ENDDATE), DRUG_TYPE = @DRUG_TYPE,
DRUG = IF(@DRUG='', NULL, @DRUG),
DRUG_NAME_POE = IF(@DRUG_NAME_POE='', NULL, @DRUG_NAME_POE),
DRUG_NAME_GENERIC = IF(@DRUG_NAME_GENERIC='', NULL,
@DRUG_NAME_GENERIC), FORMULARY_DRUG_CD = IF(@FORMULARY_DRUG_CD='',
NULL, @FORMULARY_DRUG_CD), GSN = IF(@GSN='', NULL, @GSN),
NDC = IF(@NDC='', NULL, @NDC),
PROD_STRENGTH = IF(@PROD_STRENGTH='', NULL, @PROD_STRENGTH),
DOSE_VAL_RX = IF(@DOSE_VAL_RX='', NULL, @DOSE_VAL_RX), DOSE_UNIT_RX =
IF(@DOSE_UNIT_RX='', NULL, @DOSE_UNIT_RX), FORM_VAL_DISP =
IF(@FORM_VAL_DISP='', NULL, @FORM_VAL_DISP), FORM_UNIT_DISP =
IF(@FORM_UNIT_DISP='', NULL, @FORM_UNIT_DISP), ROUTE = IF(@ROUTE='',
NULL, @ROUTE);

DROP TABLE IF EXISTS PROCEDUREEVENTS_MV;
CREATE TABLE PROCEDUREEVENTS_MV (          -- rows=258066 ROW_ID
MEDIUMINT UNSIGNED NOT NULL,
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT UNSIGNED NOT
NULL, ICUSTAY_ID MEDIUMINT UNSIGNED, STARTTIME DATETIME NOT NULL,
ENDTIME DATETIME NOT NULL,
ITEMID MEDIUMINT UNSIGNED NOT NULL, VALUE FLOAT NOT NULL,
VALUEUOM VARCHAR(255) NOT NULL,    -- max=4 LOCATION VARCHAR(255), --
max=24 LOCATIONCATEGORY VARCHAR(255),      -- max=19 STORETIME DATETIME
NOT NULL,
CGID SMALLINT UNSIGNED NOT NULL,
ORDERID MEDIUMINT UNSIGNED NOT NULL,
LINKORDERID MEDIUMINT UNSIGNED NOT NULL, ORDERCATEGORYNAME VARCHAR(255)
NOT NULL, -- max=21 SECONDARYORDERCATEGORYNAME VARCHAR(255), -- max=0
ORDERCATEGORYDESCRIPTION VARCHAR(255) NOT NULL,            -- max=12
```

```
ISOPENBAG TINYINT UNSIGNED NOT NULL, CONTINUEINNEXTDEPT TINYINT
UNSIGNED NOT NULL, CANCELREASON TINYINT UNSIGNED NOT NULL,
STATUSDESCRIPTION VARCHAR(255) NOT NULL, -- max=15 COMMENTS_EDITEDBY
VARCHAR(255),                             -- max=7 COMMENTS_CANCELEDBY
VARCHAR(255),                             -- max=17 COMMENTS_DATE
DATETIME,
UNIQUE KEY PROCEDUREEVENTS_MV_ROW_ID (ROW_ID),            --
nvals=258066
UNIQUE KEY PROCEDUREEVENTS_MV_ORDERID (ORDERID)           --
nvals=258066
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/PROCEDUREEVENTS_MV.csv'
INTO TABLE PROCEDUREEVENTS_MV
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@STARTTIME,@ENDTIME,@ITEMID,@
VALUE,
@VALUEUOM,@LOCATION,@LOCATIONCATEGORY,@STORETIME,@CGID,@ORDERID,@LINKOR
DERID,
@ORDERCATEGORYNAME,@SECONDARYORDERCATEGORYNAME,@ORDERCATEGORYDESCRIPTIO
N,@ISO
PENBAG,@CONTINUEINNEXTDEPT,@CANCELREASON,@STATUSDESCRIPTION,@COMMENTS_E
DITEDB Y,@COMMENTS_CANCELEDBY,@COMMENTS_DATE)
SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ICUSTAY_ID = IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), STARTTIME =
@STARTTIME,
ENDTIME = @ENDTIME, ITEMID = @ITEMID, VALUE = @VALUE, VALUEUOM =
@VALUEUOM,
LOCATION = IF(@LOCATION='', NULL, @LOCATION),
LOCATIONCATEGORY = IF(@LOCATIONCATEGORY='', NULL, @LOCATIONCATEGORY),
STORETIME = @STORETIME,
CGID = @CGID, ORDERID = @ORDERID,
LINKORDERID = @LINKORDERID, ORDERCATEGORYNAME = @ORDERCATEGORYNAME,
SECONDARYORDERCATEGORYNAME = IF(@SECONDARYORDERCATEGORYNAME='', NULL,
@SECONDARYORDERCATEGORYNAME),
ORDERCATEGORYDESCRIPTION = @ORDERCATEGORYDESCRIPTION, ISOPENBAG =
@ISOPENBAG,
CONTINUEINNEXTDEPT = @CONTINUEINNEXTDEPT, CANCELREASON = @CANCELREASON,
STATUSDESCRIPTION = @STATUSDESCRIPTION,
COMMENTS_EDITEDBY = IF(@COMMENTS_EDITEDBY='', NULL,
@COMMENTS_EDITEDBY), COMMENTS_CANCELEDBY = IF(@COMMENTS_CANCELEDBY='',
NULL,
@COMMENTS_CANCELEDBY),
COMMENTS_DATE = IF(@COMMENTS_DATE='', NULL, @COMMENTS_DATE);

DROP TABLE IF EXISTS PROCEDURES_ICD;
CREATE TABLE PROCEDURES_ICD (          -- rows=240095 ROW_ID MEDIUMINT
UNSIGNED NOT NULL,
```

```
SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID MEDIUMINT UNSIGNED NOT
NULL, SEQ_NUM TINYINT UNSIGNED NOT NULL, ICD9_CODE VARCHAR(255) NOT
NULL, -- max=4
UNIQUE KEY PROCEDURES_ICD_ROW_ID (ROW_ID) -- nvals=240095
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/PROCEDURES_ICD.csv' INTO
TABLE
PROCEDURES_ICD
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
  (@ROW_ID,@SUBJECT_ID,@HADM_ID,@SEQ_NUM,@ICD9_CODE) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID, SEQ_NUM
= @SEQ_NUM, ICD9_CODE = @ICD9_CODE;

DROP TABLE IF EXISTS SERVICES;  CREATE TABLE SERVICES ( -- rows=73343
ROW_ID MEDIUMINT UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT
NULL, HADM_ID MEDIUMINT UNSIGNED NOT NULL, TRANSFERTIME DATETIME NOT
NULL, PREV_SERVICE VARCHAR(255),       -- max=5
CURR_SERVICE VARCHAR(255) NOT NULL,            -- max=5
UNIQUE KEY SERVICES_ROW_ID (ROW_ID)            -- nvals=73343
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/SERVICES.csv' INTO TABLE
SERVICES
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
  (@ROW_ID,@SUBJECT_ID,@HADM_ID,@TRANSFERTIME,@PREV_SERVICE,@CURR_SERVI
CE) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
TRANSFERTIME = @TRANSFERTIME,
PREV_SERVICE = IF(@PREV_SERVICE='', NULL, @PREV_SERVICE), CURR_SERVICE
= @CURR_SERVICE;

DROP TABLE IF EXISTS TRANSFERS;
CREATE TABLE TRANSFERS (           -- rows=261897 ROW_ID MEDIUMINT
UNSIGNED NOT NULL, SUBJECT_ID MEDIUMINT UNSIGNED NOT NULL, HADM_ID
MEDIUMINT UNSIGNED NOT NULL, ICUSTAY_ID MEDIUMINT UNSIGNED,
DBSOURCE VARCHAR(255),             -- max=10
```

```
EVENTTYPE VARCHAR(255),            -- max=9 PREV_CAREUNIT VARCHAR(255),
                                   -- max=5 CURR_CAREUNIT VARCHAR(255),
                                   -- max=5 PREV_WARDID TINYINT UNSIGNED,
CURR_WARDID TINYINT UNSIGNED,
INTIME DATETIME, OUTTIME DATETIME, LOS FLOAT,
UNIQUE KEY TRANSFERS_ROW_ID (ROW_ID)          -- nvals=261897
)
CHARACTER SET = UTF8;

LOAD DATA LOCAL INFILE
'C:/Users/dmbru/Desktop/PhD/Dissertation/MIMIC/TRANSFERS.csv' INTO
TABLE
TRANSFERS
FIELDS TERMINATED BY ',' ESCAPED BY '\\' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES

(@ROW_ID,@SUBJECT_ID,@HADM_ID,@ICUSTAY_ID,@DBSOURCE,@EVENTTYPE,@PREV_CA
REUNIT
,@CURR_CAREUNIT,@PREV_WARDID,@CURR_WARDID,@INTIME,@OUTTIME,@LOS) SET
ROW_ID = @ROW_ID, SUBJECT_ID = @SUBJECT_ID, HADM_ID = @HADM_ID,
ICUSTAY_ID = IF(@ICUSTAY_ID='', NULL, @ICUSTAY_ID), DBSOURCE =
IF(@DBSOURCE='', NULL, @DBSOURCE), EVENTTYPE = IF(@EVENTTYPE='', NULL,
@EVENTTYPE),
PREV_CAREUNIT = IF(@PREV_CAREUNIT='', NULL, @PREV_CAREUNIT),
CURR_CAREUNIT = IF(@CURR_CAREUNIT='', NULL, @CURR_CAREUNIT),
PREV_WARDID = IF(@PREV_WARDID='', NULL, @PREV_WARDID), CURR_WARDID =
IF(@CURR_WARDID='', NULL, @CURR_WARDID), INTIME = IF(@INTIME='', NULL,
@INTIME),
OUTTIME = IF(@OUTTIME='', NULL, @OUTTIME), LOS = IF(@LOS='', NULL,
@LOS)
```

Appendix C – Stop Words

i	while	hers	until
me	of	herself	should
my	at	it	now
myself	by	its	
we	for	itself	
our	with	they	
ours	about	them	
ourselves	against	their	
you	between	theirs	
your	into	themselves	
yours	through	what	
yourself	during	which	
yourselves	before	who	
he	after	whom	
him	above	this	
		that	
his	below	these	
himself	to	those	
she	from	am	
her	up	is	
as	don	because	
down	are	each	
in	was	few	
out	were	more	
on	be	most	
off	been	other	
over	being	some	
under	have	such	
again	has	no	
further	had	nor	
then	having	not	
once	do	only	
here	does	own	
there	did	same	
when	doing	so	
where	a	than	
why	an	too	
how	the	very	
all	and	s	
any	but	t	
both	if	can	
just	or	will	

Appendix D – Software Version

Machine Learning Environment: Amazon Linux AMI 2018.03.0 (HVM), SSD Volume Typem4.16xlarge – 96 cores 364gig memory, Scikit-Learn, NLTK, Gensim, Plotly, Flask

Data Analysis Environment: Microsoft Windows 10 64-bit Microsoft Excel 2016, MYSQL 6.2

Documentation Environment: Microsoft Windows 10 64-bit Build, Microsoft Word

2016, Zotero Bibliography manager

Appendix E – Setup Virtual Environment

Steps for EC2 Setup

1. Spin up new EC2 instance of Amazon Linux AMI 2018.03.0 (HVM), SSD Volume Typem4.16xlarge
– 96 cores 364gig memory
2. Use current MIMIC PEM Key

3. Enable port 8888 for inbound traffic
4. Use Putty to SSH into VM
5. Download Anaconda 3 installer by typing this command:
a. wget https://repo.continuum.io/archive/Anaconda3-4.4.0-Linux-x86_64.sh
6. Install Anaconda3 by type:
a. bash Anaconda3-4.4.0-Linux-x86_64.sh
7. At the end you'll be prompted to include Anaconda3 into your .bashrc PATH. Make sure to type "yes"
8. Set Anaconda3 as your default Python environment.
a. which python /usr/bin/python
b. source .bashrc
9. Setup Password for Jupyter
a. Type python in command
b. from IPython.lib import passwd
c. passwd()
d. set password and copy and paste SHA has in script below
10. Configure Jupyter server – quit() python
a. jupyter notebook --generate-config
b. mkdir certs
c. cd certs
d. sudo openssl req -x509 -nodes -days 365 -newkey rsa:1024 -keyout mycert.pem -out mycert.pem
e. enter personal information for PEM
f. cd back to main directory
11. User VIM to edit config file
a. vim .jupyter/jupyter_notebook_config.py
b. Hit "I" to enter insert mode and copy and paste in info in config c = get_config()
# Kernel config
c.IPKernelApp.pylab = 'inline' # if you want plotting support always in your notebook

# Notebook config
c.NotebookApp.certfile = u'/home/ec2-user/certs/mycert.pem' #location of your certificate file
c.NotebookApp.ip = '*'
c.NotebookApp.open_browser = False #so that the ipython notebook does not open up a browser by default
c.NotebookApp.password = u'Hashcode; #edit this with the SHA hash that you generated after typing in Step 9
# This is the port we opened in Step 3. c.NotebookApp.port = 8888

1. Once you got that in, hit ESC and type ":wq" to save and quit out of vim.
2. Create folder for notebooks and starting Jupyter

a. mkdir Notebooks
b. cd Notebooks
c. jupyter notebook
3. Access Jupyter Notebooks from your browser. To get there, you'll need your Public DNS (IPv4)
a. https://ec2-54-144-47-199.compute-1.amazonaws.com:8888/
4. Conda Install genism
5. Conda Install boto3
6. Upload notebook using filezilla
7. call from S3 to pull NOTEEVENTS
8. run notebook and pull model from server with filezilla

Appendix F – Beth Israel Deaconess Medical – Approved Abbreviations

Abbreviation Term
ā before
Al aortic first sound
A2 aortic second sound
AAA abdominal aortic aneurysm
AAE active assistance exercise
AAL anterior axillary line
AAROM active assistive range of motion
Ab antibody
A/B apnea & bradycardia
ABC airway, breathing and circulation
ABCDE asymmetry, border irregularity, color variation, diameter, evolution (Dermatology)
abd abdomen, abdominal
ABE acute bacterial endocarditis
ABG arterial blood gas
ABLB alternate binaural loudness balance (test)
abn abnormality(ies)
ABO blood group system (type)
ABP arterial blood pressure
ABW actual body weight
ABx antibiotics
A1C glycosylated hemoglobin A1C
AC acromioclavicular
a.c. before meals
A/C anterior chamber of the eye
ACA anterior cerebral artery
ACBE air contrast barium enema

ACE antigrade colonic enema
ACIOL anterior chamber intraocular lens
ACL anterior cruciate ligament (knee)
ACLS advanced cardiac (cardiopulmonary) life support
ACP acid phosphatase
ACS acute coronary syndrome
ACT activated clotting time
ACTA automated computerized transverse axial tomography
ACTH adrenal corticotropic hormone
ACU Ambulatory Care Unit
ACV assist control ventilation
ADA American Diabetes Association
ADA American Dental Association
ADD Attention-Deficit Disorder
ADE adverse drug event
ADH antidiuretic hormone
ADHD attention-deficit hyperactivity disorder
ADL activities of daily living
ad lib as desired
adm admission
ADR adverse drug reaction
AE above elbow (amputation)
AEB as evidenced by
AED automated external defibrillator
AED antiepileptic drug
AER auditory (acoustic) evoked response
AFB acid fast bacilli
AF/FL atrial fibrillation / flutter
AFI amniotic fluid index
A fib atrial fibrillation
AFO ankle fixation orthotic
AFP alpha fetoprotein
Ag antigen
A/G albumin to globulin ratio
AGA appropriate for gestational age
AGN acute glomerulonephritis
AHG antihemophilic globulin
AHHD arteriosclerotic hypertensive heart disease
AHI apnea-hypopnea index
AI aortic insufficiency
AICD automatic implantable cardioverter/defibrillator
AID artificial insemination by donor
AIDS acquired immune deficiency syndrome
AIH artificial insemination by husband

AIN anterior interosseous nerve
AIVR accelerated idioventricular rhythm
AJ ankle jerk
AK actinic keratosis
A/K above the knee
AKA above the knee amputation
AKI acute kidney injury
ALB albumin
A-line arterial catheter
ALK alkaline
Alk-Phos alkaline phosphatase
ALL acute lymphocytic leukemia
ALMI anterolateral myocardial infarction (location)
ALS amyotrophic lateral sclerosis
ALT alanine transaminase (serum glutamate pyruvate)
a.m. morning
AMA against medical advice
amb ambulate, ambulatory, ambulation
AMI acute myocardial infarction
AML acute myelogenous leukemia
AMML acute myelomonocytic leukemia
amp ampul
amp amplitude (unique to ventilator - High Frequency Ossicilator [HFO]
amnio amniocentesis
amt amount
AMY amylase
ANA antinuclear antibody
ANC absolute neutrophil count
anes anesthesia
ANF antinuclear factor
angio angiogram
ANLL acute nonlymphoblastic leukemia
ANS autonomic nervous system
ant anterior
ante before
anti-HBc antibody to Hepatitis B core antigen
anti-HBs antibody to Hepatitis B surface antigen
anti-T. cruzi antibody to Trypanosoma cruzi
A/O alert and oriented
Ao aorta
AOCKD acute-on-chronic kidney disease
AOD arterial occlusive disease
AODM adult onset diabetes mellitus
AOM acute otitis media

AP anteroposterior (x-ray)
A&P assessment and plans
APB abductor pollicus brevis
APD afferent pupillary defect
APD anteroposterior diameter
APG Affiliated Physicians, Inc.
APGAR appearance, pulse, grimace, activity and respiration
APL abductor pollicis longus
AP&L anteroposterior and lateral
appt. appointment
aPPT activated partial thromboplastin time
APPY appendectomy
APR abdominal perineal resection
AQ accomplishment quotient
AR aortic regurgitation
A-R apical radial (pulses)
ARBF awaiting return of bowel function
ARC AIDS related complex
ARDS adult respiratory distress syndrome
ARDSnet Acute Respiratory Distress Syndrome Network
AREDS age-related eye disease study
ARF acute renal failure
ARM artificial rupture of membranes
ARMD age related macular degeneration
AROM active range of motion
ARROM active resistive range of motion
art arterial
AS aortic stenosis
ASA American Society of Anesthesiologists
ASCVD arteriosclerotic cardiovascular disease
ASD atrial septal defect
ASHD arteriosclerotic heart disease
ASIS anterior superior iliac spine
ASMI anteroseptal myocardial infarction (location)
ASO antistreptolysin-O titer
ASPVD arteriosclerotic peripheral vascular disease
AST aspartate transaminase
astig astigmatism
as tol as tolerated
AT atrial tachycardia
ATFL anterior talofibular ligament
ATN acute tubular necrosis
ATNB auriculo temporal nerve block
ATP adenosine triphosphate

ATPS ambient temperature / pressure, saturated w/ water vapor
ATV all-terrain vehicle
aud auditory
aud comp auditory comprehension
aus auscultation
AV arterio-venous (or) arterioventricular
AVB atrioventricular block
1° AVB first degree atrioventricular block
AVF arteriovenous fistula
AVM arteriovenous malformation
avoc avocation
AVR aortic valve replacement
AVS artiovenous shunt
AVSS afebrile, vital signs stable
A&W alive and well
ax axillary
axB axillary block
A-Z test Aschheim-Zondek test (pregnancy test)
B black
Ba barium
Bab Babinski
bact bacteria
BAERs brainstem auditory evoked responses
bal balance
band band neutrpohil (stab)
baso basophil
BaS barium swallow
BBB bundle branch block
BBBB bilateral bundle branch block
BBS bilateral breath sounds
BC bone conduction
B&C bed and chair
BC/BS Blue Cross/Blue Shield
BCC basal cell carcinoma
BCE basal cell epithelioma
BCM birth control method
BCP birth control pills
BE below elbow
BEAM brain electrical activity mapping
BEE basal energy expenditure
BEI butonal extractable iodine
BF breast feeding
bHCG beta human chorionic gonadotropin (pregnancy test)
BIBA brought in by ambulance

b.i.d. twice a day
BIDMC Beth Israel Deaconess Medical Center
BIDPO Beth Israel Deaconess Physicians Organization
bil bilateral
BiPAP bilevel positive airway pressure
BiPD biparietal diameter
BI-RADS Breast Imaging Reporting & Data System (American College of Radiology)
BJ biceps jerk
BJM bones, joints, muscles
BK below the knee
BKA below the knee amputation
bl cult blood culture
BLE both lower extremities
BLS basic life support
BM bowel movement
BMH bone marrow harvest
BMI body mass index
BMJ bones, muscles, joints
BMR basal metabolic rate
BMT bone marrow transplant
BNP brain natriuretic peptide
BOM bilateral otitis media
BOS base of support
BP blood pressure
BPC bronchoprovocation
BPD bronchopulmonary dysplasia
BPH benign prostatic hypertrophy
BPM beats per minute
BPP biophysical profile
BPV benign positional vertigo
brady bradycardia
BRAO branch retinal artery occlusion
BRB bright red blood
BRBPR bright red bleeding per rectum
BRP bathroom privileges
BS blood sugar
BSA body surface area
BSE breast self examination
BSER brainstem evoked response
BSO bilateral salpingo-oophorectomy
BSOM bilateral serous otitis media
BSW Bachelor of Social Work
BT bleeding time
BTB back to bed

BTL bilateral tubal ligation
BTP breakthrough pain
BTPS body temperature pressure saturated
BUE both upper extremities
BUN blood urea nitrogen
BVM bag valve mask
BVO branch vein occlusion
BW birth weight
Bx biopsy
C celsius
C1-C7 cervical vertebrae 1 through 7
Ca calcium
CABG coronary artery bypass graft
CAD coronary artery disease
CAH chronic active hepatitis
cal calorie
CALGB Cancer and Leukemia Group B
Calc LDL calculated low density lipoprotein
cal ct calorie count
CAO chronic airway obstruction
CaO2 arterial oxygen content
CAP community-acquired pneumonia
CAPD continuous ambulatory peritoneal dialysis
CAT Children's ApperceptionTest
cath catheter, catheterization
C&B chair and bed
CBC complete blood count
CBD common bile duct
CBDE common bile duct exploration
CBI continuous bladder irrigation
CBS chronic brain syndrome
CC with correction (with glasses)
CCA common carotid artery
CCE cyanosis, clubbing, edema
CCHD critical congenital heart disease
CCO continuous cardiac output
CCR counter clockwise rotation
CCRN Certified Critical Care Registered Nurse
CCS Certified Coding Specialist
CCU Coronary Care Unit
CD closed drainage
C/D cup to disc ratio
CDB cough, and deep breath
CD&I clean, dry, and intact

C Dif Clostridium difficile
CEA carcinoembryonic antigen
CEA carotid endarterectomy
CESI cervical epidural steroid injection
CF cystic fibrosis
CF count fingers
CFL calcaneofibular ligament
CFS chronic fatigue syndrome
CFT complement fixation test
CG contact guarding
CGL chronic granulocytic leukemia
cGy centigray
CHB complete heart block
CHD congenital heart disease
chemo chemotherapy
CHF congestive heart failure
CHL conductive hearing loss
CHO carbohydrate
chol cholesterol
chr chronic
CI cardiac index
CIN cervical intraepithelial neoplasia
CINA Clinical Institute Narcotic Assessment Scale for Withdrawal Symptoms
CIWA Clinical Institute Withdrawal Assessment for Alcohol
CK creatine kinase
CK-MB creatine kinase MB fraction (primarily in cardiac muscle)
CKD chronic kidney disease
Cl chloride
CLIA Clinical Laboratory Improvement Amendments
CLL chronic lymphocytic leukemia
cl liq clear liquid
cm centimeter
cm2 square centimeter
CMC carpal metacarpal
CME cystoid macular edema
CMG cystometrogram
CML chronic myeloid leukemia
CMO comfort measures only
CMP cardiomyopathy
CMR cardiovascular magnetic resonance
CMS Centers for Medicare and Medicaid Services
CMV cytomegalovirus
CN cranial nerve
CN1 CN12

CNM Certified Nurse Midwife
CNS central nervous system
C.N.S.C. Certified Nutrition Support Clinician
C.N.S.D. Certified Nutrition Support Dietitian
CNVM choroidal neovascular membrane
CO cardiac output
CO carbon monoxide
C/O complained/complaint of, complaints
CO2 carbon dioxide
COAD chronic obstructive airway disease
cog test cognitive testing
COHb carboxyhemoglobin
COLD chronic obstructive lung disease
COM chronic otitis media
conc concentrate
conj conjunctiva
cont continuous
contu contusion
COPD chronic obstructive pulmonary disease
CP cerebral palsy
CPAP continuous positive airway pressure
CPB cardiopulmonary bypass
CPIP chronic pulmonary insufficiency of prematurity
CPK creatine phosphokinase
CPK-MB creatine phosphokinase of muscle band
CPM continuous passive motion
CPM counts per minute
CPNP Certified Pediatric Nurse Practitioner
CPPV continuous positive pressure ventilation
CPR cardiopulmonary resuscitation
CPS counts per second
CPT chest physiotherapy
CPX choroid plexus
CR complete remission
Cr creatinine
CRAO central retinal artery occlusion
CRBBB complete right bundle branch block
CrCl creatinine clearance
CREF closed reduction, external fixation
CREST calcinosis, Raynaud's disease, esophageal dysmotility, sclerodactyly, telangiectasia
CRF chronic renal failure
cric cricothyroidotomy / cricothyrotomy
CRL crown rump length
CRNA Certified Registered Nurse Anesthetist

CRP C-reactive protein
CRRT continuous renal replacement therapy
CRTT Certified Respiratory Therapy Technician
CRVO central retinal vein occlusion
C&S culture and sensitivity
C-section cesarean section
CSE combined spinal / epidural
CSF cerebrospinal fluid
CSLTM KayPENTAX Computerized Speech Lab
CSM carotid sinus massage
CSM circulation, sensation, movement
CSO Board Certified Specialists in Oncology Nutrition
CSOM chronic serous otitis media
C-spine cervical spine
CSRU Cardiac Surgical Recovery Unit
cSt centistoke
CTA clear to auscultation
cTnl cardiac troponin I
CTR carpal tunnel release
CTS carpal tunnel syndrome
CT scan computerized tomography scan
CT surgery cardiothoracic surgery
CTV clinical target volume
CV cardiovascular
CVA cerebrovascular accident
CVC central venous catheter
CVD cardiovascular disease
CVF central visual field
CVICU Cardiovascular Intensive Care Unit
CVL central venous line
CVO central vein occlusion
CvO2 mixed venous oxygen content
CVP central venous pressure
CVS clean voided specimen
CVST cerebral venous sinus thrombosis
CVVH continuous venovenous hemofiltration
C/W consistent with
CWR clockwise rotation
CX cervix
CXR chest x-ray
cyl cylinder
cysto cystoscopy
DA degenerative arthritis
DAPT Draw-A-Person Test

DAT direct antiglobulin test - Coombs
dB decibel
DB&C deep breathing and coughing
D bili direct bilirubin
D&C dilatation and curettage
DCCV direct current cardioversion
dCHF diastolic congestive heart failure
D&E dilation and evacuation
DEA# Drug Enforcement Administration number (physician's federal narcotic number)
derm dermatology
DF dorsiflexion
DFE dilated fundus examination
D fib defibrillation
DHL diffuse hystocytic lymphoma
DI diabetes insipidus
DIC disseminated intravascular coagulation
DIEP deep inferior epigastric perforator (flap)
diff differential white cell count
DIP distal interphalangeal (joints)
disch discharge
DJD degenerative joint disease
DKA diabetic ketoacidosis
dL deciliter
DLCOcorr single breath diffusion capacity corrected for hemoglobin
DLCO/SB diffusion capacity of carbon monoxide, single breath
DL/VA/SB/Hgb diffusion
capacity/alveolar
volume single breath
hemoglobin Pulmonary page 33
DM diabetes mellitus
dmax depth of maximum dose
DME durable medical equipment
DMH Department of Mental Health
DNA deoxyribonucleic acid
DNAR do not attempt to resuscitation (formerly DNR)
DNI do not intubate
DO diet order
D.O. Doctor of Osteopathy
DOA dead on arrival
DOB date of birth
DOE dyspnea on exertion
DOL day of life
DPAP diastolic pulmonary artery pressure
DPH Department of Public Health

DPL diagnostic peritoneal lavage
DPP dorsalis pedis pulse
DPP duration of positive pressure
DPT diphtheria, pertussis, tetanus (immunization)
DR diabetic retinopathy
DRG diagnostic related group
DRR digitally reconstructed radiograph
DSA digital subtraction angiography
DSD dry sterile dressing
DSG dressing
DSM Diagnostic and Statistical Manual of Mental Disorders
DSS Department of Social Services
DST dexamethasone suppression test
DT delirium tremens
DTR deep tendon reflexes
DTV due to void
DU duodenal ulcer
DUB dysfunctional uterine bleeding
DUVC double lumen umbilical venous catheter (NICU)
DVA distance visual acuity
DVT deep vein thrombosis
DVT Px deep vein thrombosis prophylaxis
Dx diagnosis
EAC external auditory canal
EBL estimated blood loss
EBM expressed breast milk
EBV Epstein-Barr virus
ECA external carotid artery
ECC endocervical curettage
ECCE extracapsular cataract extraction
ECF extended care facility
ECG electrocardiogram
echo echocardiogram
ECMO extracorporeal circulation membrane oxygenation
ECOG Eastern Cooperative Oncology Group
ECT electroconvulsive therapy
ED Emergency Department
ED&C electrodessication and curettage
EDD estimated date of delivery
EDV end diastolic volume
EDVi end diastolic volume index
EEG electroencephalogram
EENT eyes, ears, nose and throat
EEP end expiratory pressure

EF ejection fraction
EFM external fetal monitor
EFW estimated fetal weight
e.g. for example
EGA estimated gestational age
EGD esophagogastroduodenoscopy
eGFR estimated glomerularfiltration rate
EIP Early Intervention Program
ELF elective low forceps
EM electron microscopy
eMAR electronic medication administration record
EMC endometrial curettage
EMD electromechanical dissociation
EMG electromyograph
EMS electronic muscle stimulation
EMT Emergency Medical Technician
ENG electronystagmogram
ENT ears, nose, throat
EOB edge of bed
EOM extraocular movement
EOMI extraocular muscles intact
eos eosinophil
EPAP expiratory airway pressure
EPB extensor pollicis brevis
EP dept Electrophysiology Department
epi epinephrine
epis episiotomy
epith epithelial
EPL extensor pollicis longus
EPS electrophysiologic study
EPSS E point septal separation
ERA estrogen receptor assay
ERCP endoscopic retrograde choliangiopancreatography
ERE external rotation in extension
ERF external rotation in flexion
ERG electroretinogram
ERM epiretinal membrane
ERV expiratory reserve capacity
ESBL extended-spectrum beta-lactamases
ESI epidural steroid injection
ESR erythrocyte sedimentation rate
ESRD end-stage renal disease
est estimated
ESU Electrosurgical Unit

ESWL extracorporeal shock wave lithotripsy
ESV end systolic volume
ET eustachian tube
EtCO2 end tidal carbon dioxide
ETT endotracheal tube
ETT exercise tolerance test
EtOH alcohol
EUA exam under anesthesia
EUS endoscopic ultrasonography
eval evaluation
EVD external ventricular drain
ex excision
exp expired
ext extension
F fahrenheit
FA fatty acid
FAP familial adenomatous polyposis
FAST focused assessment by sonography in trauma
FB foreign body
FBS fasting blood sugar
FCE functional capacity evaluation
FDA Food and Drug Administration
Fe iron
F/E flexion/extension
FEF max maximum forced expiratory flow
fem femoral
fem-pop femoral popliteal (bypass)
FEN fluid, electrolytes, nutrition
FEV forced expiratory volume
FEV1 forced expiratory volume in 1 second
FEV3 forced expiratory volume in 3 seconds
FFA free fatty acid
FFP fresh frozen plasma
FHPA functional health pattern assessment
FHR fetal heart rate
FHx family history
fib fibrillation
fib fibula
FiO2 fraction of inspired oxygen
FeCO2 fraction of expired carbon dioxide
FeO2 fraction of expired oxygen
FL femoral / femur length
FLACC facial expression, leg movement, activity, crying, consolability (pain assessment scale)
flex sig flexible sigmoidoscopy

FM fetal movement (replaced AFM active fetal movement)
FN finger-to-nose (test)
FNF finger-nose-finger (neurological test)
FOB fecal occult blood
FOB father of baby
FOBT fecal occult blood test
FOS force of stream (Urology)
FQ frequency
Fr French (catheter gauge)
FRC functional residual capacity
FROM full range of motion
FS frozen section
FSBG fingerstick blood glucose
FSD focal skin distance
FSH follicle stimulating hormone
FTN finger to nose
FTT failure to thrive
F-tube feeding tube
F/U follow up
FUO fever of unknown origin
FVC forced vital capacity
FWB full weight bearing
Fx fracture
g gram
GA gestational age
GATB General Aptitude Test Battery
GB gallbladder
GBS gallbladder series
GC gonococci (gonorrhea)
GCS Glasgow Coma Scale
Gd gadolinium
GDM gestational diabetes mellitus
GERD gastroesophageal reflux disease
GETA general endotracheal anesthesia
GFR glomerular filtration rate
GG gamma globulin
GH growth hormone
GI gastrointestinal
GIFT gamete intrafallopian transfer
GL glaucoma
gluc glucose
GM + gram positive
GM - gram negative
gm % grams per hundred milliliters

GnRH gonadotropin releasing hormone
GONB greater occipital nerve block
gonio gonioscopy
GOO gastric outlet obstruction
GRS gender reassignment surgery
GSW gun shot wound
GSV great saphenous vein
GTPAL gestation, term, preterm, abortion and living)
GTT glucose tolerance test
G-tube gastrostomy tube
GTV gross tumor volume
GU genitourinary
GYN gynecology
H hydrogen
HA headache
HAAb Hepatitis A antibody
HAL hyperalimentation
HAP hospital-acquired pneumonia
HASCVD hypertensive arteriosclerotic cardiovascular disease
HBcAg Hepatitis B core antigen
HBeAb Hepatitis Be antibody (antigen)
HBIG Hepatitis B immune globulin
HBP high blood pressure
HBcAb Hepatitis B core antibody
HBsAg Hepatitis B surface antigen
HBV Hepatitis B virus
HC head circumference
HCA Healthcare Associates
HCAb Hepatitis C antibody
hCG human chorionic gonadotropin
HCO3 bicarbonate
Hct hematocrit
HCV Hepatitis C virus
HCVD hypertensive cardiovascular disease
HD hemodialysis
HD hospital day
HDL high density lipoprotein
HDR hemodynamic response
HEENT head, ears, eyes, nose, throat
HeliOx helium and oxygen mix
Hem/Onc Hematology/Oncology
HEP home exercise program
HF high frequency
HFNC high flow nasal cannula

HFOV high frequency oscillatory ventilation
HFpEF heart failure with preserved ejection fraction
HFrEF heart failure with reduced ejection fraction
Hg mercury
Hgb hemoglobin
HH hiatal hernia
H&H hemoglobin and hematocrit
HHA Home Health Aid
HHC home health care
HHD hypertensive heart disease
HHN hand held nebulizer
5-HIAA 5-hydroxyindoleacetic acid
HIV human immunodeficiency virus
HL heparin lock
HLA human leukocyte antigen
HMD hyaline membrane disease
HMF human milk fortifier
HMFP Harvard Medical Faculty Physicians
HNP herniated nucleus pulposus
HNPCC hereditary non polyposis colorectal cancer
H2O water
HO House Officer
H/O history of
HOB head of bed
HOH hard of hearing
hosp hospitalization
H&P history and physical
HPF high power field
HPI history of present illness
hr hour
HR heart rate
HRS hepatorenal syndrome
HSV herpes simplex virus
ht height
HTLV III human T-cell lymphotrophic virus type III
HTN hypertension
HTO high tibial osteotomy
HTS heel-to-shin (test)
HVD hypertensive vascular disease
HVMA Harvard Vanguard Medical Associates
Hx history
hypo hypodermic injection
hyst hysterectomy
Hz hertz frequency

IABP intra-aortic balloon pump
IADHS inappropriate antidiuretic hormone syndrome
IADL instrumental activities of daily living
IASD interatrial septal defect
IBC iron binding capacity
IBCLC International Board-Certified Lactation Consultant
IBD inflammatory bowel disease
I bili indirect bilirubin
IBS irritable bowel syndrome
IBW ideal body weight
IC inspiratory capacity
ICA internal carotid artery
ICCE intracapsular cataract extraction
ICD implantable cardioverter-debifrillator
ICD-9 International Statistical Classification of Diseases, 9th revision
ICF intracellular fluid
ICP intracranial pressure
ICS intercostal space
ICU Intensive Care Unit
I&D incision and drainage
ID identification
IDDM insulin dependent diabetes mellitus
IDM infant of diabetic mother
I/E ratio inspiratory to expiratory time ratio
IFUP Infant Follow-up Program
Ig immunoglobulin
IgA immunoglobulin A
IgE immonoglobulin E
IgG immunoglobulin G
IgM immunoglobulin M
IGRT image guided radiation therapy
IHI Institute for Healthcare Improvement
IHSS idiopathic hypertrophic subaortic stenosis
IJ internal jugular
ILA inferior lateral angle
ILBBB incomplete left bundle branch block
ILM internal limiting membrane
ILMI inferolateral myocardial infarct
IM intramuscular
IMA inferior mesenteric artery
IM internal mammary artery
IME independent medical exam
IMI inferior myocardial infarction (location)
imp impression

IMRT intensity modulated radiation therapy
IMV intermittent mandatory ventilation
in inches
inf inferior
ing inguinal
ingred ingredient(s)
inj injection
INO inhaled nitric oxide
INR international normalized ratio
insp/min breaths per minute
int internal
int-rot internal rotation
intub intubation
inver inversion
IO inferior oblique
I&O intake and output
IOFB intraocular foreign body
IOL intraocular lens
IOP intraocular pressure
IP inpatient
IPAP inspiratory positive airway pressure
IPPB intermittent positive pressure breathing
IQ intelligence quotient
IR Interventional Radiology
I/R/B/A Indications/Risks/Benefits/Alternatives
IRBBB incomplete right bundle branch block
IRE internal rotation in extension
IRF internal rotation in flexion
irreg irregular
irrig irrigation
IRV inspiratory reserve volume
IS incentive spirometer
ITP idiopathic thrombocytopenic purpura
IUC intrauterine catheter (indwelling)
IUD intrauterine device
IUFD intrauterine fetal demise
IUGR intrauterine growth restriction
IUP intrauterine pregnancy
IUPC intrauterine pressure catheter
IV intravenous
IVC inferior vena cava
IVCD interventricular conduction delay/defect
IVD intravenous drip
IVDA intravenous drug abuse

IVF in vitro fertilization
IV fluids intravenous fluids
IVP intravenous pyelogram
IVPB intravenous piggyback
JODM juvenile onset diabetes mellitus
JP Jackson-Pratt (drain)
JR junctional rhythm
J-tube jejunostomy tube
JVD jugular venous distention
JVP jugular venous pressure (pulse)
K thousand
K+ potassium
kcal kilocalorie
kg kilogram
kHz kilohertz
KI knee immobilizer
KJ knee jerk
km/h kilometers per hour
KUB kidneys, ureters, bladder
KVO keep vein open
K-wire Kirschner wire
L liter
L1 L5
LA left atrium
LA4ch left atrial 4 chamber
LAA left atrial appendage
lab laboratory
lac laceration
LAD left anterior descending (coronary vessels)
LAH left anterior hemiblock
lam laminectomy
LAN lymphadenopathy
LAO left anterior oblique
LAP left atrial pressure
LAP leukocyte alkaline phosphatase
lap chole laparoscopic cholecystectomy
lat lateral
lb pound
LBBB left bundle branch block
LBP low back pain
LBQC large based quad cane
LC Lactation Consultant
LCA left coronary artery
LCX left circumflex coronary artery

L&D labor and delivery
LDH lactic dehydrogenase
LDL low density lipoprotein
L.D.N. Licensed Dietitian Nutritionist
LE lower extremity
LEA lumbar epidural analgesia
LE prep lupus erythematous preparation
LESI lumbar epidural steroid injection
LF low forceps
LFT liver function tests
LGA large for gestational age
LH luteinizing hormone
LHF left heart failure
Li lithium
lig ligament
LIH left inguinal hernia
LIMA left internal mammary artery (graft)
LINAC linear accelerator
liq liquid
LL long leg
LL lower limb
Llat left lateral
LLB long leg brace
LLC long leg cast
LLE left lower extremity
LLL left lower lobe (lung)
LLQ left lower quadrant (abdomen)
LMA laryngeal mask airway
LME left mediolateral episiotomy
L/min liters per minute
LML left middle lobe
LMP last menstrual period
LMWH low molecular weight heparins
LOA leave of absence
LOB loss of balance
LOC level of consciousness
LOM loss of motion
LOQ lower outer quadrant
LOS length of stay
LP lumbar puncture
LPHB left posterior hemiblock
L.P.N. Licensed Practical Nurse
LPO left posterior oblique
LR lactated ringers (IV solution)

LS lumbosacral
LSB left sternal border
LSB lumbar sympathetic block
LSCTA lung sounds clear to auscultation
LSD lysergide
LSO left salpingo-oophorectomy
L-spine lumbar spine
LT Levin tube
LTAC long term acute care
LTG long term goal
LTH luteotropic hormone
LTL laparoscopic tubal ligation
LTT lactose tolerance test
LUE left upper extremity
LUL left upper lobe (lung)
LUOQ left upper outer quadrant
LUQ left upper quadrant
LV left ventricle / ventricular
LVAD left ventricular assist device
LVEDD left ventricular end diastolic diameter
LVEDP left ventricular end diastolic pressure
LVEF left ventricular ejection fraction
LVH left ventricular hypertrophy
LVMI left ventricular mass index
LVOT left ventricular outflow tract
L&W living and well
LWBS left without being seen
LY30 part of thromboelastogram measurement (represents clot lysis)
lymphs lymphocytes
lytes electrolytes (Na, K+, Cl etc.)
m meter
m2 height in meters squared
MA maximum amplitude
MA Medical Assistant
MAC monitored anesthesia care
mammo mammogram
MAOI monoamine oxidase inhibitor
MAP mean airway pressure
MAPSE mitral annular plane systolic excursion
MAR medication administration record
MAS meconium aspiration syndrome
MAT Miller-Abbot tube
MAWP mean airway pressure
max maximal / maximum

max A maximal assistance
max PF maximum peak flow
MBD minimal brain damage
MBS modified barium swallow
MCA middle cerebral artery
mcg microgram
MCH mean corpuscular hemoglobin
MCHC mean corpuscular hemoglobin concentration
mCi millicurie
mcL microliter
MCL medial collateral ligament
MCP metacarpophalangeal joint
MCT medium chain triglyceride
MCV mean corpuscular volume
mcV microvolt
MD muscular dystrophy
M.D. Medical Doctor
MDI manic-depressive illness
MDRD Modification of Diet in Renal Disease (study)
ME Medical Examiner
mec meconium
med medial
meds medications
MELD model for end-stage liver disease (score)
mEq milliequivalent
met metastasis
METS metabolic equivalents
MeV million (1,000,000) electron volts
mEq/L milliequivalents per liter
MFAT multifocal atrial tachycardia
MFR myofascial release
Mg magnesium
mg milligram
MH marital history
MHT malignant hypertension
MI myocardial infarction
MIC minimum inhibitory concentration
MICU Medical Intensive Care Unit
min minute
min A minimal assistance
mL milliliter
MLC multileaf collimator
MM mucous membrane
mm millimeter

MMEFR maximal mid-expiratory flow rate
MMF mean maximum flow
mmHg millimeters of mercury
MMM moist mucous membrane
mmol millmole
MMPI Minnesota Multiphasic Personality Inventory
MMR measles, mumps and rubella
Mn manganese
MNCV motor nerve conduction velocity
mod moderate
mod A moderate assistance
mono mononucleosis
mOsmol milliosmole
mph miles per hour
MPAP mean pulmonary artery pressure
MPI myocardial perfusion imaging
MR mitral regurgitation
MR mental retardation
M&R measure and record input/output
MRG murmurs, rubs and gallops
MRI Magnetic Resonance Imaging
MRN medical record number
MRSA methicillin resistant staphylococcus aureus
MS Master of Science
MSE Mental Status Examination
msec millisecond
MSK musculoskeletal
MSL midsternal line
MSSA methicillin susceptible staphylococcus aureus
M.S.W. Master of Social Work
M/T myringotomy with tubes
MTA metatarsal abduction
MTP metatarsophalangeal
mu monitor unit
mU milliunits
MV megavolt
MVA motor vehicle accident
MVC motor vehicle collision
MVI multiple vitamin injection
MVP mitral valve prolapse
MVR mitral valve replacement
Na+ sodium
N/A not applicable
NABS normoactive bowel sounds

NAD no acute distress
NAS no added salt
NAT nucleic acid test
NB newborn
NBS normal bowel sounds
NC no complaints
N/CAN nasal cannula
NCS nerve conduction studies
NCSE nonconvulsive status epilepticus
neb nebulizer (hand held)
NED no evidence of disease
neg negative
NEOB New England Organ Bank
Neuro Neurology / neurologic / neurological
NFPEx nutrition-focused physical examination
ng nanogram
NG nasogastric
NGT nasogastric tube
NH nursing home
NIBP non-invasive blood pressure
NICU Neonatal Intensive Care Unit
NIDDM non-insulin dependent diabetes mellitus
NIF negative inspiratory force
NJT nasojejunal tube
NKA no known allergies
NKDA no known drug allergies
NKFA no known food allergies
NL normal
NLP no light perception
NMJ neuromuscular junction
NND neonatal death
NNP Neonatal Nurse Practitioner
noct nocturnal
non-std not standard (used with TPN solutions)
NOS not otherwise specified
NP nasopharyngeal
N.P. Nurse Practitioner
NPA nasal pharyngeal airway
NPH neutral protamine hagedorn isophane insulin
NPO nothing by mouth
NR no refill
NRB nonrebreather (oxygen mask)
NSAID nonsteroidal anti-inflammatory drug
NSD normal spontaneous delivery

NSG nursing
NSR normal sinus rhythm
NSS neurological signs stable
NST nonstress test
NSTEMI non ST segment elevation myocardial infarction
NSVD normal spontaneous vaginal delivery
NT nasotracheal
NT/ND non-tender / non-distended
NT-proBNP N-terminal brain natriuretic peptide
NV neurovascular
N&V nausea and vomiting
NVA near visual acuity
NVD normal vaginal delivery
NWB non-weight bearing
O2 oxygen
OA osteoarthritis
OA occiput anterior
OB obstetrics
OBS organic brain syndrome
obj objective
O-CAT oral care assessment tool
occ occasionally
OD right eye
OE otitis externa
OG oral gastric (feeding)
OGT orogastric tube
oint ointment
OKInt okay to intubate
OKN opticokinetic nystagmus
OM otitis media
OMR on-line medical record
OOB out of bed
OP outpatient
OP oropharyngeal
op operation
OPA oral pharyngeal airway
OPD outpatient department
OR operating room
ORIF open reduction-internal fixation
Ortho orthopedic / orthopaedic
OS left eye
OSA obstructive sleep apnea
OSH outside hospital
O.T. occupational therapy / Occupational Therapist

OTA open to air
OTC over the counter (sold without prescription)
OU both eyes
OV office visit
oz ounce
P phosphorus
p after
P2 pulmonic second heart sound
P.A. Physician Assistant
PA pulmonary artery
P&A percussion and auscultation
PAC premature atrial contraction
PACEN paracentesis
PaCO2 partial pressure (tension) of carbon dioxide, artery
PACS picture archiving and communications systems (Radiology)
PACU Post-Anesthesia Care Unit
PADSS post anesthesia discharge scoring system
PAF paroxysmal atrial fibrillation
PAINAD pain assessment in advanced dementia (pain assessment scale)
PA line pulmonary artery line
palp palpation
PAML pre-admission medication list
PaO2 arterial oxygen pressure
PAP pulmonary arterial pressure
pap smear papanicolaou smear
PAR post anesthetic recovery
para number of pregnancies producing viable offspring (parity)
PASP pulmonary artery systolic pressure
PAT preadmission testing
path pathology
PAW peak airway pressure
PAWP pulmonary artery wedge pressure
PB barometric pressure
PbtO2 brain tissue partial pressure of oxygen
PC pressure control
p.c. after meals
PCA patient controlled analgesia
PCB paracervical block
PCCU Post Coronary Care Unit
PCI percutaneous coronary intervention
PCIOL posterior chamber intraocular lens
PCNT percutaneous nephrostomy tube
PCO2 partial pressure (tension) of carbon dioxide, artery
PCP Primary Care Physician

PCR polymerase chain reaction
PCT Patient Care Technician
PCTA percutaneous coronary transluminal angiography
PCV packed cell volume
PCW pulmonary capillary wedge
PCWP pulmonary capillary wedge pressure
PCXR portable chest x-ray
PD peritoneal dialysis
PDA patent ductus arteriosus
PDR Physicians' Desk Reference
PE pulmonary embolism
PEA arrest. pulseless electrical activity
PeCO2 mixed expired carbon dioxide tension
Peds pediatrics
PEEP positive end expiratory pressure
PEF peak expiratory flow
PEFR peak expiratory flow rate
PEG percutaneous endoscopic gastrostomy
PEmax maximum expiratory pressure
PERLA pupils equally reactive to light and accommodation
PERRL pupils equal, round, and reactive to light
PERRLA pupils equal, round, reactive to light and accommodation
PET positron emission tomography
PetCO2 peak end tidal carbon dioxide
PEx physical examination
PF peak flow
PF plantar flexion
PFO patent foramen ovale
PFSH past, family and social history
PGY post graduate year
PFT pulmonary function test
pH hydrogen ion concentration (degree of acidity)
PH pinhole
PHVA pinhole visual acuity
phaco phacoemulsification
PHx past history
PI present illness
PICC peripherally inserted central catheter
PID pelvic inflammatory disease
PIH pregnancy induced hypertension
PImax maximum inspiratory pressure
PIN posterior interosseous nerve
PIP peak inspiratory pressure
PIPJ proximal interphalangeal joint

PIV peripheral intravenous
PKU phenyl ketonuria
Plt platelet
PLTc platelet count
PM postmortem
p.m. evening
PMA post menstrual age
PMHx past medical history
PMI point of maximal impulse
PMN polymorphonuclear leukocyte
PMR polymyalgia rheumatic
PM&R physical medicine and rehabilitation
PMS premenstrual syndrome
PMT premenstrual tension
PN parenteral nutrition
PNA pneumonia
PNB premature nodal beat
PND paroxysmal nocturnal dyspnea
PNP peak negative pressure
Pnx pneumothorax
PO by mouth
PO2 partial pressure (tension) of oxygen, artery
PO4 phosphate
POC point-of-care
POC product of conception
POD post operative day
POE provider order entry
POI point of interest
POLY polymorphonuclear leukocyte
PONV postoperative nausea and vomiting
poplit popliteal
pos positive
post after
post-op after surgery
POV privately owned vehicle
PP postpartum
PPD purified protein derivative (of tuberculin)
PPF plasma protein fraction
PPFT post pyloric feeding tube
PPI proton pump inhibitor
PPLAT plateau pressure
PPN peripheral parenteral nutrition
PPS post perfusion syndrome
PPV positive pressure ventilation

PR per rectum
PRBC packed red blood cells
pre-op before surgery
prep prepare / preparation
primip primipara (1st pregnancy)
p.r.n. as often as necessary
PROM premature rupture of membranes
PROM passive range of motion
pron pronation
PRP pan-retinal photocoagulation
PS pulmonary stenosis
PS pressure support (ventilator mode)
PSA prostate specific antigen
PSG polysomnogram, polysomnography
PSHx past surgical history
PSI pounds per square inch
PSP phenosulfonphthalein
PST posterior sub-tenon (capsule)
PSVT paroxysmal supraventricular tachycardia
P.T. physical therapy /Physical Therapist
PT prothrombin time
PT preterm
PTA prior to admission
PTB patellar tendon bearing
PTCA percutaneous transluminal coronary angioplasty
PTFE polytetrafluoroethylene
PTH parathyroid hormone
PTP posterior tibial pulse
PTPN peripheral total parenteral nutrition
PTS patella tendon suspension
PTT partial thromboplastin time
PTV planning target volume
PUBS percutaneous umbilical blood sampling
PUD peptic ulcer disease
pul pulmonary
P&V percussion and vibration
PVB premature ventricular beat
PVC premature ventricular contraction
PVD peripheral vascular disease
PVD posterior vitreous detachment
PVR proliferative vitreoretinopathy
PVR post voiding residual
PVR pulmonary vascular resistance
PVT paroxysmal ventricular tachycardia

PWB partial weight bearing
Px prophylasis
PXAT paroxysmal atrial tachycardia
Q every
QC quality control
QB blood flow
q.i.d. four times a day
QP/QS ratio of pulmonary blood to systemic blood flow
QS sufficient quantity
qt quart
quad quadriceps
R respiration
+R Rinne test positive
-R Rinne test negative
RA rheumatoid arthritis
RA right atrium
RAD right axis deviation
RAM rapid alternating movements
RAO right anterior oblique
RAPD relative afferent pupillary defect
RASS Richmond Agitation-Sedation Scale
RBBB right bundle branch block
RBC red blood cell (count)
RCA right coronary artery (coronary vessels)
RCM right costal margin
R.C.P. Respiratory Care Practitioner
RD radial deviation
RD retinal detachment
R.D. Registered Dietitian
RDA recommended daily allowance
RDI respiratory disturbance index
RDS respiratory distress syndrome
RDW red (cell) distribution width
REE resting energy expenditure
ref referred
REDF reverse end - diastolic flow
rehab rehabilitation
REM rapid eye movement
RER renal excretion rate
RES reticuloendothelial system
resp respiratory
retic reticulocyte
RF rheumatoid factor
RF rheumatic fever

Rh Rhesus blood factor
RHD rheumatic heart disease
RIA radioimmunoassay
RIH right inguinal hernia
RIMA right internal mammary artery
RISS regular insulin sliding scale
RLF retrolental fibroplasia
RLS ringer lactate solution
RLE right lower extremity
RLL right lower lobe
RLQ right lower quadrant
RML right middle lobe
RMR resting metabolic rate
R.N. Registered Nurse
RNA ribonucleic acid
R/O rule out
ROI region of interest
ROM range of motion
ROM rupture of membranes
ROS review of systems
rot rotator
RP retinitis pigmentosa
RPE rating of perceived exertion
RPM revolutions/rotations per minute
RPO right posterior oblique
RPR rapid plasma reagin (syphilis)
RQ respiratory quotient
RR respiratory rate
RRE round, regular and equal (pupils)
RRR regular rhythm and rate
RSBI Rapid Shallow Breathing Index
RSD reflex sympathetic dystrophy
RSI rapid sequence induction/intubation
RSO right salpingo-oophorectomy
RSR regular sinus rhythm
RT radiation therapy
RTW return to work
RUE right upper extremity
RUL right upper lobe
RUOQ right upper outer quadrant
RUQ right upper quadrant
RV right ventricle / ventricular
RV residual volume
RVEF right ventricular ejection fraction

RVOT right ventricular outflow tract
RVG radionuclide ventriculography
RVH right ventricular hypertrophy
RW rolling walker
RWMA regional wall motion abnormalities
Rx prescription
S1 first heart sound
S2 second heart sound
S3 third heart sound (ventricular filling gallop)
S4 fourth heart sound (atrial gallop)
S1 ... S5 sacral vertebrae
S/A sugar and acetone
SAB spontaneous abortion
SAD source axis distance
SAH subarachnoid hemorrhage
SAM systolic anterior motion
SAN sinoatrial node
sang sanguinous
SaCO arterial saturation of carbon monoxide
SaO2 arterial oxygen percent saturation
SASC Skills and Simulation Center
sat saturation
SB small bowel
SB scleral buckling
SBE subacute bacterial endocarditis
SBFT small bowel follow through
SBO small bowel obstruction
SBP systolic blood pressure
SBQC small based quad cane
SBS small bowel series
SBT spontaneous breathing test/trial
SC subclavian
sc without correction (without glasses)
SCC squamous cell carcinoma
sCHF systolic congestive heart failure
SCT Sentence completion test
S/D systolic/diastolic ratio
SCD sequential compression device
SDH subdural hematoma
SDS same day surgery
sec second
sed rate sedimentation rate
segs segmented neutrophils
SEM systolic ejection murmur

SEMI subendocardial myocardial infarction (type)
SFA superficial femoral artery
SG specific gravity
SGA small for gestational age
SGB stellate ganglion block
SGC Swan-Ganz catheter
SGOT serum glutamic oxaloacetic transaminase
SGPT serum glutamate pyruvate transaminase
SHx social history
SI suicidal ideation
SIADH syndrome of inappropriate antidiuretic hormone secretion
sib sibling
SICU Surgical Intensive Care Unit
SIDS sudden infant death syndrome
SIG let it be marked (appears on prescription before direction for patient)
SIJB sacroiliac joint block
SIL squamous intraepithelial lesion
SIMV synchronized intermittent mandatory ventilation
SIRS systemic inflammatory response syndrome
SK-SD streptokinase-streptodornase
S/L slit lamp
SLB short leg brace
SLC short leg cast
SLE systemic lupus erythematosus
SLex sialyl Lewis x (antigen)
SLP speech language pathologist
SLR straight leg raising
SMA superior mesenteric artery
SNF Skilled Nursing Facility
SNHL sensineural hearing loss
SO superior oblique
S-O salpingo-oophorectomy
SOAP subjective, objective, assessment and plan
SOB shortness of breath
sol solution
SOM serous otitis media
SONB supra orbital nerve block
sono sonogram
S/P status post
SPA salt poor albumin
SPCO saturation of hemogloben by pulse oximetry
spec specimen
SPECT single-photon emission computed tomography
SpG specific gravity

sph sphere
SPL sound pressure level
SPO2 oxygen saturation by pulse oximeter
spont spontaneous
SPP suprapubic prostatectomy
SP tube suprapubic tube
SR sinus rhythm
SROM spontaneous rupture of membrane
SRT speech reception threshold
S&S signs and symptoms
SSCP substernal chest pain
SSD source to skin distance
SSE soapsuds enema
SSI Supplemental Security Income
SSN Social Security number
SSS sick sinus syndrome
SSV small saphenous vein
ST sinus tachycardia
staph staphylococcus aureus
stat immediately (statim)
STD sexually transmitted disease
STEMI ST elevation myocardial infarction
STH somatotrophic hormone
STPD standard temperature and pressure - dry
strep streptococcus
STS serologic test for syphilis
STSG split thickness skin graft
subcut subcutaneous
subl sublingual
SUC urinary catheter - straight
supp suppository
SV stroke volume
SV seminal vesicle
SVC superior vena cava
SVC slow vital capacity
ScvO2 central venous oxygen saturation
SVD spontaneous vaginal delivery
SVG saphenous vein graft
SVI stroke volume index
SvO2 mixed venous oxygen saturation
SVR systemic vascular resistance
SVT supraventricular tachycardia
SW standard walker
SZ seizure

T temperature
T3 triiodothyronine
T4 thyroxine
T1 T12
T&A tonsillectomy and adenoidectomy
TAA thoracic aortic aneurysm
TAB therapeutic abortion
TAH total abdominal hysterectomy
TAHBSO total abdominal hysterectomy, bilateral salpingo-oophorectomy
TAM total active motion
T APPL applanation tonometry
TAPSE tricuspid annular plane systolic excursion
TAR total ankle replacement
TAT Thematic Apperception Test
TAVR transcatheter aortic valve replacement
TB tuberculosis
TBG thyroxine binding globulin
TBI total body irradiation
T bili total bilirubin
tbl tablespoon
TCA tricyclic antidepressant
TCC transitional cell carcinoma
TCDB turn, cough and deep breath
T Chol total cholesterol
TcPO2 transcutaneous oxygen
TCO2 total carbon dioxide
TCU Transitional Care Unit
TD tumor dose
TDI tissue Doppler imaging
TDWB touch down weight bearing
tE total expiratory time
TEC total eosinophil count
TEE transesophageal echocardiography
TENS transcutaneous electrical nerve stimulation
TER total elbow replacement
tert tertiary
TESI thoracic epidural steroid injection
TFB trifascicular block
TFCC triangular fibrocartilage complex
TFESI transforaminal epidural steroid injection
TFTs thyroid function tests
TG triglycerides
TGA transposition of the great arteries
TGV transposition of great vessels

T&H type and hold
THCEN thoracentesis
th-cult throat culture
ther ex therapeutic exercise
THP total hip prosthesis
THR total hip replacement
Ti titanium
TIA transient ischemic attack
tib tibia
TIBC total iron-binding capacity
t.i.d. three times a day
TIMI thrombolysis in myocardial infarction
tinc tincture
TJ triceps jerk
TKO to keep open
TKR total knee replacement
TL tubal ligation
Tl thalium
TLC total lung capacity
TLD tubes, lines & drains
TLS-spine thoracic, lumbar, sacral spine
TM tympanic membrane
TMA transmetatarsal amputation
Tmax temperature maximum
TMB transient monocular blindness
TMJ temporomandibular joint
TMP transmembrane pressure
TMT tarsometatarsal
TNI total nodal irradiation
TNM primary tumor, regional lymph nodes, and distant metastasis
TnT troponin T
TOCO tocodynamometer
TOF tetralogy of Fallot
tol tolerate
TOLAC trial of labor after cesarean delivery
tomo tomography
TORCH toxoplasmosis, other, rubella, cytomegalovirus, and herpes simplex
TORP total ossicular replacement prosthesis
TOS thoracic outlet syndrome
TOXO toxoplasmosis antibody
TP total protein
TPIT trigger point injection therapy
TPN total parenteral nutrition
TPR temperature, pulse, and respiration

tr trace
trach tracheostomy
TRH thyrotropin releasing hormone (thyroid)
Trig triglycerides
T3RU triiodothyronine (T3) resin uptake
T&S type and screen
TSH thyroid stimulating hormone
TSICU Trauma Surgical Intensive Care Unit
tsp teaspoon
T-spine thoracic spine
TSR total shoulder replacement
TSS toxic shock syndrome
TT thrombin time
TTN transient tachypnea of the newborn
TTP thrombotic thrombocytopenic purpura
TUR transurethral resection
TURB transurethral resection of the bladder
TURBT transurethral resection of bladder tumor
TURP transurethral resection of prostate
TURV transurethral resection valves
TV tricuspid valve
TVR tricuspid valve replacement
TWE tapwater enema
UA urinalysis
UAC umbilical artery catheter
UBW usual body weight
UC ulcerative colitis
UCG test urinary chorionic gonadotropins test
u-cult urine culture
UE upper extremity
UF ultrafiltration
UFR ultrafiltration rate
UFV ultrafiltration volume
UGI upper gastrointestinal series
UIQ upper inner quadrant
UL upper limb
U-line umbilical line
UMN upper motor neuron (disease)
ung ointment (unguentum)
UNOS United Network for Organ Sharing
UO urinary output
UOQ upper outer quadrant
Ur Ac uric acid
URI upper respiratory infection

UROL urology
URQ upper right quadrant
URR urea reduction ratio
US ultrasonography
USN ultrasonic nebulizer
USO unilateral salpingo-oophorectomy
UTI urinary tract infection
UUN urinary urea nitrogen
UV ultraviolet
UVC umbilical vein catheter
V minute volume (cardiac output)
VA visual acuity
VAD vascular (venous) access device
vag vagina
VAP ventilator associated pneumonia
VAS Visual Analog Scale
VA/SB alveolar volume single breath
VATS video-assisted thoracoscopic surgery
VB venous blood
VBAC vaginal delivery after cesarean delivery
VBG venous blood gas
VC vital capacity
VCT venous clotting time
VCO2 carbon dioxide output
VD venereal disease
vdg voiding
VDRL Venereal Disease Research Laboratory (test for syphilis)
VD/VT dead space to tidal volume ratio
Ve minute ventilation (pulmonary function test)
VE vaginal exam
VEA ventricular ectopic activity
vent ventilator
VER visual evoked responses
VF ventricular fibrillation
VI inspired volume
VICU Vascular Intermediate Care Unit
vit vitamin
VLDL very low density lipoprotein
VMA vanillymandelic acid
VMI visual motor integration
VN visiting nurse
VNA Visiting Nurses' Association
VO verbal order
VO2 oxygen consumption

vol volume
VOR vestibule-ocular reflex
VP venous pressure
VPB ventricular premature beat
VRE vancomycin resistant enterococci
VS vital signs (temperature, pulse and respiration)
VSD ventricular septal defect
VSS vital signs stable
Vt tidal volume
V tach ventricular tachycardia
VTI velocity time integral
VTE venous thromboembolism
w/ with
WAIS Wechsler Adult Intelligence Scale
WAP wandering atrial pacemaker
Wass Wasserman test
WB whole blood
WBAT weight bearing as tolerated
WBC white blood cell (count)
WC wheelchair
W/D warm and dry
WFL within functional limits
WIC Women, Infants, and Children (program)
WINROP Weight gain, Insulin-like growth factor-1, Neonatal ROP
WISC Wechsler Intelligence Scale for Children
WISC-R Wechsler Intelligence Scale for Children - Revised
wk week
WMA wall motion motility
WMS Wechsler Memory Scale
WN well nourished
WNL within normal limits
WNV West Nile Virus
w/o without
WOB work of breathing
WPW Wolff-Parkinson-White (syndrome)
WRAT Wide Range Achievement Test
WRISS Weapon-Related Injury Surveillance System
wt weight
W-T-D wet to dry
W/U work up
X times
XM cross match
X-ray Radiology image
XRT radiation therapy

YO years old
yr year
YTD year to date