DEVELOPMENT AND EVALUATION OF A MACHINE LEARNING ALGORITHM TO MAP MEDICAL CONDITIONS AND PROCEDURES FROM REAL-WORLD DATA

By

Rupa Makadia, MS, PhD (candidate)

A Dissertation Proposal 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

February 2019

RUTGERS
School of Health Professions

Final Dissertation Defense Approval Form

DEVELOPMENT AND EVALUATION OF A MACHINE LEARNING ALGORITHM TO MAP MEDICAL CONDITIONS AND PROCEDURES FROM REAL-WORLD DATA

By

Rupa Makadia, MS, PhD (candidate)

Dissertation Committee:

Dr. Shankar Srinivasan, PhD

Dr. Patrick Ryan, PhD

Dr. Fredrick Coffman, PhD

Dr. Christian Reich, MD, PhD

Approved by the Dissertation Committee:

_____ **Date:**_____

_____ **Date:**_____

_____ **Date:**_____

_____ **Date:**_____

ABSTRACT

DEVELOPMENT AND EVALUATION OF A MACHINE LEARNING ALGORITHM TO MAP MEDICAL CONDITIONS AND PROCEDURES FROM REAL-WORLD DATA

Rupa Makadia, MS, PhD (candidate)

Background: Ontologies characterize complex and detailed information and are extensively used in healthcare research. Medical information (textbooks, expert opinions, clinical evidence) has information on conditions and its corresponding procedures (treatments), but this information is not captured or structured in any ontology. The objective of the research is to create a condition-procedure ontology from real world data to be utilized in observational research or electronic health record (EHR) system.

Methods: Predictive models are developed to learn from five datasets (administrative claims, hospital charge data) to generate two algorithms (diagnostic and therapeutic) to predict condition-procedure relationships in the SNOMED-CT vocabulary. A reference set with 100 positive pairs per algorithm, and 32,132 negative pairs were developed. Predictive models were constructed by designing 51 possible covariates that describe condition-procedure pairs from Optum© De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum) dataset and determining which covariates discriminated between the positive and negative controls, as measured by Area Under Receiver Operator Characteristic Curve (AUC). External validation of the final algorithms was performed on 4 other databases. The final algorithms were applied across the universe of condition and procedure pairs in all five databases to construct the full condition-procedure ontology, and the ontology was evaluated for validity and coverage

of condition and procedure concepts from the set of identified condition-procedure pairs. An additional analysis was trained to classify diagnostic vs. therapeutic intervention based on the overlap of pairs within the two algorithms.

Results: Algorithms include the following covariates: condition-procedure occurring together, relative risk, support and sensitivity. Both algorithms had AUCs greater than .90, and external validation also showed similar results. In Optum, 98% of conditions and 63% of procedure codes had at least one relationship identified in the ontology. The intervention type analysis resulted in an AUC of 0.79.

Conclusions: Real-world data can be utilized to construct a medical ontology of condition-procedure relationships with strong performance and good coverage. These results can be utilized to fuel research efforts in healthcare such as cohort generation and computer provider order entry systems by understanding conditions and procedures and their application to diagnose or treat a patient.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my employer Janssen, a Johnson and Johnson family of companies to allowing me access to the data to make this work and research possible.

An incredible thank you to my advisor Dr. Srinivasan and Dr. Coffman at Rutgers for the support and guidance through my time at Rutgers.

A huge thank you to Dr. Reich for your guidance and comments on this work.

To my family and friends, I am so grateful to have you in my life for support, love and care that has shaped me to the be person I am today.

To my colleagues at Janssen and elsewhere, you have been the pillar in my growth as a researcher and would have never gotten to this place without your support.

Lastly, I would like to thank Dr. Patrick Ryan for everything he has done for me, providing direction and leadership in this work and all my other research endeavors. You have inspired me in the work that we do every day and hope that I can always imbibe your passion for helping patients through research. I truly can’t thank you enough.

DEDICATION

To my parents…my first leaders and always in my heart in every step I take…

&

To Arushi & Anjali…may you both reach for the moon and attain the stars…

TABLE OF CONTENTS

LIST OF TABLES 11

LIST OF FIGURES 13

I. CHAPTER 1 INTRODUCTION 14

1.1. BACKGROUND 14

1.2. OBSERVATIONAL DATA AND DATA STANDARDIZATION 17

1.3. STRUCTURED HEALTHCARE ONTOLOGIES 18

1.4. OBJECTIVES AND HYPOTHESES 21

1.5. SIGNIFICANCE OF RESEARCH 23

II. CHAPTER 2 LITERATURE REVIEW 25

2.1. BIOMEDICAL ONTOLOGIES 25

2.2. PROCEDURAL ONTOLOGIES 29

2.3. SNOMED-CT VOCABULARY 31

2.4. OMOP COMMON DATA MODEL 33

2.5. UTILIZATION OF LARGE DATA SOURCES FOR KNOWLEDGE GENERATION 37

III. CHAPTER 3 METHODS 41

3.1. OVERALL STUDY DESIGN 41

3.2. STUDY POPULATION 43

3.2.1.DATA AND DATABASES 43

3.2.2.INCLUSION CRITERIA 45

3.2.3.EXCLUSION CRITERIA 46

3.3. DATA ANALYSIS 46

3.3.1.SOFTWARE FOR DATA ANALYSIS 46

3.3.2.PROCEDURAL VOCABULARY ASSESSMENT 46

3.3.3.GENERATION OF REFERENCE SET 50
3.3.4.COVARIATE DERIVATION 106
3.3.5.PREDICTIVE MODELS 116
3.3.6.EXTERNAL VALIDATION 118
3.4. ALGORITHM APPLICATION 119
3.5. INTERVENTION TYPE ALGORITHM 121
IV. CHAPTER 4 RESULTS 123
4.1. PROCEDURAL VOCABULARY ASSESSMENT RESULTS 123
4.2. UNIVARIATE STATISTICS OF PARAMETERS 136
4.3. MODEL STATISTICS 142
4.3.1.SINGLE VARIABLE MODEL AUC'S 142
4.3.2.FULL MODEL (DIAGNOSTIC ALGORITHM) 145
4.3.3.FULL MODEL (THERAPEUTIC ALGORITHM) 149
4.3.4.STEPWISE SELECTION (DIAGNOSTIC ALGORITHM) 152
4.3.5.STEPWISE SELCTION (THERAPEUTIC ALGORITHM) 154
4.3.6.LASSO REGRESSION FOR FEATURE SELECTION 156
4.4.FINAL ALGORITHM SELECTION (DIAGNOSTIC AND THERAPEUTIC) . 157
4.5. EXTERNAL VALIDATION (DIAGNOSTIC AND THERAPEUTIC) 161
4.6. ALGORITHM APPLICATION 164
4.7. ALGORITHM DIFFERENTIATION 175
V. CHAPTER 5 DISCUSSION 178
5.1. LIMITATIONS & NEXT STEPS 181
VI. BIBLIOGRAPHY 183
APPENDIX A: NEGATIVE CONTROL LIST 186

LIST OF TABLES

Table 1. Overview of databases 45

Table 2. Positive controls by algorithm type includes SNOMED-CT concept ids and their descendants for both condition and procedures. Source codes counts for conditions (ICD9 Diagnosis and ICD10 Diagnosis) and source code counts for procedures (CPT-4, ICD9Proc and HCPCS) and validation for each condition/procedure pair. 53

Table 4. Confusion matrix for co-occurrence statistics 113

Table 5. Test and train split of dataset for each algorithm..... 118

Table 6. Data available in all databases for condition-procedure mapping..... 119

Table 7. Source code counts by vocabulary and counts of codes by data source..... 123

Table 8. SNOMED-CT mapping to standard vocabulary terms from the vocabulary database..... 124

Table 9. Summary table of the percentage of missing codes for by databases and by vocabulary..... 126

Table 10. Orphan codes by database..... 128

Table 11. SNOMED-CT mapping of descendants and source codes by database 134

Table 12. Parameter statistics for each parameter calculated in Optum (continued)..... 137

Table 13. AUC's for single parameter model's 142

Table 14. Confusion matrix for diagnostic algorithm using test data..... 146

Table 15. True conditions and procedures pairs for diagnostic algorithm 146

Table 16. Confusion matrix for therapeutic algorithm using test data 149

Table 17. True conditions and procedures pairs for diagnostic algorithm 150

Table 18. Stepwise model selection for diagnostic algorithm 152

Table 19. Confusion matrix for diagnostic algorithm using stepwise selection..... 153

Table 20. Stepwise model selection for therapeutic algorithm..... 154

Table 21. Confusion matrix for therapeutic algorithm using stepwise selection..... 155
Table 22. Confusion matrix for therapeutic algorithm using stepwise selection..... 156
Table 23. Diagnostic algorithm performance summary 158
Table 24. Therapeutic algorithm selection summary..... 160
Table 25. Confusion matrix for therapeutic algorithm using stepwise selection..... 161
Table 26. AUC's by database for both diagnostic and therapeutic algorithm..... 162
Table 27 . Overall summary for diagnostic algorithm..... 166
Table 28. Overall summary for therapeutic algorithm..... 167
Table 29. Adjudication results..... 173
Table 30. Overlap of pairs 173
Table 31. Algorithm overlap for p >0.5..... 175

LIST OF FIGURES

Figure 1. Structured knowledge diagram for the condition appendicitis 19

Figure 2. MedDRA hierarchy terms 28

Figure 3. Representation of SNOMED-CT design[37] 31

Figure 5. High level representation of the OMOP common data model[42] 33

Figure 6. CDM 5.0 schema[13] 35

Figure 8. Diagram of full analysis plan by aim 42

Figure 9. Negative control matrix 52

Figure 10. Full diagnostic algorithm positive predictive values < 0.5 for positive controls 147

Figure 11. ROC curve for test data on full diagnostic algorithm 148

Figure 12. Full therapeutic algorithm positive predictive values < 0.5 for positive controls 150

Figure 13. ROC curve for full therapeutic algorithm in Optum Extended SES 151

Figure 14. External validation ROC curves by database 163

Figure 15. Number of SNOMED-CT condition codes mapped in algorithms 165

Figure 16. Number of SNOMED-CT procedure codes mapped in algorithms 165

Figure 17. Distribution of diagnostic algorithm probabilities greater than 0.1 by database 169

Figure 18. Distribution of diagnostic algorithm probabilities greater than 0.1 by database 169

Figure 19. Coverage of codes and data for conditions by database and probability value 170

Figure 20. Coverage of codes and data for procedures by database and probability value 171

Figure 22. ROC curve for step wise selection to determine intervention type 176

Figure 23. ROC curve for LASSO feature selected covariates 177

I. CHAPTER 1 INTRODUCTION

1.1. BACKGROUND

The healthcare landscape has changed greatly in the United States. With the advent of electronic data capture, the healthcare system has evolved to collect data elements from all aspects of patient engagement.[1,2] The increase in data capture has influenced how research is conducted and utilized. Today, many groups are leading efforts to process and utilize the data to help directly impact the healthcare system. Utilizing healthcare data to answer questions about patient safety, utilization and to define patient populations which feed into comparative studies are areas of research that can influence how healthcare is conducted. To conduct these studies, ontologies are necessary to characterize complex and detailed information clearly.

Due to the increase in technology, the types of the health care data that can be collected can vary with the different participants or "players" in the healthcare system. The central participant is the patient, and the subsequent participants are payers/providers, hospitals, and other research entities such as outpatient facilities and laboratories all collect data about a patient. This data can be called real-world data, as it can describe what happens in practice during interactions with the healthcare system. In some cases, this data is unified into a single source such as a claim or a record in an EHR (Electronic healthcare record). EHR systems capture data in real-time and can record many aspects of care received. Items such as arrival time of a patient, vitals taken, diagnosis given, medications prescribed and if any procedures performed.[3] The data are collected by private insurers, EHR providers and the government for billing purposes and

can also be used for research.[4] The amount of data being collected via these processes amounts to millions of data points based on the system via data collection happens.

However, this health care data does come with many limitations and considerations when utilized for research. The data comes in various forms, and some in unstructured text formats; additionally, the format from a private insurer is different from a private practice that may have an EHR system implemented. This makes working with the data and understanding the data very difficult for researchers. Another problem that researchers face is that data may not be standardized for research purposes. There are coding systems in place, such as International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes used in the United States, while systems in the United Kingdom might use READ codes to identify diagnosis.[5] Structured data, which are standardized and share a common meaning can be transcribed upon source data to conducting precise and replicable research.

One advantage of data that is structured with standardized coding systems or terminologies is that researchers can call upon formalized knowledge to help interpret the data. An ontology is based upon relationships of concepts in a specific domain such as healthcare.[6] The primary goal for developing an ontology are to formalize the concepts of domain-specific information, and to allow the reuse and distribution of knowledge.[7]

For example, the types of data that can be found in ontologies are structured definitions around conditions, laboratory procedures and drugs. Resolving differences between these coding systems and performing ‘semantic harmonization’ is an area of

active research and many groups have implemented different approaches to normalize data.[8]

Traditionally, ontologies can be curated through algorithms or manual curation. The type knowledge that are created through manual curation are relationships about two entities such as a drug and an indication. For example, understanding what drugs are indicated for heart failure, the relationship that is being asserted is the indication between two entities in this case the drug and its treatment or indication. This information can be collected manually by curating the product label information and putting the information into an ontology. Health care data can be utilized to assert various relationships, possible reasons why a person is seeking care or why a laboratory test is performed, and this data can be stored in an ontology to be utilized in future research.

Healthcare research has recently adopted techniques such as data mining and modeling to answer some difficult medical questions. These questions are answerable with advanced methods to analyze the data but require the use of ontologies to organize information, utilize relationships to help facilitate good research. The generation of automated ontologies using health care data is an area that can be further developed. By incorporating information from healthcare claims and applying machine learning techniques such as logistic regression, relationships between two entities can be created to increase the amount and richness of the information by providing contextual information that can be utilized in healthcare research.

1.2. OBSERVATIONAL DATA AND DATA STANDARDIZATION

Data used in observational research is routinely and systematically collected for patients.[9] This is in contrast to randomized clinical trials, that data is often limited to the types of data points that are captured, usually only in relevance to the clinical question at hand, and is limited to the number of patients that are captured.[10] The use of the EMR allows for longitudinal data capture and the ability to follow patients for the duration of disease or treatment. Hospital charge level data is another rich source of data because it can have very detailed information about a patient's visit which usually involves treatment for a disease in the form of a procedure that can only be performed in a hospital setting which is line level detail about each hospital transaction.[9] Claims data, which are standardized transactions that happen between the insurance company and the provider (physician, laboratory facilities etc.) can provide data about patients encounters for the care they receive.

There have been many efforts to create a "common language" for all health care data. Many data models have been proposed to standardize data structure into a common format for research purposes. The Sentinel Initiative, an effort led by Harvard Pilgrim for the Food and Drug Administration (FDA) to standardize health care data in a common data format that can be used for research purposes.[11] OMOP (Observational Medical

Outcomes Partnership) was supported in part by the FDA and developed a common data model (CDM) that can house all types of healthcare data.[12] The OMOP CDM also includes an extensive collection of ontologies which are used in the CDM to create consistent mappings for conditions, drugs, procedures and more. For the purpose of this research, the CDM provides the essential components to conduct methodological research and has the ability to harness all types of data.[13]

The data model retains the original data, as well as the mapped data to the various ontologies such as SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) or RxNorm.[13] This enables a researcher to utilize relationships within terminologies to help answer research questions.

1.3. STRUCTURED HEALTHCARE ONTOLOGIES

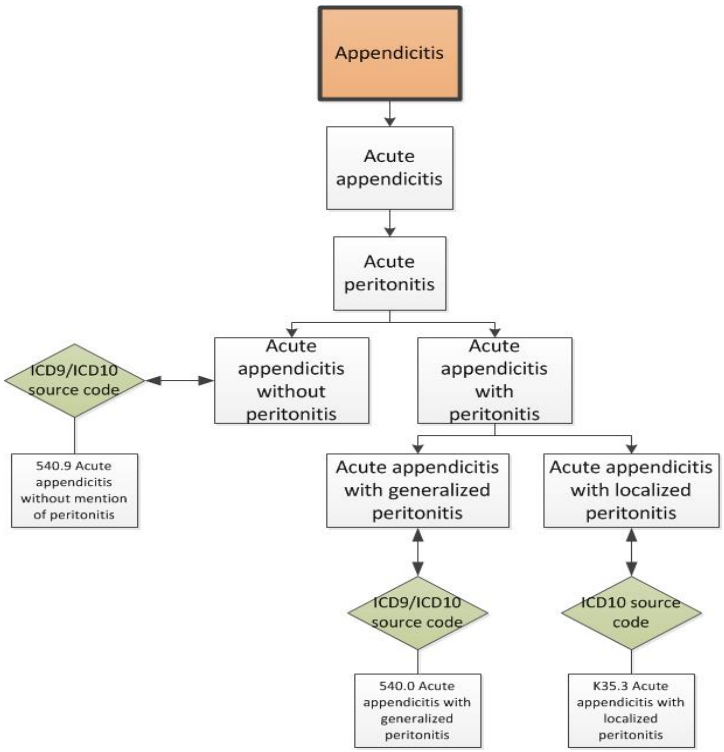
There are many entities that curate ontologies for research purposes. The National Institutes of Health have created an entity, UMLS (Unified Medical Language System), which has a relational database of various medical terms, including diagnosis coding systems, procedures and drug information.[14] The purpose of creating UMLS was to have a central integrated repository of curated knowledge in the form of ontologies that can help answer research questions.[14] The ontologies can then be used to map other source data to those terms which can be useful to increase the breath of coverage for an ontology. A common concern of converting source data to standardized terminologies is loss of data. Reich et al. have evaluated the use of standard coding systems mapped to standardized vocabularies to determine if there was a loss of data with the standardization effort and concluded that the loss was minimal.[15]

Often the creation of new ontologies requires substantial human work but with the availability of various data resources and advanced methods allows for incorporation of knowledge from other sources such as healthcare data.

Structured knowledge can add substantial knowledge to a “flat” data element. For example, understanding that a patient has “acute peritonitis” is informative but one might want to be able to group that information into a higher classification such as “Appendicitis”.

Figure 1 demonstrates the complexity and dimension data points can have by applying structured knowledge on top of existing data. This type of structured knowledge can be useful when conducting analysis or population level analyses.

Figure 1. Structured knowledge diagram for the condition appendicitis



A popular classification system is SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) was developed by combining terminologies by National Health Service (NIH) and the College of American Pathologists (CAP) but has evolved with the collaboration of many individuals from various disciplines.[16] Over the course of a few decades SNOMED-CT has been developed and started with over 120,000 healthcare concepts.[16] The collaborators are made up of nurses, physicians, healthcare professionals, informaticists and others in the United States and the United Kingdom.[16] The collaboration of many individuals and creation of sub-working groups helps provide specific input on the creation of the terminology and ensures quality assurance. [16]

Manual curation has its advantages because it's often thorough and has gone through expert review, but the drawbacks are that it time-consuming and expensive. Data mining "is the analysis of (often large) observational data sets to find unknown relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."[17] Thus, using existing healthcare data such as claims or EHR to create ontologies and relationships has been researched and gaining popularity since knowledge creation and data mining algorithms can be developed and tested with millions of data points to find novel relationships.

The SNOMED-CT terminology has been linked to other terminologies such as LOINC (Logical Observation Identifiers Names and Codes) and ICD9-CM diagnosis codes.[15,18] The SNOMED-CT terminology has been widely used to standardize data within electronic medical record system. The representation of condition codes has been established using many different terminologies such as ICD9-CM diagnosis codes,

MedDRA (Medical Dictionary for Regulatory Activities), SNOMED-CT and many others. Procedures have been standardized through CPT-4 (Current Procedural Terminology), HCPCS (Healthcare Common Procedure Coding System), ICD9-CM procedure codes, SNOMED-CT and others. The principle of having the two very well associated aspects of clinical care (a condition and a procedure) have a relationship within two terminologies is a current gap within healthcare ontologies.

CPT-4 coding terminology is maintained by the AMA (American Medical Association) and is used to describe medical, surgical and diagnostic procedures [19]. The CPT-4 terminology also includes codes to describe services rendered, lab tests, and quality metrics. These codes are mainly utilized for reimbursement purposes. CPT-4 codes exist in three categories, category I, II and III. These categories divide the codes into clusters based on the type of procedure or service rendered. HCPCS codes, which are maintained CMS (Centers for Medicaid and Medicare) are based on one of levels the CPT-4 terminology structure but defined as various levels (I, II, III). [20] The ICD9-CM Procedure terminology is a subset of the full version of ICD9-CM codes [21]. These terminologies are the base of the research and the additional knowledge of relationships that distinguish diagnostic or therapeutic procedures which could be very powerful to help expand how procedures are utilized in research.

1.4. OBJECTIVES AND HYPOTHESES

The intent of the research is to learn from healthcare claims data which conditions and procedure are related to each other. There is a distinction in clinical care between these relationships because some procedures are used to diagnose a condition, while

others are used to treat a condition. Currently, there is no knowledge base that asserts diagnostic and therapeutic relationships between conditions and procedures utilizing healthcare claims.

Medical information (textbooks, expert opinions, clinical evidence) has information on conditions and its corresponding procedures (treatments), but this information is not captured or structured in any ontology. The research aims to utilize observational data (real world data) to construct relationships between conditions and procedures to create an ontology therefore addressing a gap in biomedical science.

The specific aims of this research are:

1. Measure the frequency and coverage of the procedural ontologies (CPT-4, ICD9-CM Procedure codes, HCPCS, SNOMED-CT) and assess any gap in the existing procedural ontologies.

2. Develop an algorithm by using controls and real-world data (claims) and evaluate the performance of the algorithm by AUC for diagnostic and therapeutic ontologies and preform external validation for both the algorithms.

3. Apply the algorithm on the universe of condition-procedure pairs; and evaluate the distribution of probabilities for all condition-procedure relationships found and evaluate the coverage of codes by domain and the amount of data covered from each dataset.

The hypothesis of the overall research is as follows:

Null Hypothesis H_0: Observations or attributes in observational data between conditions and procedures cannot be used to identify the clinical relationship between conditions and procedures

Alternate Hypothesis H_1: Observations or attributes in observational data between conditions and procedures can be used to identify the clinical relationship between conditions and procedures.

1.5. SIGNIFICANCE OF RESEARCH

The research aims to utilize of real-world data to develop algorithms that define the type of relationship (diagnostic or therapeutic) and find condition-procedure relationships. By utilizing real world data to develop a method and algorithm to build relationships from existing ontologies provides a base for additional exploration of other types of relationships (i.e. laboratory tests and conditions) utilizing a similar methodology. The validated knowledge base can also provide inputs for CPOE (computer physician order entry) or EMR systems and can influence how patient care and engagement is conducted.[22] By enabling learning from the real-world data, the results provide current standards of care. Questions about clinical care that we could not have without this relationship such as what procedure is typically administered to those that have a corresponding condition. Therefore, this information can be utilized in observational data research. Cohort generation and phenotyping can utilize this relationship to find patients who have a condition by adding an additional relationship in their cohort derivation. For example, to find a cohort of breast cancer patients we can utilize the diagnosis of breast cancer, but with a condition-procedure ontology we can

utilize the information of a mastectomy or chemotherapy administration to further refine our cohort. By utilizing ontologies, the re-use and sharing of patient level data can occur. Ontologies also provide semantic-based criteria (such as relationships) and can provide support to different statistical aggregations for cohort creation, patient safety reporting and comparative effectiveness reporting. Finally, ontologies provide integration of data and knowledge that can influence how healthcare is conducted and evaluated.

Another benefit is using the part of the information for a proxy. For example, some healthcare data may be limited in the information it has, for example if it only has access to diagnosis data and no coded procedure information, an inference the potential procedures can be made by using the relationship to find procedure counterpart. Lastly benefits in utilizing this information for healthcare utilization, quality of care and healthcare economics research.

Construction of two algorithms; diagnostic and therapeutic are based on the premise that these two relationships are inherently different from one another from utilization to frequency. The diagnosis of condition could require multiple test that vary in the length of time, in which some tests are generalized to many conditions and then more specific tests may be utilized. Procedures that are treatments are also very different as they may only occur once or occur often such as the administration of chemotherapy as well as the total number of specific therapeutic treatments for a treatment of a disease are relatively limited.

Using knowledge to help facilitate meaningful analysis to draw conclusions about utilization of procedures or conditions within real world data as evidence is an integral

piece of this research as it will aid in its application to healthcare research and overall generalizability to various research efforts.

II. CHAPTER 2 LITERATURE REVIEW

The research question will utilize an algorithm to develop condition procedure relationships. Since the area of generating this type of knowledge is extensively diverse, the complexity of performing a complete literature review has its limitations. Below, an overview of the standard procedural vocabularies is described. Additionally, a review of how SNOMED-CT terms are organized, and their creation will be described to provide perspective on the choice of using this ontology to perform research. A review of the CDM and its structure is described as it is the data model utilized for the research. Lastly, key works of automated knowledge generation or ontology creation and studies that utilize predictive models are reviewed to gain an understanding of methods and respective results.

2.1. BIOMEDICAL ONTOLOGIES

The advent of knowledge bases and ontologies has become popular as data and reporting needs have increased.[23] An ontology in healthcare is utilized to represent medical data in a structured manner by representing complex terms into concepts and

allowing for interoperability between systems.[24] Many researchers recognized the need to represent structured knowledge for use within the research community.[25] In health care, the knowledge that needs to be represented is complex and has many layers. For example, a patient and its characteristics, along with the provider's information, the symptoms and conditions that are diagnosed, procedures and test performed, and many other criteria are often stored to conduct research. These events occur at different times and places and may or may not be related to each other, which is important to understand why a patient may be seeking care or receiving a procedure.

The idea of representing medical knowledge in a structured form is a relatively new concept itself in observational research as compared to other disciplines. Cimino condenses how one can start to think about representing clinical knowledge in a structured format, and what should be contained in biomedical terminologies, specifically calling that terminologies should capture what is known about the patient, have the ability to store and retrieve data and the ability to be reused and to aggregate knowledge.[26]

The Unified Medical Language System (UMLS), which was started as a project in 1987 by the National Library of Medicine enables the development and distribution of various knowledge sources and software tools to create, integrate and aggregate biomedical and health information for informatics research.[14] Today, the UMLS houses 100s of different vocabularies in different languages and domains. It has the capability of linking information to aid in reporting health statistics, linking various providers, and different data mining efforts.[27] With the creation of the UMLS, utilizing ontologies has become streamlined and efforts to incorporate them in data systems becoming more

popular. Amongst the vocabularies that are housed in UMLS, some notable derivations and current use are described for the LOINC and MedDRA terminology.

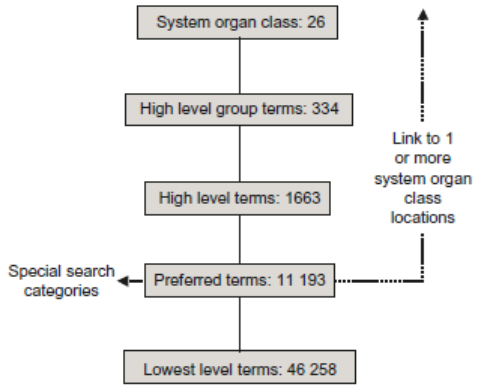
The LOINC vocabulary was developed to capture laboratory tests and have the ability to exchange information between providers and organizations; the goal of the research was to create a structure to codes that could present all aspects of the laboratory measurement.[28,29] The development of the LOINC vocabulary arose from the need to implement a system from the users perspective.[29] By creation of t

his ontology, the inoperability increased between various systems and by improves the management of care. The interesting aspect of this work is that the research utilized various inputs from external sources to create the LOINC ontology.[29] The curation was done manually with the intent to start at the top of the classification and work their way down to more granular terms, the small committee had created more than 10,000 terms in the less than 3 years. The vocabulary itself has gained much popularity as it has been accepted by many organizations and lab vendors and used by most insurance providers to bill for laboratory tests.[29] To date there have been many relationships built from the integration of standard LOINC terms. In 2013 formally LOINC terms began to be mapped into SNOMED-CT terms to help facilitate the use of research.[30]

The MedDRA vocabulary has been used in various studies with different types of data, from adverse event reporting to observational studies. MedDRA, a standardized medical terminology has been developed to assess the needs of regulatory purposes and for pharmaceutical industry.[31] The intent of the MedDRA vocabulary is to cover diagnosis, symptoms, adverse drug reactions, surgical and medicinal procedures and

medical/social history.[32] There are five levels within MedDRA: system organ class, high level group terms, high level terms, preferred terms and lowest level terms.[31]

Figure 2. MedDRA hierarchy terms



The preferred terms within MedDRA allow specific searches on diseases and then link up to higher level categories. This terminology has been used extensively in reporting for potentially safety issues of drug products for the Food and Drug Administration (FDA) in a system called FAERS (FDA Adverse Event Reporting System).[33] MedDRA was designed by ICH (International Conference on Harmonisation) and officially released in its current format in 1999.[31] The MedDRA vocabulary was developed by ICH and includes the following features: updates as required by the needs of the users, new clinically validated terms, developments in responses to new scientific knowledge and advances in the field of medicine.[31]

Understanding the framework of how these two important but distinctly different terminologies of MedDRA and LOINC play an important piece in how knowledge is generated in various ways and its diverse use amongst a community of researchers. Vocabularies have their individual purpose and intent but can also be mapped into existing vocabularies to indicate synonymous meanings. Mapping existing vocabularies to others has been an effort that has been done to create synergy between vocabularies that are intended for different purposes. This type of knowledge integration can help facilitate research and derivations of new knowledge. An analysis done by Reich et al. examines the use of MedDRA, SNOMED-CT and ICD9-CM to evaluate how mapping different vocabularies to data can affect the detection of drug and outcome associations[15]. The analysis focused on mapping source data, ICD9-CM to SNOMED-CT or MedDRA and understanding the coverage of codes and the impact of using alternative vocabularies to represent the same information.[15] This analysis highlights of having integrity and interoperability of vocabularies is important when utilizing this information. In this aspect various terminologies have been "discovered" creating interoperability between sources and vocabularies.

2.2. PROCEDURAL ONTOLOGIES

This research intends to pursue the derivation of relationships between conditions and procedures by utilizing three main procedural vocabularies in the US: CPT-4, HCPCS, and ICD9-CM Procedure codes.

A CPT-4 code describes mainly medical, surgical and diagnostic services that have been performed or rendered.[19] CPT-4 codes have been developed by the American

Medical Association.[34] It was developed to help physicians report procedures to providers, hospitals and insurers to consolidate and initiate payment for procedures. A CPT-4 Editorial panel helps maintain and ensure that all codes are up to date and has an open process to facilitate the addition and revision of codes. [34] CPT-4 codes can be split into multiple categories: Category I, II, III. Category I are composed of medical procedure or services. Category II are composed of performance measurement codes such as physical examination performed. Category III codes describe emerging technology and when data is being collected for Food and Drug Administration (FDA) approval processes.[34] In healthcare practice, utilization of these codes will vary as reimbursement will be dependent on which codes are claimed. For this reason, the use of codes will vary among different institutions such as inpatient hospitals and outpatient providers.

ICD-9-CM Procedure codes is a system that groups procedural into comprehensive categories and was developed and maintained by CMS.[21] These sets of procedure codes are a part of the ICD-9-CM system which gets updated every few years. The codes are mainly used to report inpatient procedures in US hospital systems.[35] The codes that are comprised of procedures are much fewer than CPT-4 codes.

HCPCS codes are separated into two levels, level I is represented by CPT-4 codes, while the level II set of codes was developed for submitting claims to Medicare for procedures and services and devices such as durable medical equipment, orthotics and prosthetics.[36]

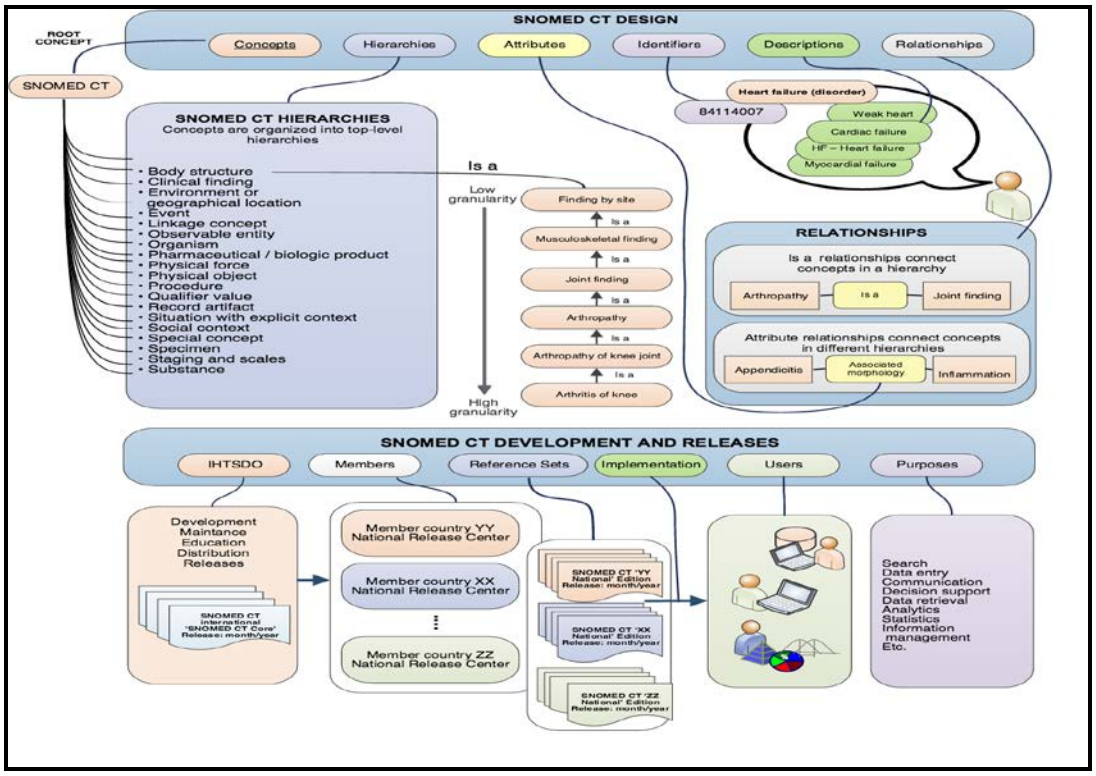
The composition and use of these codes can affect how relationships between conditions and procedures are formed based on their use and frequency in real-world data.

2.3. SNOMED-CT VOCABULARY

The SNOMED-CT vocabulary was first released in January 2003 and versions are updated two times a year.[18] Over the years, vocabularies and their respective relationships have been added to the vocabulary. SNOMED-CT is one of the most comprehensive and multilingual terminologies in the world and utilized in over 50 countries[37]. The vocabulary allows for identification of developing health issues and can be used to monitor population and evaluation of clinical practice. SNOMED-CT can be utilized in EHR systems to support the collection of standardized record keeping of patient data[37]. The EHR data can then be used to support clinical research and contribute to evidence generation.[38]

The SNOMED-CT vocabulary is grouped into three different components: concepts, descriptions and relationships. The total number of concepts available has over 340,000 and are broken down into the following concept categories: clinical finding concepts, procedures, body structures, pharmaceutical or biological products, and organisms. One of the key attributes of the SNOMED-CT vocabulary is the use of the current relationships and hierarchies that are available within the vocabulary. Figure 3 defines the various hierarchies and relationships.

Figure 3. Representation of SNOMED-CT design[37]



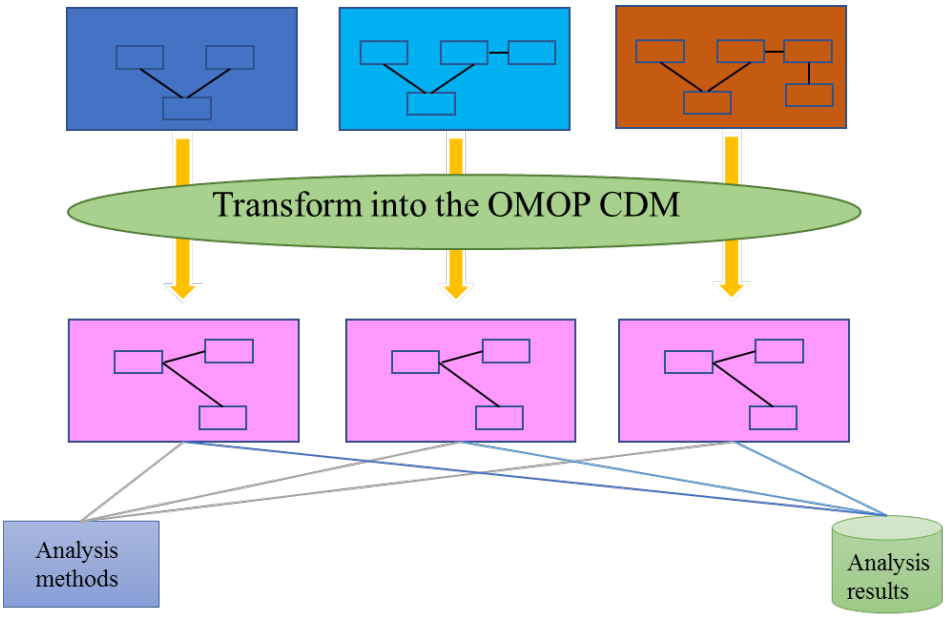
A powerful attribute within the SNOMED-CT terminology is that is that it harnesses semantic properties of relationships within the concepts such as “is-a” or “same as” and uses description-to-concept maps.[16] Thus far most of what has been described here are manually curated and standardized terminologies that are widely used. SNOMED-CT can be used to help aid in cope systems.[39] Lee et al. preformed a systematic literature review to understand the use of SNOMED-CT in published works in PubMed and Embase databases published in 2001 to 2012 and found 488 papers. The literature was then classified into groups describing the type of research utilized with the SNOMED-CT vocabulary.[40] The authors concluded that most studies fall into evaluated the terminology and determining if its use would be appropriate to adopt for their purpose. The main limitation in the study is that they evaluated published literature so many implementations may not be published[40]. SNOMED-CT can be used to support clinical

decision support systems in EHR's or computerized physician order entry systems[41]. Healthcare systems such as Kaiser Permanente utilize the SNOMED-CT vocabulary in its EHR system to provide clinical support for patients.[41] By creating another relationship within the SNOMED-CT vocabulary to determine which procedure is related to which condition this new relationship could be utilized in decision support and to preform meaningful observational research.

2.4. OMOP COMMON DATA MODEL

The OMOP common data model is a person-centric model to house large scale observational data.[42] The tables within the data model define things such as: patients, conditions, procedures, laboratory measurements and their results, and provider information.[13] The CDM has the ability to take various input for data sources and transform them into a common format (Figure 5).[42] The unifying force that keeps the data "talking in the same language" is standardizing to the vocabulary that is utilized as a part of the transformation process.

Figure 4. High level representation of the OMOP common data model[42]

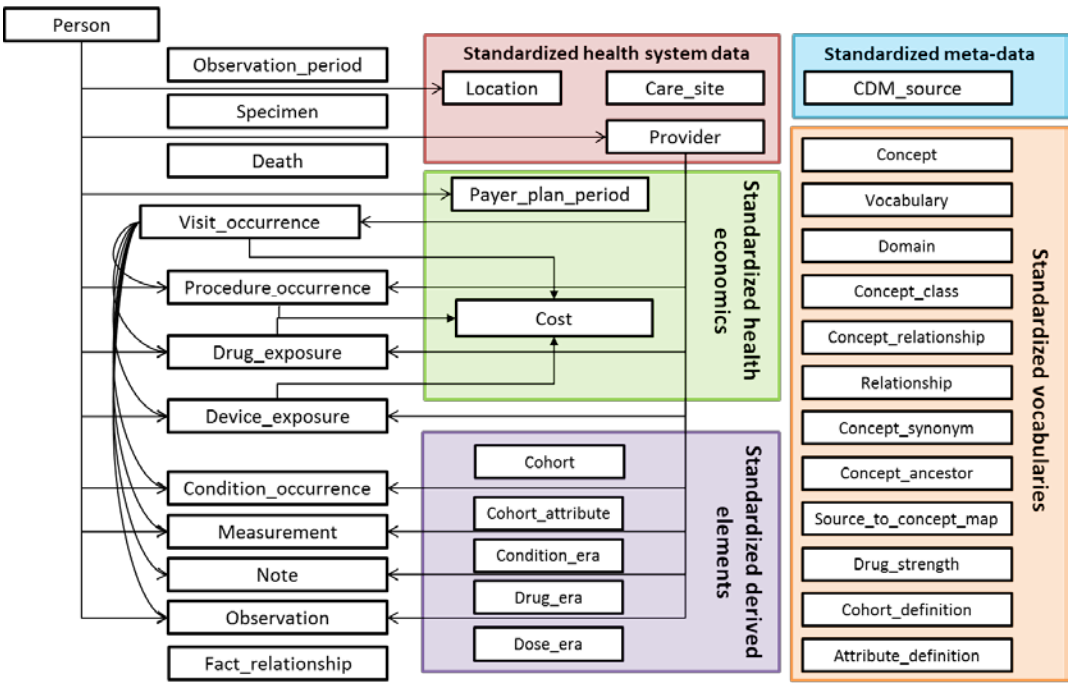


Once the data are transformed into the correct table, the appropriate vocabulary are applied based on the source data type. Not all source data can be mapped into a standard vocabulary, those data elements are preserved and mapped to a standard concept of 0.[13] As part of the CDM implementation, a version of the vocabulary can be obtained with tables that describe concepts, their relationships and source data maps.[13] The current version of the *CONCEPT* table has over 74 distinct vocabularies and over 5 million concepts, many of which have been carried through a period to include codes and concepts that are no longer in use as a method to track evolution of codes and vocabularies.

The 5.2 version of the CDM has 13 core tables that use the person id as the identifier to link the tables together. Information about visits, conditions, drugs, procedures and devices are recorded.[13] Along with this information, observations that occur during a visit can be recorded, such as height or weight. Additional health system

data utilization data can be recorded such as location, site and provider. Vocabularies have hierarchies that can be mapped and those are stored in *CONCEPT_ANCESTOR* table. Each concept can have different relationships to other concepts, for example an ICD-9-CM diagnosis code is related to a SNOMED-CT concept and each of those relationships are stored in the *CONCEPT_RELATIONSHIP* table.[13] Vocabularies and relationships can be added to the concept table for any data source. Most data points can have a place in this schema including information such as survey data, or self-reported data.[13] The figure 6 below is a schematic of the CDM version 5.0 tables.

Figure 5. CDM 5.0 schema[13]



The driver in applying a common data format allows the user to only understand the schema of the CDM and the idiosyncrasies of individual data sources are minimized. There is a perception that there can be loss of data by converting data into a common data model because one is taking many data points and trying to fit them into a pre-specified model. This claim has been refuted by the many number of transformations that have been done with a large variety of data sources. A study by Voss et al. examines the transformation of 6 de-identified large scale observational databases and the loss of data along with the application of CDM data to a standard epidemiological protocol. The databases that were converted ranged from claims data, to hospital and EMR data.[43] Another key component of conversion is that the data quality can be maintained by cleaning up any obvious data issues. For the purposes of the transformation conducted by Voss et al, persons with multiple genders or age changes greater than 2 years were removed from the data. The results showed that all 6 of the databases show more than

89% of records in the databases mapped.[43] Source code mapping does remain to be a big drawback to transforming data within the common data model, but with the advent of new vocabularies and mapping tables to standard terminologies the data can become more operational for research.

Converting various types of databases each has its own limitations but by conversion some of the technical and nuances of the data are eliminated. A study by Makadia and Ryan illustrate the effectiveness of conversion of a hospital billing system, Premier[44]. This data has robust information about a patient's visit and has detailed information about a condition that a person visited the hospital for and any associated procedures that were performed at the time of the visit. One of the biggest drawbacks of the data itself is the extensive source codes that in the database, but the conversion to the common data model allowed for over 90% of utilized codes to be mapped.[44] The advantages of having the ability to run multiple analyses with one set of code can save time and enable faster research and results.

2.5. UTILIZATION OF LARGE DATA SOURCES FOR KNOWLEDGE GENERATION

Knowledge generation with the use of large data sources has been explored greatly in recent years. Large data sources in health care can range from existing ontologies, biomedical literature or knowledge and patient data. The techniques can also vary from hand curation to complex and automated methods for knowledge generation.

A study conducted by Jiang et al utilized MedDRA terms within the AERS system to generate a knowledge based of severe adverse event events that can utilized to

assist in use with current algorithms to detect pharmacovigilance signals.[45] The methodologies applied were to utilize various standardized databases to create an ontological framework to then apply to the AERS dataset and create a linked dataset with serious AERS and outcomes. The conditions were standardized into MedDRA terms, drugs in RxNorm and with the use additional input datasets they were able to create a dataset that had a grading system, drug, ADE (adverse drug event) [45]. Today the AERS system is a main source utilized in drug safety surveillance and that there is a need for the use of clinical data to augment ADE reporting and preform signal detection. An automated method of knowledge generation are methods that require little manual intervention and can greatly speed the process of generation information.

There have been many automated generation techniques that have built upon associations between a disease or condition to drugs, laboratory measures, indications, medications and genetic information. Wang et al describe a classification method to separate noise from signal rather than utilizing arbitrary cut-off points by recognizing that manual curation is an expensive task to create an association between conditions and medications.[46] To boost the disease-medication associations, they utilized a combination of biomedical literature with clinical data. The utilized simple regression models and found that performance of the model was better with the use of clinical data. To test the generalizability of their model they applied these models to predict new diseases.

Wright et al. use association rule mining techniques to find medication-problem and laboratory-problem pairs/linkage for 10,000 patients.[47] The authors used various statistics such as support, confidence, chi square, interest and conviction to compare to a gold

standard.[47] Association rule mining was the best performing method to join both medications and problem lists and laboratory and problem lists.[47] The results of the study found that a high proportion of pairs were found. This technique utilized the input as a patient level model and used various co-occurrence statistics to use in the model which yielded successful results, the authors also used manual review as a part of their validation process.

Another a popular linkage or relationship that is often derived is the medication and indication linkage. A paper by Burton et al looks to address the question of populating a problem list by reviewing medications and indications. They used the Regenstrief data for 1.6 million de-identified patients with the goal to automatically link problems from the problem list to medication orders and dispense records.[48] Medication terms were mapped to RxNorm concepts and diagnosis were mapped to SNOMED-CT terms. The linkage between medications and problems was determined using NDF-RT relationship of "may treat".[48] Sensitivity and specificity were determined by randomly choosing 1,000 drug and indication pairs that were manually reviewed by two Board Certified Internists. Concordance and Kappa coefficients were calculated for both raters.[48] A total of 24, 398 problem and medication pairs were mapped, the concordance was 85.9% with a Kappa coefficient of 0.66 with an overall sensitivity and specificity for the adjusted term pairs at 67.5% and 86.0%. The researchers found that the low sensitivity may be attributed to the experts' perception of "reasonably indicated" and the NDF-RT concept of "may treat" which both have definitions of varying degrees of ambiguity in the definition.

Predictive models provide a valuable source of curating new information. Adverse drug events are something that usually is recorded after the events happen. By utilizing prediction models and utilizing drug safety information, the authors are able to create a model to predict ADE's.[49] A dataset of drug-ADE associations for 809 drugs and 852 ADE's were created, and a logistic regression model applied to predict unknown drug-ADE associations. The model was evaluated using AUC, which was 0.87 indicating that the model was able to predict ADE's. The focus of the work is to demonstrate the use of this type of data network to preform predictive modeling. This network is a collection of data from various sources databases that feed into the model.[49] The authors were able to demonstrate the use of complex data, from various sources to generate new knowledge that can utilized to predict future adverse drug events or generate a relationship between a drug and condition.

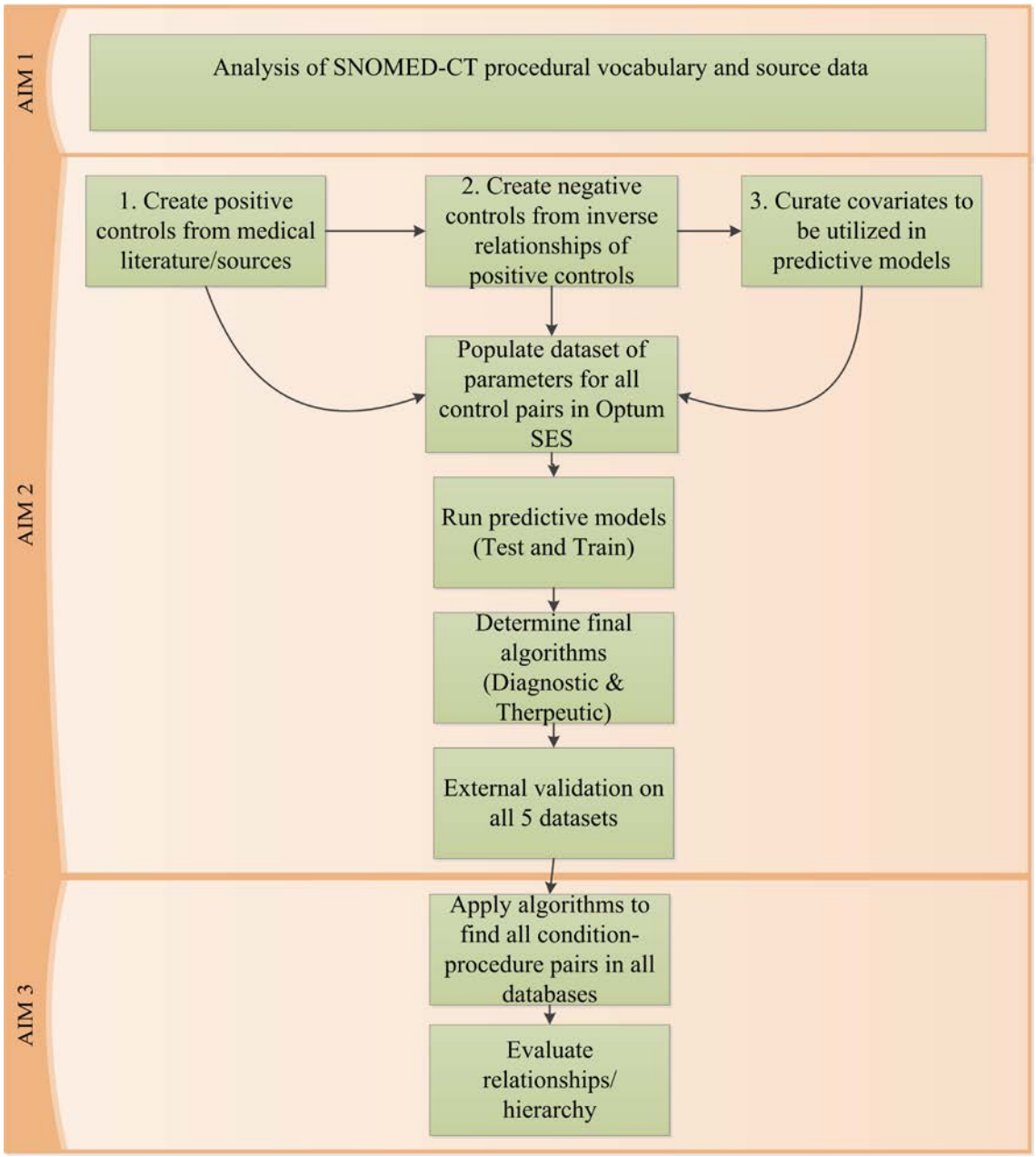
The derivation of relationships or knowledge by utilizing an algorithm is the common theme amongst the works discussed. The methods vary from logistic regression models to natural language processing and data mining techniques. Each work has its limitations and relies on existing data elements that are mapped to standardized terminologies and utilize that knowledge as input into their respective algorithms. The ability to use large datasets and statistical algorithms to generate knowledge is possible.

III. CHAPTER 3 METHODS

3.1. OVERALL STUDY DESIGN

This research aims to create relationships through predictive models on a reference set and preform internal and external validation to evaluate the algorithm's performance. The algorithm will then be applied to all datasets to derived condition-procedure relationships. The final diagnostic and therapeutic ontology will be evaluated for number of SNOMED-CT codes covered, data coverage and number of condition-procedure relationships found. A landscape analysis is conducted to assess the procedural vocabulary (CPT-4, ICD9-CM Procedure, HCPCS) and the mappings to SNOMED-CT. A summary of the overall research plan in shown in Figure 8.

Figure 6. Diagram of full analysis plan by aim



3.2. STUDY POPULATION

3.2.1. DATA AND DATABASES

The data used in the study are be both claims data and hospital charge data in the United States. The IBM® MarketScan® Commercial Claims and Encounters Database (formally Truven) is an administrative health claims database for active employees, early retirees, COBRA continues, and their dependents insured by employer-sponsored plans (individuals in plans or product lines with fee-for-service plans and fully capitated or partially capitated plans). This dataset contains person-specific clinical utilization, expenditures, and enrollment across inpatient, outpatient, prescription drug, and carve-out services. It also includes results for outpatient lab tests processed by large national lab vendors.[43,50]

The IBM® MarketScan® Medicare Supplemental and Coordination of Benefits Database (formally Truven MDCR) is an administrative health claims database for Medicare-eligible active and retired employees and their Medicare-eligible dependents from employer-sponsored supplemental plans (predominantly fee-for-service plans). Only plans where both the Medicare-paid amounts and the employer-paid amounts were available and evident on the claims were selected for this database. The IBM® MarketScan® Multi-State Medicaid Database (formally Truven MDCD) contains inpatient services, prescription drug claims, enrollment, long term care and other medical

care for a pooled population of Medicaid enrollees from multiple states[43,50]. The Premier Healthcare Database (PHD) hospital dataset contains anonymized hospital transactional database from over 500 hospitals from 2000-present day includes inpatient, outpatient and emergency room visits. The database is a visit-oriented database, with each visit having its own unique id. Conditions are coded as ICD9-CM codes and procedures are coded both in ICD9-CM, CPT-4 and HCPCS procedure codes. Drugs, labs, and other procedures are coded as a standard charge code and occur as a transaction.[44] The Optum© De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (SES). An administrative health claims database for members of United Healthcare, who are fully insured in commercial plans or in administrative services only (ASOs) (after May, 2000, 66M) and Medicare (after January, 2006, 6.4M). The data captures person-specific clinical utilization, expenditures, and enrollment across inpatient, outpatient, prescription drug information and results for outpatient lab tests. The Optum population is geographically diverse, spanning all 50 states. It is considered broadly representative of the US population enrolled in commercial health plans. Members maintain their same identifier even if they leave the system.[43,50]

The various databases represent different populations and therefore different types of procedures will be captured in the data. The hospital charge database represents a visit centric database that will mainly contain hospital visits thus having the potential to represent procedures that are administered in the hospital and those diagnoses are captured in the data. The demographic breakdown and sizes of the databases should help yield some variability in the types of conditions and procedures that are available in the data.

The use of OPTUM, PREMIER, IBM data sets (Formally Truven CCAE, MDCD, and MDCR) was reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research.

All of the data used in this analysis has been transformed into the Common Data Model V5.2.[13] The vocabulary version used for the analysis is: VOCABULARY_20171201 (v5.0 01-DEC-17). A brief overview of datasets is shown in Table 1.

Table 1. Overview of databases

	IBM CCAE (formally Truven)	**IBM MDCD (formally Truven)**	**IBM MDCR (formally Truven)**	**OPTUM EXTENDED SES**	**PREMIER**
Patients	135.2MM	25.5MM	9.8MM	79.6MM	163.2MM
Coverage	2000-2017	2006-2016	2000-2017	2000-2017	12/1998-2017
In/Out patient	IN/OUT PATIENT	IN/OUT PATIENT	IN/OUT PATIENT	IN/OUT PATIENT	IN/OUT PATIENT
Type	Claims (mostly private)	Claims (select state Medicaid plans)	Claims (private insurance and Medicare claims)	Claims (private)	Claims
Source	US only	US only	US only	US only	US only
Version	V697	V699	V698	V694	V696

3.2.2. INCLUSION CRITERIA

All person records have valid observations in the years 2011 to September 30th, 2015 are included in the study, all ages are included. All persons must have at least 180 days prior

enrollment from the first condition or procedure. All patient records have at least procedure in their observation period are included in the cohort from the year 2011 to September 30th, 2015. Data does not extend beyond October 1st, 2015 due to the introduction of ICD-10DM codes in the databases and lack of mapping to the SNOMED-CT vocabulary for procedural data. Observation period criteria are not applied to the Premier dataset as that database does not support longitudinal data capture.

3.2.3. EXCLUSION CRITERIA

Patient records that do not have at least 180 days of enrollment prior to the first condition or procedure in the record will be excluded from the analysis, except for the Premier database.

3.3. DATA ANALYSIS

3.3.1. SOFTWARE FOR DATA ANALYSIS

The software used to complete the analysis is R version 3.4.2.

3.3.2. PROCEDURAL VOCABULARY ASSESSMENT

To assess the coverage and mapping of the procedural vocabulary to SNOMED-CT and the source vocabularies: ICD9-CM Procedure, CPT-4, and HCPCS. The assessment will be conducted by analyzing the breadth of the vocabulary and its mapping and quantifying the utilization of codes in all the databases including the amount of missing codes and code capture. Each analysis includes information from both the vocabulary and

utilization in each of five datasets. This analysis quantifies the data we have available to construct condition-procedure pairs through the SNOMED-CT vocabulary.

1. **Source code assessment and utilization by data source**

The count of codes in each procedural vocabulary and the number of codes utilized in the database (all time). All concepts are standard concepts, in the procedural domain and are valid concepts. The count of codes utilized and percentage in each database are calculated.

Statistic	Description
NUMBER OF SOURCE CODES	The total count of distinct terms for each source code for all time in the vocabulary
N	The total count of unique codes utilized all time by database
PERCENTAGE	The percentage of the number of unique source codes divided by the total number of codes in the vocabulary.

2. **SNOMED-CT mapping**

The objective is to understand the mapping from the SNOMED-CT vocabulary to the source codes, the number of codes that are mapped in procedural vocabulary. All concepts analyzed are standard concepts in the procedural domain and are valid concepts. The count of codes utilized with a valid SNOMED-CT equivalent map and percentage are calculated. To understand the ratio of codes mapped to SNOMED-CT terms, an average map ratio, dividing the number of source codes over standard SNOMED-CT

codes to determine how many source codes map to a standard SNOMED-CT concept id. The utilization of the SNOMED-CT mapped terms in the data are evaluated to determine to coverage of the mapped terms in the data. Only valid and standard concepts are evaluated.

Statistic	Description
NUMBER OF SOURCE CODES	The total count of distinct terms for each source code for all time in the vocabulary
NUMBER OF MAPPED SOURCE CODES	The total count of mapped source codes in the vocabulary to a SNOMED-CT equivalent concept
PERCENTAGE	The percentage of the number of mapped source codes divided by the total number of codes in the vocabulary.
NUMBER OF MAPPED SNOMED-CT CODES	The total number of SNOMED-CT codes that are represented by the source codes by vocabulary
AVERAGE MAP RATIO	The number of source codes/SNOMED-CT codes, a ratio of 1 indicates that 1:1 ratio between source code and SNOMED map, a ratio > 1 represents more than 1 source code to every SNOMED-CT concept.

3. Unused and orphan codes

3a. Unused codes

Unused codes are those codes that exist within the vocabulary but are not utilized by 1 or more datasets. When understanding how unused codes from the vocabulary can be viewed, visually confirmation indicates a code is not something that would likely be utilized in a database for research purposes but also understand how frequently it appears in more than one database. The unused codes from each vocabulary and each database are saved into a temp SQL table with a field to indicate the database. A count of codes by

vocabulary and database were evaluated. Manual review of these concepts was also done to ensure that the unused concepts do not amount to a substantial amount of codes and that they do not appear to have significant value to the overall dataset.

Field	Description
COUNT OF CODES	The count of unique codes that are missing from each vocabulary from all databases (Optum, CCAE, MDCD, MDCR, Premier)
PERCENTAGE	The percentage as the number of unused codes divided by the total codes by vocabulary.
DATABASES MISSING CODES	The count of databases that are missing codes, 1 indicates that it is just missing from any one of the database, 5 indicates that is not utilized in of the databases

3b. Orphan codes

The number of "orphan" source codes or those codes that do not have a map to a standard SNOMED-CT concept are captured by database. The percentage of orphan codes is captured by dividing the count for an orphan code in each source vocabulary by the total number of records in the entire procedure occurrence table all time, with no restrictions. The codes that represent more than 1% of the utilized codes are captured, or the top 10 occurring codes by vocabulary and by database.

Field	Description
PERCENTAGE	The percentage as the number of orphan codes divided by the total occurrences in the database by vocabulary.
CONCEPT NAME	The top five concepts by database

4. SNOMED-CT vocabulary hierarchy and code coverage

The SNOMED-CT vocabulary has a procedural hierarchy within its vocabulary. The objective is to quantify the number of source terms within the procedural vocabulary, and the levels of the hierarchy and quantify where source codes are mapped within the SNOMED-CT vocabulary. The distribution of codes will give a perspective of the use of the procedural hierarchy and the number of descendant concepts and the amount of data represented starting from the term “Procedure”. The total number of codes are all records in the procedure occurrence table all time with no restrictions except a valid mapping to SNOMED-CT.

Field	Description
NUMBER OF SNOMED-CT TERMS	The total count of SNOMED-CT procedural terms, standard valid codes are utilized
SNOMED-CT LEVEL FROM PROCEDURE	The name of the SNOMED-CT concept that are the children of the term “Procedure” in the SNOMD vocabulary.
NUMBER OF SOURCE CODES	The number of source codes (CPT-4, ICD9Proc, HCPCS) that have mapped terms to each level of the SNOMED-CT vocabulary.
PERCENTAGE OF CODES UTILIZED IN DATABASE	The percentage of records with mapped source codes utilized by database divided by the total number of codes.

3.3.3. GENERATION OF REFERENCE SET

The reference set that will be utilized as the input for various predictive models is a collection of positive and negative associations of condition-procedure pairs. The compilation of negative and positive controls will serve as the “ground truth” dataset for this research.

Since no gold standard exists in the literature for conditions and procedures, the set used in the experiment will be derived or hand curated from medical references. Medical textbooks will cite how clinicians diagnose problems and have possible procedural information regarding diagnoses and possible treatments. In addition to medical textbooks, medical guides such as WebMD, or Mayo clinic will be utilized to extract information about conditions and related diagnostic procedures or recommended therapeutic procedures.

Those key "concepts" or words for each condition or procedure described are represented as SNOMED-CT terms for both conditions and procedures by searching the vocabulary for the appropriate term. Concepts are specific to the condition and procedure without being generic but also consider the constraints of the vocabulary i.e. codes must appear in the database for Optum SES and have utilization. The concepts must be valid, and all relationships must also be valid. Both algorithms will require distinct lists of positive controls, while negative controls can be shared to be utilized in both algorithms. In the experiment, 100 positive controls are generated for each algorithm. The summary table of positive controls, number of source codes, SNOMED-CT concepts and their descendants are presented in Table 2.

A negative control is generated by inverting the matrix of all controls from the list of positive controls. The product of all combinations results in 32,923 number of new combinations which are all potential candidates for a negative control. A good negative control from a condition and procedure which should have no association, a cross check

against literature will be conducted to ensure that the condition and procedure are not related.

Figure 9 illustrates the transformation of how a negative control is generated from the set of positive controls. For the example below, each one of the pairs has an associated positive control and every other cell in the matrix is a negative control except for any relationship that could be related, such as neoplasms and various treatments or diagnostics such as radiation or magnetic resonance imaging utilized that could apply to both.

Figure 7. Negative control matrix

		Procedures			
		Radiography of ankle	Transfusion of blood product	Magnetic resonance imaging	Radiation oncology AND/OR radiotherapy
Conditions	Sprain of ankle	P	N	N	N
	Hemolytic anemia	N	P	N	N
	Malignant tumor of breast	N	N	D	P
	Neoplasm of pituitary gland	N	N	P	D

Legend: P=Positive control; N=negative control; D=deleted control

Final review from the potential list of negative controls generates 32,132 negative controls (Appendix A)

Table 2. Positive controls by algorithm type includes SNOMED-CT concept ids and their descendants for both condition and procedures. Source codes counts for conditions (ICD9 Diagnosis and ICD10 Diagnosis) and source code counts for procedures (CPT-4, ICD9Proc and HCPCS) and validation for each condition/procedure pair.

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Abnormal cervical Papanicolaou smear	8	5	7	Cervical biopsy	21	15	5	0	Website. Webmd	https://www.webmd.com/women/tc/abnormal-pap-test-topic-overview#2
D	Acute heart disease	128	59	41	Heart disease screening	4	0	0	0	Website. American Heart Association	https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/heart-health-screenings
D	Advanced maternal age gravida	2	4	0	Amniocentesis	8	5	2	0	Bornstein E, Lenchner E, Donnenfeld A, Barnhard Y, Seubert D, Divon MY. Advanced maternal age as a sole indication for genetic amniocentesis; risk-benefit	https://www.ncbi.nlm.nih.gov/pubmed/18999912

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
										analysis based on a large database reflecting the current common practice. Journal of perinatal medicine 2009;**37**(2):99-102 doi: 10.1515/jpm.2009.032[published Online First: Epub Date	
D	Allergic rhinitis	33	6	7	Hyposensitization to allergens	22	18	3	0	Website. Webmd	https://www.webmd.com/allergies/understanding-hay-fever-diagnosis-and-treatment#2
D	Angina pectoris	29	5	46	Echocardiography	56	36	6	6	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	116
D	Anomaly of chromosome pair	365	12	49	Chromosome analysis	8	0	0	0	Website. Lab tests	https://labtestsonline.org/tests/chromosome-analysis-karyotyping

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Benign neoplasm of colon	19	1	10	Screening for malignant neoplasm of colon	1	0	0	0	Website. Center for Disease Control	https://www.cdc.gov/cancer/colorectal/basic_info/screening/index.htm
D	Benign neoplasm of skin	256	17	31	Skin disease screening	6	0	0	0	Website. American Academy of Dermatology	https://www.aad.org/public/spot-skin-cancer/programs/screenings/what-to-expect-at-a-screening
D	Breast lump	581	21	119	Mammography	59	3	5	3	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	603
D	Cerebral infarction	53	13	110	Computerized axial tomography of brain	16	6	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	746

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Chest pain	122	14	20	Plain chest X-ray	32	12	13	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	186
D	Chest pain	122	14	20	Continuous ECG monitoring	1	2	0	0	Website. Harvard	https://www.health.harvard.edu/pain/when-chest-pain-strikes-what-to-expect-at-the-emergency-room
D	Chlamydial infection	61	19	24	Screening for Chlamydia trachomatis	2	0	0	0	Website. Webmd	https://www.webmd.com/sexual-conditions/guide/chlamydia#1
D	Chronic obstructive lung disease	50	9	13	Diagnostic radiography of chest, PA	1	7	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	184

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Chronic pancreatitis	13	1	2	Ultrasonography of abdomen	260	44	6	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	270
D	Cirrhosis of liver	60	4	12	Percutaneous liver biopsy	11	2	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	276
D	Clavicle fracture	40	24	198	Radiologic examination of clavicle, complete	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/broken-collarbone/symptoms-causes/syc-20370311
D	Cor pulmonale	4	1	5	Echocardiography	56	36	6	6	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	194

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Coronary arterioscleros is	44	10	4	Exercise tolerance test	14	5	3	0	Website. Healthline	https://www.healthline.com/health/exercise-stress-test#uses
D	Coronary arterioscleros is	22	5	8	Radionuclide myocardial perfusion study	26	9	0	0	Website. Ottawa Heart Institute	https://www.ottawaheart.ca/heart-condition/coronary-artery-disease-atherosclerosis
D	Crohn's disease	28	5	20	Colonoscopy	20	10	4	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	264
D	Cyst of breast	36	4	8	Diagnostic aspiration of breast cyst	1	3	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/breast-cysts/diagnosis-treatment/drc-20370290

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Cyst of ovary	50	8	14	Imaging of genitourinary system	478	111	36	3	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/ovarian-cysts/diagnosis-treatment/drc-20353411
D	Cystic fibrosis	13	6	7	Cystic fibrosis screening	1	0	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/cystic-fibrosis/diagnosis-treatment/drc-20353706
D	Deep venous thrombosis of lower extremity	37	10	76	Ultrasound scan of lower limb arteries	4	7	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	578
D	Degeneration of lumbosacral intervertebral disc	2	2	2	Imaging of spine	334	79	12	1	Website. Webmd	https://www.webmd.com/back-pain/tc/degenerative-disc-disease-topic-overview#2

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Depressive disorder	165	29	30	Depression screening	4	0	0	1	Website. National Institutes of Health	https://www.nimh.nih.gov/health/topics/depression/index.shtml?
D	Developmental delay	18	2	0	Developmental handicap screening	1	0	0	0	Wesite. University of Michigan Medicine	http://www.med.umich.edu/yourchild/topics/devdel.htm
D	Diabetes mellitus	132	47	203	Diabetes mellitus screening	1	0	0	0	Website. Cleveland Clinic	https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview
D	Diabetic retinopathy	77	9	62	Diagnostic procedure on retina	1	5	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	210
D	Diverticular disease	96	19	45	Colonoscopy	20	10	4	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	628

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Dysphagia	15	8	14	Endoscopy	681	631	107	3	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	250
D	Dyspnea	50	2	5	Plain chest X-ray	32	12	13	0	Website. Webmd	https://www.webmd.com/lung/shortness-breath-dyspnea#2-4
D	Dysuria	3	1	3	Screening for malignant neoplasm of prostate	1	0	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087
D	Eustachian tube disorder	34	18	60	Tympanometry testing	2	3	0	0	Website. Healthline	https://www.healthline.com/health/tympanometry#results
D	Excessive and frequent menstruation	2	2	2	Hysteroscopy	13	11	6	0	Website. Webmd	https://www.webmd.com/women/guide/what-is-hysteroscopy#1

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Female infertility	86	16	12	Transvaginal echography	2	7	0	0	Website. Healthline	https://www.healthline.com/health/transvaginal-ultrasound#purpose
D	Fracture at wrist and/or hand level	345	52	2302	Radiography of wrist	36	5	3	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/broken-wrist-broken-hand/symptoms-causes/syc-20353169
D	Fracture of calcaneus	9	2	281	Diagnostic radiography of calcaneus	1	2	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	544
D	Fracture of fibula	56	19	1269	Tibia and/or fibula X-ray	15	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/broken-ankle-broken-foot/diagnosis-treatment/drc-20355498

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Fracture of humerus	76	34	1402	Radiography of humerus	14	2	0	0	Website. Healthline	https://www.healthline.com/health/humerus-fracture
D	Fracture of rib	47	26	50	Radiography of ribs	15	6	3	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/do-i-have-a-broken-rib#1
D	Galactosemia	5	1	1	Galactosemia screening	1	0	0	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/do-i-have-a-broken-rib#1
D	Glaucoma	111	53	279	Gonioscopy	5	2	0	0	Website. Webmd	https://www.webmd.com/children/gonioscopy#1
D	Glomerulonephritis	134	19	16	Kidney biopsy	31	11	4	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	310

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Hearing loss	158	28	48	Hearing examination	126	53	9	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/hearing-loss/diagnosis-treatment/drc-20373077
D	Heart failure	198	56	80	Echocardiography	56	36	6	6	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	134
D	Heart valve disorder	802	46	110	Cardiac catheterization	23	27	5	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/cardiac-catheterization/about/pac-20384695
D	Hemorrhoids	84	20	26	Anoscopy	1	6	3	0	Website. Webmd	https://www.webmd.com/digestive-disorders/understanding-hemorrhoids-symptoms

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Human papilloma virus infection	50	7	6	Screening for malignant neoplasm of cervix	1	0	0	1	Website. MedlinePlus	https://medlineplus.gov/hpv.html
D	Hyperlipidemia	54	5	6	Hyperlipidemia screening	2	0	0	0	Website. Webmd	https://www.webmd.com/cholesterol-management/hyperlipidemia-overview#1
D	Hypothyroidism	95	7	12	Thyroid disorder screening	2	0	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/hypothyroidism/symptoms-causes/syc-20350284
D	Incontinence of feces	17	3	3	Anal sphincter manometry	1	4	0	0	Website. Cleveland Clinic	https://my.clevelandclinic.org/health/diagnostics/12760-anorectal-manometry/additional-details

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Intracranial aneurysm	34	2	2	Magnetic resonance imaging	675	107	9	27	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/mri/details/why-its-done/icc-20235706
D	Kidney stone	36	4	6	Intravenous pyelogram	5	5	3	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/kidney-stones/symptoms-causes/syc-20355755
D	Laryngitis	75	26	12	Laryngoscopy	43	45	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/laryngitis/diagnosis-treatment/drc-20374267

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Late effects of poliomyelitis	2	1	1	Polio screening	1	0	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/polio/diagnosis-treatment/drc-20376517
D	Low back pain	24	1	5	Intra-articular injection	104	36	9	0	Website. Healthline	https://www.healthline.com/health/facet-arthropathy#diagnosis
D	Malignant lymphoma	713	##	476	Bone marrow biopsy, needle or trocar	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/bone-marrow-biopsy/basics/why-its-done/prc-20020282
D	Malignant neoplasm of uterus	294	10	14	Hysterectomy	139	92	48	0	Website. Webmd	https://www.webmd.com/women/guide/hysterectomy#1

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Meningitis	167	53	47	Diagnostic lumbar puncture	2	2	3	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/lumbar-puncture/about/pac-20394631
D	Missed miscarriage	1	1	1	Ultrasonography of uterus	63	26	3	0	Website. March of Dimes	https://www.marchofdimes.org/pregnancy/ultrasound-during-pregnancy.aspx
D	Mitral valve disorder	310	32	38	Echocardiography	56	36	6	6	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	144
D	Multiple sclerosis	16	1	1	MRI of head and neck	142	27	4	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/multiple-sclerosis/diagnosis-treatment/drc-20350274

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Myopia	20	2	10	Comprehensive eye examination	1	6	4	0	Website. Webmd	https://www.webmd.com/eye-health/nearsightedness-myopia#1
D	Neck pain	23	1	2	Range of motion testing	2	3	3	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/neck-pain/symptoms-causes/syc-20375581
D	Neoplasm of bladder	230	15	18	Transurethral cystoscopy	6	44	3	0	Website. Webmd	https://www.webmd.com/cancer/bladder-cancer/do-i-have-bladder-cancer#1
D	Neoplasm of brain	494	20	38	Computerized axial tomography of brain	16	6	0	0	Website. Webmd	https://www.webmd.com/cancer/brain-cancer/brain-cancer-diagnosis
D	Neoplasm of colon	271	12	28	Sigmoidoscopy	13	11	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press	616

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
										2017.	
D	Neoplasm of esophagus	155	11	9	Fiberoptic esophagoscopy	16	25	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	618
D	Neoplasm of pituitary gland	59	3	6	Magnetic resonance imaging	675	107	9	27	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	234
D	Neoplasm of skin	1192	##	286	Biopsy of skin	27	3	0	0	Website. Healthline	https://www.healthline.com/health/skin-neoplasm#next-steps
D	Neoplasm of stomach	290	13	18	Gastrointestinal tract endoscopy	180	123	32	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	619

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Neoplasm of uterine cervix	144	9	15	Screening for malignant neoplasm of cervix	1	0	0	1	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501
D	Osteoporosis	44	6	5	Bone density study, dual photon absorptiometry	1	5	0	1	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/osteoporosis/diagnosis-treatment/drc-20351974
D	Pain in lower limb	88	2	34	Ultrasonography of lower limb	111	13	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/knee-bursitis/diagnosis-treatment/drc-20355506

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Peripheral neuritis	64	17	19	Electromyography	22	21	6	1	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/peripheral-neuropathy/diagnosis-treatment/drc-20352067
D	Peripheral vertigo	17	10	21	Vestibular function test	23	13	6	0	Website. American Hearing Association	https://www.american-hearing.org/disorders/vestibular-testing/
D	Pneumonia	448	##	166	Plain chest X-ray	32	12	13	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	724
D	Pneumonia	448	##	166	Bronchoscopy	37	31	8	0	Website. Webmd	https://www.webmd.com/lung/interventional-pulmonology-uses-effects#1

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Poisoning	3778	##	3746	Chemical/poison screening	1	0	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	840
D	Polyneuritis	11	2	10	Nerve conduction study	8	12	0	0	Website. Hopkins Medicine	https://www.hopkinsmedicine.org/healthlibrary/test_procedures/neurological/nerve_conduction_velocity_92,P07657
D	Polyp of intestine	55	3	10	Barium enema	12	5	3	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/barium-enema/about/pac-20393008
D	Psychotic disorder	157	78	51	Psychiatric interview and evaluation	10	5	3	0	Website. Webmd	https://www.webmd.com/schizophrenia/guide/mental-health-psychotic-disorders#3-7

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Pulmonary embolism	120	68	40	Pulmonary ventilation study	19	5	0	0	Website. MedlinePlus	https://medlineplus.gov/ency/article/003828.htm
D	Pulmonary embolism	120	68	40	Computerized tomography without IV contrast followed by IV contrast and more sections	1	37	0	0	Website. Webmd	https://www.webmd.com/lung/tc/pulmonary-embolism-topic-overview#2
D	Retinal detachment	79	28	76	Ophthalmoscopy with medical evaluation, extended, for retinal detachment mapping	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/retinal-detachment/symptoms-causes/syc-20351344

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Rheumatoid arthritis	48	3	476	Rheumatoid arthritis screening	1	0	0	0	Website. Spine Health	https://www.spine-health.com/treatment/physical-therapy/manual-physical-therapy-pain-relief
D	Sciatica	14	2	8	Radiography of spine	43	20	7	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/sciatica/diagnosis-treatment/drc-20377441
D	Second trimester pregnancy	15	3	364	Antenatal screening	16	0	0	0	Website. Health of Childern	http://www.healthofchildren.com/A/Antenatal-Testing.html

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Seizure	61	5	8	Electroencephalogram	40	28	5	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/seizure/diagnosis-treatment/drc-20365730
D	Sialoadenitis	28	1	5	Sialogram	22	3	0	0	Website. Healthline	https://www.healthline.com/health/sialogram
D	Sleep apnea	9	4	5	Polysomnography	1	9	3	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/polysomnography/about/pac-20394877
D	Sprain of ankle	21	6	50	Radiography of ankle	7	3	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/sprained-ankle/diagnosis-treatment/drc-20353231

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Subarachnoid hemorrhage	44	20	54	Computerized axial tomography of brain	16	6	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	476
D	Syncope and collapse	2	1	0	Tilt table test	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/tilt-table-test/about/pac-20395124
D	Transient cerebral ischemia	15	4	10	Angiography	1433	177	49	16	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	476
D	Tuberculosis	240	##	68	Tuberculosis screening	1	0	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/tuberculosis/symptoms-causes/syc-20351250

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
D	Ulcerative colitis	18	7	22	Colonoscopy	20	10	4	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	262
D	Urinary incontinence	68	24	24	Ultrasound procedure on urinary AND/OR male genital system	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/urinary-incontinence/diagnosis-treatment/drc-20352814
T	Abscess	531	72	192	Incision and drainage of abscess	63	51	2	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/abscess#1
T	Acquired trigger finger	7	1	18	Release of trigger finger	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/trigger-finger/diagnosis-treatment/drc-20365148

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Actinic keratosis	16	1	1	Destruction of lesion of skin	43	40	0	0	Yeung H, Baranowski ML, Swerlick RA, et al. Use and cost of actinic keratosis destruction in the medicare part b fee-for-service population, 2007 to 2015. JAMA Dermatology 2018 doi: 10.1001/jamadermatol.2018.3086[published Online First: Epub Date]|	https://jamanetwork.com/journals/jamadermatology/fullarticle/2701728
T	Anal fissure	4	2	4	Anal fissurectomy	4	8	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/anal-fissure/diagnosis-treatment/drc-20351430
T	Angle-closure glaucoma	22	6	22	Laser iridotomy	3	2	0	0	Website. Webmd	https://www.webmd.com/eye-health/acute-angle-closure-glaucoma#2

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Anorexia nervosa	10	1	4	Psychotherapy	222	28	19	7	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	759
T	Aortic valve disorder	162	16	17	Aneurysmectomy	63	11	3	0	Website. Medical Dictionary	https://medical-dictionary.thefreedictionary.com/aneurysmectomy
T	Appendicitis	32	6	8	Appendectomy	13	6	7	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	608
T	Arthropathy of knee joint	365	82	122 24	Arthroplasty	414	146	43	1	Website. Webmd	https://www.webmd.com/osteoarthritis/osteoarthritis-knee-replacement-surgery#2

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Arthropathy of spinal facet joint	34	1	0	Facetectomy of vertebra	2	16	0	0	Website. Emedicine	https://emedicine.medscape.com/article/1890471-overview
T	Ascites	29	4	3	Abdominal paracentesis	1	3	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	764
T	Burn of skin	328	##	1182	Skin grafting	164	108	27	0	Website. Webmd	https://www.webmd.com/pain-management/guide/pain-caused-by-burns
T	Callosity	30	1	0	Corn and callus procedures	6	4	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/corns-and-calluses/diagnosis-treatment/drc-20355951

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Cardiac arrest	34	4	26	Cardiopulmonary resuscitation	3	2	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	894
T	Carpal tunnel syndrome	1	1	4	Neuroplasty and transposition of median nerve at carpal tunnel	1	2	0	0	Website. Webmd	https://www.webmd.com/pain-management/carpal-tunnel/carpal-tunnel-syndrome#1
T	Cataract	300	##	278	Cataract surgery	45	7	12	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/cataracts/symptoms-causes/syc-20353790
T	Cataract	300	##	278	Insertion of intraocular lens	21	5	5	0	Website. Webmd	https://www.webmd.com/eye-health/cataracts/default.htm

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Cerebral edema	12	1	2 2	Craniectomy	93	13 7	12	0	Website. Webmd	https://www.webmd.com/brain/brain-swelling-brain-edema-intracranial-pressure#3
T	Cervicitis and endocervicitis	1	1	0	Electrocautery operation	7	7	0	0	Davis C. The cautery treatment of chronic endocervicitis. Journal of the American Medical Association 1926;86(23):1763-65 doi: 10.1001/jama.1926.02670490025010[published Online First: Epub Date]|.	https://jamanetwork.com/journals/jama/article-abstract/241179
T	Chronic hepatitis C	5	1	1	Transplantation of liver	8	9	4	1	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/hepatitis-c/diagnosis-treatment/drc-20354284

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Chronic sinusitis	14	7	8	Frontal sinusectomy	6	2	3	0	Website. Webmd	https://www.webmd.com/allergies/sinusitis-do-i-need-surgery#1
T	Closed fracture of phalanx of finger	42	1	70	Splinting of finger	4	5	0	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/broken-finger#1
T	Congestive heart failure	29	11	16	Automatic defibrillator procedure	33	3	11	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	770
T	Cyst of breast	36	4	8	Excision of cyst of breast	1	2	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/breast-cysts/diagnosis-treatment/drc-20370290

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Cyst of ovary	50	8	14	Excision of cyst	111	79	4	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/ovarian-cysts/diagnosis-treatment/drc-20353411
T	Cystocele	21	2	4	Cholecystectomy	19	15	6	0	Website. Webmd	https://www.webmd.com/urinary-incontinence-oab/bladder-prolapse-surgery#2
T	Degeneration of cartilage AND/OR meniscus of knee	37	11	125	Arthroscopic meniscectomy	12	3	0	0	Website. Webmd	https://www.webmd.com/pain-management/knee-pain/meniscus-tear-surgery#1

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Degeneration of lumbosacral intervertebral disc	2	2	2	Decompression of lumbar spine	12	24	0	1	Website. Webmd	https://www.webmd.com/back-pain/tc/degenerative-disc-disease-topic-overview#2
T	Developmental speech disorder	34	7	14	Speech therapy	224	3	12	2	Webiste. American Speech and Hearing Association	https://www.asha.org/public/speech/disorders/
T	Deviated nasal septum	3	1	1	Nasal septoplasty	17	10	3	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/septoplasty/about/pac-20384670
T	Disorder of acromioclavicular joint	31	4	71	Acromioplasty of shoulder	4	6	0	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/subacromial-smoothing-and-acromioplasty-for-rotator-cuff-disorders

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Disorder of bursa of shoulder region	9	2	14	Arthroscopic shoulder decompression	1	2	0	0	Website. Medicine Net	https://www.medicinenet.com/shoulder_bursitis/article.htm#are_there_home_remedies_for_shoulder_bursitis
T	Disorder of pericardium	176	22	31	Pericardiocentesis	6	5	2	0	Website. Hopkins Medicine	https://www.hopkinsmedicine.org/healthlibrary/test_procedures/cardiovascular/pericardiocentesis_135,361
T	Ectopic pregnancy	23	16	8	Laparotomy	22	16	5	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	606
T	Ectropion	23	6	42	Repair of ectropion	18	5	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press	192

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
										2017.	
T	Effusion of joint	53	22	99	Arthrocentesis	4	6	3	0	Website. Medicine Net	https://www.medicinenet.com/joint_aspiration/article.htm#what_is_the_purpose_of_joint_aspiration_arthrocentesis_and_when_is_it_performed
T	Excessive and frequent menstruation	2	2	2	Endometrial ablation	16	5	3	0	Website. Webmd	https://www.webmd.com/women/endometriosis/what-is-endometrial-ablation#1
T	Female infertility	86	16	12	Intrauterine artificial insemination	5	2	0	1	Website. American Pregnancy Association	http://americanpregnancy.org/infertility/intrauterine-insemination/

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Fetus with chromosomal abnormality	7	2	0	Genetic counseling	3	0	0	0	Website. March of Dimes	https://www.marchofdimes.org/pregnancy/genetic-counseling.aspx
T	Fracture of forearm	144	57	3309	Application of plaster cast to upper limb	15	6	0	0	Website. Hopkins Medicine	https://www.hopkinsmedicine.org/healthlibrary/conditions/orthopaedic_disorders/ulna_and_radius_fractures_forearm_fractures_22,UlnaAndRadiusFractures
T	Gangrenous disorder	120	20	55	Amputation	168	72	35	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	660
T	Gastroesophageal reflux disease	10	1	3	Fundoplication	13	11	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press	624

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
										2017.	
T	Generalized anxiety disorder	1	1	1	Psychotherapy	222	28	19	7	Website. Webmd	https://www.webmd.com/anxiety-panic/guide/anxiety-disorders#2-7
T	Gigantism and acromegaly	9	1	1	Procedure on pituitary gland	52	5	17	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	238
T	Heart failure	198	56	80	Transplantation of heart	9	6	5	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/heart-transplant/about/pac-20384750

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Hemolytic anemia	148	21	28	Transfusion of blood product	52	15	16	1	Website. Webmd	https://www.webmd.com/a-to-z-guides/tc/blood-transfusion-overview#1
T	Hemorrhoids	84	20	26	Hemorrhoidectomy	12	14	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	632
T	Hepatorenal syndrome	5	1	2	Transplantation of liver	8	9	4	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	275
T	Hernia of abdominal cavity	253	58	68	Hernia repair	266	92	60	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	612

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Hidradenitis	9	1	1	Incision and drainage of abscess	63	51	2	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/hidradenitis-suppurativa/symptoms-causes/syc-20352306
T	HPV - Human papillomavirus test positive	6	3	8	Colposcopy	2	11	0	0	Website. Webmd	https://www.webmd.com/cancer/cervical-cancer/do-i-need-colposcopy-and-cervical-biopsy#1
T	Hyperaldosteronism	18	5	9	Adrenalectomy	14	5	6	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	228
T	Impacted cerumen	1	1	5	Ear care	1	2	0	0	Website. Cleveland Clinic	https://my.clevelandclinic.org/health/diseases/14428-ear-wax-buildup--blockage

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Infection of nail	51	4	2	Total avulsion of nail plate	1	3	0	0	Website. Memorial Sloan Kettering Cancer Center	https://www.mskcc.org/cancer-care/patient-education/about-your-nail-avulsion-01
T	Injury of spleen	28	13	31	Splenectomy	12	13	7	0	Website. Webmd	https://www.webmd.com/digestive-disorders/splenectomy#1
T	Intestinal obstruction	143	32	30	Excision of intestinal structure	335	120	80	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	610
T	Intestinal volvulus	9	1	1	Sigmoidoscopy	13	11	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	611

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Kidney stone	36	4	6	Extracorporeal shockwave lithotripsy	23	4	8	1	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/kidney-stones/symptoms-causes/syc-20355755
T	Leukemia	175	##	146	Hemopoietic stem cell transplant	27	6	12	2	Website. Cancer	https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/stem-cell-transplant/types-of-transplants.html
T	Malignant tumor of breast	712	28	162	Partial mastectomy	23	12	7	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	602

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Malignant tumor of breast	712	28	162	Radiation oncology AND/OR radiotherapy	380	129	24	4	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	602
T	Malignant tumor of ovary	124	3	10	Omentectomy	6	17	0	0	Website. Moffitt Cancer Center	https://www.moffitt.org/cancers/ovarian-cancer/omentectomy-orlando/
T	Malignant tumor of pituitary gland	13	1	2	Transsphenoidal hypophysectomy	7	2	6	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	234

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Metatarsus primus varus	1	1	1	Correction of metatarsus varus	1	5	0	0	Zöch K. The surgical treatment of metatarsus primus varus in the adult. Archives of Orthopaedic and Trauma Surgery 1989;108(6):346-48 doi: 10.1007/bf00932443[published Online First: Epub Date]|.	https://rd.springer.com/article/10.1007/BF00932443
T	Mitral valve disorder	310	32	38	Repair of mitral valve	30	4	5	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	144
T	Morbid obesity	3	2	2	Laparoscopic sleeve gastrectomy	1	0	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	626

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Myocardial infarction	83	42	20	Cardiac pacemaker procedure	185	93	45	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	132
T	Neoplasm of lung	429	9	37	Lobectomy of lung	25	12	4	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	176
T	Neoplasm of pancreas	257	11	13	Pancreaticoduodenectomy	6	5	2	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	271
T	Neoplasm of rectum	125	5	12	Radiation oncology AND/OR radiotherapy	380	129	24	4	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	631

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Neoplasm of tongue	141	11	13	Removal of lesion of tongue	4	6	2	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/tongue-cancer/symptoms-causes/syc-20378428
T	Non-Hodgkin's lymphoma	419	82	259	Administration of antineoplastic agent	23	20	6	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/non-hodgkins-lymphoma/diagnosis-treatment/drc-20375685
T	Obesity	51	20	23	Bypass gastroenterostomy	21	24	6	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	626

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Obstructive sleep apnea syndrome	5	1	1	Palatopharyngoplasty	1	2	0	0	Website. Rush University Medical School	https://www.rush.edu/services/test-treatment/palatopharyngoplasty-1
T	Open wound of hand, excluding finger(s)	4	4	0	Simple repair of wounds of extremities	1	7	0	0	Website. Webmd	https://www.webmd.com/first-aid/does-this-cut-need-stitches#1
T	Otitis media	87	43	154	Tympanostomy	2	2	3	0	Website. Emedicine	https://emedicine.medscape.com/article/1890757-overview
T	Pain in thoracic spine	1	1	1	Therapeutic mechanical traction	1	2	0	0	Website. Healthline	https://www.healthline.com/health/spinal-traction
T	Pericarditis	115	18	14	Thoracentesis	8	5	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	192

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Polyp of larynx	3	1	1	Diagnostic endoscopic examination of larynx using rigid instrument	1	2	0	0	Website. Cleveland Clinic	https://my.clevelandclinic.org/health/diseases/15424-vocal-cord-lesions-nodules-polyps-and-cysts
T	Pressure ulcer	54	17	250	Debridement of soft tissue	26	43	3	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/bed-sores/diagnosis-treatment/drc-20355899
T	Primary malignant neoplasm of ovary	82	1	2	Salpingo-oophorectomy	30	39	9	0	Website. Emedicine	https://emedicine.medscape.com/article/1894587-overview
T	Primary malignant neoplasm of prostate	36	1	0	Unilateral orchidectomy	7	4	2	0	Website. Webmd	https://www.webmd.com/prostate-cancer/orchiectomy-surgery

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Rectal prolapse	9	1	2	Rectosigmoid ectomy	6	2	3	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	630
T	Renal failure syndrome	117	30	15	Dialysis procedure	59	13	8	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	306
T	Ruptured aortic aneurysm	11	4	4	Laparotomy	22	16	5	1	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	606
T	Sciatica	14	2	8	Manipulation of spine	19	6	0	0	Website. Webmd	https://www.webmd.com/back-pain/guide/sciatica-symptoms

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Single live birth	3	18	18	Episiotomy	9	10	10	0	Website. Mayo clinic	https://www.mayoclinic.org/healthy-lifestyle/labor-and-delivery/in-depth/episiotomy/art-20047282
T	Single live birth	3	18	18	Cesarean section	20	13	9	0	Website. Mayo clinic	https://www.mayoclinic.org/tests-procedures/c-section/about/pac-20393655
T	Single live birth	3	18	18	Vaginal delivery of fetus	80	9	42	0	Website. Parents magazine	https://www.parents.com/pregnancy/giving-birth/labor-and-delivery/the-stages-of-labor-and-birth-in-a-vaginal-delivery/
T	Skin tag	14	1	1	Excision of skin tag	2	3	3	0	Website. Medical News Today	https://www.medicalnewstoday.com/articles/67317.php

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Spinal stenosis of lumbar region	7	2	1	Foraminotomy	2	11	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/spinal-stenosis/symptoms-causes/syc-20352961
T	Substance abuse	102	47	156	Substance abuse counseling	4	0	3	1	Website. Webmd	https://www.webmd.com/mental-health/addiction/substance-abuse#1
T	Tobacco dependence syndrome	6	1	3	Smoking cessation assistance	1	5	0	0	Website. MedlinePlus	https://medlineplus.gov/quittingsmoking.html
T	Tonsillitis	29	3	11	Tonsillectomy	18	13	5	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/elbow-dislocation#1

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Torsion of testis	3	3	3	Simple orchiectomy with placement of testicular prosthesis by scrotal approach	1	2	0	0	Wilkinson I. *Oxford handbook of clinical medicine*. 10th ed. Oxford: : Oxford University Press 2017.	652
T	Traumatic dislocation of elbow joint	33	13	5 8	Closed reduction of dislocation of elbow	2	4	3	0	Website. Webmd	https://www.webmd.com/a-to-z-guides/elbow-dislocation#1
T	Type 2 diabetes mellitus	18	4	2 4	Nutrition therapy	19	4	0	2	Website. Cleveland Clinic	https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview/management-and-treatment

Algorithm (T or D)	Condition Name	SNOMED-CT Count	ICD9 Count	ICD10 Count	Procedure Name	SNOMED-CT count	CPT-4 Count	ICD9 Proc Count	HCPCS Count	Reference	Page number/ URL
T	Umbilical hernia	20	4	5	Repair of umbilical hernia	27	19	8	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/umbilical-hernia/diagnosis-treatment/drc-20378689
T	Urinary incontinence	68	24	24	Bladder outlet operations	26	8	2	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/urinary-incontinence/diagnosis-treatment/drc-20352814
T	Verruca vulgaris	39	5	5	Diathermy of wart	2	3	0	0	Website. Mayo clinic	https://www.mayoclinic.org/diseases-conditions/common-warts/diagnosis-treatment/drc-20371131

3.3.4. COVARIATE DERIVATION

Covariate derivation to build these algorithms required the transformation of the patient level data within the CDM. The covariates require the normalization of the information to indicate if a relationship is positive or negative between two terms. Most covariates utilized are aggregated by counting the total number of conditions or procedures for those that meet the inclusion criteria. These aggregations can be used to create ratios that gives relation between two amounts (total counts) showing the number of times one value is contained within the other thus inferring a relationship. A ratio is created by taking the total conditions divided by total procedures; a ratio of 1 may indicate that both happen equally which could indicate a plausible relationship. Other methods of derivation of covariates include statistical measures such as relative risk ratios, sensitivity and specificity all derived from a 2x2 table which also imply different types of relationships. The rationale behind the parameters chosen relies on general medical knowledge which recognizes that procedures usually occur at around the time of a diagnosis due to common billing practices in the United States. Reimbursement will not occur unless there a reason or diagnosis to perform a diagnostic or therapeutic procedure.

The qualifying cohort in Optum SES is saved as two datasets, condition and procedure that satisfy the inclusion and exclusion criteria. The data elements that are saved off are: condition code (SNOMED-CT code), procedure code (source code transformed to SNOMED-CT),condition_occurrence_id,person_id,procedure_occurrence_id,

condition_concept_id,procedure_concept_id,procedure_date,condition_occurrence_start_date,condition_type_concept_id,visit_occurrence_id,observation_period_start_date,observation_period_end_date,gender_concept_id, year_of_birth, visit_concept_id. From this dataset 51 covariates are created and split into 3 categories: demographic and count variables, treatment utilization variables and co-occurrence variables for the conditions-procedure pairs in "ground truth" dataset. The covariate names, a description and any relevant formula/logic and justification are included below.

I. Demographic and count parameters

1. ***COND_PROC_RATIO***—the ratio of the total count conditions over the total count of procedures for each condition-procedure pair. A ratio of 1 indicates that the number of conditions and procedures are the same. No restrictions on time, or observation periods. The rationale behind parameter selection is the assumption that acute conditions and procedures occur together and at the same frequency if they are positively related.

$$Cond\_proc\_ratio = \frac{Count\ of\ conditions}{Count\ of\ procedures}$$

2. ***COND_PROC_PERSONS***—the ratio of the count of distinct persons with the condition over the count of distinct persons with the procedures. A ratio of 1 indicates that the number of people with conditions and procedures are the same. No restrictions on time, or observation periods. The rationale behind parameter selection is the assumption that the same persons have conditions and procedures. The same distribution of patients with the condition and procedure should be the same.

$$Cond\_proc\_persons = \frac{Count\ of\ persons\ with\ conditions}{Count\ of\ persons\ with\ procedures}$$

3. ***ABS_FEMALE***—the absolute value of (number of persons identified female with the conditions/total persons with condition) divided by (number of persons identified female/total procedures with procedure). No restrictions on time, or observation periods. The rationale behind parameter selection is to understand if gender plays a role in conditions and procedure pairs; certain conditions and procedures may be gender dependent.

$$Abs\_Female = Abs\left(\frac{\frac{Females\ with\ condition}{Total\ persons\ with\ condition}}{\frac{Females\ with\ procedure}{Total\ persons\ with\ procedure}}\right)$$

4. ***AVERAGE_AGE_RATIO***—The ratio of (average age during the time of the condition/ average age during the time of the procedure). The average accounts for persons with multiple diagnoses or procedures codes over the years. No restrictions on time, or observation periods. The rationale behind parameter selection is that assumption that the distribution of age amongst a true condition procedure pair would be the same when the person gets diagnosed with the condition and when they receive the procedure.

$$\text{Avg\_age }\_ratio = \frac{Average\ age\ condition}{Average\ age\ procedure}$$

5. ***AVG_PER_PERSON***—The number of times the condition or procedure code occurs by person. The ratio (Average number of condition/person) / (Average number of procedure/person). No restrictions on time or observation periods. The rationale behind parameter selection is the assumption that for conditions that

require multiple procedures, such as chemotherapy for neoplasms will have both the condition and procedure coded each time when they receive treatment.

$$Average\_per\_person = \frac{Average\ number\ of\ condition/person}{Average\ number\ of\ procedure/person}$$

6. ***AVG_OBS_TIME_DAYS***—A numeric value of average observation time within observation window, calculated per person (longest per/person) that have both the condition and procedure in the same observation period. If the procedure and condition do not occur in the same observation period window, a value of 0 will be assigned. The rationale behind parameter selection is the assumption that both the condition and procedure occur in the same observation window for positive relationships.

$$Avg\_Obs\_Time\_Days = Avg(Longest\ observation\_period\ length\ (days)\ for\ both\ condition\ and\ procedure)$$

7. ***RATIO_INPT***—The count of conditions that occur in an inpatient setting/ Count of procedures that occur in an inpatient setting. No restrictions on time, or observation periods. The rationale behind parameter selection is that assumption if a condition or procedure are a likely pair then the ratio of the place of the visit will be the same.

$$Ratio\_Inpt = \frac{Count\ of\ conditions\ in\ inpatient\ setting}{Count\ of\ procedures\ in\ inpatient\ setting}$$

8. ***RATIO_OUTPT***—The count of conditions that occur in an outpatient setting/ Count of procedures that occur in an outpatient setting. No restrictions on time, or observation periods. The rationale behind parameter selection is that assumption

if a condition or procedure are a likely pair then the ratio of the place of the visit will be the same.

$$Ratio\_Outpt = \frac{Count\ of\ conditions\ in\ outpatient\ setting}{Count\ of\ procedures\ in\ outpatient\ setting}$$

9. ***NUM_DAYS_BTW***—A numeric absolute value of number of days between condition and procedure (first occurrence of each)/ number of patients with condition. No restrictions on time or observation periods. The rationale behind parameter selection is that the number of days between a condition and procedure affects the relationship, a greater number of days between indicates a negative relationship.

$$Num\_Days\_Btw = abs(number\ of\ days\ between\ condition\ and\ procedure)$$

10. ***AVG_TIME_COND_DAYS***—Average number of times the condition appears before procedure (calculated per person). The rationale behind parameter selection is that the number of times the condition appears before the procedure can indicate that the condition is related to procedure for non-acute condition procedure relationships.

$$Avg\_Time\_Cond\_Days = Avg\left(\frac{Total\ count\ of\ conditions\ before\ procedure}{Total\ with\ condition\ and\ procedure}\right)$$

11. ***COND_PROC_SAME***—The number of conditions and procedures that occur on the same day, requires a person has both the condition and procedure to occur on the same day. The rationale behind parameter selection is that coding practices can influence how diagnosis and procedures are recorded. A procedure may not be billed without an appropriate diagnosis code.

$$Cond\_proc\_same = \frac{Count\ of\ conditions\ -procedure\ pair\ occur\ same\ day}{Count\ of\ total\ condition - procedure\ pair}$$

12. ***COND_PROC_FSAME***—The number of conditions and procedures that occur on the same day but first occurrence of each, requires that a person has both the condition and procedure. The rationale behind parameter selection is that acute conditions need prompt attention and likely resolved as the first diagnosis of the condition appears. This contrasts with conditions in oncology and central nervous system disorders that may require many visits before an appropriate intervention is determined.

$$Cond\_proc\_fsame = \frac{Count\ of\ conditions\ -procedure\ pair\ occur\ same\ day,\ first\ \ occurrence}{Count\ of\ total\ condition - procedure\ pair}$$

13. ***COUNT_PROC_FIRST***—The number of times that the procedure that occur first prior to the condition, requires that a person has both the condition and procedure. The rationale behind parameter selection is to understand the influence of procedures occurring prior to a condition, the expectation is that procedures typically do not occur prior to a diagnosis of a condition.

$$Proc\_first = \frac{Count\ of\ condition\&procedure\ pair\ (procedure\ first)}{Count\ of\ total\ condition\&procedure\ pair}$$

14. ***COUNT_COND_FIRST***—The number of times that the condition occurs prior to the procedure, requires that a person has both the condition and procedure. The rationale behind this parameter is that usually a condition diagnosis is made prior to administering a procedure.

$$Cond\_first = \frac{Count\ of\ condition\&procedure\ pair\ (condition\ first)}{Count\ of\ total\ procedure}$$

15. ***AVG_VISITS_BETWEEN***—the average number of visits between the condition and procedure. The rationale behind parameter selection is to understand the influence of time between a condition and visits regarding visits, a person with chronic conditions may have more visits between before an appropriate procedure determined.

$$Avg\_visit\_ct = Avg(Visits\ between\ condition\ and\ procedure)$$

II. Utilization parameters

The utilization of various components in the CDM allow us to look at drug use, measurements, procedures and conditions all at various points in time. As a patient receives care they should have some stability between the amount of utilization they receive at the time of the condition and procedure.

16., 17., 18. Drug utilization is measured in three time windows, 0-60 days, 61-180 days and 181 to 365 days. The ratio of the count of number of prescriptions in the time window for the condition / the count of the number of prescriptions in the time window for the procedure. The diagnosis of the condition or procedure is time 0. The rationale behind this parameter selection is that the number of prescriptions at the time of diagnosis remain relatively constant at the time of the procedure. RX_60, RX_180, RX_365 are the parameters calculated.

$$Rx = \frac{Count\ of\ rx's\ within\ \textbf{time window in days}\ of\ condition}{Count\ of\ rx's\ within\ \textbf{time window in}\ days\ of\ procedure}$$

19., 20, 21. The count of conditions that occur relative to the condition/procedure pair would be another aspect of a person's care we would assume to have stability between how many conditions occurs around the time that the condition and procedure are

diagnosed. The count of conditions are measured in the three time windows 0-60 days, 61-180 days, and 181-365 days. The time that the condition or procedure occurs is time 0. The three parameters are: COND_60, COND_180 and COND_365. For example, the condition-procedure pair of Neoplasm of breast and mammogram, the occurrence of the condition would be time 0, for each person that had the condition, a count of all conditions that happened 0 to 60 days would be the numerator.

$$Cond = \frac{Count\ of\ conditions\ within\ time\ window\ in\ days\ of\ condition}{Count\ of\ conditions\ within\ time\ window\ in\ days\ of\ procedure}$$

22., 23., 24. The count of distinct measurements within each time period for the condition over the count of distinct procedures within each time period for the procedure. The count of conditions or procedures are measured in the three time windows 0-60 days, 61-180 days, and 181-365 days. The time that the condition or procedure occurs is time 0. The three parameters are MEAS_60, MEAS_180 and MEAS_365.

$$Meas = \frac{Count\ of\ \mathbf{measurements}\ within\ time\ window\ in\ days\ of\ condition}{Count\ of\ \mathbf{measurements}\ within\ time\ window\ in\ days\ of\ procedure}$$

25., 26.,27. The count of number of distinct procedures within time period (SNOMED-CT concept ids) from condition/Count of number of distinct procedures within time period (SNOMED-CT concept ids) from procedure. The three parameters are: PROC_60, PROC_180 and PROC_365

$$Proc = \frac{Count\ of\ \mathbf{procedures}\ within\ time\ window\ in\ days\ of\ condition}{Count\ of\ \mathbf{procedures}\ within\ time\ window\ in\ days\ of\ procedure}$$

III. Co-occurrence statistics

The co-occurrence statistics are based off a 2x2 table for each condition-procedure pair at various time values:

Table 3. Confusion matrix for co-occurrence statistics

	CONDITION (YES)	CONDITION (NO)	TOTAL
PROCEDURE (YES)	Count of person with both condition and procedure (A)	(A+B)-(A)=(B)	The total number of persons with the procedure in dataset (A+B)
PROCEDURE (NO)	(A+C)-(A)=C	(C+D)-C=D	N-(A+B)=(C+D)
TOTAL	The total number of persons with condition in dataset (A+C)	(B+D)	Total number of unique patients from both the condition and procedure dataset (N)

The total (N) is calculated by taking the aggregate number of unique persons in both the condition and procedure cohort. The (A+C) is the total number of persons with the condition in the cohort, and (A+B) is the total number of persons with the procedure in the cohort. The value that changes in the confusion matrix is the time frame of when the condition and procedure occur or (A) from Table 3. If the absolute value of the time between the condition start date and procedure is 0, then it is considered time 0, if the absolute value is between 1 and 90 then it is considered time 90, if the absolute value is between 91 and 365 then it is considered time 365. If at any time the condition and procedure occur together we consider that all time. The remainder of the numbers are calculated from the equations in Table 3. The six statistics are relative risk, odds ratio, support, misclassification, sensitivity and specificity.

28., 29., 30., 31. Relative risk is the measure of association between a treatment and an outcome [51]. In the context of determining if a condition or procedure are related, we can

assume the treatment is the procedure and the outcome is the condition. A relative risk score of 1 means that there is no relationship between condition and procedure, a value less than one means that risk of having the procedure is decreased by the condition and a value greater than one means that the risk of having the procedure is increased by the condition. Relative risk is calculated at time 0, time 90, time 365 and all time.

$$Relative\ risk = \mathrm{A}/(\mathrm{A}+\mathrm{B})/\ \mathrm{C}/(\mathrm{C}+\mathrm{D})$$

32., 33., 34., 35. Odds ratio are the ratio of the odds of an event in the treatment group compared to the odds in the control group. The odds of an event are the ratio of events over the number of non-events. The values will be calculated at time 0, time 90, time 365 and all time. The odds ratio is calculated as:

$$Odds\ Ratio = (\frac{A}{B})/(\frac{C}{D})$$

36., 37., 38., 39. Support are the count of the co-occurrence of condition and procedure. The count will be a large or small number depending on how frequently the two items co-occur. The values will be calculated at time 0, time 90, time 365 and all time.

$$Support = A$$

40., 41., 42., 43. Misclassification is the rate of all false or incorrectly identified pairs over the total. The ratio will be small if there are less incorrectly classified while it will be large for pairs that have more misclassified pairs. The values will be calculated at time 0, time 90, time 365 and all time.

$$Misclassification = B + C\ /(A + B + C + D)$$

44., 45., 46., 47. Sensitivity or the true positive rate is the proportion of actual positive relationships found. The ratio will be between 0 and 1. The values will be calculated at time 0, time 90, time 365 and all time.

$$Sensitivity = A/(A + C)$$

48., 49., 50., 51. Specificity or the true negative rate is the proportion of negatives that are correctly identified. The ration will be between 0 and 1. The values will be calculated at time 0, time 90, time 365 and all time.

$$Specificity = B/(B + D)$$

3.3.5. PREDICTIVE MODELS

The outcome we are interested in is classification of a negative or positive value, thus binary models are appropriate. The 3 types of regression models are utilized to generate the algorithm: logistic (single variable models and full model), step-wise and LASSO (utilized only for covariate selection). The logistic regression model will return the logit(p) which is transformed to provide a probability from 0 to 1. Values close to 1 indicate a probability of being associated, or pairs that are positive relationships. The equation for a logistic regression model is below[52]:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

The formula to transform the logit(p) to a probability is:

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Stepwise regression uses a combination of forward selection and backward selection to add variables to the regression model, and then test if all the variables are significant and if not, they are removed. Variables are entered into the model based upon a probability or cut-off value and before another variable is added or entered the significance of the model is checked. A null model will be entered into the model and then variables will be added or removed based upon Chi-square test. A model with significant variables and a low Akaike information criterion (AIC) will be selected as the final model using the stepwise procedure.[53] AIC criterion is an estimator of quality of statistical models used in the step-wise selection model.[53] LASSO a regularized regression technique that is primarily used for feature selection technique that will shrink maximum likely hood ratio to zero thus eliminating the variable. The regression technique will be utilized only to understand feature selection and find significant coefficients.[54] The purpose for utilizing LASSO is mainly to evaluate which covariates are significant, 51 covariates results in a complex model that would be difficult to interpret and utilize in the practice.

Using these regression techniques to train various models in the following order:

1. Single covariate logistic regression models for diagnostic and therapeutic algorithms for all 51 variables.
2. Full model using all 51 covariates with GLM in R
3. Stepwise regression
4. LASSO model

The predictive models are developed using “test-train” 80:20 split cross validation method.

Table 4. Test and train split of dataset for each algorithm

Dataset	Positive	Negative
Test	20	6,506
Train	80	25,626

The goal is to build an algorithm that has good performance by a high AUC and has variables that are generalizable to other datasets. An AUC of 0.5 would indicate that the parameter by itself is predictive the same random guessing while a lower AUC indicates that the regression model does worse than random guessing[52].

A final algorithm that is parsimonious would be advantageous as applying the algorithm to large variety of healthcare data would become less labor intensive and easier to understand. Additional statistics such as sensitivity, specificity, AIC and the confusion matrix will be reported for selected models to provide evidence for model selection but not utilized in selection of the final algorithm.

3.3.6. EXTERNAL VALIDATION

Generalizability is important when evaluating how well a algorithm preforms and can discriminate positive and negative relationships when applied to other datasets not used for developing the model.[55] Final algorithms are applied to the IBM CCAE, MDCD and MDCR and Premier dataset along with the Optum dataset. A dataset will be created by data source that generates the positive predictive values for all 32,323 condition-procedure pairs. The AUC’s will be calculated for each algorithm for evaluation.

The final algorithm AUC's will be compared against the database the algorithm was trained on to conduct external validation.

3.4. ALGORITHM APPLICATION

The final algorithm will be applied on the universe of condition-procedure codes by calculating the positive predictive values (PPV) for each pair. Table 6, shows the available data in each database and number of possible pairs to be calculated starting from 2011 to September 30th, 2015.

Table 5. Data available in all databases for condition-procedure mapping

	Database					
	Vocabulary	Optum SES	IBM CCAE	IBM MDCD	IBM MDCR	Premier
Number of SNOMED-CT condition codes	104,310	13,130	11,760	12,843	10,800	13,402
Number of SNOMED-CT procedure codes	13,939	12,515	12,257	12,205	12,400	9,963

The positive predictive values will be plotted by database for all condition-procedure pairs. The number of codes covered, percentage of data by database will be evaluated.

1. Coverage of condition and procedure codes from overall database

The algorithm results produce a set of condition-procedure pairs from the database, in which some condition and procedure concepts will not be mapped. An evaluation of the number of condition and procedure concepts mapped, and percentage of the data by database will be evaluated for both the procedure and condition domain.

2. Overall summary by algorithm type and probability

The overall summary will count the number of pairs found by database and for each algorithm type. A distribution will be generated for codes by database for probabilities between 0.1 and 1.0 for each algorithm.

3. Coverage of data and codes by domain and probability

The percentage of SNOMED-CT codes covered by domain are counted by probability values ($p > 0.5$, $p > 0.6$, $p > 0.7$, $p > 0.8$, $p > 0.9$) for conditions and procedures. Using those codes we calculated the percentage of data captured at each threshold and the number of codes that are encompassed within each probability bin.

4. Adjudication of condition-procedure pairs

The evaluation of how many codes fall into each probability bin do not account for the quality or integrity of the found pairs, a review of the top 100 codes by algorithm and database to classify if values are a good match, a possible match or not a match. The criteria to classify the pairs are if from common knowledge or from the text they belong together the pair would get a yes, if they refer to two very distinctly different items, the pair would be designated as a no, and if no determination can be made then it would be classified as a maybe.

5. Overlap of pairs by database

By completing the exercise of external validation on our originally developed algorithm, the determination was made that the algorithm is generalizable to more than the database

it was developed on, keeping that principle in mind we can test if our pairs exist in more than one database at the same probability bin.

6. Overlap of pairs in all databases by probability

The algorithm was produced on the same dataset but with different parameters thus, the algorithms could select the same pairs and classify them similarly or differently. The evaluation of the overlap of algorithms by the number of pairs that occur in a single algorithm, both and the percentage of overlap are calculated.

7. Overlap of pairs within algorithms

3.5. INTERVENTION TYPE ALGORITHM

The original intent of the analysis was to develop two algorithms, due to the understanding that there may inherent differences between how diagnostic relationships are developed compared to therapeutic relationships. The covariates were similar between the two algorithms and inspection of the condition-procedure pairs in both algorithms show great amounts of overlap between the two algorithms. This finding let to the exploration of a two-staged approach which would first find all condition-procedure relationships and then apply a secondary model to determine the intervention type, diagnostic or therapeutic. To test and explore the second hypothesis of the ability to classify therapeutic or diagnostic condition-procedure pairs, the positive controls were used to generate another algorithm. The positive controls from that were calculated from the Optum dataset (100 therapeutic and 100 diagnostic) were recoded into another dataset of 200 pairs total, 0 indicating a diagnostic pair and 1 indicating a

therapeutic pair. Then a similar split of 80:20 for train and test was applied. The following models were run to determine a final algorithm to determine intervention:

1. Full model with 51 variables
2. Step wise procedure
3. LASSO regression for feature selection
4. Logistic regression with selected covariates

The primary measure to calculate how well the model does will be the comparison of the AUC values among the 4 models.

IV. CHAPTER 4 RESULTS

4.1. PROCEDURAL VOCABULARY ASSESSMENT RESULTS

The purpose of analyzing the procedural vocabulary is to assess the overall vocabulary structure and determine if SNOMED-CT as a vocabulary would be a viable choice in moving forward with the logistic regression with SNOMED-CT as terms used to build relationships.

1. Source code assessment and utilization by data source

Table 6. Source code counts by vocabulary and counts of codes by data source

		Optum SES		**Truven CCAE**		**Truven MDCD**		**Truven MDCR**		**Premier**	
		Utilized codes		Utilized codes		Utilized codes		Utilized codes		Utilized codes	
Vocab-ulary	Total codes	N	%	N	%	N	%	N	%	N	%
CPT-4	8524	8002	93.8	8008	93.9	7866	92.2	7835	91.9	7811	91.6
HCPCS	764	626	81.9	597	78.1	448	58.6	559	73.1	320	41.8
ICD9DM-Proc	4651	4636	99.6	3950	84.9	3879	83.4	3857	82.9	4042	86.9
SNOMED-CT	48203	248	0.51	240	0.50	239	0.50	240	0.50	30	0.06

The code coverage by database for CPT-4 codes is above 90% for all databases, therefore more than 90% of the CPT-4 codes are being used in the data. The code coverage by database for ICD9-CM Procedure results in over 80% of the overall code library being utilized. The use of HCPCS varies greatly amongst each of the 5 databases ranging from 41.88% to 81.94%. The use of SNOMED-CT terms to represent a procedural concept

value is very low, about .50% in the 4 claims database, and 0.06% in Premier the hospital dataset. Overall, from the amount of procedure codes available most of codes are utilized in the data except for SNOMED-CT codes. The CDM convention does allow any of the four to represent the standard concept for procedures. By analyzing the universe of codes available in the vocabulary and mapped we understand that the depth and breadth of the vocabulary must account for a large volume of codes.

2. SNOMED-CT mapping

Table 7. SNOMED-CT mapping to standard vocabulary terms from the vocabulary database.

Source Codes Mapped	Total Codes	Percent Mapped	Relationship Id	Mapped SNOMED codes	Average map ratio
7143	8524	83.80%	CPT-4 – SNOMED-CT eq	6227	1.147
247	764	32.33%	HCPCS – SNOMED-CT proc	183	1.350
4466	4651	96.02%	ICD9P – SNOMED-CT eq	3411	1.309

The relationship will allow us to utilize the relationships that have already been created in SNOMED-CT to the three source vocabularies (CPT-4, ICD9-CM Procedure, and CPT-4). For the purposes of understanding the vocabulary, and to utilize the hierarchical nature of SNOMED-CT, the controls and eventual relationship map will pair two SNOMED-CT codes to each other. Amongst the procedural source codes, the map to SNOMED-CT represents over 80% of the CPT-4 codes having a map to a standard SNOMED-CT term, 32.33% of HCPCS codes have a map to a standard SNOMED-CT concept and over 96% of ICD9-CM Procedure codes have a map to standard SNOMED-

CT code in the vocabulary. The amount of codes that are represented by SNOMED-CT vary by code type, and the average map ratio for each code type is over 1.1 indicating that more than 1 standard procedural code will map into a single SNOMED-CT concept. To understand how many mapped codes are utilized, the counts/percentages by database are shown below for the codes that have do not have a map into SNOMED-CT, this will allow us to assess the gaps where there is no relationship between a SNOMED-CT procedure code and source code.

3. Un-utilized SNOMED-CT codes and orphan codes

3a. Unused codes

From Table 6, there is evidence that there are some source codes that aren't utilized in the database to varying degrees. For CPT-4 codes, we see approximately 92% to 94% utilization of codes, ICD9-CM Procedure codes approximately 83% to 99% codes utilization and HCPCS from 42%, and from Table 6 we are aware that range of mapped codes varies from 32.33% and 96.02% which leaves two gaps, gaps of codes that don't ever get utilized and are they not utilized across the databases or are they only not utilized in a single source. Second, we understand that SNOMED-CT has some gaps amongst mapping source codes to standard SNOMED-CT concepts which we will identify as "orphan codes". Amongst the codes that are not utilized in any given database, the summary of how many databases do not show utilization in the data source by source code and number of databases. For CPT-4 codes, the number of unique codes amongst all the databases is 924, and 52.81% of the codes don't appear in all 5 databases, while there are 32.25% of codes that missing from 1 database but appear in the other 4. For HCPCS

codes, the total number of unique codes is 496 and 20.36% of codes are missing from all 5 databases and 35.08% are missing from only 1 database. For ICD9-CM Procedure codes the total number of unique codes is 871, 64.18% of codes are missing from 4 of databases but appear in one of the databases.

Table 8. Summary table of the percentage of missing codes for by databases and by vocabulary

Vocabulary	Count of codes	Percentage	Databases missing codes
CPT-4	298	32.25%	1
CPT-4	80	8.66%	2
CPT-4	32	3.46%	3
CPT-4	26	2.81%	4
CPT-4	488	52.81%	5
HCPCS	174	35.08%	1
HCPCS	115	23.19%	2
HCPCS	63	12.70%	3
HCPCS	43	8.67%	4
HCPCS	101	20.36%	5
ICD9Proc	133	15.27%	1
ICD9Proc	33	3.79%	2
ICD9Proc	137	15.73%	3
ICD9Proc	559	64.18%	4
ICD9Proc	9	1.03%	5

Orphan codes or those codes that have no map to a SNOMED-CT procedure concept are represented in Table 9 by database, and by source code vocabulary. The codes that represent at least 1% of the total data are recorded and included the top 10 from each database is represented below. The most frequent orphan codes in the CPT-4 vocabulary in all databases are those that define office visits making up 12.05% of Optum, 13.19% of CCAE, 10.92% of MDCD, 11.39 of MDCR and 0.80% of Premier. The codes that are not mapped to a SNOMED-CT procedure concept are those that are more administrative or

non-specific in nature that a corresponding condition pair would be difficult to address. The same observation is made for HCPCS codes and ICD9 procedure codes.

Table 9. Orphan codes by database

		Optum SES (n=2,778,686,534)	Truven CCAE (n=4,271,516,644)	Truven MDCD (n=1,113,264,671)	Truven MDCR (n=1,004,499,958)	Premier (n=3,146,944,195)
Code Type	**Concept Name**	**Percent**	**Percent**	**Percent**	**Percent**	**Percent**
CPT-4	Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low	12.05%	13.19%	10.92%	11.39%	0.80%
	Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity.	7.45%	6.72%	5.59%	8.06%	0.33%
	Manual therapy techniques (eg, mobilization/ manipulation, manual lymphatic drainage, manual traction), 1 or more regions, each 15 minutes	2.11%	2.71%	0.24%	1.71%	0.52%
	Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A problem focused history; A problem focused examination; Straightforward medical decision	1.95%	2.21%	2.20%	2.14%	0.59%

	making.					
	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	1.57%	1.69%	0.94%	0.82%	0.10%
	Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: An expanded problem focused interval history; An expanded problem focused examination;	1.28%	0.72%	2.08%	3.32%	0.06%
	Application of a modality to 1 or more areas; electrical stimulation (unattended)	0.65%	1.29%	0.06%	0.19%	0.18%
	Application of a modality to 1 or more areas; electrical stimulation (manual), each 15 minutes	0.32%	0.42%	0.04%	0.17%	0.05%
	Spirometry, including graphic record, total and timed vital capacity, expiratory flow rate measurement(s), with or without maximal voluntary ventilation	0.18%	0.18%	0.13%	0.15%	0.08%
	Removal impacted cerumen requiring instrumentation, unilateral	0.11%	0.10%	0.09%	0.16%	0.01%

	Subsequent nursing facility care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: A problem focused interval history; A problem focused examination; Straightforward medical decision making.	0.07%	0.00%	0.17%	0.16%	0.00%
	Inpatient consultation for a new or established patient, which requires these 3 key components: A comprehensive history; A comprehensive examination; and Medical decision making of high complexity. Counseling and/or coordination of care with other physician	0.06%	0.07%	0.08%	0.20%	0.00%
	Therapeutic procedure, 1 or more areas, each 15 minutes; aquatic therapy with therapeutic exercises	0.04%	0.05%	0.04%	0.07%	0.05%
	Diffusing capacity (eg, carbon monoxide, membrane) (List separately in addition to code for primary procedure)	0.03%	0.02%	0.03%	0.05%	0.03%
	Ultrasound, breast, unilateral, real time with image documentation, including axilla when performed; limited	0.03%	0.03%	0.02%	0.01%	0.04%
	Critical care, evaluation and management of the critically ill or critically injured patient; each additional 30 minutes (List separately in addition	0.02%	0.01%	0.03%	0.03%	0.01%

	to code for primary service)					
	Attendance at delivery (when requested by the delivering physician or other qualified health care professional) and initial stabilization of newborn	0.00%	0.01%	0.04%	0.00%	0.01%
	Opioids and opiate analogs; 1 or 2	0.00%	0.00%	0.03%	0.00%	0.00%
	Pregabalin	0.00%	0.00%	0.03%	0.00%	0.00%
	Alcohols	0.00%	0.01%	0.04%	0.00%	0.04%
	Fresh frozen plasma, thawing, each unit	0.00%	0.00%	0.00%	0.00%	0.04%
HCPCS	Travel allowance one way in connection with medically necessary laboratory specimen collection drawn from home bound or nursing home bound patient; prorated trip charge	0.11%	0.00%	0.03%	0.04%	0.06%
	Screening papanicolaou smear; obtaining, preparing and conveyance of cervical or vaginal smear to laboratory	0.10%	0.11%	0.04%	0.09%	0.00%
	Travel allowance one way in connection with medically necessary laboratory specimen collection drawn from home bound or nursing home bound patient; prorated miles actually travelled	0.08%	0.00%	0.04%	0.02%	0.01%

	Direct skilled nursing services of a registered nurse (rn) in the home health or hospice setting, each 15 minutes	0.06%	0.01%	0.04%	0.03%	0.00%
	Influenza virus vaccine, split virus, when administered to individuals 3 years of age and older, for intramuscular use (fluvirin)	0.04%	0.01%	0.02%	0.05%	0.00%
	Influenza virus vaccine, split virus, when administered to individuals 3 years of age and older, for intramuscular use (fluzone)	0.04%	0.01%	0.01%	0.05%	0.00%
	Drug test presump not opt	0.03%	0.02%	0.07%	0.00%	0.01%
	Hit antibiotic q24h diem	0.02%	0.02%	0.00%	0.02%	N/A
	Trimming of dystrophic nails, any number	0.02%	0.00%	0.01%	0.04%	0.00%
	Drug test def 1-7 classes	0.02%	0.01%	0.02%	0.00%	0.01%
	Mental health service plan development by non-physician	0.01%	0.01%	0.10%	0.00%	0.00%
	Administration of oral, intramuscular and/or subcutaneous medication by health care agency/professional, per visit	0.00%	0.00%	0.07%	0.00%	N/A
ICD9Proc	Other fetal monitoring	0.01%	0.01%	0.03%	0.00%	0.08%
	Non-invasive mechanical ventilation	0.00%	0.00%	0.01%	0.01%	0.03%

	Other physical therapy	0.00%	0.00%	0.00%	0.00%	0.02%
	Other diagnostic procedures on fetus and amnion	0.00%	0.00%	0.00%	0.00%	0.01%
	Insertion of indwelling urinary catheter	0.00%	0.00%	0.00%	0.01%	0.02%
	Fetal EKG (scalp)	0.00%	0.00%	0.01%	N/A	0.01%
	Occupational therapy	0.00%	0.00%	0.00%	0.00%	0.02%
	Open and other right hemicolectomy	0.00%	0.00%	0.00%	0.00%	0.01%
	Other active musculoskeletal exercise	0.00%	0.00%	0.00%	0.00%	0.00%
	Consultation, described as comprehensive	0.00%	0.00%	0.01%	0.00%	0.00%
	Consultation, described as limited	0.00%	0.00%	0.01%	0.00%	0.00%
	Microscopic examination of specimen from unspecified site, culture	0.00%	0.00%	0.00%	0.00%	0.00%

4. SNOMED-CT vocabulary assessment

To understand how the hierarchical system within SNOMED-CT maps procedure codes and how many source codes are associated to it, we can evaluate the hierarchy within 1 level from the term “Procedure” in SNOMED-CT. The number of descendants from that procedure code, the number of mapped codes (ICD9-CM Procedure codes, CPT-4 and HCPCS) and the percentage of data it represents within the 5 databases is shown below in Table 10.

Table 10. SNOMED-CT mapping of descendants and source codes by database

			Databases				
			Optum	CCAE	MDCD	MDCR	Premier
SNOMED Concept Name	SNOMED-CT descen-dants	Mapped source codes (N)	% of data	% of data	% of data	% of data	% of data
Procedures relating to mobility	64	0	0.00	0.00	0.00	0.00	0.00
Laboratory procedure	650	186	1.60	1.76	1.02	1.82	3.72
Procedure by method	43602	11983	66.83	67.38	57.36	71.75	45.68
Regimes and therapies	938	396	8.86	11.08	7.12	7.43	5.14
Procedures relating to eating and drinking	12	0	0.00	0.00	0.00	0.00	0.00
Procedure by priority	143	26	0.02	0.04	0.06	0.04	0.04
Social service procedure	29	1	0.00	0.00	2.23	0.00	0.00

Provider-specific procedure	613	188	5.73	5.94	4.82	8.45	3.40
Nuclear medicine procedure	442	133	0.21	0.15	0.12	0.43	0.20
Outpatient procedure	1	4	0.21	0.09	0.33	0.47	0.06
Procedure in coronary care unit	1	0	0.00	0.00	0.00	0.00	0.00
Procedure related to anesthesia and sedation	598	336	0.84	0.91	0.89	1.15	0.21
Procedure by intent	4652	1072	10.77	12.32	12.26	7.67	5.94
Procedure by site	36380	10912	31.33	30.76	23.46	33.83	29.24
Procedure categorized by device involved	9964	3391	6.21	5.57	3.81	7.95	6.56
General treatment	1	0	0.00	0.00	0.00	0.00	0.00
Procedure related to breastfeeding	25	0	0.00	0.00	0.00	0.00	0.00
Procedure with a procedure focus	1794	429	7.50	7.08	5.64	4.99	10.51
Procedure with a clinical finding focus	1350	239	0.41	0.38	3.21	0.39	0.37
Procedure on ganglion cyst	37	5	0.01	0.01	0.00	0.00	0.00
Obstetric procedure	346	150	0.22	0.33	0.66	0.00	0.47
Postoperative procedure	4	0	0.00	0.00	0.00	0.00	0.00
Preoperative procedure	31	1	0.00	0.00	0.00	0.00	0.00
Determination of information related to transfusion	10	0	0.00	0.00	0.00	0.00	0.00

Procedure on wound	345	282	0.24	0.22	0.21	0.24	0.41
Treatment of comorbid condition	1	0	0.00	0.00	0.00	0.00	0.00
Promotion	46	0	0.00	0.00	0.00	0.00	0.00

The term “procedure by method” utilizes the highest number of codes and utilizes the largest volume of data. There are many categories that are not utilized and not mapped to any descendant codes such as “procedures relating to mobility, “procedure related to eating and drinking”, “general treatment”, “procedures related to breastfeeding” and “treatment of comorbid conditions”. In some cases, these categories would likely not be candidates in finding a condition-procedure relationship. The codes that were determined to be orphan codes in Table 9 could be evaluated to map into these SNOMED-CT categories, such as the generalized codes for outpatient encounter could be placed in “general treatment” as they may apply to any procedure or treatment and is a non-specific service that is received. The structure of the first level of procedures allows us to understand the hierarchy and distribution of codes to help us better evaluate and understand the types of procedures that can be mapped to a corresponding condition.

4.2. UNIVARIATE STATISTICS OF PARAMETERS

To determine a algorithm that can accurately assess positive procedural relationships we can analyze the univariate statistics. The minimum, median, maximum, and average value for each parameter are shown in Table 10 for both algorithms, therapeutic and diagnostic and for each set of controls. By analyzing the univariate statistics, we can see that for most variables we see complete separation in values for the positive and negative controls for each algorithm type.

Table 11. Parameter statistics for each parameter calculated in Optum (continued)

		cond_proc_ratio	cond_proc_persons	abs_female	avg_age_ratio	avg_per_person	avg_obs_time_days	ratio_inpt	ratio_outpt	num_days_btw	avg_time_cond_days	cond_proc_same
Therapeutic Controls	MIN	0.01	0.01	0.32	0.22	0.27	719.00	0.00	0.01	1.00	0.00	0.00
	MAX	985.45	507.40	2.26	3.20	15.86	997.00	780.21	2660.74	268.00	27.00	1.00
	MEDIAN	3.65	2.60	1.03	1.03	1.50	887.00	3.07	5.30	80.00	2.00	0.70
	AVG	55.32	21.17	1.08	1.05	2.36	874.31	27.24	190.51	85.60	2.89	0.64
Diagnostic Controls	MIN	0.01	0.00	0.48	0.20	0.19	0.00	0.00	0.01	0.00	0.00	0.00
	MAX	12749.22	2140.91	261.85	19.00	13.47	935.00	6897.65	52965.81	252.00	27.00	0.99
	MEDIAN	1.71	0.76	0.98	1.02	2.28	888.00	2.10	1.72	90.50	1.50	0.60
	AVG	146.61	29.54	3.60	1.25	3.01	865.28	282.68	547.48	95.36	2.25	0.57
Negative controls	MIN	0.00	0.00	0.00	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	81937.73	23519.85	337.87	37.00	22.49	1733.00	249069.00	251693.55	1276.00	115.00	1.00
	MEDIAN	3.45	1.71	1.00	1.04	1.97	909.00	3.46	3.83	246.00	2.00	0.02
	AVG	208.03	68.34	2.84	1.45	2.83	884.43	368.07	579.71	233.28	2.74	0.07

		cond_proc_fsame	proc_first	cond_first	avg_visit_ct	rx_60	rx_180	rx_365	cond_60	cond_180	cond_365	meas_60
Therapeutic Controls	MIN	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
	MAX	0.94	1.00	28.05	119.00	535.71	780.41	745.97	615.62	839.71	833.91	1378.10
	MEDIAN	0.21	0.15	1.10	15.00	2.56	2.81	2.63	2.37	2.21	1.85	3.09
	AVG	0.28	0.24	1.94	20.12	22.63	31.78	30.36	25.15	27.93	27.77	44.84
Diagnostic Controls	MIN	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
	MAX	0.90	1.50	22.29	81.00	1528.23	2473.32	2439.00	2190.51	1627.28	1781.66	1814.57
	MEDIAN	0.29	0.22	1.10	17.50	0.94	0.85	0.81	0.90	0.70	0.66	0.60
	AVG	0.33	0.27	1.67	20.31	28.09	34.82	33.95	32.27	24.67	25.65	25.03
Negative controls	MIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	1.00	83.50	166.00	405.00	56760.88	44201.33	34386.27	46591.66	42658.07	43995.96	66191.70
	MEDIAN	0.01	0.26	0.92	37.00	2.03	2.08	1.96	1.89	1.56	1.44	1.78
	AVG	0.03	0.46	1.46	38.73	98.79	93.30	87.35	97.27	82.12	81.90	80.24

		meas_180	meas_365	proc_60	proc_180	proc_365	rr_ratio_0	rr_ratio_90	rr_ratio_365	rr_ratio_all_time	Odds_ratio_0	Odds_ratio_90
Therapeutic Controls	MIN	0.00	0.00	0.00	0.00	0.00	-1.58	-0.57	-0.98	-0.13	-497.31	-581.45
	MAX	708.78	696.58	105.70	116.14	107.96	4.05	3.96	3.20	4.23	2.78	1.80
	MEDIAN	2.18	2.02	0.35	0.38	0.40	1.73	1.85	1.15	2.20	-0.37	-0.61
	AVG	27.71	26.73	2.23	2.32	2.24	1.77	1.85	1.23	2.25	-7.19	-7.79
Diagnostic Controls	MIN	0.00	0.00	0.00	0.00	0.00	-0.95	-0.59	-0.44	-0.23	-928.18	-1810.82
	MAX	1103.88	748.71	174.89	193.77	223.93	2.75	3.10	2.07	3.14	1.50	1.11
	MEDIAN	0.66	0.64	1.22	1.24	1.26	0.94	0.93	0.70	1.21	-2.24	-1.95
	AVG	18.59	13.70	7.02	7.18	7.53	0.91	1.00	0.73	1.28	-32.02	-41.07
Negative controls	MIN	0.00	0.00	0.00	0.00	0.00	-3.88	-3.16	-2.87	-2.57	-14321.64	-11887.61
	MAX	39500.68	30801.75	4993.17	5534.71	4181.27	3.04	4.39	3.14	4.57	1.39	1.75
	MEDIAN	1.47	1.38	0.54	0.49	0.51	-0.72	-0.07	0.00	0.25	-3.42	-3.17
	AVG	70.09	65.83	12.96	13.91	14.26	-0.78	-0.10	-0.01	0.23	-50.29	-53.80

		odds_ratio_365	odds_ratio_all_time	support_0	support_90	support_365	Support_all_time	misclassification_0	misclassification_90	misclassification_365	misclassification_all_time
Therapeutic Controls	MIN	-539.97	-635.44	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00
	MAX	0.37	3.06	146841.00	82095.00	55702.00	150205.00	0.11	0.11	0.11	0.11
	MEDIAN	-1.47	0.01	2237.00	2173.00	774.00	3054.00	0.01	0.01	0.01	0.01
	AVG	-8.34	-7.77	10171.39	8523.17	4256.29	12857.94	0.02	0.02	0.02	0.02
Diagnostic Controls	MIN	-1141.47	-3509.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	0.43	1.67	188545	154847.00	113493.0	279109.0	0.29	0.29	0.29	0.28
	MEDIAN	-2.57	-1.55	5357	5198.00	3428.00	8184.50	0.04	0.04	0.04	0.03
	AVG	-31.31	-67.36	20678.21	17942.76	11815.13	30191.55	0.05	0.06	0.06	0.05
Negative controls	MIN	-11901.84	-25994.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	0.62	1.98	202372.	252834.00	238257.0	319252.00	0.39	0.35	0.35	0.33
	MEDIAN	-3.09	-2.82	6.00	91.00	118.00	211.00	0.02	0.02	0.02	0.02
	AVG	-52.11	-56.99	517.34	1934.95	2066.45	3362.00	0.04	0.04	0.04	0.04

		sensitivity_0	specificity_0	sensitivity_90	specificity_90	sensitivity_365	specificity_365	sensitivity_all_time	specificity_all_time
Therapeutic Controls	MIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	1.00	0.11	0.98	0.11	0.70	0.11	1.00	0.11
	MEDIAN	0.29	0.01	0.27	0.01	0.12	0.01	0.50	0.01
	AVG	0.38	0.01	0.39	0.02	0.19	0.02	0.50	0.01
Diagnostic Controls	MIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	0.97	0.27	0.93	0.27	0.72	0.27	0.98	0.27
	MEDIAN	0.10	0.01	0.11	0.01	0.06	0.01	0.18	0.01
	AVG	0.18	0.03	0.18	0.03	0.11	0.03	0.26	0.02
Negative controls	MIN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAX	0.96	0.30	0.98	0.28	0.81	0.28	0.99	0.28
	MEDIAN	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
	AVG	0.01	0.02	0.03	0.02	0.03	0.02	0.05	0.02

4.3. MODEL STATISTICS

4.3.1. SINGLE VARIABLE MODEL AUC'S

Single parameter logistic regression models were run with each individual parameter with test and train split (80:20). The single parameter model AUC's are shown below for both the diagnostic and therapeutic algorithm.

Table 12. AUC's for single parameter model's

Therapeutic Logistic Regression Model	Therapeutic algorithm AUC	Diagnostic Logistic Regression Model	Diagnostic algorithm AUC
negative control~rr_ratio_all_time	0.9627	negative control~cond_proc_fsame	0.9717
negative control~rr_ratio_365	0.9531	negative control~cond_proc_same	0.9680
negative control~rr_ratio_90	0.9465	negative control~support_0	0.9460
negative control~cond_proc_same	0.9439	negative control~rr_ratio_0	0.9359
negative control~rr_ratio_0	0.9313	negative control~sensitivity_0	0.9269
negative control~sensitivity_0	0.9309	negative control~rr_ratio_90	0.9042
negative control~cond_proc_fsame	0.9284	negative control~num_days_btw	0.9003
negative control~num_days_btw	0.8893	negative control~rr_ratio_all_time	0.9001
negative control~sensitivity_90	0.8726	negative control~support_90	0.8781
negative control~support_0	0.8697	negative control~rr_ratio_365	0.8689
negative control~sensitivity_all_time	0.8610	negative control~support_all_time	0.8567
negative control~avg_visit_ct	0.8430	negative control~sensitivity_90	0.8360
negative control~sensitivity_365	0.8275	negative control~support_365	0.8263
negative	0.8024	negative	0.8245

control~odds_ratio_all_time		control~avg_visit_ct	
negative control~odds_ratio_90	0.7848	negative control~sensitivity_all_time	0.8058
negative control~support_90	0.7525	negative control~sensitivity_365	0.7652
negative control~avg_obs_time_days	0.7457	negative control~avg_obs_time_days	0.7216
negative control~odds_ratio_365	0.7290	negative control~avg_time_cond_days	0.6336
negative control~support_all_time	0.7150	negative control~rx_365	0.6319
negative control~odds_ratio_0	0.7032	negative control~cond_proc_persons	0.6317
negative control~proc_first	0.6919	negative control~rx_180	0.6300
negative control~support_365	0.6774	negative control~misclassification_0	0.6295
negative control~cond_first	0.6501	negative control~meas_60	0.6285
negative control~misclassification_all_time	0.6160	negative control~meas_365	0.6210
negative control~proc_365	0.5967	negative control~rx_60	0.6200
negative control~proc_180	0.5951	negative control~cond_60	0.6181
negative control~misclassification_90	0.5949	negative control~meas_180	0.6173
negative control~misclassification_365	0.5843	negative control~cond_180	0.6170
negative control~proc_60	0.5802	negative control~cond_365	0.6166
negative control~avg_time_cond_days	0.5673	negative control~cond_proc_ratio	0.6129
negative control~avg_per_person	0.5515	negative control~misclassification	0.6067

		_365	
negative control~ratio_inpt	0.5458	negative control~misclassification_90	0.5958
negative control~specificity_all_time	0.5307	negative control~misclassification_all_time	0.5812
negative control~misclassification_0	0.5264	negative control~specificity_0	0.5524
negative control~avg_age_ratio	0.5151	negative control~avg_per_person	0.5310
negative control~specificity_90	0.5140	negative control~odds_ratio_all_time	0.5297
negative control~specificity_365	0.5120	negative control~ratio_inpt	0.5286
negative control~specificity_0	0.4433	negative control~odds_ratio_90	0.5144
negative control~rx_60	0.4228	negative control~cond_first	0.5028
negative control~cond_60	0.4155	negative control~specificity_365	0.4924
negative control~rario	0.4141	negative control~avg_age_ratio	0.4919
negative control~cond_proc_persons	0.4066	negative control~odds_ratio_365	0.4822
negative control~abs_female	0.4050	negative control~proc_first	0.4793
negative control~rx_365	0.4043	negative control~specificity_90	0.4665
negative control~rx_180	0.4028	negative control~abs_female	0.4639
negative control~cond_180	0.4003	negative control~odds_ratio_0	0.4606
negative control~cond_365	0.3993	negative control~specificity_all_time	0.4351
negative control~meas_180	0.3930	negative control~ratio_outpt	0.3768
negative control~meas_365	0.3915	negative control~proc_60	0.3726
negative	0.3861	negative	0.3695

control~ratio_outpt		control~proc_180	
negative control~meas_60	0.3762	negative control~proc_365	0.3682

The single logistic regression models provide the basis for analyzing the most parsimonious algorithm, a single variable algorithm. The AUC's range from 0.97 to 0.36 for the diagnostic algorithm, to 0.96 to 0.37 for the therapeutic algorithm. The variables that have the highest AUC's for both algorithms include variables that describe co-occurrence. The variables that have the lowest AUC's are those that describe utilization patterns, thus indicating that those would be less predictive than co-occurrence statistics.

4.3.2. FULL MODEL (DIAGNOSTIC ALGORITHM)

Using a generalized logistic regression model with all 51 covariates results in coefficient values for the full diagnostic algorithm:

negative_control ~ -5.135+0.003499*cond_proc_ratio+0.00325*cond_proc_persons+-1.053*abs_female+
0.08025*avg_age_ratio+0.04595*avg_per_person+0.002604*avg_obs_time_days+0.00003173*ratio_inpt+
-0.0004213*ratio_outpt+-0.007529*num_days_btw+-0.1291*avg_time_cond_days+0.9737*cond_proc_same+
0.2765*cond_proc_fsame+-0.2106*proc_first+0.1075*cond_first+-0.02241*avg_visit_ct+-0.007088*rx_60+
0.03581*rx_180+-0.02593*rx_365+0.004021*cond_60+-0.09418*cond_180+0.09988*cond_365+
-0.05226*meas_60+0.1177*meas_180+-0.1418*meas_365+0.003652*proc_60+0.01834*proc_180+
-0.04062*proc_365+2.467*rr_ratio_0+-1.138*rr_ratio_90+2.265*rr_ratio_365+-1.487*rr_ratio_all_time+
-0.005341*odds_ratio_0+-0.007402*odds_ratio_90+0.007878*odds_ratio_365+0.002723*odds_ratio_all_time+
0.0004751*support_0+-0.0001924*support_90+0.0004075*support_365+-0.000668*support_all_time+
1338*misclassification_0+-774.6*misclassification_90+630.6*misclassification_365+-1180*misclassification_all_time+
-5.474*sensitivity_0+ 247.1*specificity_0+2.59*sensitivity_90+629.2*specificity_90+-
5.213*sensitivity_365+526.5*specificity_365+6.417*sensitivity_all_time+-897.4*specificity_all_time

The full diagnostic algorithm yields an AIC of 613.13 on the train diagnostic dataset. The results of the model on test dataset produce the following confusion matrix with a positive predictive value cut off at 0.5. For the test dataset 99.9% of negative controls

were predicted correctly, while 0.1% percent of positive controls were predicted correctly for the diagnostic algorithm.

Table 13. Confusion matrix for diagnostic algorithm using test data

	False	True	Totals
Negative controls	6,502 (TN*)	4 (FN*)	6,506
Positive controls	18 (FP*)	2 (TP*)	20
Total	6,520	6	6,526

*TN=True Negative; TP=True Postive; FP=False Positive; FN=False Negative

The 4 pairs of conditions and procedures that were false negatives are described in the table below.

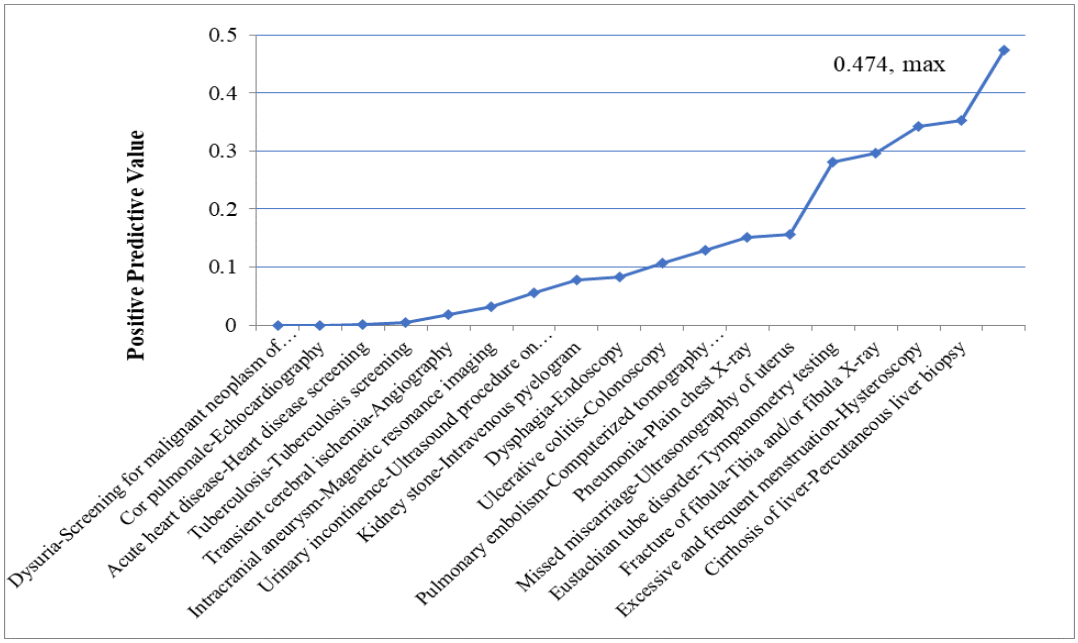
Table 14. True conditions and procedures pairs for diagnostic algorithm

Control type	Positive predictive value	Condition Name	Procedure Name	Remarks
Negative	0.6293	Benign neoplasm of colon	Endoscopy	The negative control is misclassified, and thus should be identified as a positive control.
Negative	0.5986	Neoplasm of pancreas	Splenectomy	The concepts under neoplasm of the pancreas include the spleen thus this control is misclassified.
Negative	0.9711	Renal failure syndrome	Transplantation of heart	Renal failure syndrome includes concepts for heart conditions, thus the

				control is misclassified.
Negative	0.6308	Fracture of forearm	Sialogram	Sialogram's are done in a radiology unit as are diagnostic x-rays for fracture's. The PPV is 0.6308 which is near the 0.5 cut-off meaning the model classified this incorrectly.

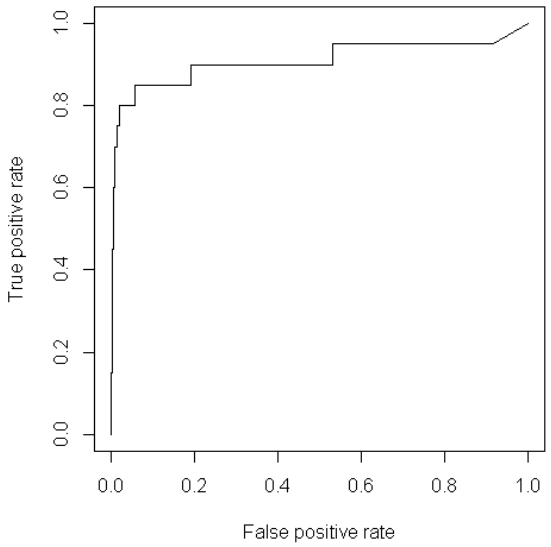
The two positive controls that were above 0.5 were for glaucoma-goinscopy and fracture of humerus-radiography of humerus. Figure 9 displays the distribution of the positive predictive values of the 18 positive controls that were less than 0.5. The maximum value is 0.474 while the lowest value is 0. This graph illustrates that the cut-off points for PPV are very low in the positive controls thus it would not make sense to lower our threshold from 0.5 to a lower value to consider good controls that were just below the cut off.

Figure 8. Full diagnostic algorithm positive predictive values < 0.5 for positive controls



The positive predictive value of the full diagnostic algorithm is .10 [(TP / (TP + FP)) = 2/(2+18)=0.10] or .10 of predicted positive cases were identified correctly. The sensitivity is 33% [(TP / (TP + FN)) 2/ (2+4) =0.33] or a 33% chance that we would identify a positive value given that the pair is positive. The specificity is 99% [(TN/(TN+FP)) = 6502/(6502+18)=0.99] or a 99% chance that a negative relationship will be classified as a negative. The AUC for the full diagnostic algorithm is 0.9093.

Figure 9. ROC curve for test data on full diagnostic algorithm



4.3.3. FULL MODEL (THERAPEUTIC ALGORITHM)

Using a generalized logistic regression model with all 51 covariates results in coefficient values

for the therapeutic algorithm:

negative_control ~ -7.105+-0.0008146*cond_proc_ratio+-0.01571*cond_proc_persons+-0.2902*abs_female+-0.224*avg_age_ratio+-0.2651*avg_per_person+0.00427*avg_obs_time_days+-0.0007029*ratio_inpt+-0.0004898*ratio_outpt+-0.01596*num_days_btw+0.0179*avg_time_cond_days+1.625*cond_proc_same+-2.109*cond_proc_fsame+-0.07654*proc_first+0.1279*cond_first+-0.001098*avg_visit_ct+-0.01739*rx_60+0.04476*rx_180+-0.008253*rx_365+0.01214*cond_60+-0.05099*cond_180+0.03109*cond_365+0.005285*meas_60+0.007886*meas_180+-0.02647*meas_365+0.02234*proc_60+0.007119*proc_180+-0.02011*proc_365+1.505*rr_ratio_0+-1.341*rr_ratio_90+1.069*rr_ratio_365+1.032*rr_ratio_all_time+-0.02954*odds_ratio_0+-0.1771*odds_ratio_90+-0.1723*odds_ratio_365+0.4434*odds_ratio_all_time+0.0008315*support_0+-0.0004072*support_90+0.001946*support_365+-0.002305*support_all_time+1401*misclassification_0+-2674*misclassification_90+5993*misclassification_365+-4685*misclassification_all_time+-2.918*sensitivity_0+1054*specificity_0+-2.361*sensitivity_90+2486*specificity_90+6.083*sensitivity_365+-2196*specificity_365+2.208*sensitivity_all_time+-1374*specificity_all_time

T

he full therapeutic algorithm yields an AIC of 439.36 on the train therapeutic dataset. The results of the model on test dataset produce the following confusion matrix with a positive predictive value cut off at 0.5. For the test dataset 99.9% of negative controls were predicted correctly, while 45% of positive controls were predicted correctly for the diagnostic algorithm.

Table 15. Confusion matrix for therapeutic algorithm using test data

	False	True	Totals
Negative controls	6,504 (TN*)	2 (FN*)	6,506
Positive controls	11 (FP*)	9 (TP*)	20
Total	6,515	11	6,526

*TN=True Negative;TP=True Postive;FP=False Positive;FN=False Negative

The 2 pairs of conditions and procedures that were false negative are described in the table

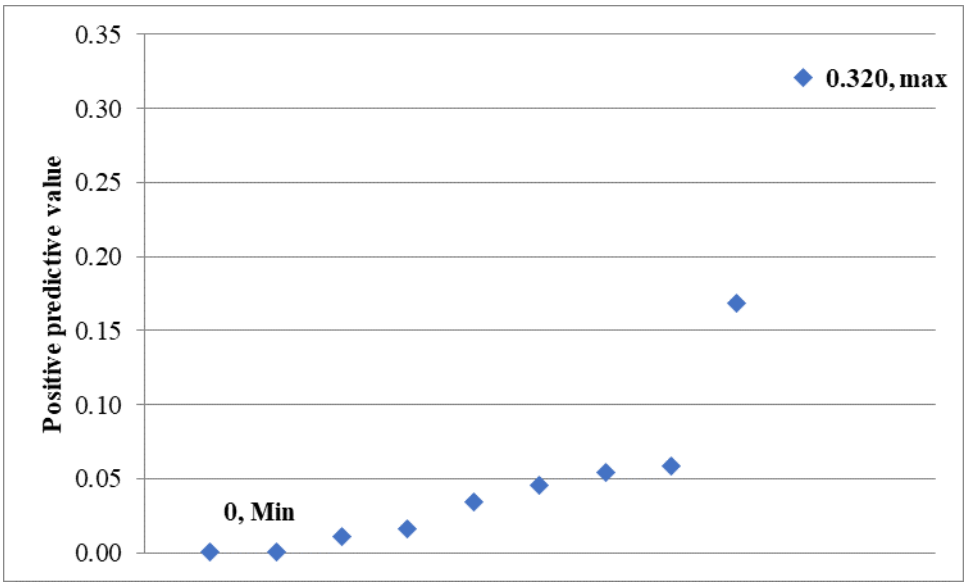
below.

Table 16. True conditions and procedures pairs for diagnostic algorithm

Control type	Positive predictive value	Condition Name	Procedure Name	Notes
Negative	0.6558	Acute heart disease	Diagnostic radiography of chest- PA	Heart disease can often manifest as pain in the chest, thus a diagnostic x-ray of the chest as a first measure, this this pair is misclassified
Negative	0.7608	Neoplasm of brain	Procedure on pituitary gland	The pituitary gland and brain are very close to each other and the procedure on the pituitary gland is a very broad concept thus this pair misclassified

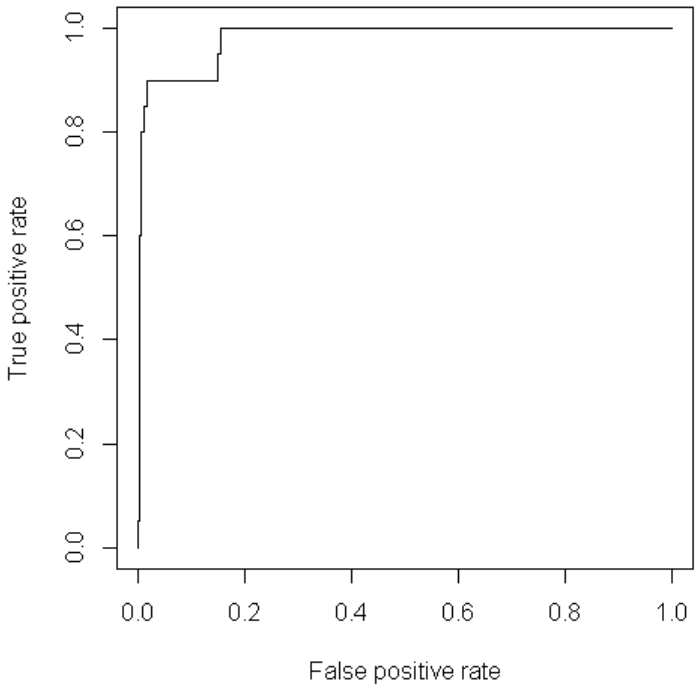
Figure 11 displays the distribution of the positive predictive values of the 11 positive controls that were less than 0.5. The maximum value is 0.320 while the lowest value is 0.

Figure 10. Full therapeutic algorithm positive predictive values < 0.5 for positive controls



The positive predictive value of the full therapeutic algorithm is .45 [(TP / (TP + FP)) = 9/(9+11)=0.45 or .45 of predicted positive cases were identified correctly.. The sensitivity is 10% [(TP / (TP + FN)) 9/ (9+2) =0.81] or a 81% chance that we would identify a positive value when given a pair that is positive. The specificity is 99% [(TN/(TN+FP)) = 6504/(6504+11)=0.99] or a 99% chance that a negative relationship will be classified as a negative. The AUC for the full therapeutic algorithm is 0.9241

Figure 11. ROC curve for full therapeutic algorithm in Optum Extended SES



4.3.4. STEPWISE SELECTION (DIAGNOSTIC ALGORITHM)

The stepwise selection process adds variables based upon Chi square test. Starting with no variables and just the response variable (negative_control) we calculate the AIC for each step (either the addition or subtraction of variables) to determine a model with the lowest AIC. Table 16 shows each step with variables and generated AIC.

Table 17. Stepwise model selection for diagnostic algorithm

Model	AIC
Null (negative_control ~1)	1085.34
negative_control ~ rr_ratio_0	715.1
negative_control ~ rr_ratio_0 + cond_proc_same	654.08
negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0	633.14

negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0 + num_days_btw	624.94
negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0 + num_days_btw + rr_ratio_365	618.07
negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0 + num_days_btw + rr_ratio_365 + odds_ratio_all_time	612.27
negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0 + num_days_btw + rr_ratio_365 + odds_ratio_all_time + abs_female	607.11
negative_control ~ rr_ratio_0 + cond_proc_same + misclassification_0 + num_days_btw + rr_ratio_365 + odds_ratio_all_time + abs_female + support_0	605.15

The final algorithm with coefficients using the stepwise procedure is:

negative_control ~ -5.015 + 0.0605*rr_ratio_0 + 3.407*cond_proc_same + 0.01169*misclassification_0 + -0.001004*num_days_btw + 1.466*rr_ratio_365 + 0.000310*odds_ratio_all_time + -0.0904*abs_female + 0.0000009360*support_0

The confusion matrix for this model utilizing the same cutoff at a positive predictive value of 0.5 is:

Table 18. Confusion matrix for diagnostic algorithm using stepwise selection

	False	True	Totals
Negative controls	6,506 (TN*)	0 (FN*)	6,506
Positive controls	18 (FP*)	2 (TP*)	20
Total	6,524	2	6,526

*TN=True Negative;TP=True Postive;FP=False Positive;FN=False Negative

The positive predictive value of the stepwise diagnostic algorithm is .10% [(TP / (TP + FP)) = 2/(2+18)=0.1], or .10 of predicted positive cases were identified correctly which is the same as the full model. The sensitivity is 10% [(TP / (TP + FN)) 2/ (2+0) =1] or a

100% chance that we would identify a positive value when given a pair that is positive. The specificity is 100% [(TN/(TN+FP)) = 6506/(6506+0)=1] or a 100% chance that a negative relationship will be classified as a negative. The AUC for this algorithm is 0.8993. The stepwise selection for the diagnostic algorithm resulted in a smaller algorithm, but with an AUC that is lower than the full model.

4.3.5. STEPWISE SELCTION (THERAPEUTIC ALGORITHM)

Using the same approach as the stepwise selection for the diagnostic algorithm, Table 18 outlines

the AIC values for algorithm selection.

Table 19. Stepwise model selection for therapeutic algorithm

Model	AIC
Null (negative_control ~1)	1085.34
negative_control ~ rr_ratio_0	441.46
negative_control ~ rr_ratio_0 + support_90	433.64
negative_control ~ rr_ratio_0 + support_90 + rr_ratio_all_time	428.09
negative_control ~ rr_ratio_0 + support_90+ rr_ratio_all_time + cond_proc_same	420.16
negative_control ~ rr_ratio_0 + support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt	412.76
negative_control ~ rr_ratio_0 + support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time	408.25
negative_control ~ rr_ratio_0 + support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt	405.40
negative_control ~ support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt	404.37
negative_control ~ support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt + rr_ratio_90	402.8
negative_control ~ support_90 + rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt + rr_ratio_90 +	401.9

avg_per_person	
negative_control ~ rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt + rr_ratio_90 + avg_per_person	400.76
negative_control ~ rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt + rr_ratio_90 + avg_per_person + sensitivity_365 + num_days_btw	400.42
negative_control ~ rr_ratio_all_time + cond_proc_same + ratio_outpt + odds_ratio_all_time + ratio_inpt + rr_ratio_90 + avg_per_person + sensitivity_365 + num_days_btw + cond_first	398.8

Using the therapeutic algorithm stepwise the final algorithm results is:

negative_control ~-6.5565965+1.0274337*rr_ratio_all_time+3.331501*cond_proc_same+-0.0010994*ratio_outpt+0.0218179*odds_ratio_all_time+-0.0018185*ratio_inpt+1.1364606*rr_ratio_90+-0.1949317*avg_per_person+2.9926269*sensitivity_365+-0.0074504*num_days_btw+0.1162868*cond_first

The confusion matrix for this algorithm is shown in Table 18.

Table 20. Confusion matrix for therapeutic algorithm using stepwise selection

	False	True	Totals
Negative controls	6,505 (TN*)	1 (FN*)	6,506
Positive controls	12 (FP*)	8 (TP*)	20
Total	6,517	9	6,526

*TN=True Negative;TP=True Postive;FP=False Positive;FN=False Negative

The positive predictive value of the stepwise therapeutic algorithm is .40 [(TP / (TP + FP)) = 8/(8+12)=0.40], or .40 of predicted positive cases were identified correctly. The sensitivity is 88% [(TP / (TP + FN)) 8/ (8+1) =0.88] or a 88% chance that we would identify a positive value when given a pair that is positive. The specificity is 99% [(TN/(TN+FP)) = 6505/(6505+12)=0.99] or a 99% chance that a negative relationship will be classified as a negative. The AUC for this algorithm is: 0.9205. Though this algorithm has

less variables the AUC is slightly lower than the full model. There is a trade-off in sensitivity and positive predictive value with this model, as our sensitivity has increased slightly, our positive predictive value is lowered with the modified algorithm.

4.3.6. LASSO REGRESSION FOR FEATURE SELECTION

Utilizing lasso techniques for feature selection for the diagnostic algorithm results in the

following model:

negative_control ~ -5.916+0.01761317*cond_proc_fsame +-0.000004427961*support_0+6.098789*sensitivity_0

When inputting the lasso regression model for the therapeutic algorithm, the model did not converge at a value for lambda so we were unable to utilize lasso regression for feature selection. The diagnostic algorithm resulted in three variables that were significant which were if the ratio of condition and procedure occurred for the first time together, support at time zero and sensitivity at time zero. Running this algorithm as a logistic regression model resulted in the following model:

negative_control ~ -6.631+ 5.552*cond_proc_fsame+0.000008313*support_0+5.809*sensitivity_0

Table 21. Confusion matrix for therapeutic algorithm using stepwise selection

	False	True	Totals
Negative controls	6,505 (TN*)	1 (FN*)	6,506
Positive	19	1	20

controls	(FP*)	(TP*)	
Total	6,524	2	6,526

The positive predictive value of the stepwise therapeutic algorithm is .50 [(TP / (TP + FP)) = 1/(1+19)=0.50], or .50 of predicted positive cases were identified correctly. The sensitivity is 10% [(TP / (TP + FN)) 1/ (1+1) =0.50] or a 50% chance that we would identify a positive value when given a pair that is positive. The specificity is 99% [(TN/(TN+FP)) = 6505/(6505+1)=0.99] or a 99% chance that a negative relationship will be classified as a negative. The AUC for this algorithm is 0.9725.

4.4. FINAL ALGORITHM SELECTION (DIAGNOSTIC AND THERAPEUTIC)

To determine the final algorithm to utilized for both diagnostic condition-procedure pairs and therapeutic condition-procedure pairs, a summary of all model building techniques are compared for each algorithm type. The main statistic used to compare the algorithms, is AUC's or area under the curve to determine how well a model does. The number of variables in the algorithm range from 51 variables to 1 (single model prediction). The balance between variable selection and AUC to obtain a clinically relevant and generalizable algorithm would guide one to choose a algorithm that is simple yet can predict a condition-procedure pair correctly.

The diagnostic algorithm summaries are displayed in Table 21. The AUC for the final algorithm is the highest and utilizes three variables, making it more parsimonious than the stepwise or full model selection. The variables chosen to make logical sense for the

diagnostic algorithm as the first-time occurrence variable was chosen, along with sensitivity at time 0 and support at time 0. In an essence inferring that when the first encounters are most significant which makes logical sense as a diagnosis is made first before treatments can be applied, be it procedures or drugs.

Table 22. Diagnostic algorithm performance summary

Model selection technique	Algorithm	# of variables	AUC	Sensitivity	Specificity	PPV
Single variable model	negative_control~ variable	1	0.3682-0.9716	N/A	N/A	N/A
Full model	Full model	51	0.9093	10%	99%	0.33
Step-wise	negative_control ~ -5.015 + -0.06.05***rr_ratio_0** + 3.407***cond_proc_same** + 0.01169***misclassification_0** + -0.001004***num_days_btw** + 1.466***rr_ratio_365** + -0.000310***odds_ratio_all_time** + -0.0904***abs_female** + 0.0000009360***support_0**	8	0.8993	100%	100%	0.10
Lasso	negative_control ~ -5.916 + -0.01761317***cond_proc_fsame**+4.427***support_0** +6.098789***sensitivity_0**	3	0.9698	N/A	N/A	N/A
Lasso coefficients (represented as a logistic	negative_control ~ -6.631+ 5.552***cond_proc_fsame** +0.000008313***support_0**+5.809***sensitivity_0**	3	0.9725	50%	99%	0.50

regressio n)						
Final diagnosti c model	negative_control ~ -6.631+ 5.552***cond_proc_fsame** +0.000008313***support_ 0**+5.809***sensitivity_0**	3	0.97 25	50%	99%	0.50

There are trade-offs made as we move through each model selection technique with sensitivity and specificity and positive predictive value. The therapeutic algorithm selection utilizes univariate model statistics to create a similarly parsimonious algorithm such as the diagnostic algorithm but because the added benefit of conducting the lasso regression thus we rely on the results of the single variable modes to determine the final algorithm. Clinically, the variables that are most predictive from the single variables models are relative risk measures and whether or not the condition and procedure occur at the same time or on the same day (irrespective of number of times). A single variable model to determine if a condition procedure pair are therapeutic would be the most parsimonious but considering that this algorithm would be applied to a variety of datasets and considering more than one variable would inform of a relationship the two variables relative risk and condition and procedure occurring on the same day were choose, the AUC of this algorithm is 0.9604 which is slightly less than 0.9627 which is the single variable algorithm for relative risk all time. Table 22 describes algorithm generation for the therapeutic algorithm.

Table 23. Therapeutic algorithm selection summary

Model selection technique	Model	# of variables	AUC	Sensitivity	Specificity	PPV
Single variable model	negative_control~ variable	1	0.3762 - 0.9627	N/A	N/A	N/A
Full model	Full model	51	0.9241			
Step-wise	negative_control ~ -6.55+ 1.027***rr_ratio_all_time** + 3.331***cond_proc_same** + -0.001***ratio_outpt** + 0.0218***odds_ratio_all_time** + -0.001***ratio_inpt** + 1.136***rr_ratio_90** + -0.194***avg_per_person** + 2.99*s**ensitivity_365** + -0.0074***num_days_btw** + 0.116***cond_first**	10	0.9205	88%	99%	0.44
Lasso	*Model convergence not achieved, not generated*			N/A	N/A	N/A
Lasso coefficeints (represented as a logistic regression)	*Model convergence not achieved, not generated*			N/A	N/A	N/A
Final Model	negative_control ~ -8.9867+ 2.4146***rr_ratio_all_time** + 4.3575***cond_proc_same**	2	0.9604	96%	100%	0.40

The final model statistics are based on the following confusion matrix:

Table 24. Confusion matrix for therapeutic algorithm using stepwise selection

	False	True	Totals
Negative controls	6,506 (TN*)	0 (FN*)	6,506
Positive controls	12 (FP*)	8 (TP*)	20
Total	6,518	8	6,526

The positive predictive value of the stepwise therapeutic algorithm is .40 [(TP / (TP + FP)) = 8/(8+12)=0.40], or the algorithm has a .40 chance to correctly identify a positive value. The sensitivity is 10% [(TP / (TP + FN)) 8/ (8+0) =1] or a 100% chance that we would identify a positive value when given a pair that is positive. The specificity is 99% [(TN/(TN+FP)) = 6505/(6505+1)=0.99] or a 99% chance that a negative relationship will be classified as a negative.

4.5. EXTERNAL VALIDATION (DIAGNOSTIC AND THERAPEUTIC)

The final algorithms were run on data from 5 databases (CCAE, MDCD, MDCR, Optum and Premier) were for the known set of positive and negative controls or ground truth datasets. The algorithm was applied on 5 databases and logit odds obtained, which is transformed into a positive predictive value from 0 to 1. The AUC's are calculated for each algorithm/database combination. Table 24 shows the AUC's for all 5 databases for each algorithm type.

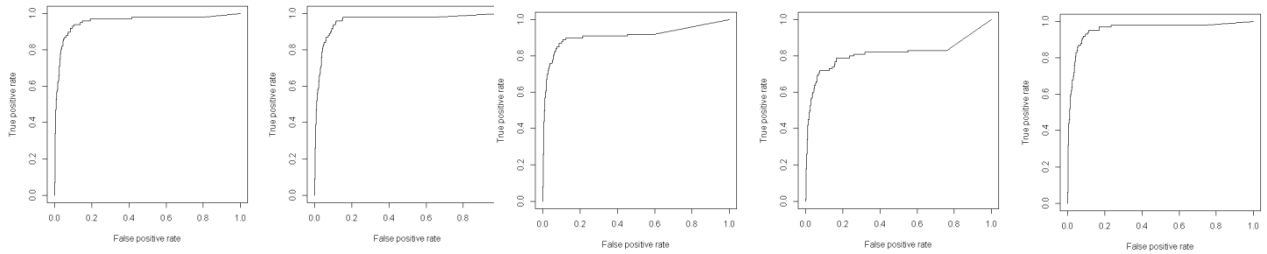
Table 25. AUC's by database for both diagnostic and therapeutic algorithm

Database	Algorithm Type	AUC	AUC from algorithm developed in Optum	Algorithm performance
CCAE	Diagnostic	0.9570	0.9725	Lower than developed algorithm
CCAE	Therapeutic	0.9825	0.9604	Higher than developed algorithm
MDCD	Diagnostic	0.9589	0.9725	Lower than developed algorithm
MDCD	Therapeutic	0.9735	0.9604	Higher than developed algorithm
MDCR	Diagnostic	0.9123	0.9725	Lower than developed algorithm
MDCR	Therapeutic	0.9563	0.9604	Lower than developed algorithm
Optum	Diagnostic	0.9576	0.9725	Lower than developed algorithm
Optum	Therapeutic	0.9796	0.9604	Higher than developed algorithm
Premier	Diagnostic	0.8128	0.9725	Lower than developed algorithm
Premier	Therapeutic	0.9801	0.9604	Higher than developed algorithm

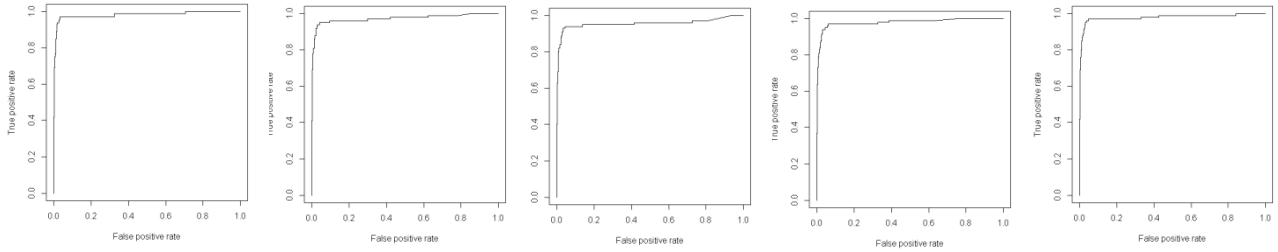
Overall the algorithm's all result in an AUC higher than .90, except for Premier for the diagnostic algorithm. In some cases the algorithm performs better than the test dataset in Optum, for the diagnostic algorithm the algorithm performs lower than the trained dataset while the therapeutic algorithm does better in all datasets except for MDCR.

Figure 12. External validation ROC curves by database

Diagnostic algorithm					
Database	CCAE	MDCD	MDCR	Premier	Optum
AUC	0.9570	0.9589	0.9123	0.8128	0.9576



Therapeutic algorithm					
Database	CCAE	MDCD	MDCR	Premier	Optum
AUC	0.9825	0.9735	0.9563	0.9801	0.9796



4.6. ALGORITHM APPLICATION

The range of probabilities can help determine what condition-procedure pairs could be negative relationships and which could positive relationships. The analysis will focus on how the pairs are distributed, coverage of data, adjudication of results, and algorithm separation.

To understand the coverage of codes that the algorithms utilized to generate a condition-procedure pair, we examine the number of codes identified by database for conditions and procedures. The coverage of codes estimates the number of codes that are utilized in the algorithm for any pair, thus providing information about which codes remain and did not get utilized in the algorithm. The condition codes (Figure 14) show that roughly over 90% of codes are utilized in a mapping to a procedure in all databases, MDCR having the least coverage by codes. The condition concepts that are missing include codes that are utilized very rarely in the database and include terms such as: “Indeterminate leprosy”, “Infantile botulism”, and “Poisoning by diphtheria vaccine”. The procedure codes (Figure 15) show a different story, as more codes are not utilized, roughly over 60% of SNOMED-CT procedure codes are utilized from the overall SNOMED-CT procedural vocabulary in the algorithms. Codes may be missing due to mapping from the source vocabulary not adequately being able to be mapped to a SNOMED-CT procedure term and the also due to non-specific terms such as “Risk assessment”, “Child examination – birth”, and “History and physical examination, insurance”. The starting population of condition and procedure codes that are utilized in

the algorithm show that the representation of the codes is adequate and expansive, further evaluations are conducted to understand the quality of the mapping.

Figure 13. Number of SNOMED-CT condition codes mapped in algorithms

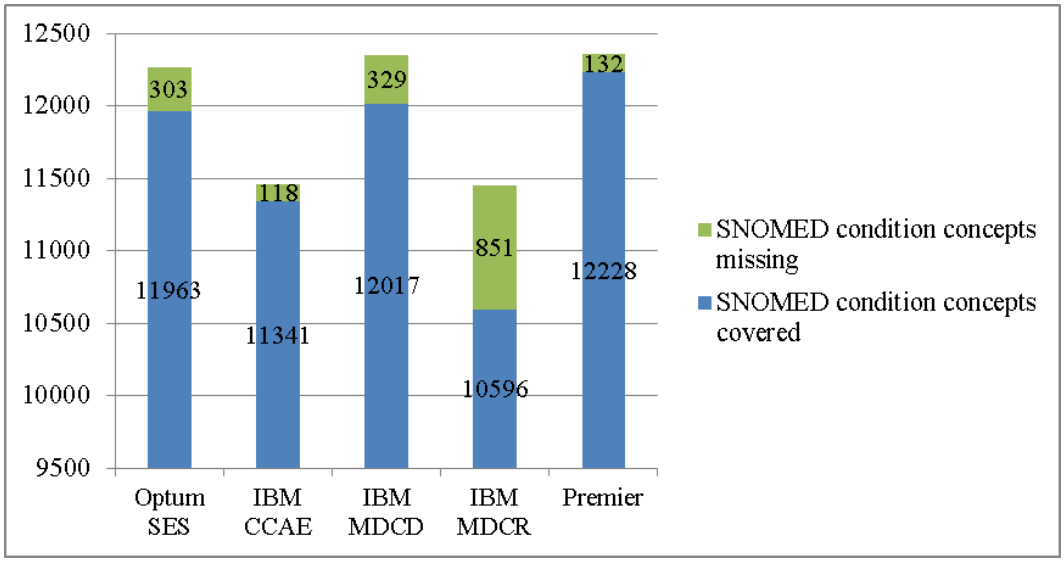


Figure 14. Number of SNOMED-CT procedure codes mapped in algorithms

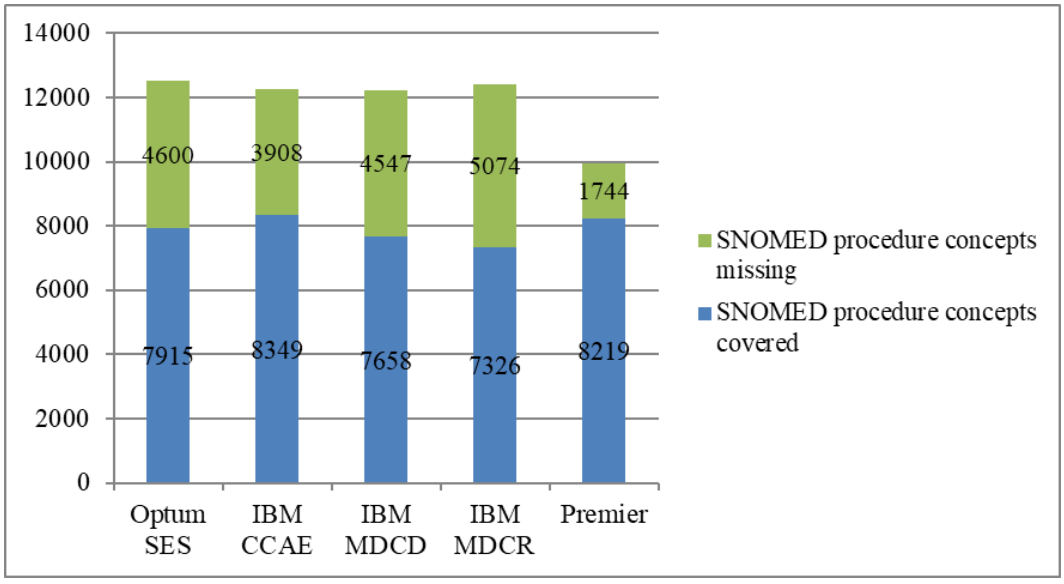


Table 26 . Overall summary for diagnostic algorithm

	Diagnostic Algorithm									
	Optum		CCAE		MDCD		MDCR		Premier	
	N	%	N	%	N	%	N	%	N	%
Total pairs	4805178	100	5454826	100	3803763	100	2605088	100	8186908	100
Number of pairs with p > 0.5	11778	0.24	6734	0.12	11066	0.2909	13144	0.5046	3740	0.04
p < 0.1	4457209	92.75	5097312	93.44	3559800	93.5863	2425215	93.0953	7998379	97.69
p >= 0.1 and p <=0.2	11956	0.24	14096	0.25	8709	0.2290	6911	0.2653	4095	0.05
p >0.2 and p <=0.3	275513	5.73	306039	5.61	186124	4.8932	126547	4.8577	167731	2.04
p > 0.3 and p <=0.4	39928	0.83	25338	0.46	30212	0.7943	25529	0.9800	10453	0.12
p > 0.4 and p <=0.5	8794	0.18	5307	0.09	7852	0.2064	7742	0.2972	2510	0.03
p > 0.5 and p <=0.6	5461	0.11	3146	0.05	4947	0.1301	5980	0.2296	1833	0.02
p > 0.6 and p <=0.7	582	0.01	436	0.008	448	0.0118	399	0.0153	53	0.00006
p > 0.7 and p <=0.8	3001	0.06	1632	0.029	2755	0.0724	3285	0.1261	800	0.00098
p > 0.8 and p <=0.9	2502	0.05	1337	0.024	2749	0.0723	3235	0.1242	985	0.012
p > 0.9	232	0.004	183	0.003	167	0.0044	245	0.0094	69	0.00008

Table 27. Overall summary for therapeutic algorithm

	Therapeutic Algorithm									
	Optum		CCAE		MDCD		MDCR		Premier	
	N	%	N	%	N	%	N	%	N	%
Total pairs	4805178	100	5454826	100	3803763	100	2605088	100.	8186908	100
Number of pairs with p > 0.5	177402	3.69	219738	4.02	140259	3.68	72242	2.77	121683	1.48
p < 0.1	4290162	89.28	4822634	88.41	3414454	89.76	2388272	91.67	7806441	95.35
p >= 0.1 and p <=0.2	155764	3.24	191464	3.51	114301	3.00	67294	2.58	123389	1.50
p >0.2 and p <=0.3	82247	1.71	100187	1.83	60833	1.59	35143	1.34	62875	0.76
p > 0.3 and p <=0.4	56350	1.17	67952	1.24	41760	1.09	23956	0.91	41269	0.50
p > 0.4 and p <=0.5	43253	0.90	52851	0.96	32156	0.84	18181	0.69	31251	0.38
p > 0.5 and p <=0.6	36702	0.76	44204	0.81	27214	0.71	15000	0.57	25905	0.31
p > 0.6 and p <=0.7	32937	0.68	39912	0.73	24927	0.65	13262	0.50	22423	0.27
p > 0.7 and p <=0.8	31274	0.65	38509	0.70	24031	0.63	12667	0.48	21075	0.25
p > 0.8 and p <=0.9	32812	0.68	40547	0.74	26288	0.69	13531	0.51	21859	0.26
p > 0.9	43677	0.90	56566	1.03	37799	0.99	17782	0.68	30421	0.37

The overall summary for the diagnostic algorithm over 4 million pairs identified by the algorithm and approximately.24% in Optum and .04% in Premier of those pairs have a probability over 0.5. The distribution of the probabilities (Figure 16) show a spike in the number of pairs identified between 0.1 and 0.2 and this due to a large proportion of pairs having at least one coefficient value equal to 0, thus the probability is the result of the intercept term plus one or two variables rather than the 3 terms plus the intercept term. The terms in the diagnostic algorithm require the "first" exposures for both the condition and procedure and the count of the occurrence of both at time 0 (support) and the sensitivity at time 0.

The overall summary for the therapeutic algorithm show approximately 3% of the total pairs having a probability greater than 0.5 in the claims databases (Optum, CCAE, MDCD, and MDCR) and Premier at 1.48% (Figure 17). The overall distribution of pairs shows a high number closer to probabilities 0.10 and a downward trend and small increase closer to $p > 0.9$. There are a large number of pairs that are located between $p > 0.2$ and $p < 0.8$ that do not necessarily fall into a positive or negative category would need further evaluation to determine a proper cut off point.

From the visual evaluation of the pairs that fall into either the diagnostic or therapeutic algorithm, the therapeutic algorithm has more of the expected pattern, large amounts of pairs on either end with some level of uncertainty in the middle. The diagnostic algorithm has some limitations as it seems that there are many covariates with a value close to or equal to 0 which cause many pairs to have a lower probability than if

they were excluded from the algorithm all together. Overall the therapeutic algorithm has a greater number of pairs with a probability over 0.5.

Figure 15. Distribution of diagnostic algorithm probabilities greater than 0.1 by database

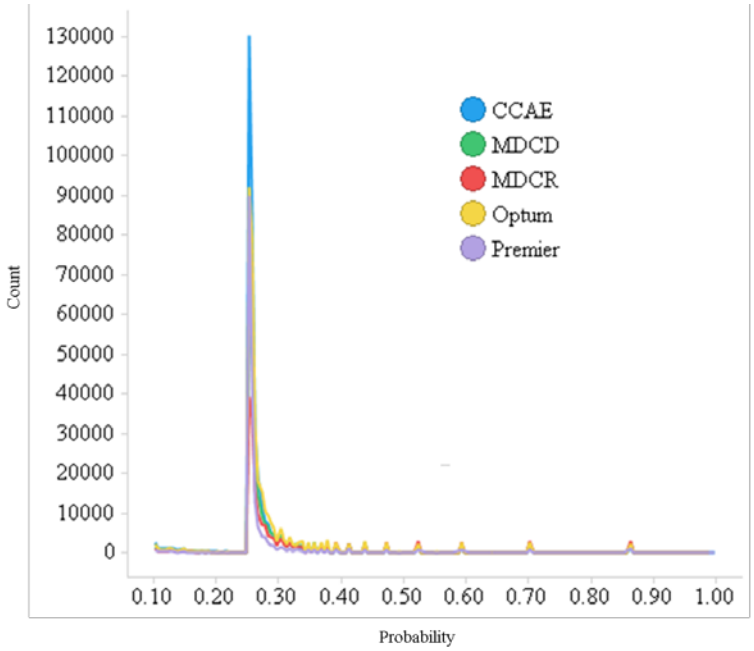


Figure 16. Distribution of diagnostic algorithm probabilities greater than 0.1 by database

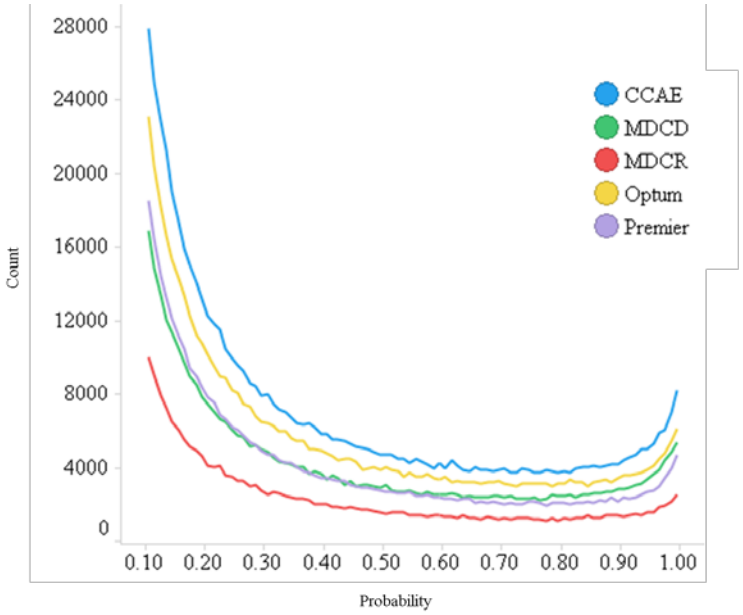


Figure 17. Coverage of codes and data for conditions by database and probability value

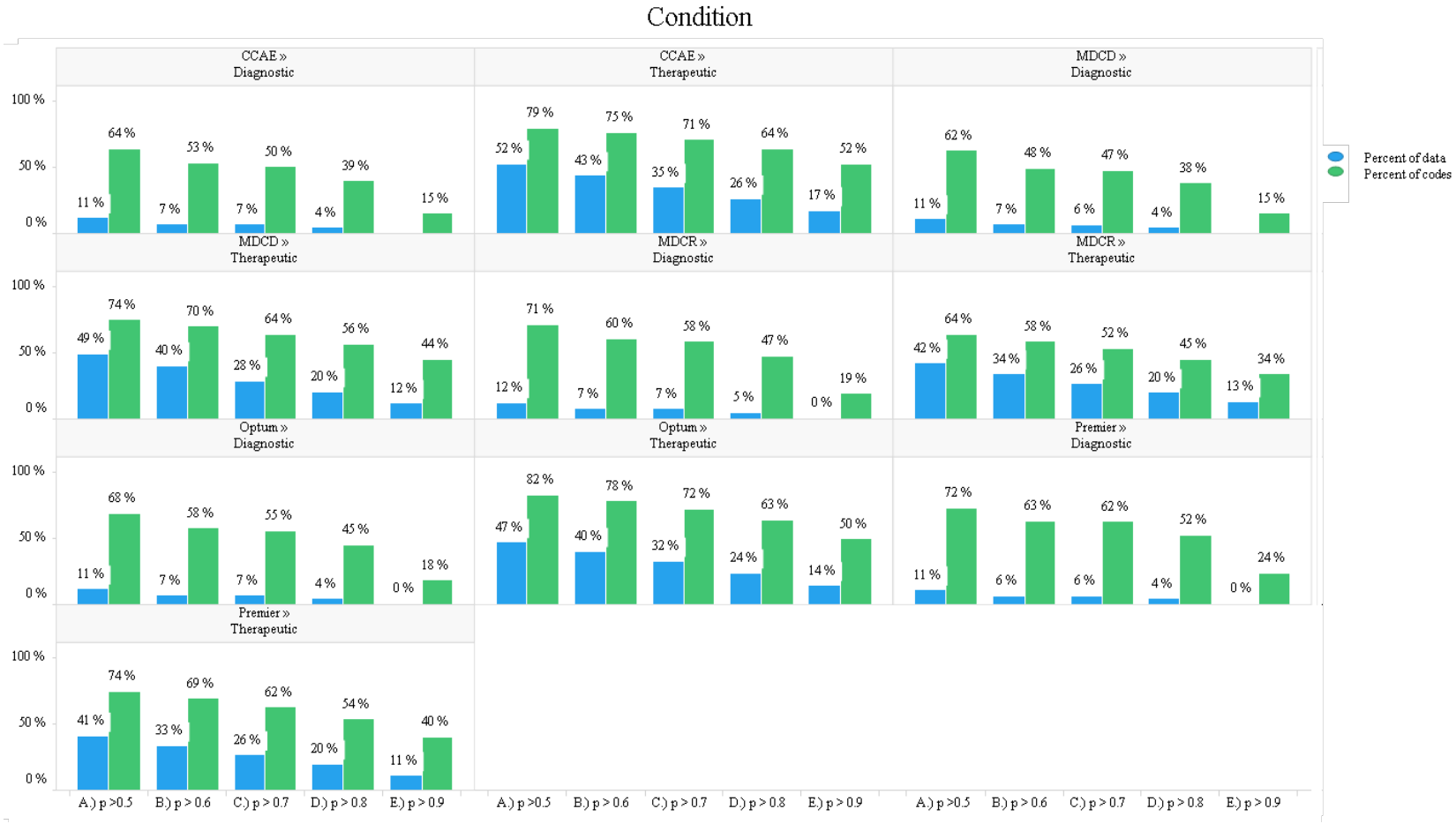
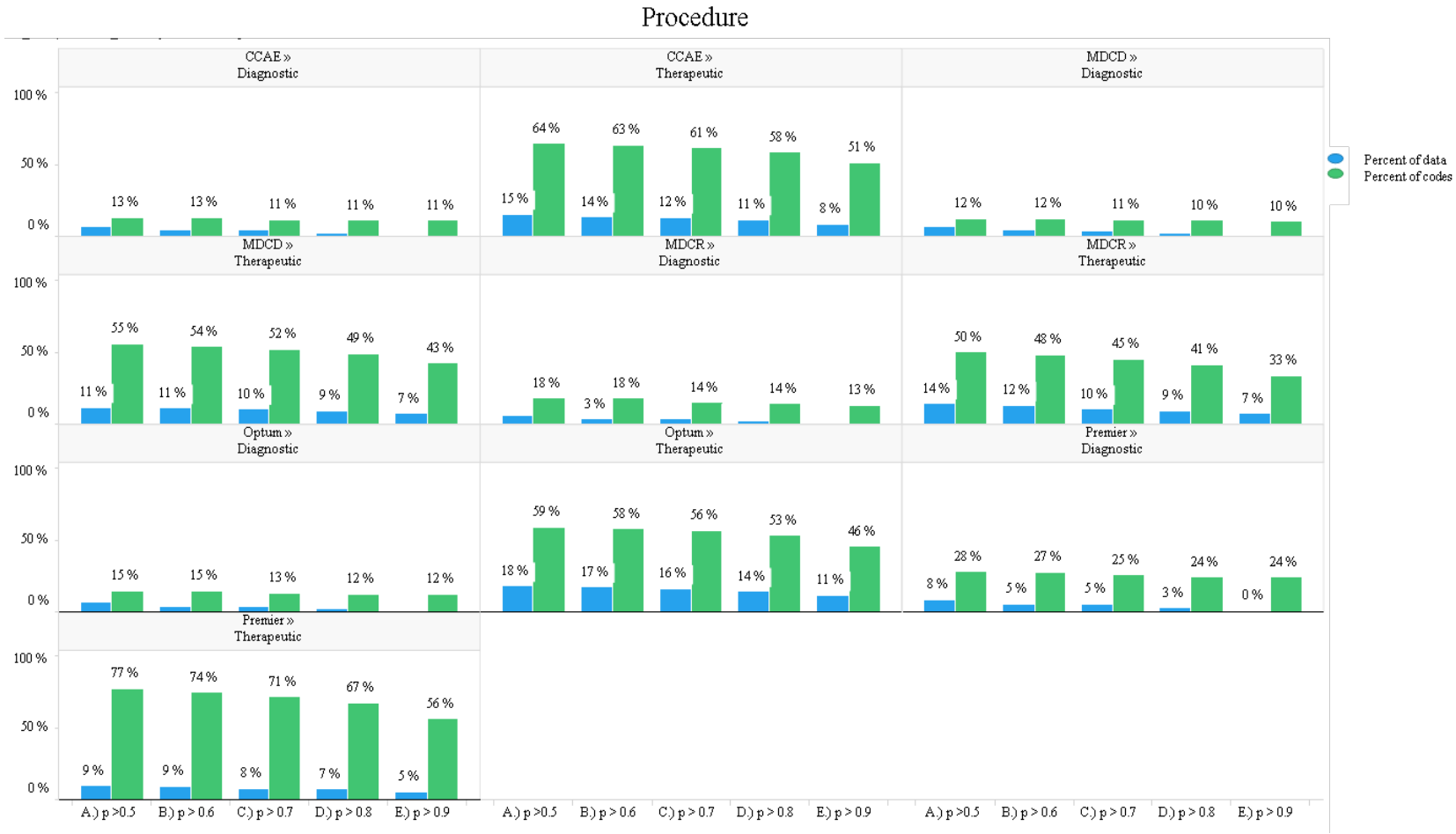


Figure 18. Coverage of codes and data for procedures by database and probability value



The coverage of codes and data represented by domain for both algorithms is evaluated but determining the percentage of codes utilized from the overall number of codes available and the amount of data those codes represent. For conditions (Figure 18), Optum has over 64% of condition codes with a probability of greater than 0.5 for the diagnostic algorithm. Overall over 60% of condition codes are utilized by both algorithms and represent different proportions of the data based on the algorithm type. The therapeutic algorithm covers over 40% of overall data for all databases for those pairs with $p > 0.5$.

The coverage of codes and data for the procedure domain vary compared to conditions, as there are many procedure codes that can billed for a visit that do not have a logical map to a condition. Over 55% of codes are utilized in the therapeutic algorithm and represents 9-18% of the data.

An adjudication of the top 100 pairs for each algorithm and database is conducted to understand how well the pairs have been mapped. From textual verification of a condition and procedure pair a ‘yes’ is given if the words/phrases are about each other, for example the conditions of chalazion and the associated procedure is: chalazion removal or common knowledge can applied to understand the relationship. A ‘maybe’ is indicated for those pairs that cannot be logically verified without the appropriate clinical background, and a ‘no’ is indicated for those pairs that at face-value do not belong together, for example if they are two different body structures (i.e. a one stage myelotomy to treat anxiety disorder, dysuria to diagnose a procedure on bone). Table 29 presents the results of the adjudication by algorithm for each database.

Table 28. Adjudication results

		Database				
Diagnostic Algorithm		Optum	CCAE	MDCD	MDCR	Premier
	Total	100	100	100	100	69
	Yes	38	54	30	46	28
	No	9	10	6	1	0
	Maybe	53	36	64	53	41
Therapeutic Algorithm		Optum	CCAE	MDCD	MDCR	Premier
	Total	100	100	100	100	100
	Yes	31	30	31	58	39
	No	1	1	1	2	1
	Maybe	68	69	68	40	60

On overage from both algorithms, approximately 30% of pairs are classified “Yes”, while ~ 60-70% are “Maybe” and ~1-10% are classified as “No”. While some proportion of pairs lend themselves well textually and through common knowledge to figure out if the pair is a good match a large proportion of pairs are unknown and would need further adjudication by a clinician.

To understand how much concordance the pairs have within various databases we compare the overlap of pairs in multiple databases and their associated probabilities. For example, if the same pair has a probability of greater than 0.9 and occurs at that level in all 5 databases, it provides greater certainty of a valid pair. Table 30 presents the algorithms and overlap of pairs by database.

Table 29. Overlap of pairs

	Number of Database	**Total number of pairs**	**p >0.5**	**p >0.6**	**p >0.7**	**p >0.8**	**p >0.9**

Diagnostic Algorithm	1	9,506,799	41,840	22,482	21,417	10,892	678
	2	838,247	984	478	348	127	25
	3	882,462	382	235	157	70	32
	4	1,023,971	357	228	147	87	18
	5	1,385,840	16	8	1	-	-
Therapeutic Algorithm	1	9,506,799	513,887	413,859	323,164	234,940	138,766
	2	838,247	32,147	25,539	19,358	13,456	7,551
	3	882,462	19,405	14,975	11,108	7,674	4,217
	4	1,023,971	13,302	10,118	7,581	5,162	2,809
	5	1,385,840	8,344	6,393	4,662	3,152	1,698

For all pairs found, 69.71% appear in only 1 database, 6.15% appear in two, 6.47% appear in three, 7.51% appear in 4 and 10.16% appear in all five databases. In both the therapeutic and diagnostic algorithm less than 1% of pairs appear in all 5 databases at the same probability level. Overlap of pairs by database shows less concordance overall and by probability level which is likely due to variation in coding practices and populations that utilizing the healthcare system differently.

To understand the distinction between the two algorithms and if the algorithms produce two distinct algorithms that could distinguish between diagnostic condition-procedure pairs and therapeutic pairs, the overlap between the algorithms was assessed.

For pairs that have a p greater than 0.5, the number of pairs that are distinctly diagnostic or therapeutic or appear in both. The percentage of pairs that are in both algorithms relative to the number of the pairs in the algorithm is compared.

Table 30. Algorithm overlap for p >0.5

Database	Only Diagnostic	Only Therapeutic	Both	% of diagnostic pairs that appear in both algorithms (Both/Total diagnostic pairs p >0.5)	% of therapeutic pairs that appear in both algorithms (Both/Total diagnostic pairs p >0.5)
Optum	3845	169469	7933	67.35%	4.47%
CCAE	1414	214418	5320	79.00%	2.42%
MDCD	3253	132446	7813	70.60%	5.57%
MDCR	5531	64629	7613	57.92%	4.29%
Premier	1340	119283	2400	64.17%	1.35%

The percentage of diagnostic pairs that overlaps in the therapeutic algorithm is greater than 57% in all databases. These results indicate that despite having a large proportion of pairs identified in the data, it may be difficult to determine if a pair is diagnostic or therapeutic. An additional sensitivity analysis to take the positive control data and run a logistic regression model to determine the binary variable of D, or diagnostic or T, therapeutic.

4.7. ALGORITHM DIFFERENTIATION

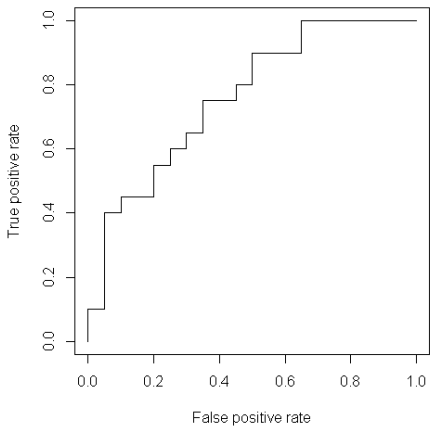
The full model results from inputting all 51 variables into the model results in perfect separation of the outcome variables. Model separation can occur for many reasons such as the predictor in some form or fashion is a covariate or sample size is very small. Penalized regression models such as step-wise or LASSO should produce a value.

The stepwise procedure procedures the following model:

model_numeric ~-6.559433+3.07256*rr_ratio_all_time+0.030643*num_days_btw+-0.003301*ratio_inpt+-43.454089*misclassification_all_time+5.14732*proc_first+-1.684642*rr_ratio_365+-0.027213*abs_female+26.829212*specificity_0

The AUC for this model is 0.760.

Figure 19. ROC curve for step wise selection to determine intervention type



Using LASSO regression to determine feature selection, the following 14 variables remained in the model: cond_proc_ratio, abs_female,avg_age_ratio, avg_per_person, ratio_inpt, num_days_btw, avg_time_cond_days, cond_proc_fsame, rx_60,

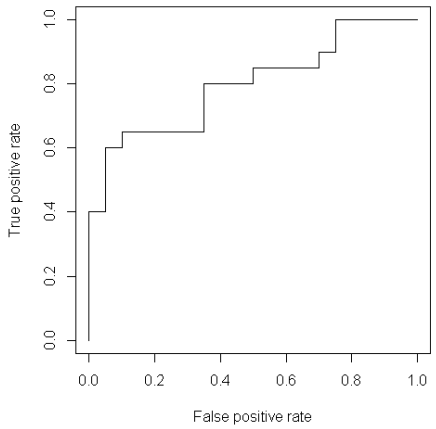
rr_ratio_all_time, support_365, misclassification_90, misclassification_all_time, sensitivity_all_time.

Using logistic regression the following model was produced:

model_numeric ~-3.469+-0.000299*cond_proc_ratio+-0.02126*abs_female+-0.2211*avg_age_ratio+-0.372*avg_per_person+-0.0009093*ratio_inpt+0.02068*num_days_btw+0.2483*avg_time_cond_days+-0.3148*cond_proc_fsame+-0.005092*rx_60+1.596*rr_ratio_all_time+-0.00002218*support_365+11.43*misclassification_90+-25.47*misclassification_all_time+1.078*sensitivity_all_time

The AUC for this model is: 0.7975.

Figure 20. ROC curve for LASSO feature selected covariates



The AUC's from all four model selected techniques resulted in AUC's greater .60, and when using the covariates selected from the LASSO regression the AUC is almost a .80, which gives confidence in the ability of these variables to predict the type of pair once a model selected the condition-procedure pairs. This analysis challenges the single model derivation for determining both a relationship and model type and instead

presumes that the determination of the relationship is agnostic to the type of procedure and that becomes a second step in the process or another relationship type in the context of an ontology.

V. CHAPTER 5 DISCUSSION

This research represents the journey of using predictive modeling techniques and real-world data to develop knowledge in the medical domain. The knowledge of condition procedure pairs was something that mainly exists in sources of literature such as textbooks or in the minds of clinicians. Developing the relationship of conditions and procedures from real world data provides the user with reliable and usable information because these are the conditions and procedures that millions of patients are being seen for and being treated for. This research applied a method to derive a solution that can be updated in the future by reapplying the algorithms quickly and efficiently to new sources and update existing data.

The evaluation of the SNOMED-CT terminology gives a high level look at how well the vocabulary can map the source codes in the procedure domain and amongst those codes that do not have maps or appear in multiple databases, and if that would be a limitation for the final algorithm. The total number of codes utilized in the databases ranged widely depending on the type of code, CPT-4 codes being the most utilized and HCPCS being utilized the least which is what we would have expected in a claims database since most encounters are outpatient. The evaluation of the mapping that already

exists between source code terminologies and SNOMED-CT also has a wide range depending on the source code type with approximately 80% of CPT-4 codes having a valid SNOMED-CT map. Amongst the missing codes, the unmapped codes were mainly composed of codes that are administration codes. Missing codes, or those that aren't utilized in the database are important to understand the limitation of the algorithm Understanding how missing codes may affect the generalizability when eventually applying the algorithm, the evaluation of how often its missing from multiple databases shows that more than 50% of CPT-4 codes don't appear in any of the 5 databases indicated that they may be underutilized across multiple databases or not used at all. If a source code is missing in more than 3 databases, an assumption is made that the source code may never be utilized but if codes are missing in only 1 database we may have codes that could be likely candidates for a condition-procedure relationship and the distribution may affect the overall algorithms due to the inconsistencies of missing codes. The evaluation of the source codes and SNOMED-CT vocabularies indicates that the mapping will be minimally affected by limitations in these datasets from utilization of codes, missing codes and unmapped codes.

The construction of the positive controls and covariates in the model produced sizable data to train and test the model. The effort to collate knowledge from literature sources was a time consuming, and labor-intensive task that resulted in 100 positive controls for each model type. The covariates derived for the algorithm utilized many various aspects of data elements from drugs, measurements, and co-occurrence statistics. From applying various model selection techniques, a final diagnostic and therapeutic model were selected by evaluating AUC. The evaluation of various predictive modeling

techniques showed a wide variance of AUC among the models trained/tested. The diagnostic algorithm utilized 3 variables: first occurrence of the condition and procedure, support at time 0, and sensitivity at time 0. The algorithm covariates utilize first occurrences which makes clinical sense as a diagnosis are made when the patient first interacts with the healthcare system. The therapeutic algorithm resulted in two variables: relative risk ratio all time and count of condition-procedures that occur on the same day. The therapeutic algorithm utilizes variables that take in account variables that look throughout time, clinically therapies can be applied at various time intervals within a patient's journey after diagnosis. The algorithms were developed on the Optum SES database and external validation was performed to see if the algorithms are generalizable to other datasets and the AUC's were all above .90 expect for Premier.

The algorithm was then applied on all the data available and evaluated for the number of pairs determined at various p cut-offs. Both algorithms covered over 60% of the condition codes and approximately 50% of procedure codes. Adjudication of pairs revealed that some clinical expertise is necessary to understand the quality of the mappings. Finally, the overlap in pairs amongst the two algorithms prompted a secondary analysis to determine intervention type. By fitting a model for intervention type with the existing dataset, an AUC of .79 was achieved. Now we can generate condition-procedure pairs and then determine the intervention type rather than having preconceived relationships about condition-procedure pairs being diagnostic or therapeutic. This would develop a secondary relationship of intervention type in addition to condition-procedure pair.

This knowledge can be utilized to fuel research efforts in healthcare, when understanding conditions and their most common procedures and their application to diagnose or treat a patient. The data can be utilized in observational research for phenotyping a definition or cohort construction.

5.1. LIMITATIONS & NEXT STEPS

The biggest limitation in the research relies on the fact that the pairs will eventually need some manual intervention to classify and truly determine if they are in fact related. But the ability to use the data to generate relationships that could be verified by a clinician is a great step forward to the alternative of selecting a condition and then finding its associated procedure from a vast list of procedures that would have no context in their use in the real world. Another limitation is that most databases utilized in this research are typical US claims datasets and its application on EHR datasets is limited. The use of Premier which is primarily a hospital dataset showed results that were less favorable when external validation was conducted with AUC's lower than .90 for the diagnostic algorithm As well as overall the number of viable pairs in this dataset is limited in comparison to the claims, and part of the reason is due to data availability it does not have the longitudinally as claims or other EHR's may have so testing the algorithm externally could be another validation.

Additionally, we can test our hypothesis to see we can further create one parsimonious algorithm to determine the relationship between condition and procedure and then apply a second algorithm to determine if those pairs are diagnostic or therapeutic. The two algorithms resulted in very similar covariates as well as the overlap

of almost 60% of the diagnostic algorithm in the therapeutic algorithm indicates that it may be useful to use all the controls to create a single algorithm, and then determine additional parameters determine the type of pair, diagnostic or therapeutic.

From the data already collected an ideal next step would be to come up with a method to quickly annotate through the pairs found and build a formal ontology that can identify the relationship between the condition and procedure. Also determine methods to update the pairs as new data is available and incorporate them into the corpus of knowledge. Future research can include applying this method to derive other relationships such as conditions and measurements.

VI. BIBLIOGRAPHY

1. Finney Rutten LJ, Vieux SN, St Sauver JL, et al. Patient perceptions of electronic medical records use and ratings of care quality. *Patient related outcome measures.* 2014;5:17-23.
2. Krist AH, Woolf SH. A vision for patient-centered health information systems. *Jama.* 2011;305(3):300-301.
3. Services USCfMM. Electronic Health Records. 2012. Accessed 10/12/2017, 2017.
4. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clinical research in cardiology : official journal of the German Cardiac Society.* 2017;106(1):1-9.
5. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International journal of epidemiology.* 2015;44(3):827-836.
6. TR G. Toward principles for the design of ontologies used for knowledge sharing? . *Int J Hum Comput Stud.*43(5-6):907-928.
7. Kramer F, Beissbarth T. Working with Ontologies. *Methods in molecular biology (Clifton, NJ).* 2017;1525:123-135.
8. Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: a case study. *Journal of biomedical informatics.* 2007;40(3):353-364.
9. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *The New England journal of medicine.* 2010;363(6):501-504.
10. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Science translational medicine.* 2010;2(57):57cm29.
11. Xu Y, Zhou X, Suehs BT, et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. *Drug safety.* 2015;38(8):749-765.
12. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association : JAMIA.* 2012;19(1):54-60.
13. Reich C RP, Belenkaya R, Natarajan K, Blacketer C OMOP Common Data Model v5.2 Specifications. 2017; https://github.com/OHDSI/CommonDataModel/wiki. Accessed 9/1/2017, 2017.
14. Humphreys BL, Lindberg DA, Hole WT. Assessing and enhancing the value of the UMLS Knowledge Sources. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care.* 1991:78-82.
15. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of biomedical informatics.* 2012;45(4):689-696.

16. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. *Proceedings / AMIA Annual Symposium AMIA Symposium.* 2001:662-666.
17. Hand D, Mannila, H., Smyth, P. *Principles of data mining.* MIT; 2001.
18. Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearbook of medical informatics.* 2018;27(1):129-139.
19. Hirsch JA, Leslie-Mazwi TM, Nicola GN, et al. Current procedural terminology; a primer. *Journal of neurointerventional surgery.* 2015;7(4):309-312.
20. Nusgart M. HCPCS Coding: An Integral Part of Your Reimbursement Strategy. *Advances in wound care.* 2013;2(10):576-582.
21. Organization WH. History of the development of the ICD http://www.who.int/classifications/icd/en/. Accessed 4/30/2018, 2018.
22. Weiner M. Computerized Physician Order Entry. In: Liu L, ÖZsu MT, eds. *Encyclopedia of Database Systems*. Boston, MA: Springer US; 2009:432-437.
23. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. *Yearbook of medical informatics.* 2006:124-135.
24. Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. *Journal of innovation in health informatics.* 2015;22(2):309-315.
25. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical-concept representation language. The Canon Group. *Journal of the American Medical Informatics Association : JAMIA.* 1994;1(3):207-217.
26. Cimino JJ. In defense of the Desiderata. *Journal of biomedical informatics.* 2006;39(3):299-306.
27. Health NIo. UMLS Quick Start Guide. 2017; https://www.nlm.nih.gov/research/umls/quickstart.html. Accessed 12/15/2017.
28. Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical chemistry.* 1996;42(1):81-90.
29. Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. *Journal of the American Medical Informatics Association : JAMIA.* 1998;5(3):276-292.
30. Inc. RI. Regenstrief and the SNOMED International are working together to link LOINC and SNOMED CT. 2013; https://loinc.org/news/new-regenstrief-and-ihtsdo-agreement-to-make-emrs-more-effective-at-improving-health-care/, 2018.
31. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug safety.* 1999;20(2):109-117.
32. Wood KL CR, Wood SM. *MEDDRA: the basis for the new international medical terminology for regulatory purposes.* 2 ed: ESRA Rapporteur; 1995.
33. FDA US Food and Drug Administration (2015) Adverse Event Reporting System. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. Accessed 5/11/2015, 2015.
34. Association AM. CPT® Purpose & Mission. 2018; https://www.ama-assn.org/practice-management/cpt-purpose-mission.

35. Medicare CoMa. ICD-10-CM, ICD-10-PCS, CPT, AND HCPCS CODE SETS. 2018; https://www.cms.gov/Medicare/Coding/ICD10/. Accessed December, 2018.
36. Medicare CfMa. HCPCS Level II Coding Process & Criteria. 2018; https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/HCPCSCODINGPROCESS.html. Accessed September 15, 2018.
37. (IHTSDO) IHSDO. The SNOMED-CT Starter Guide. 2014; www.snomed.org/doc, 2018.
38. Willett DL, Kannan V, Chu L, et al. SNOMED CT Concept Hierarchies for Sharing Definitions of Clinical Conditions Using Electronic Health Record Data. *Applied clinical informatics.* 2018;9(3):667-682.
39. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics.* 2006;121:279-290.
40. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association : JAMIA.* 2014;21(e1):e11-19.
41. Organisation IHTSD. Decision Support with SNOMED CT. 2018; http://snomed.org/cds.
42. Informatics OHDSa. Data Standardization. 2017; https://www.ohdsi.org/data-standardization/. Accessed 8/23/2017, 2017.
43. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association : JAMIA.* 2015;22(3):553-564.
44. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *EGEMS (Washington, DC).* 2014;2(1):1110.
45. Jiang G, Wang L, Liu H, Solbrig HR, Chute CG. Building a knowledge base of severe adverse drug events based on AERS reporting data using semantic web technologies. *Studies in health technology and informatics.* 2013;192:496-500.
46. Wang L, Haug PJ, Del Fiol G. Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository. *Journal of biomedical informatics.* 2017;69:259-266.
47. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics.* 2010;43(6):891-901.
48. Burton MM, Simonaitis L, Schadow G. Medication and indication linkage: A practical therapy for the problem list? *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium.* 2008:86-90.
49. Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. *Science translational medicine.* 2011;3(114):114ra127.
50. Voss EA, Ma Q, Ryan PB. The impact of standardizing the definition of visits on the consistency of multi-database observational health research. *BMC medical research methodology.* 2015;15:13.

51. Rothman K, Greenland, S., & Lash, TL. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
52. Hosmer DW Jr LS, Sturdivant RX *Applied Logistic Regression.* 3 ed. New Jersey: John Wiley & Sons; 2013.
53. James G, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert. *An Introduction to Statistical Learning with Application in R.* New York: Springer; 2013.
54. R. T. Regression shrinkage and selection via the Lasso. *J R Statist Soc B.* 1996;58(1):267-288.
55. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association : JAMIA.* 2018;25(8):969-975.

APPENDIX A: NEGATIVE CONTROL LIST