AN INTEROPERABLE ELECTRONIC MEDICAL RECORD-BASED PLATFORM
FOR PERSONALIZED PREDICTIVE ANALYTICS

Hamed Abedtash

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

July 2017
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Josette F. Jones, RN, PhD, Chair

Jon D. Duke, MD, MS

Jennifer Wessel, PhD, MPH

Xiaochun Li, PhD

Richard J. Holden, PhD

May 31, 2017
DEDICATION

To my wife, Shirin who has always supported me with love and encouragement
ACKNOWLEDGEMENTS

I would like to acknowledge the assistance of Regenstrief Institute and Mr. Andrew Martin for providing HL7 CCD files to this project. My thanks go to Dr. Jennifer Wessel, Dr. Xiaochun Li, and Dr. Richard Holden for their comments. My special thanks are also due to my supervisors Dr. Jon D. Duke and Dr. Josette Jones for their support and continuous guidance.
Precision medicine refers to the delivering of customized treatment to patients based on their individual characteristics, and aims to reduce adverse events, improve diagnostic methods, and enhance the efficacy of therapies. Among efforts to achieve the goals of precision medicine, researchers have used observational data for developing predictive modeling to best predict health outcomes according to patients’ variables.

Although numerous predictive models have been reported in the literature, not all models present high prediction power, and as the result, not all models may reach clinical settings to help healthcare professionals make clinical decisions at the point-of-care. The lack of generalizability stems from the fact that no comprehensive medical data repository exists that has the information of all patients in the target population. Even if the patients’ records were available from other sources, the datasets may need further processing prior to data analysis due to differences in the structure of databases and the coding systems used to record concepts.

This project intends to fill the gap by introducing an interoperable solution that receives patient electronic health records via Health Level Seven (HL7) messaging standard from other data sources, transforms the records to observational medical outcomes partnership (OMOP) common data model (CDM) for population health research, and applies predictive models on patient data to make predictions about health outcomes.

This project comprises of three studies. The first study introduces CCD-TO-OMOP parser, and evaluates OMOP CDM to accommodate patient data transferred by HL7 consolidated continuity of care documents (CCDs). The second study explores how to adopt predictive model markup language (PMML) for standardizing dissemination of OMOP-based predictive models. Finally, the third study introduces Personalized Health Risk Scoring Tool (PHRST), a pilot, interoperable OMOP-based model scoring tool that processes the embedded models and generates risk scores in a real-time manner.
The final product addresses objectives of precision medicine, and has the potentials to not only be employed at the point-of-care to deliver individualized treatment to patients, but also can contribute to health outcome research by easing collecting clinical outcomes across diverse medical centers independent of system specifications.

Josette F. Jones, RN, PhD, Chair
# TABLE OF CONTENTS

List of Tables................................................................................................................................. xii
List of Figures ................................................................................................................................. xiv
List of Appendices ............................................................................................................................ xvi
List of Abbreviations......................................................................................................................... xvii
Chapter 1. Introduction ....................................................................................................................... 1
  1.1. Background ............................................................................................................................... 1
  1.2. Statement of the Problem ......................................................................................................... 3
  1.3. Conceptual Framework of the Study ....................................................................................... 3
  1.4. Purpose of the Study ............................................................................................................... 6
  1.5. Research Questions ................................................................................................................. 8
  1.6. Definition of Terms ................................................................................................................. 8
  1.7. Significance of the Study ........................................................................................................... 9
  1.8. Organization of the Study ........................................................................................................ 10
Chapter 2. OMOP Common Data Model Accommodates HL7 Consolidated CDA-based CCD Data for Population Health Research 11
  2.1. Introduction ............................................................................................................................. 11
    2.1.1. Problem statement .............................................................................................................. 11
    2.1.2. HL7 continuity of care document ....................................................................................... 12
    2.1.3. The Observational Medical Outcomes Partnership (OMOP) ............................................. 12
    2.1.4. Objectives .......................................................................................................................... 14
  2.2. Methods and Materials ............................................................................................................. 17
    2.2.1. HL7 Continuity of Care Documents (CCDs) ................................................................. 17
    2.2.2. OMOP vocabulary ............................................................................................................ 17
    2.2.3. Overview of the ETL package ............................................................................................ 17
    2.2.4. Extraction of HL7 CCD data ............................................................................................. 19
    2.2.5. Transformation of HL7 CCD data to OMOP CDM .......................................................... 19
      2.2.5.1. Mapping source values to OMOP concepts ................................................................. 20
      2.2.5.2. Person table .................................................................................................................. 21
      2.2.5.3. Observation Period table ............................................................................................. 23
      2.2.5.4. Visit Occurrence table .................................................................................................. 24
      2.2.5.5. Condition Occurrence table .......................................................................................... 26
2.2.5.6. Condition Era table ................................................................. 30
2.2.5.7. Procedure Occurrence table .................................................. 31
2.2.5.8. Drug Exposure table ............................................................... 34
2.2.5.9. Drug Era table ........................................................................ 41
2.2.5.10. Measurement table ................................................................. 42
2.2.5.11. Observation table .................................................................... 47
2.2.6. Performance assessment of ETL pipeline ..................................... 51
  2.2.6.1. Evaluation of data extraction .................................................. 51
  2.2.6.2. Evaluation of concept mapping ................................................ 51
  2.2.6.3. Evaluation of derived elements ............................................... 52
  2.2.6.4. Evaluation of calculated data fields ......................................... 52
  2.2.6.5. Evaluation of data loading .................................................... 52
2.2.7. Statistical methods ..................................................................... 53
2.3. Results ........................................................................................... 54
  2.3.1. Performance of data extraction pipeline ..................................... 54
  2.3.2. Performance of data transformation pipeline .............................. 54
    2.3.2.1. Standardized derived elements ............................................. 54
  2.3.3. Mapping performance of CCD data to OMOP CDM .................. 55
    2.3.3.1. Conditions and diagnoses .................................................... 55
    2.3.3.2. Drugs ................................................................................ 55
    2.3.3.3. Procedures ......................................................................... 56
    2.3.3.4. Measurements and clinical evaluations ................................. 56
    2.3.3.5. Observations ...................................................................... 56
    2.3.3.6. CDC race and ethnicity value set ......................................... 57
  2.3.4. Accuracy of concept mapping .................................................... 59
  2.3.5. Completeness of CCD data elements required by OMOP CDM .... 60
2.4. Discussion ....................................................................................... 62
  2.4.1. Accommodation of CCD data in OMOP CDM ............................ 62
  2.4.2. Strengths .................................................................................. 63
  2.4.3. Challenges ................................................................................ 64
  2.4.4. Limitations of the study ............................................................ 65

Chapter 3. A Standard for Disseminating Health Risk Prediction Models to
Support Clinical Decision-Making

ix
3.1. Background .................................................................66
3.2. Problem statement ..........................................................66
3.3. Overview of PMML ............................................................67
3.4. The Structure of PMML .......................................................67
3.5. Objectives .................................................................69
3.6. Methods ........................................................................70
  3.6.1. The structure of OMOP-compliant PMML .........................70
  3.6.2. The OMOP-compliant PMML scoring engine .....................80
  3.6.3. Case study: CVD risk scoring model ................................81
      3.6.3.1. Data source .........................................................82
      3.6.3.2. The Framingham 10-year risk of CVD scoring model ........82
      3.6.3.3. The specifications of O-PMML containing Framingham
                algorithms ...............................................................83
      3.6.3.4. Performance assessment of Framingham O-PMML .......88
3.7. Results ..........................................................................88
  3.7.1. Patient characteristics ..................................................88
  3.7.2. Risk scores .............................................................89
3.8. Discussion .......................................................................90
  3.8.1. An overview of O-PMML ..............................................90
  3.8.2. Advantages of O-PMML ................................................90
  3.8.3. Limitations ..................................................................91

Chapter 4. The Interoperable System For Delivering Personalized Health

Outcome Predictions 92

4.1. Background .................................................................92
4.2. Objectives .................................................................93
4.3. Methods .......................................................................93
  4.3.1. The personalized health outcome prediction framework .......93
  4.3.2. Architecture and dataflow of PHRST ................................95
  4.3.3. Case study: Framingham risk functions in action .............97
      4.3.3.1. Data source .........................................................97
      4.3.3.2. Predictive models ..................................................97
4.4. Results .........................................................................97
  4.4.1. PHRST development ....................................................97
4.4.2. Deployment of Framingham risk functions on PHRST............................99

4.5. Discussion ........................................................................................................99
  4.5.1. An overview ..................................................................................................99
  4.5.2. Advantages of the framework .......................................................................99
  4.5.3. Similar studies .............................................................................................100
  4.5.4. Limitations ..................................................................................................100

Chapter 5. Conclusion ..............................................................................................101
  5.1. OMOP CDM Accommodates HL7 Consolidated CCD Data ......................101
  5.2. A New Standard Enables Sharing Health Risk Prediction Models .............101
  5.3. An Interoperable System Delivers Personalized Health Outcome
         Predictions .......................................................................................................101
  5.4. Future Work ....................................................................................................101

Appendices .............................................................................................................103
References ............................................................................................................149

Curriculum Vitae
LIST OF TABLES

Table 1. A summary of document-level template of HL7 C-CDA Consolidated Continuity of Care Document (CCD) Release 1.1 and corresponding Object identifiers (OID). ................................................................. 15

Table 2. Mapping HL7 C-CDA CCD data to Person table: Applied rules and corresponding sections. .......................................................... 22

Table 3. Matching OMOP concepts of HL7 administrative gender codes ........................................ 23

Table 4. Matching OMOP concepts of CDC race and ethnicity value set .................................... 23

Table 5. Applied rules to build Observation Period table .................................................................. 24

Table 6. Mapping HL7 C-CDA data to Visit Occurrence table: Applied rules and corresponding sections .................................................................. 25

Table 7. Mapping HL7 C-CDA CCD data to Condition Occurrence table: Applied rules and corresponding sections ............................................ 28

Table 8. Applied rules to build Condition Era table ........................................................................ 30

Table 9. Mapping HL7 C-CDA CCD data to Procedure Occurrence table: Applied rules and corresponding sections ............................................. 32

Table 10. Corresponding quantity units to CCD quantity values in Drug Exposure table .................................................. 37

Table 11. Mapping HL7 C-CDA CCD data to Drug Exposure table: Applied rules and corresponding sections .................................................................. 38

Table 12. Applied rules to build Drug Era table ............................................................................. 41

Table 13. Mapping HL7 C-CDA CCD data to Measurement table: Applied rules and corresponding sections .................................................................. 44

Table 14. Matching OMOP concepts of HL7 smoking status value set ........................................... 48

Table 15. Mapping HL7 C-CDA CCD data to Observation table: Applied rules and corresponding sections .................................................................. 49

Table 16. An excerpt of records in Condition Era table ................................................................... 54

Table 17. An excerpt of records in Drug Era table ......................................................................... 55

Table 18. Mapping performance of source codes to standard concepts of OMOP CDM vocabulary .................................................................. 58

Table 19. Accuracy assessment of ETL pipeline to map source codes to standard concepts .................................................................. 59
Table 20. The number of patients and records for which CCDs carried data. ..........61
Table 21. The criteria applied to collect values of participating variables in
   Framingham 10-year risk of cardiovascular disease........................................84
Table 22. The characteristics of patient records for which Framingham 10-
   year risk score of cardiovascular disease was generated ..................................88
LIST OF FIGURES

Figure 1. The scope of the project is to deploy predictive models on patient data exchanged via HL7 messaging standard. The project involves standardizing transferred patient information into OMOP CDM, introducing a new standard for disseminating OMOP-based predictive models, and building a pilot solution for deployment of health risk scoring models. ................................................................. 8

Figure 2. OMOP Common Data Model Version 5.1 conceptual model (83) ................. 13

Figure 3. A schematic of CCD-TO-OMOP package ............................................ 18

Figure 4. Overall mapping performance of concepts and records to OMOP CDM vocabulary by domain ................................................................. 57

Figure 5. The general structure of a PMML document (left), and the schema of sections under Regression Model (right) ........................................... 69

Figure 6. The XML schema of the general architecture of O-PMML ......................... 71

Figure 7. The XML schema of Header section in O-PMML ................................... 73

Figure 8. The XML schema of Data Dictionary section and Data Field elements in O-PMML ........................................................................... 74

Figure 9. The XML schema of Transformation Dictionary section, Local Transformations section, Derived Field elements, and acceptable data transformation expressions in O-PMML ....................................................... 75

Figure 10. The new XML schema of Mining Build Task section and the required elements and attributes in O-PMML ........................................... 77

Figure 11. The XML schema of Regression Model section in O-PMML ...................... 78

Figure 12. The XML schema of Mining Schema section and Mining Field elements under Regression Model in O-PMML ........................................... 79

Figure 13. The XML schema of Regression Table section under Regression Model ...................................................................................... 79

Figure 14. The XML schema of Output section and Output Field elements in O-PMML ...................................................................................... 80

Figure 15. A schematic of O-PMML scoring engine .................................................. 81

Figure 16. Framingham 10-year risk of CVD for men (105) ....................................... 82

Figure 17. Framingham 10-year risk of CVD for women (105) ................................. 82
Figure 18. An excerpt of the Mining Build Task section of Framingham O-PMML .................................................................83

Figure 19. The Data Dictionary section of Framingham O-PMML ..................85

Figure 20. The Transformation Dictionary section of Framingham O-PMML ........86

Figure 21. The Regression Model section of Framingham O-PMML ...............87

Figure 22. The timeline of estimated 10-year risk score of cardiovascular disease of patients that had full set of required values to generate scores. ..................................................................................................................89

Figure 23. A schematic of the interoperable framework for delivering personalized health outcome predictions .................................................................94

Figure 24. A diagram of the architecture and data flow of personalized health outcome prediction framework .................................................................................96

Figure 25. The view of PHRST application that allows users to select patients and order risk score estimates. ........................................................................98

Figure 26. The view of PHRST application that displays more details of the selected patient’s risk factor data and the estimated risk score. ..................98
LIST OF APPENDICES

Appendix 1. Template IDs of C-CDA Continuity of Care Document (CCD)
   Release 1.1 templates used by CCD parser to locate entries ..................103
Appendix 2. SQL query to map source codes to OMOP source concepts ..........105
Appendix 3: SQL query to map source codes to OMOP standard concepts .......106
Appendix 4. Mapped CVX codes to OMOP standard Concept IDs ..................107
Appendix 5: SQL query to find ingredients of OMOP standard drug concepts ...111
Appendix 6. SQL query to map measurement observations to the corresponding clinical evaluation and measurement value concepts ..........112
Appendix 7. SQL script to add table constrains to OMOP CDM ..................113
Appendix 8. The XML schema of O-PMML standard ................................116
Appendix 10. The O-PMML containing predictive model for estimating Framingham 10-year risk of cardiovascular disease for women ..........140
Appendix 11. Estimated 10-year risk score of cardiovascular disease for 56 records of 8 unique patients. F: Female, M: Male, TCL: Total cholesterol level, HDL: High-density lipoprotein cholesterol level, SBP:
   Systolic blood pressure. ........................................................................146
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application programming interface</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and regression tree</td>
</tr>
<tr>
<td>CCD</td>
<td>Continuity of care</td>
</tr>
<tr>
<td>C-CDA</td>
<td>Consolidated clinical document architecture</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical document architecture</td>
</tr>
<tr>
<td>CDC</td>
<td>Centers for disease control and prevention</td>
</tr>
<tr>
<td>C-CDA</td>
<td>Consolidated clinical document architecture</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical document architecture</td>
</tr>
<tr>
<td>CDC</td>
<td>Centers for disease control and prevention</td>
</tr>
<tr>
<td>CDM</td>
<td>Common data model</td>
</tr>
<tr>
<td>CDS</td>
<td>Clinical decision support</td>
</tr>
<tr>
<td>CPT</td>
<td>Current procedural terminology</td>
</tr>
<tr>
<td>DDD</td>
<td>Defined daily dose</td>
</tr>
<tr>
<td>EHR</td>
<td>Electronic medical records</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract transform load</td>
</tr>
<tr>
<td>HCPCS</td>
<td>Healthcare common procedure coding system</td>
</tr>
<tr>
<td>HDL</td>
<td>High density lipoprotein</td>
</tr>
<tr>
<td>HIE</td>
<td>Health information exchange</td>
</tr>
<tr>
<td>HL7</td>
<td>Health level seven</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>The international classification of diseases, ninth revision, clinical modification</td>
</tr>
<tr>
<td>ICD-10-CM</td>
<td>The international classification of diseases, tenth revision, clinical modification</td>
</tr>
<tr>
<td>IT</td>
<td>Information technology</td>
</tr>
<tr>
<td>LOINC</td>
<td>Logical observation identifiers names and codes</td>
</tr>
<tr>
<td>NDC</td>
<td>National drug code</td>
</tr>
<tr>
<td>NDF-RT</td>
<td>The national drug file – reference terminology</td>
</tr>
<tr>
<td>OMOP</td>
<td>Observational medical outcomes partnership</td>
</tr>
<tr>
<td>O-PMML</td>
<td>OMOP-compliant predictive model markup language</td>
</tr>
<tr>
<td>PHRST</td>
<td>Personalized health risk scoring tool</td>
</tr>
<tr>
<td>PMML</td>
<td>Predictive model markup language</td>
</tr>
<tr>
<td>REST</td>
<td>Representational state transfer</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>RIM</td>
<td>Reference information model</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>SBD</td>
<td>Semantic brand drug</td>
</tr>
<tr>
<td>SBP</td>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>SCD</td>
<td>Semantic clinical drug</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>Systematized nomenclature of medicine – clinical terms</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>VA</td>
<td>Veterans affairs</td>
</tr>
<tr>
<td>WHO</td>
<td>World health organization</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible markup language</td>
</tr>
</tbody>
</table>
CHAPTER 1. INTRODUCTION

1.1. Background

This project intended to develop an interoperable information technology (IT) solution that deploys health outcome predictive models at the point-of-care for healthcare providers use. The final product not only will address objectives of precision medicine to deliver individualized care to patients, but also has the potentials to be employed in clinical decision support systems.

Personalized medicine also known as “precision medicine” refers to the delivering of customized treatment to patients based on their individual characteristics (1, p. 125), and aims to reduce adverse events, improve diagnostic methods, and enhance the efficacy of therapies. Among efforts to achieve the goals of precision medicine, scientists and researchers build predictive modeling using population health data to best predict health outcomes according to patients’ variables.

Although numerous electronic medical records (EHR)-based predictive models have been reported in the literature, not all models present high prediction power, and in rare cases they reach clinical settings to help healthcare professionals make clinical decisions at the point-of-care (2, 3). The lack of precision stems from the fact that developed predictive models are highly centralized to the data warehouse that trained the model. This is because no medical data repository exists that stores all health records of a patient from other sources, such as hospitals, insurance companies, outpatient pharmacies; thus, predictive models cannot predict the outcome as expected when applied outside of the training system. As the result, it may fail to be tested and be used in medical practice. Even if the full records were available to the researchers, the data may need further processing prior to data analysis because they may differ in how the concepts were coded, what coding systems were used, and how the databases are structured.

There are only few reports of deployment of health outcome predictive models, and no report exist that applies or evaluates the models across multiple data warehouses for clinical effectiveness and cost efficiency. For example, Hu et al., 2015
reported an online application for predicting the next 6-month healthcare resource utilization by chronic disease patients. The prediction system could make real-time risk assessment using a health information exchange (HIE) electronic health records warehouse; however, it was not evaluated in other HIE networks. In another effort in Canada, Khazaie et al., 2015 (5) proposed a cloud-based Analytics-as-a-Service framework for real-time patient monitoring in the clinical edition and retrospective health analytics in the research edition across multiple EHR systems. They deployed an algorithm to identify septic neonates in an intensive care unit; however, the report did not provide any information about data exchange standards or data transformation processes (e.g., inter-vocabulary mapping of concept code). Although no predictive model was deployed, this study sheds light on novel approaches of integrating advanced analytics in healthcare system. Toerper et al., 2015 (6) also developed a web-based application that predicts daily admission bed needs based on EHR data to improve patient flow management. Despite providing a real-time tool for monitoring and forecasting patient flow in a hospital setting, the proposed system is not interoperable to work across other centers to retrieve new incoming patients’ data for better prediction performance.

Another objective of the project is to provide a decision-making tool for computerized clinical decision support (CDS) systems. CDS systems have demonstrated advantages for improving patient health and safety through helping providers in diagnosing diseases (7, 8), reducing medication errors (9, 10), improving patient throughput (11), and healthcare cost reduction (12). The CDS systems also support evidence-based practice via delivering recommendations based on evidences from clinical research regarding the patient’s conditions, treatments, and adverse health events (13, 14). The CDS systems generate alerts based on the embedded rule engine and available knowledge base to help healthcare providers make better clinical decisions; however, the providers often ignore the messages and override about 96% of the generated alerts (15-17), mainly because the alerts are nonspecific, irrelevant, vague, or shown repeatedly that make the system’s user fatigued (15, 18, 19). This “alert fatigue” phenomenon may cause prescribers to ignore important alerts that ultimately will pose further adverse events and life-threatening risks to the patients (20, 21). For that reason, many reports have underscored the urgent need for effective solutions to trigger more relevant and accurate alerts (18, 21-23).
1.2. Statement of the Problem

The existing healthcare IT solutions have not leveraged forecasting capabilities of predictive models in medical practice to provide clinicians with personalized recommendations about patients’ health status. No system also exists that can populate and transform patient information for population health research from different data repositories with diverse coding systems. This project intends to fill the gap via designing, evaluating, implementing an interoperable solution that receives patient electronic health records via Health Level Seven (HL7) messaging standard, transforms the records to a common data model for population health research. Another solution receives predictive model in PMML format, and applies the models on patient information to make predictions about health outcomes, and delivers the results to healthcare professionals.

1.3. Conceptual Framework of the Study

The growing availability of healthcare data in the last decade has provided tremendous opportunities for conducting large-scale clinical and population health research to support evidence-based medicine. Thanks to the Meaningful Use incentive program (24) introduced by American Recovery and Reinvestment Act of 2009 (25), the electronic health records of millions of Americans are stored in repositories throughout the nation that are being used for building new, more accurate predictive models. This Big Data from hospitals, clinics, physician offices, and insurance companies comprises patient data from hospitals and physician offices, reimbursement claims (e.g., Medicaid and Medicare, private insurance companies), genetic and genomic, payments, administrative, and care expenditure data that makes it an invaluable, comprehensive resource for healthcare organizations and care providers to advance patient safety and quality of care (26).

A wide range of predictive models have been derived from EHRs and observational data with the goal to improve patient safety, cut treatment costs through early detection of diseases, and assist clinicians to make accurate clinical decisions. Typical examples of predictive models include predicting prognostic risk of health events (27-29), disease screening (30-32), and managing short- and long-term complications (33-36). In one study on senior patients in Veterans Affairs (VA) nursing homes, the researchers could estimate personalized change in functional
loss and recovery after hospitalization with 84-92% accuracy using random forest approach (37). In another study, the use of EHR data could refine prediction of 30-day hospital readmission risk after percutaneous coronary intervention (38).

Predictive models have also been used to support individuated decision support through estimating complications, short-term readmission, and long-term prognosis (39). Lee et al., 2015 (40) also reported a prognostic prediction model that has the potential to help clinicians design personalized treatments for ischemic stroke. Predictive modeling can also improve patients’ safety by reducing human errors. Physicians and healthcare providers are prone to cognitive biases, logical fallacies, false assumptions, and other reasoning failures when making clinical decision about diagnosis or treatment (41); therefore, predictive analytics tools may help them fill the gap by providing prediction estimates about the patients’ conditions.

The literature has numerous examples of predictive modeling methodologies to build individualized models, such as logistic regression to predict prognosis of health outcomes (38-40, 42), Cox proportional hazards model for estimating opioid dose-related risk of injuries in older adults (43) and survival analysis (44), linear regression for exploring outcome predictors (45), and personalizing medicine dosage (46) and risk estimations (39), and random forest for individualized medicine doses (37, 47). Researchers have also examined several machine learning methods to improve the accuracy and generalizability of the developed predictive models, such as support vector machine (SVM) for early detection of myocardial infarction (47) and mortality risk of radical cystectomy (48), Markov Decision Process for predicting mortality and length of hospitalization in septic patients (49), k-nearest neighbor for warfarin dosage estimation (50), naïve Bayes network for cardiovascular disease risk (51), classification and regression tree (CART) for heart failure patients’ readmission risk using EHR data (52).

Predicative models are also being considered to improve the performance of CDS systems. Despite the efforts to optimize the number and quality of alerts, such as stratifying alerts by severity (53, 54), adjusting messages based on care setting and provider’s specialty (55), introducing context-specific alerts (56), incorporating human factor principles (57), filtering overridden alerts (58), prioritizing alerts based on experts recommendation (59), no unanimous solution exists in the literature for this challenge. In recent years, the CDS systems have moved toward
using artificial intelligence and predictive models in clinical practice. This approach not only addresses the alert fatigue concern, but also promotes evidence-based medicine via providing personalized estimations and recommendations to clinicians for better care quality and more effective treatments. For example, Levin et al., 2012 (60) reported that implementation of a predictive model in an intensive care unit CDS system could improve patient flow management by providing real-time, personalized length of stay estimations. In addition, machine learning approaches could also improve patient outcome through generating personalized risk estimates and medicine dosages based on patient’s history of medications and comorbidities (61, 62). Artificial intelligence may similarly help clinicians in interpreting medical images to make more accurate clinical decisions (63).

Nevertheless, predictive modeling has limitations. The use of predictive models is often limited to the system that has provided the training set for model development due to data model and coding constrains. There are CDS systems equipped with predictive models to estimate personalized survival and risk of adverse events based on individual patient data (64); however, accurate predictions are only available if the new set of data is formatted the same as the training dataset. In fact, the variables within the predictive model do not match with incoming data from other sources. Therefore, predictive models are not designed to perform predictions on all patients’ data accessible through health information exchange networks.

Local and commercial EHR systems use diverse data models to store health records; consequently, medical concepts and measurements are stored using different coding systems that impedes direct exchange of information between systems for administration, reimbursement, or medical transactions. Likewise, not all predictive models are trained using the same terminologies; thus, similar predictor and outcome variables are defined differently across models. For example, there are different coding systems for diagnoses (e.g., SNOMED, ICD-9-CM, and ICD-10-CM), medicines (e.g., NDC, NDF-RT, and RxNorm), laboratory tests and observations (e.g., LOINC, CPT), and procedures (e.g., CPT, ICD-9-CM) that need further processes to map the concepts across different coding vocabularies for intersystem communications. This diversity of ‘languages’ and the need for exchanging data have been the chief motive to develop common data models, such as
HL7 Reference Information Model (RIM) (65) and the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) (66).

Adding new functionalities to EHR systems is often accompanied with deployment costs (e.g., hardware and software), modifications in the system’s architecture, and extra maintenance expenses (67). Cloud computing is an emerging solution that not only reduces the costs, but also offers rapid scaling, improved accessibility, and better adaptability to new systems and workflows (68-72). According to the National Institute of Standards and Technology (73), cloud computing technology refers to “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” Moving the computing engine to the cloud will ultimately allow the EHR vendors to expand their systems’ functionality in a shorter time with minor changes in the architecture.

1.4. Purpose of the Study

The scope of this project is to design and develop a pilot solution that standardizes HL7 CCD data into OMOP CDM, receives predictive models in OMOP-compliant PMML standard format, and deploys the model on patient data (Figure 1). The OMOP CDM was chosen over other common data model such as Mini-Sentinel (74, 75) and PCORnet (76) mainly because the OMOP CDM consists a vocabulary that provides mapping relationships to standard concepts. For example, all condition concepts whether from ICD-9 or ICD-10, they are all mapped to SNOMED CT concepts; LOINC is also the reference for standard concepts of laboratory measurement codes, and medication codes are mapped to RxNorm vocabulary. This is very helpful to achieve the goals of this project, in particular the “plug-and-play” approach of deploying predictive models.

This project consists of three studies that follow three objectives to achieve:

(a) Developing and testing an interoperable module that transforms patient electronic medical records to a common data model for population health research

(b) Developing a standard for disseminating health-related predictive models
(c) Developing an interoperable prediction system that accepts patients’ data with diverse data models to provide real-time, personalized estimations about health associated risks or outcomes.

The first study aims to design, develop, and evaluate a module that transforms patient electronic medical records transferred through HL7 consolidated-clinical document architecture (C-CDA) standard to OMOP (Observational Medical Outcomes Partnership) common data model (CDM) for population health research. This project focuses on one of HL7 C-CDA document types called continuity of care (CCD) that encapsulates summary of patient medical records. Using the HL7 C-CDA-based CCD parser called “CCD-TO-OMOP” will enable precision medicine research including predictive modeling studies to get access to a larger pool of medical information of patients.

The second study intends to develop a new standard based on the existing predictive model markup language (PMML) for sharing health outcome risk scoring models that are generated using OMOP CDM data. This new OMOP-compliant PMML (O-PMML) standard was used in developing the final interoperable solution of predictive analytics.

Ultimately, the third study aims to develop a proof-of-concept Personalized Health Risk Scoring Tool (PHRST), an interoperable OMOP-based model scoring tool that obtains risk score models in a OMOP-compliant standard format, and applies the model on patient information to deliver the personalized risk score to the end user who can be a healthcare professional.
Figure 1. The scope of the project is to deploy predictive models on patient data exchanged via HL7 messaging standard. The project involves standardizing transferred patient information into OMOP CDM, introducing a new standard for disseminating OMOP-based predictive models, and building a pilot solution for deployment of health risk scoring models.

1.5. Research Questions

This project explores answers to these questions:

- How well does OMOP CDM accommodate HL7 C-CDA-based CCD data? The answer to this question is important to choose a suitable data model that standardizes medical records for population health research purpose.
- How PMML can be adopted to share risk scoring models that are generated based on OMOP CDM?
- What are the functional requirements to develop a tool that can apply risk scoring models on OMOP CDM data to calculate the risk of health outcomes?

1.6. Definition of Terms

In this study, the term prediction or predictive refers to the prediction of any health-related event to occur in the future, risk estimation of an unobserved event that has already occurred, or prediction of an occurred event but not observed.

Population health or epidemiology is a field of study that aims to improve health-related outcomes of the population through investigating the distribution and patterns of the outcomes, and the role of interventions and policies on them (77).
Health Level Seven (HL7) consolidated-clinical document architecture (C-CDA) is a messaging standard for transferring patient medical records between systems. Continuity of Care Documents (CCD) is one of document types of HL7 C-CDA standard that contains demographic and a summary of clinical information facts about a patient's healthcare encounters (78).

A data model defines data elements, data types, coding standards, semantics, and data restrictions within a database (79). For example, Observational Medical Outcomes Partnership (OMOP) common data model (CDM) is a data model developed for population health research that allows researchers analyze observational data of dissimilar databases similarly by standardizing concepts and database structure (80).

Predictive model markup language (PMML) is an extensible markup language (XML)-based standard for inter-system sharing of predictive models and the associated data mining requirements (81).

1.7. Significance of the Study

The first study delivers CCD-TO-OMOP package that links medical data repositories to external databases to receive and transform up-to-date data of existing or new patients to a standard format, ready for population health research. This is a significant achievement as currently no links exist between repositories to share patient data that may help increase sample size of population health research, ultimately increase power of inferences. This is because of discrepancies in database structures and coding systems used to store concepts. Therefore, this module builds bridges between observational data warehouses to enhance generalizability of population health studies.

The proposed O-PMML provides an OMOP-compliant standard format for sharing health outcome risk scoring models between IT systems that defines the specifications of data mining, predictive models, and scoring process. This standard enables researchers to disseminate OMOP-based predictive models between disjoint OMOP repositories, and deploy the models on observational data of the destined database in less time but more accurately. It also provides the opportunity to disseminate models through information exchange networks.

The final interoperable system receives medical information through HL7 CCDs and predictive models via OMOP-compliant PMMLs, and displays the scoring
results to the provider through PHRST. This is a significant feature that enables the system to deploy multiple predictive models at once. This system has high potentials to act as a decision-making tool in CDS systems to improve patient health outcomes through delivering real-time clinical recommendations. The interoperable system not only makes it feasible to deploy trained predictive models in CDS systems, but also it will be accessible through health information exchange networks for distributed computing purposes with no need for further data transformation. Hence, personalized recommendations about a patient can be processed independent of the technology in CDS and the original data model used in training predictive models.

1.8. Organization of the Study

This study lays out an IT solution to deliver scoring of predictive and population health analyses on patient electronic health records in an interoperable manner where both patient information and predictive models are transferred in standard formats.

Chapter 2 asks “Does OMOP CDM accommodate health information transferred by HL7 C-CDA CCD messaging standard?” It describes the developed CCD-TO-OMOP parser that transforms patient information into OMOP common data model. It also examines the robustness of OMOP CDM to take data elements and the accuracy of concept standardizing pipeline.

Chapter 3 asks “How to adopt PMML for disseminating predictive models generated based on OMOP CDM requirements?” It explores the architecture of PMML, and describes how to adopt the language for standardizing dissemination of OMOP-based predictive models. It also introduces the scoring engine that translates OMOP-compliant PMMLs to predictive models and applies the model on patient data.

Chapter 4 explores “How to build the interoperable EHR-based predictive analytics system”. It describes how CCD-TO-OMOP parser, OMOP-compliant PMMLs, and scoring engine are assembled to build an interoperable scoring system named Personalized Health Risk Scoring Tool (PHRST) to deliver real-time risk scores.

Chapter 5 concludes with a summary of findings of this project and potential future work to improve the performance of the solution and real-world evaluations.
CHAPTER 2. OMOP COMMON DATA MODEL ACCOMMODATES HL7 CONSOLIDATED CDA-BASED CCD DATA FOR POPULATION HEALTH RESEARCH

2.1. Introduction

Personalized medicine also known as “precision medicine” refers to the delivering of customized treatment to patients based on their individual characteristics (1), and aims to reduce adverse events, improve diagnostic methods, and enhance the efficacy of therapies.

To achieve the goals of precision medicine, population health scientists and researchers extensively rely on observational data that are collected during healthcare services to study associations between health determinants and build predictive models to tailor individualized therapies for patients. However, no study findings exist that can be fully generalized to the whole target population, and not all models present high prediction power to reach clinical settings for clinical decision making at the point-of-care. The lack of precision stems from the fact that analyses are limited to the dataset of medical records that may not well represent the population.

2.1.1. Problem statement

To mitigate lack of generalizability, it is recommended to sample patients from different tiers of the target population; however, no medical data repository has the information of all patients and no repository can be found to have all health records of patients from other sources, such as hospitals, insurance companies, outpatient pharmacies. Even if the full records are available to the researchers from other data sources, it may not be straightforward to use the data for research as it is probable to see discrepancies in the coding system of concepts and how the databases are structured. This study addresses this problem by designing, evaluating, implementing an interoperable solution that receives patient electronic health records via HL7 messaging standard, and transforms the records to OMOP CDM to be used in population health research.
2.1.2. HL7 continuity of care document

The main use of CCDs is to transfer a summary of patient’s demographics information, administrative data, and clinical facts such as diagnosed disease, symptoms and signs, prescribed and administered medicines, carried out procedures, and clinical test and imaging. The document may also contain optional sections regarding care encounters, family history, immunizations, functional status, payers, and treatment plan. This study evaluated this messaging standard to be used as a data stream for epidemiological and drug safety studies. Although the architecture of generated CCDs from different institutions follow the HL7 implementation guidelines, discrepancies may exist in how clinical concepts are coded. Therefore, the encapsulated concepts need to be standardized before using the patient records in cohort selection and analyses. Table 1 shows a summary of document-level templates in HL7 version 3 consolidated clinical document architecture (C-CDA) Release 1.1.

2.1.3. The Observational Medical Outcomes Partnership (OMOP)

The Observational Medical Outcomes Partnership (OMOP) is a collaboration of FDA, pharmaceutical industry, data owners, and academia to identify informatics requirements, propose methodologies, and test the solutions for drug safety and population health research through enhancing use of observational data (i.e., medical records) (82). In population health research, it is preferable to use large, multicenter cohort of patients in the studies to minimize potential biases (such as sampling and measurement). However, medical data are stored with diverse data models and coding systems in healthcare warehouses that makes it cumbersome, often unsuccessful to integrate all data elements in one place for research. The partnership has introduced a common data model (CDM) to address this problem (80).

The OMOP CDM reorganizes the observational data elements into a format that supports population health research to explore and quantify associations between exposure and outcomes, survival analysis, and causality relationships. This project adopts the CDM to standardize all incoming data into one common format for data analyses. In this data model, different codes from various coding systems but representing one concept are linked to (i.e., unified under) a common ID called OMOP Concept ID. For example, one procedure may be coded in ICD-9-CM, ICD-10-
CM, CPT-4, or any other vocabularies in different CCDs, but it would be linked to one Concept ID in OMOP CDM. In addition, the concepts are placed under the most proper domain to conduct epidemiological research. For example, familial histories of diseases are usually coded in ICDs and may be considered conditions; however, these occurrences are fall under definition of observations. This approach not only unifies codes under one concept, but also validates whether the codes are correctly assigned to the right concept and domain.

The latest OMOP CDM version 5.1 (83) comprises six domains and 39 data tables (Figure 2). The domains are person-centric that makes the CDM an appropriate model for this project to deliver patient-specific predictions. The embedded ‘common’ vocabulary covers a complete list of standardized dictionaries and vocabularies. Each concept has a unique identifier while preserves the code from the original vocabulary for easy matching. There is also a concept relationship table that defines direct relationships between concepts to map them across vocabularies possible.

**Figure 2.** OMOP Common Data Model Version 5.1 conceptual model (83).
2.1.4. Objectives

This chapter aimed to assess feasibility and accuracy of mapping exchanged electronic health records (EHR) via HL7 C-CDA CCDs to OMOP CDM. It also validates the suitability of the model to accommodate CCD data. Finally, it delivers a validated CCD-TO-OMOP parser that will be used in the proposed interoperable predictive analytics framework.

The accommodation assessment answered the questions whether (a) OMOP CDM tables have the right data fields with appropriate data type to store CCD data, (b) OMOP CDM vocabulary covers CCD source codes, (c) OMOP CDM has the proper concept relationships to map CCD source codes, and (d) HL7 C-CDA-based CCD feeds the CDM tables with the minimum required data for epidemiological studies. A minimum required data varies one table to another, but generally it consists of the code and coding system of the event (e.g., condition report, drug administration), date of event, and associated values to the event. For example, the minimum expected data in Drug Exposure table were drug code and the respective coding system, the date of drug administered or dispensed, quantity of drug dispensed, ordered dose, and days of drug supply.
Table 1. A summary of document-level template of HL7 C-CDA Consolidated Continuity of Care Document (CCD) Release 1.1 and corresponding Object identifiers (OID).

<table>
<thead>
<tr>
<th>Section Name</th>
<th>Object Identifier (OID)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Realm Header</td>
<td>2.16.840.1.113883.10.20.22.1.1</td>
<td>Describes patient’s demographics (e.g., gender, race, marital status) and common administrative information (e.g., patient’s name and address, author, provider).</td>
</tr>
<tr>
<td>Advance Directives Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.21</td>
<td>Describes patient’s directives of living wills, resuscitation status, CPR orders, and healthcare proxies.</td>
</tr>
<tr>
<td>Allergies Section (entries required)</td>
<td>2.16.840.1.113883.10.20.22.2.6.1</td>
<td>Describes allergies to food, medicines, and other substances (e.g., latex), and reported adverse reactions.</td>
</tr>
<tr>
<td>Encounters Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.22</td>
<td>Lists the history of encounters between patient and healthcare providers for diagnosis, treatment, or evaluating medical condition purposes.</td>
</tr>
<tr>
<td>Family History Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.15</td>
<td>Describes health risk factors of patient’s biologic parents that may affect the patient’s risk of medical condition occurrences.</td>
</tr>
<tr>
<td>Functional Status Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.14</td>
<td>Describes findings and evaluation results of patient’s physical function, including basic and instrumental activities of daily living (ADLs).</td>
</tr>
<tr>
<td>Immunizations Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.2.1</td>
<td>Lists history patient’s immunization history.</td>
</tr>
<tr>
<td>Medical Equipment Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.23</td>
<td>Lists external, implanted, or durable medical devices used to treat patient’s medical condition or uphold the health status.</td>
</tr>
<tr>
<td>Medications Section (entries required)</td>
<td>2.16.840.1.113883.10.20.22.2.1.1</td>
<td>Lists patient’s history of prescribed and dispensed medications.</td>
</tr>
<tr>
<td>Section Name</td>
<td>Object Identifier (OID)</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Payers Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.18</td>
<td>Lists all financial, payment, insurance, and health plan coverage records pertinent to the healthcare services provided to the patient.</td>
</tr>
<tr>
<td>Plan of Care Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.10</td>
<td>Lists all patient’s ongoing, incomplete, pending, or unfulfilled services, orders, encounters, and procedures to be carried out in future.</td>
</tr>
<tr>
<td>Problem Section (entries required)</td>
<td>2.16.840.1.113883.10.20.22.2.5.1</td>
<td>Lists patient’s history of clinical conditions and problems.</td>
</tr>
<tr>
<td>Procedures Section (entries required)</td>
<td>2.16.840.1.113883.10.20.22.2.7.1</td>
<td>Lists patient’s history of surgical, therapeutics, and diagnostic procedures.</td>
</tr>
<tr>
<td>Results Section (entries required)</td>
<td>2.16.840.1.113883.10.20.22.2.3.1</td>
<td>Lists all results of laboratory tests, diagnostic procedures, and imaging procedures.</td>
</tr>
<tr>
<td>Social History Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.17</td>
<td>Describes patient’s social observations, e.g., smoking status, pregnancy history, tobacco use, cultural and religious practices, etc.</td>
</tr>
<tr>
<td>Vital Signs Section (entries optional)</td>
<td>2.16.840.1.113883.10.20.22.2.4</td>
<td>Describes patient’s vital signs, e.g., blood pressure, temperature, pulse rate, height, weight, etc.</td>
</tr>
</tbody>
</table>
2.2. Methods and Materials

I developed the CCD-TO-OMOP ETL module that extracts data from HL7 C-CDA-based CCDs, maps concepts to OMOP vocabulary, transforms the data into OMOP data model, and loads them into a PostgreSQL repository. The ETL pipeline located data elements in CCDs based on Template IDs and specifications defined by HL7 implementation guide document (78).

This module was developed and tested in two pilot and production phases. In the pilot phase, the module was programmed based on the HL7 implementation guide and was debugged using on 10 randomly selected CCDs. Final minor tweaks were also made in the codes to reach highest possible extraction, transformation, and loading performance. In the production phases, the module was tested for accuracy and performance on 250 CCDs, inclusive pilot CCDs.

2.2.1. HL7 Continuity of Care Documents (CCDs)

I obtained a randomly selected 250 deidentified CCD documents generated based on HL7 version 3 (V3) consolidated clinical document architecture (C-CDA) Release 1.1 standard from Regenstrief Institute, Indianapolis, IN. The Regenstrief data repository contains electronic health records of more than 2.2 million patients who have received health care in Indiana State. Each CCD document contained a summary of patient health information including demographics, medical history, diagnoses, laboratory test results, prescriptions, immunization records, and radiology reports. The sample of CCD documents were obtained randomly from all age, sex, race, ethnicity, economic and education groups.

2.2.2. OMOP vocabulary

The OMOP vocabulary was obtained on April 12, 2017 from Athena website (http://athena.ohdsi.org/), the official resource of OMOP CDM standardized vocabularies (84).

2.2.3. Overview of the ETL package

The “CCD-TO-OMOP” is a Python ETL package that extracts patient medical record from CCDs based on HL7 consolidated clinical document architecture (C-CDA) format, transforms the data into OMOP CDM, and finally loads the transformed information into staging tables to be used for further use, e.g.,
statistical analytics and health outcome predictions. The package consists of four modules (Figure 3): CCD Parser, OMOP Mapper, Loader, and Database Connector.

The *CCD Parser* module extracts demographics, medicines, conditions, care provider encounters, laboratory test results, and observations data from CCDs. The extraction pipeline points to the Template IDs specified by HL7 C-CDA (Appendix 1). It also validates the pulled data whether data elements have the correct values for the destination OMOP table.

Once CCD data are extracted, the *OMOP Mapper* module transforms the data into intermediate OMOP tables—which are instantiated from the OMOP CDM module—for further processing. The transformation process translates source codes into OMOP standard Concept IDs (e.g., mapping drug codes to a standard RxNorm code in OMOP vocabulary), and extracts values where appropriate (e.g., year and month from CCD drug dispensing date). Next, the *Loader* module loads the transformed data from the intermediate tables into an OMOP CDM database that can be accessible by the end user. Both the OMOP Mapper and Loader modules use the *Database Connector* module to lookup OMOP values and load transformed data into the database, respectively.

![Figure 3. A schematic of CCD-TO-OMOP package](image-url)
2.2.4. Extraction of HL7 CCD data

The developed ETL mainly targeted clinical data encapsulated in C-CDA-based CCD documents, including condition occurrences, allergies, drug exposures, procedures, encounters, laboratory tests, imaging reports, immunization, vital signs, and familial risk factors. The scope of this study was to explore certain tables of OMOP CDM that are key to conduct observational studies, including patient demographics (Person table), periods of observing patient health events (Observation Period table), visit encounters (Visit Occurrence table), diagnoses and health conditions (Condition Occurrence table), continuous intervals of diseases and conditions (Condition Era table), performed procedures (Procedure Occurrence table), administered medications (Drug Exposure table), continuous intervals of medication use (Drug Era table), results of medical evaluations (Measurement table), and clinical observations (Observation table).

Based on the described data requirements, the parser located the corresponding HL7 C-CDA template within the documents by Template ID to ensure targeting the right data elements. Appendix 1 presents all Template IDs that parser used to locate entries. The data was captured only if the activity was fulfilled, meaning that the activity was an event (i.e., moodCode = “EVN”) and the status was completed (i.e., Status Code = “completed”).

2.2.5. Transformation of HL7 CCD data to OMOP CDM

The parser transformed the extracted and transformed CCD data table by table in the following steps:

**Step 1)** Primary mapping source values to OMOP concepts: The parser mapped terminology codes and non-coded values to OMOP vocabulary concepts using the query in Appendix 2 against OMOP standardized vocabularies. The terminology codes represent clinical concepts, such as condition, medication, procedure, and evaluations from terminologies, including but not limited to Current Procedures Terminology (CPT), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical Observation Identifiers Names & Codes (LOINC), International Classification of Diseases (ICD), the Healthcare
Common Procedure Coding System (HCPCS), and RxNorm. The non-coded values included units of laboratory tests (e.g., mg/L) and non-numeric results of measurements (e.g., high, low, normal).

**Step 2)** Mapping source values to OMOP standard concepts: The parse also mapped terminology codes and non-coded source values to OMOP standard concepts using Appendix 3 query against OMOP vocabulary.

**Step 3)** Applying OMOP transformation rules: Each OMOP CDM table has specific business rules to transform source data to be stored in the CDM. For example, *Condition Era* table requires that the condition eras shall be built with a “Persistence Window of 30 days”, meaning that the condition era continues as long as the condition has been reported within the following 30 days after the report date (*Condition Start Date*). The parser applies these rules on a table by table basis.

**Step 4)** Appropriate data allocation: There are some instances that the extracted source data does not belong to the target table. Therefore, the ETL parser redirected the data into the proper table as required by OMOP CDM rules. For example, family history of diseases may be reported under *Problem Observation* section within the CCD document; however, these concepts are considered observations by OMOP vocabulary. Therefore, the ETL stored the concepts in *Observation* table instead of *Condition Occurrence* table.

The following sections of this manuscript describe the data transformation process more in details.

### 2.2.5.1. Mapping source values to OMOP concepts

There are two types of concepts in OMOP vocabulary: Standard and non-standard concepts. Designated standard concepts are the only concepts that can be used to represent clinical entities in standardized analytics and table fields ending with _Concept_Id. The standard concepts originate from standardized vocabularies, such as SNOMED CT for conditions and diagnoses, RxNorm for drugs, and LOINC
for observations and laboratory tests. Non-standard or source concepts are direct representations of source codes in OMOP vocabulary that need to be mapped to standard concepts to be used in analytics. OMOP table fields ending with 
(Source_Concept_Id) may contain either standard or non-standard concepts depending on whether a direct standard or source OMOP concept identifier exist for the source code.

2.2.5.2. Person table

Person table fields were sourced from recordTarget section of C-CDA documents that contains demographics data. The scope of this transformation was to process all fields of Person table except Location ID, Provider ID, and Care Site ID. Table 2 present more details of data transformation process to build OMOP Person table.

A randomly generated unique identifier was recorded in Person ID for each parsed C-CDA document to link further patient information in other tables. Gender information are located under administrativeGenderCode element in C-CDA document coded in HL7 V3 administrative gender value set, including male, female, undifferentiated. Table 3 shows the equivalent OMOP concepts of HL7 administrative gender codes.

Date of birth information was extracted from birthTime element with a precision from year to days. The first four digits of the value represent patient’s year of birth (Year of Birth field), the next two digits are month of birth (Month of Birth field), and subsequently the two next digits are days of birth (Day of Birth field).

Patient’s race and ethnicity information are coded in CDC (Centers for Disease Control and Prevention) race and ethnicity value set, and were collected from raceCode and ethnicGroupCode elements, respectively. They were mapped to OMOP concepts as shown in Table 4.
<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>person_id</td>
<td>–</td>
<td>Randomly generated unique identification per C-CDA document</td>
</tr>
<tr>
<td>gender_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of Gender Source Value</td>
</tr>
<tr>
<td>year_of_birth</td>
<td>US Realm Header: birthTime</td>
<td>Extracted year part of the date of birth</td>
</tr>
<tr>
<td>month_of_birth</td>
<td>US Realm Header: birthTime</td>
<td>Extracted month part of the date of birth</td>
</tr>
<tr>
<td>day_of_birth</td>
<td>US Realm Header: birthTime</td>
<td>Extracted day part of the date of birth</td>
</tr>
<tr>
<td>race_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of Race Source Value</td>
</tr>
<tr>
<td>ethnicity_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of Ethnicity Source Value</td>
</tr>
<tr>
<td>gender_source_value</td>
<td>US Realm Header: administrativeGenderCode</td>
<td>OMOP Concept ID of Gender Source Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Domain ID = “Gender” and Invalid Reason = Null</td>
</tr>
<tr>
<td>gender_source_concept_id</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>race_source_value</td>
<td>US Realm Header: raceCode</td>
<td>OMOP Concept ID of Race Source Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Domain ID = “Race” and Invalid Reason = Null</td>
</tr>
<tr>
<td>ethnicity_source_value</td>
<td>US Realm Header: ethnicGroupCode</td>
<td>OMOP concept id of Ethnicity Source Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If Domain ID = “Gender” and Invalid Reason = Null</td>
</tr>
<tr>
<td>ethnicity_source_concept_id</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3. Matching OMOP concepts of HL7 administrative gender codes

<table>
<thead>
<tr>
<th>HL7 V3 Administrative Gender</th>
<th>OMOP Standard Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Description</td>
</tr>
<tr>
<td>F</td>
<td>Female</td>
</tr>
<tr>
<td>M</td>
<td>Male</td>
</tr>
<tr>
<td>UN</td>
<td>Undifferentiated</td>
</tr>
</tbody>
</table>

### Table 4. Matching OMOP concepts of CDC race and ethnicity value set

<table>
<thead>
<tr>
<th>CDC Race and Ethnicity Values</th>
<th>OMOP Standard Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Description</td>
</tr>
<tr>
<td>1002-5</td>
<td>American Indian or Alaska Native</td>
</tr>
<tr>
<td>2028-9</td>
<td>Asian</td>
</tr>
<tr>
<td>2054-5</td>
<td>Black or African American</td>
</tr>
<tr>
<td>2076-8</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>2106-3</td>
<td>White</td>
</tr>
<tr>
<td>2135-2</td>
<td>Hispanic or Latino</td>
</tr>
<tr>
<td>2186-5</td>
<td>Not Hispanic or Latino</td>
</tr>
</tbody>
</table>

#### 2.2.5.3. Observation Period table

*Observation Period* table specifies the time frame in which clinical events have been continuously captured and the data is available; thus, a patient may have multiple observation periods as there are times that patients do not encounter clinical events between therapies, such as drug exposure, procedure occurrence, condition occurrence, and device exposure.

Observation periods were populated by consolidating records of visit encounters from *Visit Occurrence* table and reported clinical events from *Procedure Occurrence, Drug Exposure, Device Exposure, Condition Occurrence, and Measurement* tables. In case of inpatient visits, the observation period was set to begin at the start date of hospitalization and was limited to the clinical events...
occurred up to 30 days after discharge date to ensure follow-up clinical events were included in the same period. In case of outpatient visit occurrences, the observation period was calculated by combining continuous clinical encounters if the gaps between events were 180 days or less. This wider time windows allows to capture related outpatient follow-up events in one period. Finally, the identified periods were also consolidated if overlapped or the gap was less than 31 days. Table 5 summarizes applied rule to transform observation periods.

Table 5. Applied rules to build Observation Period table

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>observation_period_id</td>
<td>Auto-numbered unique identifier for each observation period</td>
</tr>
<tr>
<td>person_id</td>
<td>The patient’s identifier from Person table</td>
</tr>
<tr>
<td>observation_period_start_date</td>
<td>The start date of consolidated observation period</td>
</tr>
<tr>
<td>observation_period_end_date</td>
<td>The end date of consolidated observation period</td>
</tr>
<tr>
<td>period_type_concept_id</td>
<td>44814725 (Period inferred by algorithm)</td>
</tr>
</tbody>
</table>

2.2.5.4. Visit Occurrence table

All patient’s healthcare service visits are stored in Visit Occurrence table at healthcare sites, including inpatient, outpatient, emergency room, and long-term care services. The parser populates visit occurrence information from Encounter Activity entries under Encounters Section in the CCD document that lists patient’s visits with healthcare provider that triggered diagnosis, treatment, or evaluation of patient health. The CCD Encounter Activity entries provide information about the type of encounter, start and end dates of encounter, provider, care service location, reported diagnosis due to the encounter, and discharge disposition. Table 6 describes the applied rules more in details.
Table 6. Mapping HL7 C-CDA data to *Visit Occurrence* table: Applied rules and corresponding sections.

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>visit_occurrence_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each visit occurrence</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from <em>Person</em> table</td>
</tr>
<tr>
<td>visit_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of <em>Visit Source Value</em></td>
</tr>
<tr>
<td>visit_start_date</td>
<td>Encounters Section: Encounter Activity entry: <em>effectiveTime</em></td>
<td>The start date of encounter</td>
</tr>
<tr>
<td>visit_end_date</td>
<td>Encounters Section: Encounter Activity entry: <em>effectiveTime</em></td>
<td>The end date of encounter</td>
</tr>
<tr>
<td>visit_type_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of recorder encounter type at C-CDA encounter activity element:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 9201 (Inpatient Visit)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 9202 (Outpatient Visit)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 9203 (Emergency Room Visit)</td>
</tr>
<tr>
<td>visit_source_value</td>
<td>Encounters Section: Encounter Activity entry: <em>code</em></td>
<td>Specifies the care setting where encounter occurred, e.g., outpatient, inpatient, etc. If source code is a visit concept (i.e., <em>Domain ID</em> = &quot;Visit&quot; and <em>Invalid Reason</em> = <em>Null</em>).</td>
</tr>
<tr>
<td>visit_source_concept_id</td>
<td>–</td>
<td>OMOP Concept ID of <em>Visit Source Value</em></td>
</tr>
</tbody>
</table>
2.2.5.5. Condition Occurrence table

Observed diseases and symptoms during the observation period which are diagnosed by healthcare providers or reported by the patient are captured in Condition Occurrence table. The HL7 C-CDA CCD document reports medical conditions in Allergies Section and Problem Section. The data of entries were captured if the condition or allergy was truly observed (negationInd = “false” under Problem Observation).

The Allergies Section lists the current and past hypersensitivity, adverse reactions to any types of allergens, such as food, drug, and latex under Allergy Problem Act entries. The cause of the allergy is reported by Allergy-Intolerance Observation entry within Allergy Problem Act. Under this entry, allergy observation data are recorded in two parts: code element that defines the general type of allergy (e.g., allergy to drug substance), and Playing Entity element that specifies the allergen (e.g., penicillin). The parser put both data together to determine the corresponding OMOP concept. For example, if this was a report of drug allergy to penicillin, the parser mapped the finding to a condition concept of “Allergy to penicillin” (i.e., Concept ID = 4240903), and stored the record in Condition Occurrence table.

The parser populated reported clinical problems from Problem Observation entries under Problem Section that lists current and past medical conditions and diagnoses. Since Problem Observation entries may also contain observation, procedure, and measurement OMOP concepts other than conditions, the ETL pipeline was programmed to only capture the records of OMOP condition concepts (i.e., Domain ID = “Condition”), and recorded the data in Condition Occurrence table. If the record represented a concept other than condition domain, it was stored in the corresponding tables, e.g., familial disease history and history of diagnoses in Observation table, medicines in Drug Exposure, and procedures in Procedure Occurrence table. Table 7 presents more details on modeling of condition data.

In case of ongoing allergy and problem concerns (i.e., Status Code = “Active” under Problem Concern Act), condition start date was the first reporting date with unknown end date (i.e., Null). If it was a resolved problem concern (i.e., Status Code = “completed” under Problem Concern Act), both start and end dates were stored in
*Condition Occurrence* table. In case of an allergy which is no longer a concern (i.e., *Status Code = “completed”* under *Allergy Problem Act*), the record was an observation and stored in *Observation* table.
Table 7. Mapping HL7 C-CDA CCD data to *Condition Occurrence* table: Applied rules and corresponding sections.

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>condition_occurrence_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each condition occurrence</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from <em>Person</em> table</td>
</tr>
<tr>
<td>condition_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of <em>Condition Source Value</em></td>
</tr>
<tr>
<td>condition_start_date</td>
<td>Allergies</td>
<td>Allergies</td>
</tr>
<tr>
<td></td>
<td>Allergies Section: Allergy Problem Act:</td>
<td>Only if it is an ongoing allergy, i.e., <em>Status Code</em> = “active” under</td>
</tr>
<tr>
<td></td>
<td>Allergy-Intolerance Observation:</td>
<td><em>Allergy Problem Act</em></td>
</tr>
<tr>
<td></td>
<td><em>effectiveTime/low</em></td>
<td></td>
</tr>
<tr>
<td>condition_end_date</td>
<td>Problems</td>
<td>Null; an active allergy concern does not have end date.</td>
</tr>
<tr>
<td></td>
<td>Problem Section: Problem Concern Act:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problem Observation: <em>effectiveTime/high</em></td>
<td></td>
</tr>
<tr>
<td>condition_type_concept_id</td>
<td>–</td>
<td>38000245 (EHR problem list entry)</td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>–</td>
<td>The identifier of corresponding visit occurrence from Visit Occurrence table. Null if Condition Start Date did not match with any visit occurrence.</td>
</tr>
<tr>
<td>condition_source_value</td>
<td>Allergies</td>
<td>If the source code is a condition concept (i.e., Domain ID = “Condition” and Invalid Reason = Null).</td>
</tr>
<tr>
<td></td>
<td>Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: code</td>
<td>Allergies</td>
</tr>
<tr>
<td></td>
<td>Participating agent in allergies</td>
<td>No space in CDM table to record the participating agent.</td>
</tr>
<tr>
<td></td>
<td>Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: Product: Product Detail: playingEntity/code</td>
<td>The SNOMED code of the concept that represents the allergy or adverse event finding of the allergen is recorded (e.g., allergy to penicillin).</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>Problems</td>
</tr>
<tr>
<td></td>
<td>Problem Section: Problem Concern Act: Problem Observation: code</td>
<td>The source code of reported condition or diagnosis</td>
</tr>
<tr>
<td>condition_source_concept_id</td>
<td>–</td>
<td>Allergies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OMOP concept id of Condition Source Value which is the same as Condition Concept ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OMOP concept id of Condition Source Value</td>
</tr>
</tbody>
</table>
2.2.5.6. Condition Era table

This table aggregates records of diagnoses and medical conditions in consolidated periods of conditions, called condition eras, to prevent double-counting of reported conditions, and to enable following chronic conditions through disease progression and underlying treatments.

The parser built Condition Era table using the recorded reports of condition in Condition Occurrence table. After excluding the records that no equivalent OMOP concept existed for the condition code (i.e. Condition Concept ID = 0), the parser consolidated records if the gaps between the start dates of condition occurrences were up to 30 days. Table 8 summarizes applied rule to create condition eras.

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>condition_era_id</td>
<td>Auto-numbered unique identifier for each condition era</td>
</tr>
<tr>
<td>person_id</td>
<td>The patient’s identifier from Person table</td>
</tr>
<tr>
<td>condition_concept_id</td>
<td>OMOP standard concept for which condition era is built, excluding Condition Concept ID = 0</td>
</tr>
<tr>
<td>condition_era_start_date</td>
<td>The start date of consolidated condition era as long as medical condition was reported continuously in the next 30 days or less.</td>
</tr>
<tr>
<td>condition_era_end_date</td>
<td>The end date of consolidated condition era as long as medical condition was reported continuously in the next 30 days or less.</td>
</tr>
<tr>
<td>condition_type_concept_id</td>
<td>38000247 (Condition era - 30 days persistence window)</td>
</tr>
<tr>
<td>condition_occurrence_count</td>
<td>The number of condition occurrences involved in the consolidated condition era</td>
</tr>
</tbody>
</table>
2.2.5.7. Procedure Occurrence table

This table encompasses the records of diagnostic or therapeutic activities by healthcare provider or patient to identify diseases or medical conditions, to administer medicine, or to maintain treatment plan (e.g., patient education). The parser captured completed procedure activities (Status Code = “completed”) from four HL7 C-CDA templates: Procedure Activity Procedure, Procedure Activity Observation, and Procedure Activity Act under Procedures Section, and Problem Observation under Problem Section.

The Procedure Activity Procedure template delivers surgical operation data for diagnosing or treating diseases that involve physical changes in patient, such as biopsy procedure, open heart surgery, drug administration, and laparotomy. The Procedure Activity Observation template characterizes diagnostic procedures that provide new findings about patient’s disease or medical condition, such as magnetic resonance imaging (MRI), echocardiography, and colonoscopy. The Procedure Activity Act template covers other clinical activities that do not fall in the previous two categories, such as patient education and wound dressing change.

Other than these three templates, it is probable that some procedures are stored in Problem Observation entries. For example, although ICD-9-CM code of V76.10 (screening for malignant breast neoplasm, unspecified) is a procedure concept; it is listed under problem observations to describe the diagnosis of the cancer in claiming medical bills. Thus, the parser was programmed to collect all procedure concepts (i.e., Domain ID = “Procedure”) from entries under Problem Section.

As described in Table 9, transformation of data started with mapping procedures to OMOP standard concepts belonged to concept domain of “procedure” in OMOP vocabulary, and had procedure codes, such as CPT-4, ICDs, SNOMED CT, and HCPCS.

In case of a record of drug administration procedure, the procedure was only included in Procedure Occurrence table if the corresponding drug record existed in Data Exposure table.

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>procedure_occurrence_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each procedure occurrence</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from <em>Person</em> table</td>
</tr>
<tr>
<td>procedure_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of <em>Procedure Source Value</em></td>
</tr>
<tr>
<td>procedure_date</td>
<td><strong>Procedures</strong></td>
<td>The date when procedure occurred</td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity Procedure: <em>effectiveTime</em></td>
<td>If <em>statusCode</em> = “completed”</td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity Observation: <em>effectiveTime</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity Act: <em>effectiveTime</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Problems</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problem Section: Problem Concern Act: Problem Observation: <em>effectiveTime</em></td>
<td></td>
</tr>
<tr>
<td>procedure_type_concept_id</td>
<td>–</td>
<td><strong>Procedures</strong></td>
</tr>
<tr>
<td></td>
<td>Procedure Activity Procedure: 38000275 (EHR order list entry)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedure Activity Observation: 38000275 (EHR order list entry)</td>
<td></td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>–</td>
<td>The identifier of corresponding visit occurrence from Visit Occurrence table.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Null if Procedure Date did not match with any visit occurrence.</td>
</tr>
<tr>
<td>procedure_source_value</td>
<td>Procedures</td>
<td>If the source code is a procedure concept (i.e., Domain ID = “Procedure” and</td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity</td>
<td>Invalid Reason = Null), AND the service was completed (i.e., statusCode =</td>
</tr>
<tr>
<td></td>
<td>Procedure: code</td>
<td>“completed”).</td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation: code</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedures Section: Procedure Activity Act: code</td>
<td></td>
</tr>
<tr>
<td>procedure_source_concept_id</td>
<td>–</td>
<td>OMOP concept id of Procedure Source Value</td>
</tr>
</tbody>
</table>

For Problems:

- Problem Observation: 38000245 (EHR problem list entry)
- The identifier of corresponding visit occurrence from Visit Occurrence table.
2.2.5.8. Drug Exposure table

The Drug Exposure table contains records of vaccines, small-molecule, biological, and over-the-counter (OTC) medications. In the C-CDA CCD structure, current and past medication activities may be reported in two sections: Medications Section represents the history of drug therapy, and Immunization Section lists immunization history. The parser collected dispensed or administered medication records from Medications Activity template under Medications Section and Immunization Activity template under Immunizations Section. Table 11 summarizes the processing steps and applied rules to transform medication data.

Both medication and immunization records were only captured if they were supplied to the patients (i.e., dispensed or administered). The record had to meet all these criteria to be recognized as an actual medication or immunization activity rather than intended action:

1. **Immunization Activity** or **Medication Activity** was completed, i.e.,
   
   \[\text{substanceAdministration[@moodCode]} = \text{“EVN”}\] and
   \[\text{substanceAdministration/@statusCode} = \text{“completed”}\].

2. **Immunization Activity** must have \[\text{substanceAdministration[@negationInd]} = \text{“false”}\], indicating that immunization was truly administered.

3. Medication entries must have **Supply Activity** indicating that the medication was dispensed (\[\text{supply[@moodCode]} = \text{“EVN”}\] and \[\text{supply/@statusCode} = \text{“completed”}\]) as this template specifies the quantity of dispensed drug to the patient. This rule was not applied on immunization entries because vaccines are most often recorded as administered by a care practitioner.

2.2.5.8.1. Mapping drug product concepts

HL7 C-CDA CCD Release 1.1 requires medications to be reported in RxNorm concepts, such as semantic clinical drug (SCD), semantic brand drug (SBD), generic pack (GPCK), and brand pack (BPCK). The RxNorm codes were mapped to OMOP standard Concept IDs using scripts of Appendix 2 and Appendix 3.

The C-CDA standard also requires immunizations to be coded using Centers for Disease Control and Prevention (CDC) Vaccine Code (CVX); however, not only the OMOP vocabulary lacks the mappings of CVX codes to standard concepts, but...
also vaccine codes cannot be mapped explicitly to RxNorm concepts (85). Therefore, the equivalent CPT-4 codes of CVX codes were used as the surrogate standard concepts. Appendix 4 presents crossmaps of CVX codes to the representing CPT-4 codes retrieved from CDC website (86). The new mapping relationships were appended to Source-To-Concept Map table for automatic transformation of source codes to standard concepts.

The parser also captured the route of drug administration from Immunization Activity or Medication Activity templates, and stored the data in Route Source Value table field; however, this data was not used to find the standard concept of the route. Instead, the drug standard concept was the reference to match the correct route of administration.

2.2.5.8.2. Drug exposure duration

Start and end dates of drug exposure varies between immunization and medication entries. Immunization entries have only start date, while medication entries have both start and stop dates of drug treatment under the effectiveTime element identified by data type of contiguous time interval (i.e., IVL_TS). For example, “20170317” represents March 17, 2017 under Medication Activity template. However, in case of medications, the parser only captured the start date from low field under the effectiveTime element, and estimated end date (i.e., expiry date) by adding days of drug supply to the start date. Since immunizations and procedure drugs are administered in a single dose, the end date was recorded the same as administration date.

2.2.5.8.3. Days of drug supply

There is no field in CCD architecture to report for how long the supply of drug should has last (i.e., Days Supply); thus, it was calculated by dividing total amount of supplied drug by daily drug usage as shown in the following equation:

\[
\text{Days of supply} = \frac{\text{Quantity of supplied drug}}{\text{Daily dose interval} \times \text{Quantity of drug per dose}}
\]  

(1)

where quantity of supplied drug is captured from quantity[@value] element in Supply Activity under Medication Activity entries, daily dose interval captured from
**effectiveTime** identified by **operator = “A”** attribute, and the quantity of drug per dose extracted from **doseQuantity[@value]** of **Medication Activity** entries under **Medications Section**.

The unit of supplied drug quantity, such as tablet, milliliter, bottle, box, pack was recorded in **quantity[@unit]**. The quantity of supplied drug information was inconsistently represented with or without units. Therefore, supplied quantity of “2” for a tablet or capsule product could be reported with no units, or it was presented in tablet or bottle units. This was also the case for inhalers, solutions, and injections that units varied from no unit to milligram, gram, bottle, inhaler, vial, or syringe (Table 10).

To minimize the error in calculating duration of drug exposure, I developed this protocol to calculate days of drug supply:

1. Days of drug supply is calculated only if quantity of supplied drug (**quantity[@value]**) is provided.

2. If CCD record contains the unit of supplied quantity (**quantity[@value]**), the quantity value is standardized according to the proper physical units, and is stored in **Quantity** field. For example, if the record constituted 2 bottles of "Cephalexin 50 MG/ML Oral Suspension", the quantity of supplied drug would be 400 ml as each bottle contains 200 ml of the suspension. Table 10 presents the source and target units through the transformation.

3. If CCD record does not contain the unit of supplied quantity (**quantity[@unit]**), the quantity is transformed to the standard quantity unit according to Table 10. For example, if a record of "200 ACTUAT Albuterol 0.09 MG/ACTUAT Metered Dose Inhaler" has a quantity value of “1”, it is presumed that 200 actuations of the drug product were supplied.

4. If CCD record does not contain any of daily dose interval (**effectiveTime[@operator = “A”]**), dose quantity (**doseQuantity[@value]**), or the unit of dose quantity (**doseQuantity[@unit]**), the daily drug usage (the denominator of equation 1) is estimated using the World Health Organization (WHO) defined daily dose (DDD) (87). Thus, days of supply
is calculated by dividing total supplied quantity of drug by DDD. For example, if a record shows administration of 60 "valsartan 160 MG Oral Tablet", the days of supply would be 30 days as DDD of valsartan is 80 mg. The DDD was only used to test the pipeline.

(5) If CCD record has the unit of dose quantity (doseQuantity[@unit]), the dose quantity value is standardized to the proper physical units according to Table 10. For example, a dose quantity of 162 mg "Aspirin 81 MG Chewable Tablet" is transformed to 2 tablets.

(6) If a valid days of drug supply could not be calculated, the value was set to 1.

<table>
<thead>
<tr>
<th>Dose form</th>
<th>Quantity unit in CCD</th>
<th>Quantity unit in Drug Exposure table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerosol inhalation solution</td>
<td>aero, nebu</td>
<td>Actuation</td>
</tr>
<tr>
<td>Inhaler</td>
<td>aepb, inhaler, inhalatn</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Injectable solution</td>
<td>amp, soln, vial</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Injectable suspension</td>
<td>ml, vial</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Oral capsule suspension</td>
<td>caps, capsule, bottle, cpdr</td>
<td>Capsule</td>
</tr>
<tr>
<td>Oral solution</td>
<td>bottle, ml, soln</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Oral suspension</td>
<td>bottle, ml, powder</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Oral tablet</td>
<td>tabs, tablet, bottle, chew tab, tbdp</td>
<td>Tablet</td>
</tr>
<tr>
<td>Patch</td>
<td>patch</td>
<td>Patch</td>
</tr>
<tr>
<td>Prefilled syringe</td>
<td>syringe</td>
<td>Milliliter</td>
</tr>
<tr>
<td>Suppository</td>
<td>supp</td>
<td>Suppository</td>
</tr>
<tr>
<td>Topical cream</td>
<td>g, cream, gm, jar, tube</td>
<td>Gram</td>
</tr>
<tr>
<td>Topical gel</td>
<td>g, gel, jar, tube</td>
<td>Gram</td>
</tr>
<tr>
<td>Topical ointment</td>
<td>g, oint, jar, tube</td>
<td>Gram</td>
</tr>
<tr>
<td>Vaginal cream</td>
<td>tube</td>
<td>Gram</td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>drug_exposure_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each drug exposure occurrence</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from Person table</td>
</tr>
<tr>
<td>drug_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of Drug Source Value</td>
</tr>
<tr>
<td>drug_exposure_start_date</td>
<td>Immunizations</td>
<td>Medication effectiveTime[@xsi:type] = “IVL_TS”</td>
</tr>
<tr>
<td></td>
<td>Immunizations Section: Immunization Activity: effectiveTime[@value]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medications</td>
<td>Medication</td>
</tr>
<tr>
<td></td>
<td>Medications Section: Medication Activity: effectiveTime/low</td>
<td></td>
</tr>
<tr>
<td>drug_exposure_end_date</td>
<td>–</td>
<td>Immunization: Null</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medication The expiry date was calculated by adding days of drug supply to the start date.</td>
</tr>
<tr>
<td>drug_type_concept_id</td>
<td>–</td>
<td>38000177 (Prescription written)</td>
</tr>
<tr>
<td>refills</td>
<td>Medications</td>
<td>Immunization: Null</td>
</tr>
<tr>
<td></td>
<td>Medications Section: Medication Activity: repeatNumber</td>
<td></td>
</tr>
<tr>
<td>quantity</td>
<td>Medications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medications Section: Medication Activity: Supply Activity: quantity</td>
<td></td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>days_supply</td>
<td>–</td>
<td><strong>Immunizations</strong>: Null</td>
</tr>
<tr>
<td></td>
<td><strong>Medications</strong>: Using Days of Supply equation (1)</td>
<td></td>
</tr>
<tr>
<td>sig</td>
<td><strong>Medications</strong></td>
<td><strong>Medications</strong>: Using Days of Supply equation (1)</td>
</tr>
<tr>
<td></td>
<td>Medications Section: Medication Activity: text</td>
<td></td>
</tr>
<tr>
<td>route_concept_id</td>
<td>–</td>
<td>OMOP Concept ID of route of administration based on Drug Concept ID looked up in Concept Relationship table.</td>
</tr>
<tr>
<td>lot_number</td>
<td><strong>Immunizations</strong></td>
<td><strong>Immunizations</strong></td>
</tr>
<tr>
<td></td>
<td>Immunizations Section: Immunization Activity:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immunization Medication Information:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>manufacturedMaterial/lotNumberText</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Medications</strong></td>
<td>The identifier of corresponding visit occurrence from Visit Occurrence table. Null if Drug Exposure Start Date did not match with any visit occurrence.</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>–</td>
<td><strong>Immunizations</strong></td>
</tr>
<tr>
<td></td>
<td>If Immunization Activity was completed, AND Medication Dispense exists, AND the source code was a drug concept (i.e., Domain Id = “Drug”, AND Invalid Reason = Null).</td>
<td></td>
</tr>
<tr>
<td>drug_source_value</td>
<td><strong>Immunizations</strong></td>
<td><strong>Immunizations</strong></td>
</tr>
<tr>
<td></td>
<td>Immunizations Section: Immunization Activity:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immunization Medication Information:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>manufacturedMaterial/code</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Medications</strong></td>
<td><strong>Medications</strong></td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>drug_source_concept_id</td>
<td>Medications Section: Medication Activity: Medication Information: <em>manufacturedMaterial/code</em></td>
<td>If <em>Medication Activity</em> was completed, AND <em>Medication Dispense</em> exists, AND the source code was a drug concept (i.e., Domain Id = “Drug”, AND Invalid Reason = Null).</td>
</tr>
<tr>
<td>route_source_value</td>
<td>Immunizations</td>
<td>OMOP Concept ID of Drug Source Value</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
<td>If the route code does not exist, enter Null.</td>
</tr>
<tr>
<td></td>
<td>Medications Section: Medication Activity: <em>routeCode/@code</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medications</td>
<td></td>
</tr>
</tbody>
</table>
2.2.5.9. Drug Era table

The Drug Era table is derived from patient’s medication history (Drug Exposure table), and identifies the periods of time when patient was continuously exposed to medications at ingredient level. Thus, one or more individual drug exposure intervals will form a drug era. In case of a compound medication with multiple ingredients, individual exposure periods were generated with similar start and end date but representing different ingredients.

The parser derived drug eras from drug exposure records through the following steps. After identified OMOP-mapped medication intervals (i.e., Drug Concept ID > 0) from Drug Exposure table, the parser joined the drug concepts with the corresponding ingredients using the script in Appendix 5. Then, the intervals of ingredient exposures were consolidated if the gap between the previous end date and the next start date was 30 days or less. Table 12 summarizes applied rule to create drug eras.

Table 12. Applied rules to build Drug Era table

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug_era_id</td>
<td>Auto-numbered unique identifier for each condition era</td>
</tr>
<tr>
<td>person_id</td>
<td>The patient’s identifier from Person table</td>
</tr>
<tr>
<td>drug_concept_id</td>
<td>OMOP standard ingredient concept for which drug era is built, excluding Drug Concept ID = 0</td>
</tr>
<tr>
<td>drug_era_start_date</td>
<td>The start date of consolidated drug era as long as medication was supplied continuously in the next 30 days or less.</td>
</tr>
<tr>
<td>drug_era_end_date</td>
<td>The end date of consolidated drug era as long as medication was supplied continuously in the next 30 days or less.</td>
</tr>
<tr>
<td>drug_exposure_count</td>
<td>The number of drug exposure intervals that constitute the drug era.</td>
</tr>
<tr>
<td>gap_days</td>
<td>Deducted number of days covered by drug exposure intervals from number of days in the drug era</td>
</tr>
</tbody>
</table>
2.2.5.10. Measurement table

The Measurement table covers records of medical evaluations, including but not limited to laboratory tests, vital signs, imaging results, and pathology reports. HL7 C-CDA CCD documents extensively provide these data under three sections. Results Section reports the physician-generated results of diagnostic procedures, imaging, and laboratory test; Vital Signs Section describes the patient vital signs, such as blood pressure, temperature, height, weight, etc.; and Problem Observation entries under Problem Section record ICD codes, which are in fact measurement concepts. For example, V85.32 (Body Mass Index 32.0-32.9, adult) is an ICD-9-CM code to claim body weight service showing that patient had a body mass index of 32.0-32.9. In general, the parser stored records of measurement concepts in Measurement table (i.e., Domain ID = “Measurement” and Invalid Reason = Null). Table 13 summarizes the applied rules for populating Measurement table.

Results of medical evaluations such as laboratory tests (e.g., hematology, microbiology, toxicology, serology), imaging procedures (e.g., ultrasound, magnetic resonance imaging, angiography), and pathology reports (e.g., size of cancerous tumor) are stored in Result Organizer template under Results Section. The organizer groups related measurement results in the form of Result Observation templates in one place; for example, it categorizes all results of basic metabolic panel tests under code = “24321-2” LOINC code that contains Result Observation entries of anion gap, calcium, chloride, blood urine nitrogen (BUN), etc. The parser captured individual measurement records from these Result Observation entries, pointing at observation/code for the measurement code, observation/value for the measurement result, observation/effectiveTime for the date of service.

Vital Signs Section contains vital sign records within Vital Sign Observation entries under Vital Sign Organizer templates. The parser captured the vital sign exam code from observation/code element, date of service from observation/effectiveTime, and the vital sign value from observation/value. To ensure important vital sign parameters to clinical research are included in the mapping process, a custom mapping was added to Source-To-Concept Map table for some local codes such as weight, body mass index (BMI), and pulse rate.
Problem observations that represent measurement procedure were mapped to the corresponding clinical evaluation and value concepts using the script in Appendix 6, and the Concept IDs were stored in Measurement Concept ID and Value As Concept ID table fields. For example, “Basophilia” (ICD-9-CM = 288.65, OMOP Concept ID = 44831073) was mapped to “Basophil count” procedure (OMOP Concept ID = 4172647, Domain ID = “Measurement”) and “Above reference range” measurement value (Domain ID = “Meas Value”).

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>measurement_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each measurement</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from <em>Person</em> table</td>
</tr>
<tr>
<td>measurement_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of <em>Measurement Source Value</em></td>
</tr>
<tr>
<td>measurement_date</td>
<td>Vital signs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vital Signs Section: Vital Signs Organizer:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vital Sign Observation: <em>effectiveTime/@value/</em></td>
<td></td>
</tr>
<tr>
<td>measurement_type_concept_id</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vital signs: 44818701 (From physical examination)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Results: 44818702 (Lab result)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Problems: 38000245 (EHR problem list entry)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>operator_concept_id</td>
<td>–</td>
<td>Separate operator (&lt;, &gt;, ≤, or ≥) from <em>Value Source Value</em> if represents a value as number. Then, map to OMOP standard concept of Domain ID = 'Meas Value Operator'.</td>
</tr>
<tr>
<td>value_as_number</td>
<td>Vital signs</td>
<td>Vital signs</td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>Vital Signs Section: Vital Signs Organizer: <strong>Vital Sign Observation:</strong> <code>value[@value]</code></td>
<td>If the value is a physical quantity or real number, i.e., <code>value[@xsi:type] = “PQ”</code> or “REAL”.</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>Results Section: Result Organizer: Result Observation: <code>value[@value]</code></td>
</tr>
<tr>
<td>value_as_concept_id</td>
<td>–</td>
<td>If <code>Value Source Value</code> is a concept code form a coding system, e.g., SNOMED CT</td>
</tr>
<tr>
<td>unit_concept_id</td>
<td>–</td>
<td>OMOP Concept ID of <code>Unit Source Value</code></td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>–</td>
<td>The identifier of corresponding visit occurrence from <code>Visit Occurrence</code> table. Null if <code>Measurement Date</code> did not match with any visit occurrence.</td>
</tr>
<tr>
<td>measurement_source_value</td>
<td>Vital signs</td>
<td>Vital Signs Section: Vital Signs Organizer: <strong>Vital Sign Observation:</strong> <code>code</code></td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>Results Section: Result Organizer: Result Observation: <code>code</code></td>
</tr>
<tr>
<td>measurement_source_concept_id</td>
<td>–</td>
<td>OMOP Concept ID of <code>Measurement Source Value</code></td>
</tr>
<tr>
<td>unit_source_value</td>
<td>Vital signs</td>
<td>Vital Signs Section: Vital Signs Organizer: <strong>Vital Sign Observation:</strong> <code>value[@unit]</code></td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>value_source_value</td>
<td><strong>Results</strong></td>
<td><strong>Results</strong> If the value is a physical quantity, i.e., value/@xsi:type = “PQ”.</td>
</tr>
<tr>
<td></td>
<td><strong>Vital signs</strong></td>
<td><strong>Vital signs</strong> Section: Vital Signs Organizer: Vital Sign Observation: value/@value</td>
</tr>
<tr>
<td></td>
<td><strong>Results</strong></td>
<td><strong>Results</strong> Section: Result Organizer: Result Observation: value/@value</td>
</tr>
<tr>
<td></td>
<td><strong>Problems</strong></td>
<td><strong>Problems</strong> Section: Problem Concern Act: code</td>
</tr>
</tbody>
</table>
2.2.5.11. Observation table

The Observation table contains clinical facts of medical procedures and examinations that do not belong to other domains, including but not limited to personal and family medical history, social life, lifestyle, and smoking behavior. All records of observation concepts were stored in Observation table, meaning that Domain ID = “Observation” and Invalid Reason = Null.

Observation entries were extracted from Problem Section (personal and family history of diseases), Allergies Section (only history of allergies), Smoking Status Observation (smoking behavior) entries under Social History Section. Table 15 presents the rules applied to populate Observation table.

Smoking Status Observation reports the current smoking status of patient specified by SNOMED CT concepts, and the period of smoking activity is recorded under effectiveTime. If patient was a former smoker, observation date (Observation Date) would be the end of this period; however, if it was active or unknown smoking behavior, the observation date would be the starting time of smoking activity. Table 14 specifies the matching OMOP standard concepts of HL7 smoking status value set.

Observations may also be reported in Problem Section entries. For example, ICD-9-CM code of V12.71 (Personal history of peptic ulcer disease) is an observation concept listed under problem observations. Thus, the parser explored Problem Observation entries for observation concepts (i.e., Domain ID = “Observation”).

Resolved allergies observations captured in Allergies Section were also directed to Observation table. If the status of an allergy record was not active (i.e., Status Code = “active” under Allergy Problem Act), it was mapped to the respective observation concept (i.e., history of allergy).
<table>
<thead>
<tr>
<th>HL7 smoking status value</th>
<th>OMOP Standard Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Code Name</td>
</tr>
<tr>
<td>8517006</td>
<td>Former smoker</td>
</tr>
<tr>
<td>77176002</td>
<td>Smoker, current status unknown</td>
</tr>
<tr>
<td>266927001</td>
<td>Unknown if ever smoked</td>
</tr>
<tr>
<td>449868002</td>
<td>Current every day smoker</td>
</tr>
<tr>
<td>266919005</td>
<td>Never smoker (Never Smoked)</td>
</tr>
<tr>
<td>428041000124106</td>
<td>Current some day smoker</td>
</tr>
</tbody>
</table>
Table 15. Mapping HL7 C-CDA CCD data to Observation table: Applied rules and corresponding sections.

<table>
<thead>
<tr>
<th>Destination Table Field</th>
<th>Matching C-CDA section</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>observation_id</td>
<td>–</td>
<td>Auto-numbered unique identifier for each observation</td>
</tr>
<tr>
<td>person_id</td>
<td>–</td>
<td>The patient’s identifier from Person table</td>
</tr>
<tr>
<td>observation_concept_id</td>
<td>–</td>
<td>OMOP standard Concept ID of Observation Source Value</td>
</tr>
<tr>
<td>observation_date</td>
<td>Smoking status</td>
<td>If this is an active or unknow smoking behavior, the only existing time point is stored (i.e., effectiveTime/low).</td>
</tr>
<tr>
<td></td>
<td>Social History Section: Smoking Status</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation: effectiveTime/low or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effectiveTime/high</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problem Section: Problem Concern Act: Problem</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation: effectiveTime/low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allergies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allergies Section: Allergy Problem Act:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effectiveTime/low</td>
<td></td>
</tr>
<tr>
<td>observation_type_concept_id</td>
<td>–</td>
<td>Problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observation table can only store one time point, so only the start date is recorded in Observation Date field regardless it is a resolved or active problem.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smoking: 44814721 (Patient reported)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Problems: 38000276 (Problem list from EHR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allergies: 38000280 (Observation recorded from EHR)</td>
</tr>
<tr>
<td>Destination Table Field</td>
<td>Matching C-CDA section</td>
<td>Applied Rule</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>—</td>
<td>The identifier of corresponding visit occurrence from <code>Visit Occurrence</code> table. Null if <code>Condition Start Date</code> did not match with any visit occurrence.</td>
</tr>
<tr>
<td>observation_source_value</td>
<td>Smoking status:&lt;br&gt;Social History Section: Smoking Status Observation: <code>value/@code</code>&lt;br&gt;Problems&lt;br&gt;Problem Section: Problem Concern Act: Problem Observation: <code>code</code>&lt;br&gt;Allergies&lt;br&gt;Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: <code>code</code>&lt;br&gt;Participating agent in allergies&lt;br&gt;Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: Product: Product Detail: <code>playingEntity/code</code></td>
<td>If the source code is an observation concept (i.e., <code>Domain ID = “Observation”</code> and <code>Invalid Reason = Null</code>). Problems&lt;br&gt;The observation was truly observed (<code>negationInd = “false”</code> under <code>Problem Observation</code>). Allergies&lt;br&gt;OMOP concept id of <code>Condition Source Value</code> considering the participating allergen under <code>Allergy-Intolerance Observation</code></td>
</tr>
<tr>
<td>observation_source_concept_id</td>
<td>—</td>
<td>OMOP Concept ID of <code>Observation Source Value</code></td>
</tr>
</tbody>
</table>
2.2.6. Performance assessment of ETL pipeline

I assessed the accuracy of CCD-TO-OMOP ETL pipeline in two steps:

(1) Preliminary assessment that involved 10 randomly selected CCDs to optimize the pipeline, and

(2) Final assessment covering all 250 CCDs to finalize the transformation pipeline.

I manually reviewed CCD entries and generated data tables to examine the accuracy of data extraction, concept mapping, calculations (e.g., drug day supply), derived elements validity (e.g., drug and condition era constructions), and loading the data into OMOP CDM repository.

2.2.6.1. Evaluation of data extraction

Accuracy of data extraction phase plays a very important role in achieving the ultimate valid transformed dataset. The data extraction evaluation involved all standardized clinical data tables where the captured data usually reside in field ending with _Source_Value, including tables Person, Observation Period, Visit Occurrence, Procedure Occurrence, Drug Exposure, Condition Occurrence, Measurement, and Observation.

I manually examined captured data in the tables and checked with actual data values in CCD files’ structure to ensure:

(1) Data elements were correctly targeted in CCD architecture. For example, procedure records may be reported by different coding systems (e.g., ICD-9-CM, CPT-4) under different sections within CCDs, such as Problem Section and Procedures Section.

(2) Rules have been correctly applied when capturing data values from CCDs. For example, a drug record must be captured only if the drug was truly dispensed or administered to the patient, and the quantity of drug was reported.

If any discrepancy was observed, the pipeline was adjusted to capture the right value from the right location within the CCD files.

2.2.6.2. Evaluation of concept mapping

Concept mapping is part of the data transformation phase when captured concept codes are mapped to OMOP concepts. I manually reviewed all source codes and the matching Concept IDs for accuracy by fetching a distinct set of source codes.
and Concept IDs from all standardized clinical data tables to review a shorter list. The transformation involves mapping source values to source Concept IDs (table fields ending with _Source_Concept_Id) and OMOP standard concepts (table fields ending with _Concept_Id). In case a source code did not correctly matched or equal the proper OMOP Concept ID, it was flagged for further investigation to fix the glitch.

2.2.6.3. Evaluation of derived elements

OMOP CDM has specific tables called standardized derived elements that are derived from standardized clinical data tables. Two of these tables were in the scope of this study: Condition Era and Drug Era. I reviewed the populated condition eras and drug eras for accuracy to make sure the periods do not overlap and correspond the right condition occurrence or drug exposure.

2.2.6.4. Evaluation of calculated data fields

There are data fields in both standardized clinical data tables and standardized derived elements tables that are calculated as part of the data transformation process, including start and end dates of eras in Condition Era and Drug Era table, days of medication supply and drug exposure end date in Drug Exposure table, and observation periods in Observation Period table. I manually reviewed the calculated data fields, and adjusted the pipeline if needed.

2.2.6.5. Evaluation of data loading

Transformed data elements must be stored in the right place within OMOP CDM. Thus, I assessed all populated tables by these criteria to ensure correctness of the data loading process:

1. **Missing values**: All table fields with Null values were investigated if missing values are valid. Certain fields must be Null because either they were not in the scope of this study or no value was available in the CCDs. For example, Provider ID was always Null as the Provider table was out of scope, and Ethnicity Concept ID was if the CCD did not report ethnicity information.

2. **Concept domain**: All Concept IDs in the CDM tables should come from the respective domain. For example, condition records in Condition table must only have Condition Concept ID from condition domain.
validity of the concepts’ domains, I retrieved all recorded Concept IDs in the tables were retrieved and joined with their *Domain ID* in OMOP vocabulary, and manually reviewed their validity.

(3) *Foreign key references*: CDM tables are linked to each other via foreign keys that enables to retrieve associated events in other domains. For example, medication records of a patient in *Drug Exposure* table are identified by *Person ID* foreign key referring to the patient’s unique identifier in *Person* table. Therefore, other records from condition or procedure domains can also be collected associated with the drug therapy.

To ensure data fields are correctly linked, primary and foreign key constrains were added to OMOP CDM tables after data loading process. The SQL script of the constrains is provided in Appendix 7.

### 2.2.7. Statistical methods

This study mainly involves descriptive analysis of the ETL performance to transform data and the completeness of C-CDA CCD data to feed OMOP CDM.

It also reports accuracy assessment of mapped concepts. The accuracy of concept mapping process was assessed by measuring recall and precision of the ETL pipeline on 250 CCDs. Recall is the probability that a source code is correctly mapped to an OMOP concept, and was calculated by dividing number of truly mapped concepts by total number of correctly mapped and incorrectly unmapped concepts in each domain. Precision is the probability that a mapped source code is truly matched with an OMOP concept, and was calculated by dividing count of truly mapped concepts by total number of mapped concepts. A truly mapped code (true positive) is a valid source code that maps correctly to the corresponding OMOP standard concept, and a truly unmapped code (true negative) is an invalid code for which no standard concept is available in OMOP vocabulary. An incorrectly unmapped code (false negative) is a valid source code that a standard concept exists in the vocabulary but they do not match due to missing mapping relationships or a flaw in the pipeline.
2.3. Results

2.3.1. Performance of data extraction pipeline

The developed CCD-TO-OMOP data processing pipeline could successfully extract all required data elements from CCDs as it was planned. All rules were also applied properly as defined by HL7 implementation guide, such as the appropriate data element were targeted by Template IDs, it was ensured that drugs were truly administered to the patient, only occurred conditions were captured, and distinguished active versus history of allergies.

2.3.2. Performance of data transformation pipeline

The data transformation part of the ETL involved mapping source codes to standard concepts, building standardized derived elements, and calculating data fields. Performance of concept mapping is discussed in a separate section (see 2.3.3).

2.3.2.1. Standardized derived elements

The ETL created Condition Era (Table 16) and Drug Era (Table 17) tables, showing the continuous report of conditions and drug administrations, respectively. A total of 3,157 drug eras were derived from 7,859 drug exposure records, and 5,088 condition eras were formed from 12,648 condition occurrences.

<table>
<thead>
<tr>
<th>Condition Concept ID</th>
<th>Condition Name</th>
<th>Condition Era Start Date</th>
<th>Condition Era End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>4008576</td>
<td>Diabetes mellitus without complication</td>
<td>5/12/2012</td>
<td>5/12/2012</td>
</tr>
<tr>
<td>4008576</td>
<td>Diabetes mellitus without complication</td>
<td>12/28/2012</td>
<td>12/29/2012</td>
</tr>
<tr>
<td>4009042</td>
<td>Acute abscess of nasal sinus</td>
<td>11/10/2012</td>
<td>11/11/2012</td>
</tr>
<tr>
<td>4010024</td>
<td>Localized abdominal pain</td>
<td>10/7/2012</td>
<td>10/8/2012</td>
</tr>
<tr>
<td>4023183</td>
<td>Gastric reflux</td>
<td>10/8/2012</td>
<td>10/8/2012</td>
</tr>
<tr>
<td>4023183</td>
<td>Gastric reflux</td>
<td>6/8/2015</td>
<td>6/8/2015</td>
</tr>
<tr>
<td>4028741</td>
<td>Benign hypertension</td>
<td>4/16/2012</td>
<td>4/16/2012</td>
</tr>
<tr>
<td>4029305</td>
<td>Hypercholesterolemia</td>
<td>10/8/2012</td>
<td>10/8/2012</td>
</tr>
</tbody>
</table>
Table 17. An excerpt of records in Drug Era table.

<table>
<thead>
<tr>
<th>Drug Concept ID</th>
<th>Drug Name</th>
<th>Drug Era Start Date</th>
<th>Drug Era End Date</th>
<th>Drug Exposure Counts</th>
<th>Gap Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1307046</td>
<td>Metoprolol</td>
<td>9/21/2013</td>
<td>3/13/2015</td>
<td>4</td>
<td>-2</td>
</tr>
<tr>
<td>1307046</td>
<td>Metoprolol</td>
<td>4/14/2015</td>
<td>7/13/2015</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1308216</td>
<td>Lisinopril</td>
<td>9/10/2013</td>
<td>2/9/2016</td>
<td>8</td>
<td>-288</td>
</tr>
<tr>
<td>1332418</td>
<td>Amlodipine</td>
<td>9/10/2013</td>
<td>4/13/2015</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>1332418</td>
<td>Amlodipine</td>
<td>11/9/2015</td>
<td>2/7/2016</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1539403</td>
<td>Simvastatin</td>
<td>9/12/2013</td>
<td>10/12/2013</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1539403</td>
<td>Simvastatin</td>
<td>11/12/2013</td>
<td>2/10/2014</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>1539403</td>
<td>Simvastatin</td>
<td>3/19/2014</td>
<td>4/18/2014</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

2.3.3. Mapping performance of CCD data to OMOP CDM

2.3.3.1. Conditions and diagnoses

The ETL pipeline yielded very good to excellent performance in mapping source codes to OMOP standard concepts (Figure 4). A total of 12,648 records and 1,459 concepts of diagnoses and reported conditions were retrieved from the CCDs (Table 18). All extracted condition concepts were coded in SNOMED CT, of which 1,456 (99.8%) could map to OMOP standard concepts representing 12,644 (99.7%) records. These mapped source codes accurately matched the corresponding standard concepts in OMOP vocabulary.

The three unmapped source codes could not map to standard concepts because either the concept was deprecated or the relationship to standard concept was obsolete.

2.3.3.2. Drugs

A total of 7,859 drug administration and immunization records were extracted from the CCD documents, of which 596 records had empty drug code field; thus, 7,263 records were analyzed for evaluating mapping performance (Table 18). Out of extracted 770 drug concepts, 758 (98.4%) all-RxNorm codes correctly mapped to OMOP standard concepts representing 7,230 (99.5%) records.
The remaining 12 (1.6%) concepts that consisted 33 (0.5%) of the records were coded in local code \((n = 4)\), RxNorm \((n = 7)\), and CVX \((n = 1)\). They could not map to standard concepts because the relationships to standard concepts were deprecated or the corresponding OMOP concepts did not exist in the vocabulary.

2.3.3.3. Procedures

Out of 1,128 retrieved procedure records, 941 (83.4%) records correctly mapped to standard concepts (Table 18). This represents 367 (92.7%) of total extracted 396 concepts. Most of procedures codes were from ICD-9-CM Procedure (72.5%) and CPT-4 (25.5%) coding systems, and the remaining were local (1%) or SNOMED CT (1%) codes. Among these, almost all ICD-9-CM codes (99.3%) could map to standard concepts while 79.2% of CPT-4 codes matched the corresponding standard concepts in OMOP vocabulary. None of local codes and 2 (50%) out of four SNOMED CT procedure codes could also map to standard concepts.

Among 29 (7.3%) unmapped procedure codes, 21 CPT-4 and two ICD-9-CM, and two SNOMED CT codes could not map to standard concepts due to deprecated relationships. Regarding the remaining 4 local codes, there were no matching record available in OMOP vocabulary to standardize the concepts.

2.3.3.4. Measurements and clinical evaluations

A total of 36,184 records of 683 measurement concepts were extracted from CCDs (Table 18). The transformation process yielded 94.0% \((n = 642)\) correct mapping of concepts to OMOP standard concepts representing 35,724 (98.7%) records. The measurement concepts were mostly coded in LOINC \((n = 605)\), and the remaining were either local \((n = 3)\) or SNOMED CT \((n = 40)\) codes. Of these, 599 (99%) of LOINC codes could map to standard concepts, 40 (85.1%) SNOMED CT, and 3 (9.7%) local codes.

There were 41 (6%) unmapped source codes were representing 460 (1.3%) of total records. Among them, 6 LOINC and 7 SNOMED CT codes could not map to standard concepts because either the concept record or the mapping relationships were obsolete, while there was no record available in OMOP vocabulary to match 28 local codes.

2.3.3.5. Observations

All 71 observation concepts were coded in SNOMED CT, representing 280 records (Table 18). All concepts correctly mapped to OMOP standard concepts.
### 2.3.3.6. CDC race and ethnicity value set

Four race codes were captured from 74 records that all mapped to OMOP standard concepts (Table 18).

<table>
<thead>
<tr>
<th>Domains</th>
<th>Concept</th>
<th>Record</th>
<th>Concept</th>
<th>Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>99.8%</td>
<td>99.97%</td>
<td>98.4%</td>
<td>99.5%</td>
</tr>
<tr>
<td>Drug</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure</td>
<td>92.7%</td>
<td>83.4%</td>
<td>94.9%</td>
<td>98.7%</td>
</tr>
<tr>
<td>Measurement</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.** Overall mapping performance of concepts and records to OMOP CDM vocabulary by domain.
Table 18. Mapping performance of source codes to standard concepts of OMOP CDM vocabulary.

<table>
<thead>
<tr>
<th>Domain and Code System</th>
<th>Concepts</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mapped, n (%)</td>
<td>Unmapped, n (%)</td>
</tr>
<tr>
<td><strong>Condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>1,456 (99.8)</td>
<td>3 (0.2)</td>
</tr>
<tr>
<td><strong>Drug</strong></td>
<td>758 (98.4)</td>
<td>12 (1.6)</td>
</tr>
<tr>
<td>CVX</td>
<td>1 (100)</td>
<td>1</td>
</tr>
<tr>
<td>Local code</td>
<td>4 (100)</td>
<td>4</td>
</tr>
<tr>
<td>RxNorm</td>
<td>758 (99.1)</td>
<td>7 (0.9)</td>
</tr>
<tr>
<td><strong>Procedure</strong></td>
<td>367 (92.7)</td>
<td>29 (7.3)</td>
</tr>
<tr>
<td>CPT-4</td>
<td>80 (79.2)</td>
<td>21 (20.8)</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>285 (99.3)</td>
<td>2 (0.7)</td>
</tr>
<tr>
<td>Local code</td>
<td>4 (100)</td>
<td>4</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>2 (50)</td>
<td>2 (50)</td>
</tr>
<tr>
<td><strong>Measurement</strong></td>
<td>642 (94.0)</td>
<td>41 (6.0)</td>
</tr>
<tr>
<td>Local code</td>
<td>3 (9.7)</td>
<td>28 (90.3)</td>
</tr>
<tr>
<td>LOINC</td>
<td>599 (99.0)</td>
<td>6 (1.0)</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>40 (85.1)</td>
<td>7 (14.9)</td>
</tr>
<tr>
<td><strong>Observation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>71 (100)</td>
<td>71</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC Race and Ethnicity</td>
<td>4 (4)</td>
<td>4</td>
</tr>
</tbody>
</table>
2.3.4. Accuracy of concept mapping

The concept mapping pipeline yielded an overall recall of 98.5% and precision of 100% (Table 19). The recall of condition codes was 99.8% as three condition codes did not map to standard concepts due to missing mapping relationships. Of 12 unmapped drug codes, four codes truly and eight incorrectly did not map to standard concepts, resulting a recall of 99.0%. The lowest recall (93.6%) was achieved with procedure codes where 25 codes could not map to standard concept in OMOP vocabulary, mainly because the vocabulary failed to have code-to-standard relationships in place of deprecated ones for CPT-4 codes. All but 13 measurement codes did not map to standard concepts, yielding a recall of 98.0%. Although the vocabulary contained the concepts, proper mapping relationships did not exist to be used by the pipeline. The recall of pipeline to map observation codes was 100% as all could map to standard concepts.

Table 19. Accuracy assessment of ETL pipeline to map source codes to standard concepts

<table>
<thead>
<tr>
<th>Concept's Domain</th>
<th>Mapped Code</th>
<th>Unmapped Code</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct (TP)</td>
<td>Incorrect (FP)</td>
<td>Correct (TN)</td>
<td>Incorrect (FN)</td>
</tr>
<tr>
<td>Condition</td>
<td>1,456</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Drug</td>
<td>758</td>
<td>0</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Procedure</td>
<td>367</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Measurement</td>
<td>642</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Observation</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,294</strong></td>
<td><strong>0</strong></td>
<td><strong>4</strong></td>
<td><strong>49</strong></td>
</tr>
</tbody>
</table>

TP = True positive, FP = False positive, TN = True negative, FN = False negative
2.3.5. Completeness of CCD data elements required by OMOP CDM

In general, CCD documents provided minimum required data elements of patient information to feed OMOP CDM (Table 20). Out of 250 patients’ CCDs, all provided gender information and 94.8% had birth dates. However, the documents significantly lacked race and ethnicity information as only 29.6% held race and none had ethnicity information. On visit occurrence information, all captured records from CCDs reported start and end dates of encounters, while 95.6% of the records specified the type of setting (e.g., inpatient, outpatient). The 212 records not reporting visit codes were obtained from 48 CCDs. All records of conditions, procedures, and observations fully reported code and date of occurrences.

Drug exposure and measurement records provided data elements diversely. Although all drug exposure records reported drug supply date and 99.8% had quantity and days of supply information, only 92.4% of them accompanied drug codes, meaning that 596 drug exposure records could not be included in the CDM as the drug could not be recognized. Likewise, all measurement records had the code and date of service, but 3% of the records captured from 43 documents did not present measurement values.
Table 20. The number of patients and records for which CCDs carried data.

<table>
<thead>
<tr>
<th>OMOP Table and Data Elements</th>
<th>Records Providing Value</th>
<th>Records Not Providing Value</th>
<th>Involved CCDs, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count, n (%)</td>
<td>Count, n (%)</td>
<td></td>
</tr>
<tr>
<td><strong>Person (N = 250)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth date</td>
<td>237 (94.8)</td>
<td>13 (5.2)</td>
<td>13 (5.2)</td>
</tr>
<tr>
<td>Gender code</td>
<td>250 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Race code</td>
<td>74 (29.6)</td>
<td>176 (70.4)</td>
<td>176 (70.4)</td>
</tr>
<tr>
<td>Ethnicity code</td>
<td>0 (0)</td>
<td>250 (100)</td>
<td>250 (100)</td>
</tr>
<tr>
<td><strong>Visit Occurrence (N = 4,795)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visit Start Date</td>
<td>4,795 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Visit End Date</td>
<td>4,795 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Visit code</td>
<td>4,583 (95.6)</td>
<td>212 (4.4)</td>
<td>48 (19.2)</td>
</tr>
<tr>
<td><strong>Condition Occurrence (N = 12,648)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition date</td>
<td>12,648 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Condition code</td>
<td>12,648 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>Procedure Occurrence (N = 1,128)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure date</td>
<td>1,128 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Procedure code</td>
<td>1,128 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>Drug Exposure (N = 7,859)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug supply date</td>
<td>7,859 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Quantity supplied</td>
<td>7,847 (99.8)</td>
<td>12 (0.02)</td>
<td>7 (2.8)</td>
</tr>
<tr>
<td>Days of supply</td>
<td>7,847 (99.8)</td>
<td>12 (0.02)</td>
<td>7 (2.8)</td>
</tr>
<tr>
<td>Drug code</td>
<td>7,263 (92.4)</td>
<td>596 (7.6)</td>
<td>69 (27.6)</td>
</tr>
<tr>
<td><strong>Measurement (N = 36,184)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement date</td>
<td>36,184 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Measurement value</td>
<td>35,059 (97.0)</td>
<td>1,095 (3.0)</td>
<td>43 (17.2)</td>
</tr>
<tr>
<td>Measurement code</td>
<td>36,184 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>Observation (N = 280)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation date</td>
<td>280 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Observation code</td>
<td>280 (100)</td>
<td>0 (0)</td>
<td></td>
</tr>
</tbody>
</table>
2.4. Discussion

The OMOP CDM demonstrated the ability to accommodate patient electronic health records transferred by HL7 C-CDA-based CCD. Demographics, healthcare provider encounters, medication information, procedures, measurements, and diagnoses could be successfully extracted, transformed, and loaded to the CDM with high accuracy. The final dataset can be used for population health observational studies, such as drug safety surveillance and comparative effectiveness research.

2.4.1. Accommodation of CCD data in OMOP CDM

All CDM tables assessed in this study could successfully accommodated CCD data; however, some tweaks were needed in Drug Exposure and Condition Occurrence tables to optimize data transformation.

The Person table had proper fields to store source codes and standard concepts of patient demographics data elements captured from US Realm Header of CCD documents. The minimum required data elements included date of birth, gender, race, and ethnicity. The vocabulary possessed corresponding standard concepts of the source codes; however, it did not provide mapping relationships to transform CDC race and ethnicity value sets to standard concepts. Thus, new relationships were added to Source-To-Concept Map table to enable mapping of race information.

The Visit Occurrence, Measurements, and Observation tables successfully fitted source codes and standard concepts. The vocabulary had also the proper relationships to map source codes to standard concepts.

The Condition Occurrence table had also proper data fields to store both source code and standard concept of reported medical conditions, but failed to collect few elements of active allergy information. In order to map events to standard concepts in the OMOP vocabulary, the source code of the event is sufficient to lookup the matching concept. While C-CDA CCDs only use one code (e.g., ICDs, SNOMED CT) to report conditions, the allergies are presented by two codes, one for the allergy event and one for the playing entity that caused the allergy. On the other side, Condition Occurrence table has only space for the source code (one Condition Source Value and one Condition Source Concept ID fields); thus; can only take the generic allergy concept (e.g., Concept ID = 439224 if drug allergy).
The *Procedure Occurrence* table could properly accommodate the required data elements from CCDs, including procedure codes and date of service; however, the OMOP vocabulary presented moderate-to-good performance to map procedure codes to standard concepts as only supported mapping of 92.7% procedure codes to standard concepts, representing 83.4% of procedure records. Despite all unmapped codes (excluding local codes) had representing non-standard Concept IDs in the vocabulary, the code-to-standard relationships were flagged as depreciated and no updated mappings were provided. This issue was resolved by incorporating the updated mapping relationships in the CDM.

The *Drug Exposure* table could properly contain drug supply date, supplied drug quantity, and drug code extracted from CCDs; however, drug’s days of supply had to be estimated using the developed protocol because HL7 C-CDA does not provide this information explicitly. An alternative approach to compute days of supply is to use daily dose interval and quantity of drug per dose; however, this information was provided sparsely in the CCDs. Thus, I applied the WHO’s defined daily dose (DDD) as the reference for daily drug usage to benchmark the pipeline. It should be emphasized that the DDD was not for making inferences about drug exposures. The daily defined dose defines the average dose of a drug used for maintenance therapy of the main indication in adults, and it has been used for drug utilization, pharmacoepidemiologic, and pharmacoeconomic research (87). Although the DDD is a composite, standard measure of drug exposure in various countries, and has been used in many studies (88-93), using DDD as the substitute for prescribed daily dose may cause incorrect estimations about drug’s days of supply (94-96). Therefore, it is crucial to include enough data elements in the CCDs to correctly estimate drug’s days of supply if the documents are meant to be used for observational analyses.

### 2.4.2. Strengths

I found several strengths in standardizing CCD data to OMOP CDM. First, the CDM allows to gather patient electronic health records from other sources independent of the structure of database. Second, the OMOP vocabulary supports most of the coding systems that are being used in CCD architecture, and it is also scalable to cover local codes and emerging dictionaries in future. Third, the CDM provides new measures, such as drug’s days of supply to data analysts that are not
explicitly reported by CCDs. Fourth, the derived tables, such as *Drug Era* and *Condition Era* that provide timeline of therapies and condition reports for health outcome research. Fifth, the CDM standardizes diverse quantity unit of supplied drug to make drug exposure records comparable. In general, transforming CCD data with such highly complex architecture and varied elements into a common data model will allow using the data to analyze data across other observational databases.

### 2.4.3. Challenges

Missing data in the CCDs exceedingly challenged the pipeline, in particular to calculate drug’s days of supply. The CCDs also reported limited information on race and ethnicity, and 3% of the measurement records had no values. When using CCDs’ data for observational analyses, the researcher may need to ensure first that the missing values are legit after consulting with the data provider; then, alternative solutions should be followed to limit bias, such as case deletion, mean substitution, and imputation (97-100).

Although HL7 C-CDA Release 1.1 requires medications to be recorded in RxNorm codes, there were few records of medications coded using local codes or only the medication name was provided. The parser in this study captured drug information only if it was coded in a standardized vocabulary like RxNorm. Therefore, two types of records could not be extracted: The records with drug concepts coded in non-standardized vocabularies (e.g., local dictionaries), and the records that only provided the drug name not the standard concept.

Vaccines were recorded in CVX codes in CCDs, but the OMOP vocabulary did not have the proper mappings to standard concepts. To enhance the performance of concept mapping pipeline, new code-to-standard mapping relationships were added to *Source-To-Concept Map* table that mapped CVX codes to the corresponding standard CPT-4 drug Concept IDs as the surrogate concepts. Ideally, one vaccination code should map to one RxNorm concept, but it does not happen with vaccine codes because neither CVX nor CPT-4 codes specify the type or variant of vaccines, which is crucial to find the explicit RxNorm concept (85).
2.4.4. Limitations of the study

It is possible that the accommodation and accuracy assessment would yield different results when processing CCDs from other institutions as they may differ in providing non-required (i.e., recommended or optional) data elements. The OMOP CDM was evaluated on 250 randomly selected CCD files from Regenstrief Institute, and needs to be evaluated on a larger, diverse pool of CCD documents.

The assessment results are also limited to CCDs that are structured based on HL7 C-CDA Release 1.1 standard. Other versions of CCD documents may have different constrains on the data elements, and may vary in providing data fields. Therefore, it is recommended to test other versions of CCDs on OMOP CDM.

In general, observation entries can be found in different sections within the architecture of HL7 C-CDA CCD, including Family History Section, Functional Status Section, Problem Section, Allergies Section, and Social History Section. However, only Problem Section (personal and family history of diseases), Allergies Section (only history of allergies), Smoking Status Observation (smoking behavior) entries under Social History Section were analyzed in this study since other sections were optional sections and did not exist in the CCDs. Therefore, if the missing sections were provided, the CDM could yield a more comprehensive list of patient’s conditions and observations.

CCD documents generated by different systems may differ in providing data elements as that the CCD architecture delineates constrains of elements whether they are optional or recommended to be included in the documents, such as Encounters Section, Immunizations Section, Social History Section, Vital Signs Section, race and ethnicity in US Realm Header, the unit of supplied quantity and daily dose quantity in Medications Section, and procedure service date in Procedures Section. Therefore, it is important to ensure these data are included in the CCDs that exchange patient information for observational studies.
CHAPTER 3. A STANDARD FOR DISSEMINATING HEALTH RISK PREDICTION MODELS TO SUPPORT CLINICAL DECISION-MAKING

3.1. Background

A wide range of predictive models have been derived from EHRs and observational data with the goal to improve patient safety, cut treatment costs through early detection of diseases, and assist clinicians to make accurate clinical decisions. Typical examples of predictive models include predicting prognostic risk of health events (27-29), disease screening (30-32), and managing short- and long-term complications (33-36). In one study on senior patients in Veterans Affairs (VA) nursing homes, the researchers could estimate personalized change in functional loss and recovery after hospitalization with 84-92% accuracy using random forest approach (37). In another study, the use of EHR data could refine prediction of 30-day hospital readmission risk after percutaneous coronary intervention (38).

Predictive models have also been used to support individuated decision support through estimating complications, short-term readmission, and long-term prognosis (39). Lee et al., 2015 (40) also reported a prognostic prediction model that has the potential to help clinicians design personalized treatments for ischemic stroke. Predictive modeling can also improve patients’ safety by reducing human errors. Physicians and healthcare providers are prone to cognitive biases, logical fallacies, false assumptions, and other reasoning failures when making clinical decision about diagnosis or treatment (41); therefore, predictive analytics tools may help them fill the gap by providing prediction estimates about the patients’ conditions.

3.2. Problem statement

Deploying predictive models for testing in medical practice or delivering health-related predictions at the point-of-care is a costly and time-consuming process that requires teams of professionals from IT managers, computer programmers, and analysts to database administrators to deploy the models and analyze the results for accuracy. The process needs to be repeated every time a new model is deployed as predictive models differ in specifications and data
requirements. The deployment is even more challenging when implementing the predictive models across desperate databases with different structures.

3.3. Overview of PMML

Introduced in 1997 (101), Predictive Model Markup Language (PMML) is an open source that allows exchanging predictive and data mining models between data systems (81, 101). The PMML can share the specifications of predictive models, data mining process, data transformation procedure, definitions of variables, the output of the model, model scoring steps, model explanation, model validation, and target properties of outputs (81). A PMML document may contain the definition of one or more analytic models. Due to the extensible markup language (XML) based architecture, the PMML is platform-independent, human readable, and easy to implement and maintain in the existing operating systems. It also supports varieties of algorithms as the latest PMML version 4.3 covers 17 classes of models from logistic regression, linear regression, Bayesian network to Cox proportional hazards model and decision tree (102).

3.4. The Structure of PMML

The general structure of PMML version 4.3 document is composed of five sections (103): Header, Mining Build Task, Data Dictionary, Transformation Dictionary, and Model (Figure 5).

The required Header section starts the document, and provides information about copyright, a description of the model, model’s version, and the model generator software, and the timestamp when model was created.

The optional Mining Build Task section describes the specifications of data mining process, and how the model was trained. This section has no predefined structure, and can contain any mining standards, such as structured query language (SQL).

The required Data Dictionary section is shared among all models in the document, and defines the name, type (e.g., integer, double, string, date), and value of participating data fields in the model. Each Data Field element defines one continuous, categorical, ordinal, or text variable of the model.

The optional Data Transformation section specifies the procedures (e.g., normalization, discretization, value mapping, built-in functions, aggregation) to
transform input variables from *Data Dictionary* to *Derived Fields* that will be used by the model. This section also allows to define new functions to apply on variables.

The required *Model* section defines the specifications of predictive model. The elements of *Model* section may differ depending on the model type; however, these elements are shared between all types: Model name, mining function name (e.g., regression, clustering, classification), *Mining Schema*, *Output*, *Model Statistics*, *Targets*, *Local Transformations*, *Model Verification*, and *Model Explanation*. All models in a PMML document can access the elements under *Data Dictionary* and *Transformation Dictionary* sections. Every model mining process begins with the *Mining Schema* section that lists the participating variables in the under-process model as *Mining Field* elements from the pool of *Data Field* and *Derived Field* elements. The optional *Output* section lists *Output Field* elements that specify the model’s returning result values including the name, value type (i.e., continuous, categorical, ordinal, or text), data type (e.g., integer, double, date, text), and the procedure to compute final output values. The optional *Model Statistics* section contains univariate and multivariate statistics on *Data Fields*, providing further information about the nature of training dataset, and the model’s quality. The optional *Targets* section provides more information about the predicted *Data Field*, such as prior probabilities, scaling factor, minimum and maximum values of the *Data Field* in training set. The *Local Transformations* section specifies transformation procedures as *Derived Field* elements for *Mining Field* elements. The transformation types are similar to *Transformation Dictionary* but are only applied locally to the parent *Model*. The *Model Verification* section offers a sample set of inputs and validated results from the training dataset to the destination system. This helps assess the performance of model and compare accuracy of the model’s validity across datasets. While *Model Statistics* provides descriptive properties of input variables of the model, the *Model Explanation* section contains information about the quality of model, such as receiver operating characteristic (ROC), adjusted r-squared, the sum of squares regression statistics to enable assessing quality of model across datasets.

Each type of predictive model may also have specific sections. As an example, regression model has also *Regression Table* section that specifies the values of
participating predictor variables (Figure 5). If the model predicts one numeric variable (i.e., linear regression), only one Regression Table should exist, and the model will have two or more of this section if the predicted variable is categorical (i.e., logistic or polynomial regression).

![PMML Diagram](image)

**Figure 5.** The general structure of a PMML document (left), and the schema of sections under Regression Model (right).

### 3.5. Objectives

This study intended to develop a new standard based on the existing predictive model markup language (PMML) for disseminating health outcome risk scoring models that are generated based on OMOP CDM data. This new OMOP-compliant PMML (O-PMML) standard defines the specifications of models, involved variables, data mining process, scoring steps, and data transformation procedure in accordance to the CDM specifications. The O-PMML enables IT solutions to deploy predictive models on OMOP CDM repositories in a “plug-and-play” manner that not only can save implementation cost and time, but also ensures consistency of computations across systems. This paper evaluates the feasibility and performance
of using the new standard to estimate risk score Framingham 10-year risk of cardiovascular diseases.

3.6. Methods

3.6.1. The structure of OMOP-compliant PMML

This paper presents adoption of PMML version 4.3 (104) for sharing regression models to estimate disease risk scores using OMOP CDM data. The O-PMML standard was designed to fulfill specific requirements of mining OMOP CDM. These requirements included:

- **Req. 1)** The standard should enable sharing information about the model provider, the generator software, descriptions of the model, and compatible version of OMOP CDM.
- **Req. 2)** The standard must enable sharing participating dependent and independent variables and the corresponding coefficients.
- **Req. 3)** The standard must enable sharing data transformation procedures. For example, what are the cutoff points to transform a continuous measure to categorical values.
- **Req. 4)** The standard must enable sharing the mining process of CDM tables for defining variables. For example, how to determine whether the patient is diabetic or current smoker.
- **Req. 5)** The standard must specify the output values, and enable sharing the process to compute the output values. For example, how to calculate the risk score yielded from the model.

The OMOP-compliant PMML is based on PMML architecture (Figure 5) with proper modifications to satisfy the aforementioned requirements of OMOP CDM. The modifications mostly involve changes in constraints and *Extension* elements. It should be noted that all added extensions have an attribute of *extender* throughout the O-PMML that shows the change is due to OMOP requirements. The general architecture of O-PMML consists of *Header, Mining Build Task, Data Dictionary,* and *Transformation Dictionary* sections (Figure 6). All sections are required to be included in the PMML document except *Transformation Dictionary* which is added if data fields need transformation before entering the predictive model. The type of
predictive model is selected from Model Element group, which only covers regression models in our case. The full XML schema of O-PMML is provided in Appendix 8.

The Header section satisfies the requirement of sharing general information about the model, author, and CDM version (Req. 1). The Data Dictionary and Regression Model sections of the PMML fulfill Req. 2 with no further modifications needed. The Data Dictionary section specifies non-transformed data fields that will participate in the models within PMML document. The Transformation Dictionary and Local Transformations sections also satisfy Req. 3 to share data transformation processes. The Mining Build Task section addresses Req. 4 as enables sharing database mining pipelines. Finally, the Output elements under Regression Model section fulfills Req. 5 as provide the space to specify yield values and processes to calculate the final outputs (e.g., risk score).

```xml
<xs:element name="PMML">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Header"/>
      <xs:element ref="DataDictionary"/>
      <xs:element ref="TransformationDictionary" minOccurs="0"/>
      <xs:element ref="MiningBuildTask"/>
      <xs:sequence minOccurs="1">
        <xs:group ref="MODEL-ELEMENT"/>
      </xs:sequence>
    </xs:sequence>
    <xs:attribute name="version" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>

<xs:group name="MODEL-ELEMENT">
  <xs:choice>
    <xs:element ref="RegressionModel"/>
  </xs:choice>
</xs:group>
```

Figure 6. The XML schema of the general architecture of O-PMML
The *Header* section has already the placeholder for storing the information of software generating the predictive model under *Application* element, and descriptions of the model in *description* attribute under *Header* element, but an extension was added to cover information on OMOP version and model provider (Figure 7). The new extension has two required children elements of *OmopCdm* representing the version of OMOP and *author* showing the model’s author.

The *Data Dictionary* section lists data variables in *Data Field* elements that will participate in the model (Figure 8). The data fields are categorical, ordinal, or continuous values retrieved directly from OMOP CDM that may or may not undergo transformations before being used in the model. If the data is categorical or ordinal, the *value* element specifies possible values under *Data Field* elements. In case of interval data fields, the *interval* element defines the lower and upper bound closures. The data fields also allow to distinguish missing, valid, and invalid input values. For example, we can specify if ‘99’ represents missing values or a certain value like ‘1’ denotes invalid values in a field to be excluded from analyses.

The *Transformation Dictionary* and *Local Transformations* sections define transformation procedures of data fields before participating in the models (Figure 9). Under these sections, data fields can be transformed to yield *Derived Field* elements through normalization, discretization, value mapping, text indexing, aggregation (e.g., count, average, sum), lag, and applying built-in (e.g., addition, subtraction, divide, multiplication, log10, ln, conditions) or user-defined functions. In contrast to *Transformation Dictionary* that applies globally to all models within the PMML document, *Local Transformations* can only be used by the local *Regression Model* where located (Figure 11).
Figure 7. The XML schema of Header section in O-PMML
Figure 8. The XML schema of *Data Dictionary* section and *Data Field* elements in O-PMML
Figure 9. The XML schema of Transformation Dictionary section, Local Transformations section, Derived Field elements, and acceptable data transformation expressions in O-PMML.
The *Mining Build Task* section acts as the blueprint to mine *Data Field* values from OMOP CDM. It describes what criteria were applied to define the variable, and which CDM tables are involved to extract data fields. This section of PMLL has no specific structure, and it can contain any mining standards. Therefore, a new XML schema was designed for this section that not only satisfies Req. 4, but also needs minimum implementation in the existing working PMML parser modules. The new schema uses *Extension* elements to define the data mining pipelines, specified by attribute *extender* of “OMOP” (Figure 10). For each *Data Field* element, there is one *Extension* element with the same attribute *name*. The *Extension* element must have only one *Statement* element and may have *Input Parameters* section. The *Statement* element contains the SQL, R, or Java script to query the CDM tables to retrieve values of data fields. If the script requires input values, for example index date or Person ID, the placeholder value is denoted by at symbol (e.g., @PERSON_ID). The *Input Parameters* section specifies the input values of the *Statement* element’s script before querying the database. Each placeholder value is defined by an *Input Parameter* element that specifies the name and type of the input value.
Figure 10. The new XML schema of Mining Build Task section and the required elements and attributes in O-PMML
The Regression Model section (Figure 11) follows PMML standard with no change. It specifies the participating variables in Mining Schema section, the coefficient of variables in Regression Table section, optional data transformations in Local Transformations section applicable to the model, the yields of the model and the required processes to compute output values in optional Output section, statistics of training set in optional Model Stats section, optional information about the quality of model in Model Explanation section, properties of target values in optional Targets section, and a dataset for verifying the results of the model in optional Model Verification section.

```xml
<xs:element name="RegressionModel">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="MiningSchema"/>
      <xs:element ref="RegressionTable" maxOccurs="unbounded"/>
      <xs:element ref="LocalTransformations" minOccurs="0"/>
      <xs:element ref="Output" minOccurs="0"/>
      <xs:element ref="ModelStats" minOccurs="0"/>
      <xs:element ref="ModelExplanation" minOccurs="0"/>
      <xs:element ref="Targets" minOccurs="0"/>
      <xs:element ref="ModelVerification" minOccurs="0"/>
    </xs:sequence>
    <xs:attribute name="modelName" type="xs:string"/>
    <xs:attribute name="functionName" type="MINING-FUNCTION" use="required"/>
    <xs:attribute name="algorithmName" type="xs:string"/>
    <xs:attribute name="modelType" use="optional">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="linearRegression"/>
          <xs:enumeration value="stepwisePolynomialRegression"/>
          <xs:enumeration value="logisticRegression"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="targetFieldName" type="FIELD-NAME" use="optional"/>
    <xs:attribute name="normalizationMethod" type="REGRESSIONNORMALIZATIONMETHOD" default="none"/>
    <xs:attribute name="isScorable" type="xs:boolean" default="true"/>
  </xs:complexType>
</xs:element>
```

Figure 11. The XML schema of Regression Model section in O-PMML
The Mining Schema (Figure 12) describes the participating variables by
Mining Field elements where the usage type attribute specifies whether the variable
is dependent (i.e., predicted) or independent (i.e., active). The Mining Field elements
may refer to Data Field elements in Data Dictionary section, or Derived Field
elements in Transformation Dictionary or Local Transformations sections.

![XML schema of Mining Schema section and Mining Field elements](image1)

Figure 12. The XML schema of Mining Schema section and Mining Field elements under Regression Model in O-PMML

The Regression Table (Figure 13) denotes the intercept and the coefficients of
participating variables in the predictive model by elements Numeric Predictor (i.e.,
numeric independent variables), Categorical Predictor (i.e., categorical independent
variables), and Predictor Term (i.e., interactions).

![XML schema of Regression Table section](image2)

Figure 13. The XML schema of Regression Table section under Regression Model
The **Output** section (Figure 14) describes the output values of the model (e.g., estimated risk score, occurrence of event, probability) and the transformation processes required to calculate the final values in **Output Field** elements. The transformation expressions are similar to that of *Transformation Dictionary*.

![XML Schema](image)

**Figure 14.** The XML schema of **Output** section and **Output Field** elements in O-PMML

### 3.6.2. The OMOP-compliant PMML scoring engine

In order to evaluate the performance of the OMOP-compliant PMML standard to estimate risk score of diseases, a scoring engine was designed and developed that extracts the definitions of the embedded risk scoring model, queries the OMOP CDM to retrieve data values, and generates the output values of the model.

The O-PMML scoring engine is a Python module that comprises of three components: O-PMML Parser, SQL Executer, and Scorer (Figure 15). The O-PMML
Parser analyzes the OMOP-compliant PMMLs to extract the specifications of regression model, including participating variables from Data Dictionary and Mining Schema sections, the corresponding coefficients from Regression Table section, data mining scripts from Mining Build Task section, and transformation procedures from Data Transformation, Local Transformations, and Output sections. Then, the SQL Executer uses the SQL scripts of Mining Build Task, asks for input parameters in accordance to the specified Input Parameters, replaces SQL placeholders with the parameter values, and retrieves data fields from the CDM. Finally, the Scorer component applies the predictive model on the retrieved data values from database to generate the output values according to the procedures specified by the O-PMML. When multiple values of a measurement exist for one index date, the Scorer includes the average of all values on that date in calculating the outputs of the algorithm.

Figure 15. A schematic of O-PMML scoring engine

3.6.3. Case study: CVD risk scoring model

In order to evaluate the feasibility and performance of the O-PMML standard, I conducted a case study to use the standard for disseminating and scoring Framingham 10-year risk of CVD model.
3.6.3.1. **Data source**

The data pipelines and O-PMML were tested on the medical records of 250 patients obtained from Regenstrief Institute in HL7 consolidated CCDs. The content of CCDs was transformed to OMOP CDM using the CCD-TO-OMOP as described in Chapter 2.

3.6.3.2. **The Framingham 10-year risk of CVD scoring model**

D’Agostino et al., 2008 (105) generated this prediction algorithm from the original and offspring cohorts of Framingham Heart Study to estimate the 10-year risk of cardiovascular disease in men (Figure 16) and women (Figure 17) aged between 30 and 74 years old. The predictive model estimates the risk based on the patient’s age, serum total cholesterol level (mg/dL), serum high-density lipoprotein (HDL) cholesterol level (mg/dL), systolic blood pressure (mmHg), use of hypertension controlling medications, cigarette smoking status, and diabetes status.

\[
\sum \beta X = 3.06117 \times \ln \text{Age} + 1.12370 \times \ln \text{Total Cholesterol} - 0.93263 \times \ln \text{HDL} + 1.93303 \times \ln \text{SBP not treated} \\
+ 1.99881 \times \ln \text{SBP treated} + 0.65451 \times \text{Smoker} + 0.57367 \times \text{Diabetic}
\]

\[
\text{Risk of CVD in 10 years for men} = 1 - 0.88936^{\exp(\sum \beta X - 23.9802)}
\]

**Figure 16.** Framingham 10-year risk of CVD for men (105)

\[
\sum \beta X = 2.32888 \times \ln \text{Age} + 1.20904 \times \ln \text{Total Cholesterol} - 0.70833 \times \ln \text{HDL} + 2.76157 \times \ln \text{SBP not treated} \\
+ 2.82263 \times \ln \text{SBP treated} + 0.52873 \times \text{Smoker} + 0.69154 \times \text{Diabetic}
\]

\[
\text{Risk of CVD in 10 years for women} = 1 - 0.95012^{\exp(\sum \beta X - 26.1931)}
\]

**Figure 17.** Framingham 10-year risk of CVD for women (105)
3.6.3.3. The specifications of O-PMML containing Framingham algorithms

This study compiled one model per O-PMML document for simple presentation of the schema; thus, separate documents were generated for men and women algorithms of the Framingham risk scoring model as presented in Appendix 9 and Appendix 10. The two algorithms only differ in values of intercept and coefficients of variables.

The Framingham O-PMML was structured in a way that allowed to build a timeline of estimated risk scores. This was achieved by embedding an input parameter in the script to take the index date of interest, and gather values of participating variables in the algorithm according to the specified index date. The input parameter of Person ID also enabled calculating the score for one patient at a time.

The Mining Build Task section lists mining scripts to retrieve input data values from the CDM. The Input Parameters specifies the required entries to fill placeholders within the Statement script. For example, the age script needs Index Date, Person ID, and the database’s schema name where CDM is located before querying database to calculate patient’s age (Figure 18). Table 21 lists the criteria applied to collect the values of Framingham risk score variables.

```xml
<MiningBuildTask>
<Extension name="age" extender="omop">
<InputParameters>
    <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)" optype="continuous" dataType="date"/>
    <InputParameter name="PERSON_ID" displayName="OMOP Person ID" optype="continuous" dataType="bigint"/>
    <InputParameter name="SCHEMA" displayName="Database schema" dataType="string"/>
</InputParameters>
<Statement dialect="postgresql">
    select distinct extract(year from date '@INDEX_DATE') - year_of_birth as AGE from @SCHEMA.person where person_id=@PERSON_ID;
</Statement>
</Extension>
</MiningBuildTask>
```

Figure 18. An excerpt of the Mining Build Task section of Framingham O-PMML
Table 21. The criteria applied to collect values of participating variables in Framingham 10-year risk of cardiovascular disease

<table>
<thead>
<tr>
<th>Data Field Name</th>
<th>Data Field Display Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age</td>
<td>The age of patient at index date</td>
</tr>
<tr>
<td>TCL</td>
<td>Total cholesterol level</td>
<td>Serum total cholesterol level (OMOP Concept ID = 3027114) within 30 days of index date</td>
</tr>
<tr>
<td>HDL</td>
<td>HDL cholesterol level</td>
<td>Serum HDL cholesterol level (OMOP Concept ID = 3007070) within 30 days of index date</td>
</tr>
<tr>
<td>HTNTRT</td>
<td>Antihypertensive medication use</td>
<td>Use of antihypertensive medication within 30 days of index date, including angiotensin-converting-enzyme inhibitors, angiotensin receptor blockers, beta blockers, calcium channel blockers, diuretics, antiadrenergic agents, thiazides, hydrazinophthalazines, oral minoxidil, nitroprusside, pinacidil, tyrosine hydroxylase inhibitors, pargyline, and endothelin receptor antagonists</td>
</tr>
<tr>
<td>SBP</td>
<td>Systolic blood pressure</td>
<td>Systolic blood pressure within 30 days of index date</td>
</tr>
<tr>
<td>smoker</td>
<td>Cigarette smoking status</td>
<td>Report of smoking within 90 days of index date</td>
</tr>
<tr>
<td>diabetic</td>
<td>Diabetes status</td>
<td>Report of diabetic conditions prior to index date OR Report of fasting blood sugar ≥126 mg/dL within 60 days of index date OR Use of insulin or oral antidiabetic medications within 30 days of index date</td>
</tr>
</tbody>
</table>

The Data Dictionary section of the O-PMML specifies eight unprocessed data variables that will participate in the model after undergoing transformation processes described in Transformation Dictionary section. As shown in Figure 19, the inclusion of patients is limited to the age interval between 30 and 74 years old, and possible values of antihypertensive medication use (HTNTRT), smoking status (smoker), and diabetes status (diabetic) data fields are specified.
<DataDictionary numberOfFields="8">
    <DataField name="hazard" displayName="Cumulative Hazard" optype="continuous" dataType="double"/>
    <DataField name="age" displayName="Age" optype="continuous" dataType="double">
        <Interval closure="closedClosed" leftMargin="30" rightMargin="74"/>
    </DataField>
    <DataField name="TCL" displayName="Total Cholesterol" optype="continuous" dataType="double"/>
    <DataField name="HDL" displayName="HDL" optype="continuous" dataType="double"/>
    <DataField name="HTNTRT" displayName="Antihypertensive medication use (y/n)" optype="categorical" dataType="boolean">
        <Value value="1"/>
        <Value value="0"/>
    </DataField>
    <DataField name="SBP" displayName="Systolic Blood Pressure" optype="continuous" dataType="double"/>
    <DataField name="smoker" displayName="Smoker(y/n)" optype="categorical" dataType="integer">
        <Value value="1"/>
        <Value value="0"/>
    </DataField>
    <DataField name="diabetic" displayName="Diabetic(y/n)" optype="categorical" dataType="integer">
        <Value value="1"/>
        <Value value="0"/>
    </DataField>
</DataDictionary>

Figure 19. The Data Dictionary section of Framingham O-PMML

The Transformation Dictionary section specifies the transformation processes on data fields from Data Dictionary. As shown in Figure 20, age, total cholesterol level, HDL level, and systolic blood pressure (SBP) need to undergo natural logarithm transformation. It also uses conditional function to specify criteria to determine whether SBP was measured when the patient was taking antihypertensive medication. The equivalent SQL statements of the conditional processes are:

```sql
    case when HTNTRT=0 then logSBP else 0 end as logSBP_NOTTRT
    case when HTNTRT=1 then logSBP else 0 end as logSBP_TRT
```
The Framingham risk score equation was built based on Cox proportional-hazards regression. The equation (Figure 16 and Figure 17) consists of two parts: the linear regression of predictors that calculates relative hazard, and risk calculation. The two parts were respectively translated in Regression Table and Output sections of the Regression Model (Figure 21). The Mining Filed elements list the participating variables in the regression model from Data Dictionary (smoker, ...

Figure 20. The Transformation Dictionary section of Framingham O-PMML

```xml
<TransformationDictionary>
  <DerivedField name="logAge" dataType="double" optype="continuous">
    <Apply function="ln">
      <FieldRef field="age"/>
    </Apply>
  </DerivedField>
  <DerivedField name="logTCL" dataType="double" optype="continuous">
    <Apply function="ln">
      <FieldRef field="TCL"/>
    </Apply>
  </DerivedField>
  <DerivedField name="logHDL" dataType="double" optype="continuous">
    <Apply function="ln">
      <FieldRef field="HDL"/>
    </Apply>
  </DerivedField>
  <DerivedField name="logSBP" dataType="double" optype="continuous">
    <Apply function="ln">
      <FieldRef field="SBP"/>
    </Apply>
  </DerivedField>
  <DerivedField name="logSBP_NOTTRT" dataType="double" optype="continuous">
    <Apply function="if">
      <Apply function="equal" dataType="boolean">
        <FieldRef field="HTNTRT"/>
        <Constant dataType="integer">1</Constant>
      </Apply>
      <Constant dataType="integer">0</Constant>
      <FieldRef field="logSBP"/>
    </Apply>
  </DerivedField>
  <DerivedField name="logSBP_TRT" dataType="double" optype="continuous">
    <Apply function="if">
      <Apply function="equal" dataType="boolean">
        <FieldRef field="HTNTRT"/>
        <Constant dataType="integer">1</Constant>
      </Apply>
      <FieldRef field="logSBP"/>
      <Constant dataType="integer">0</Constant>
    </Apply>
  </DerivedField>
</TransformationDictionary>
```
diabetic) and Transformation Dictionary (logAge, logTCL, logHDL, logSBP_TRT, logSBP_NOTTRT) fields. The Output Field elements specify the computation processes to calculate ultimate output values (i.e., risk score). The Scorer component only delivers the outputs that are specified as final results (i.e., isFinalResult = "true").

```xml
<RegressionModel modelName="framingham10ycvdmen" functionName="regression" algorithmName="Cox proportional-hazards regression" isScorable="true">
  <MiningSchema>
    <MiningField name="hazard" usageType="predicted"/>
    <MiningField name="logAge" usageType="active"/>
    <MiningField name="logTCL" usageType="active"/>
    <MiningField name="logHDL" usageType="active"/>
    <MiningField name="logSBP_TRT" usageType="active"/>
    <MiningField name="logSBP_NOTTRT" usageType="active"/>
    <MiningField name="smoker" usageType="active"/>
    <MiningField name="diabetic" usageType="active"/>
  </MiningSchema>
  <Output>
    <OutputField name="hazard" optype="continuous" dataType="double" feature="predictedValue" isFinalResult="false"/>
    <OutputField name="hazard_ratio" optype="continuous" dataType="double" feature="transformedValue" isFinalResult="false"/>
      <Apply function="exp">
        <FieldRef field="hazard"/>
      </Apply>
    </OutputField>
    <OutputField name="risk" optype="continuous" dataType="double" feature="transformedValue" isFinalResult="true"/>
      <Apply function="-">
        <Constant>1.0</Constant>
        <Apply function="pow">
          <Constant>0.95012</Constant>
          <FieldRef field="hazard_ratio"/>
        </Apply>
      </Apply>
    </OutputField>
  </Output>
  <RegressionTable intercept="-26.1931">
    <NumericPredictor name="logAge" coefficient="2.32888"/>
    <NumericPredictor name="logTCL" coefficient="1.20904"/>
    <NumericPredictor name="logHDL" coefficient="-0.70833"/>
    <NumericPredictor name="logSBP_TRT" coefficient="2.82263"/>
    <NumericPredictor name="logSBP_NOTTRT" coefficient="2.76157"/>
    <CategoricalPredictor name="smoker" value="1" coefficient="0.52873"/>
    <CategoricalPredictor name="diabetic" value="1" coefficient="0.69154"/>
  </RegressionTable>
</RegressionModel>
```

**Figure 21.** The Regression Model section of Framingham O-PMML
3.6.3.4. Performance assessment of Framingham O-PMML

The scoring engine was programmed to generate 10-year risk score of cardiovascular disease using the develop Framingham O-PMMLs at every visit occurrences (i.e., index date). The risk scores were calculated only if all variables were available at the index date. The estimated scores were manually assessed later for accuracy.

3.7. Results

3.7.1. Patient characteristics

Out of 250 studied patients, 226 (82 men and 144 women) were included in the final analysis as 24 patients did not have any information about date of birth, or visit occurrence. The analysis of visit occurrence records of 226 patients yielded 4,472 index dates. The initial pool of patient records aged from 0 to 87 years old, but the risk score was generated for the records between 30 and 67 years old. Table 22 shows detail information about the characteristics of included patient records.

Table 22. The characteristics of patient records for which Framingham 10-year risk score of cardiovascular disease was generated

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Analyzed Patient Records (N = 4,472)</th>
<th>Records with Risk Scores (N = 56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>1,154 (25.8)</td>
<td>11 (19.6)</td>
</tr>
<tr>
<td>Women</td>
<td>3,318 (74.2)</td>
<td>45 (80.4)</td>
</tr>
<tr>
<td>Age, range</td>
<td>0 – 87 years</td>
<td>30 – 67 years</td>
</tr>
<tr>
<td>Total cholesterol level, range</td>
<td>87 – 298 mg/dL</td>
<td>87 – 219 mg/dL</td>
</tr>
<tr>
<td>HDL cholesterol level, range</td>
<td>25 – 65 mg/dL</td>
<td>25 – 47.5 mg/dL</td>
</tr>
<tr>
<td>Systolic blood pressure, range</td>
<td>106 – 162 mmHg</td>
<td>106 – 162 mmHg</td>
</tr>
<tr>
<td>Antihypertensive medications user, n (%)</td>
<td>537 (12.0)</td>
<td>31 (55.4)</td>
</tr>
<tr>
<td>Current smoker, n (%)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Diabetic, n (%)</td>
<td>709 (15.9)</td>
<td>4 (7.1)</td>
</tr>
</tbody>
</table>
3.7.2. Risk scores

A total of 56 risk scores were calculated for 8 unique patients. These patients had the full set of required data values to generate the risk score. Most of the analyzed records lacked one or more data values; thus, they were excluded from the scoring step. The risk scores ranged from 1.11% to 36.37%. The lowest calculated risk score belonged to a 30-year old, non-smoker, and non-diabetic woman with total cholesterol level of 184 mg/dL, HDL of 31 mg/dL, no antihypertensive medication use, and SBP of 106 mmHg. The patient with the highest risk score was diabetic, non-smoker, 64-year old man with total cholesterol level of 87 mg/dL, HDL of 25 mg/dL, taking antihypertensive medications, and SBP of 149.5 mmHg. All generated risk scores were valid after inspected manually. The full list of estimated risk scores is available in Appendix 11.

The architecture of O-PMML also allowed to capture the timeline of risk scores. Figure 22 depicts the timeline of risk score for the 8 patients who had the required values of algorithm’s variables at the index dates.

![Figure 22](image_url)

**Figure 22.** The timeline of estimated 10-year risk score of cardiovascular disease of patients that had full set of required values to generate scores.
3.8. Discussion

3.8.1. An overview of O-PMML

This study demonstrated that the adopted O-PMML can be used for delivering estimations and predictions about health-related events of patients to care providers to support clinical-decision making and precision medicine. The adoption process intended to keep the original structure of PMML intact, and only involved a new structure for the Mining Build Task section; thus, the O-PMML will work with the already-developed PMML parsers. However, minor modifications in the parsers are needed to process the Mining Build Task section that contains scripts to properly mine OMOP CDM. The study also found the O-PMML as a feasible tool to deploy models on OMOP CDM, and score the algorithms through the developed scoring engine.

The O-PMML was designed to satisfy the five requirements of sharing OMOP-compatible predictive models, metadata, model’s specifications, database mining pipelines, data transformation procedures, and model’s output. However, the PMML allows sharing further capabilities to evaluate the quality of models, share a sample of training dataset for cross-validation, and specify the properties of output values using optional sections Model Stats, Model Explanation, Targets, and Model Verification. This paper did not discuss these sections in detail, but their compliance with OMOP CDM can be subjects of new research in future.

3.8.2. Advantages of O-PMML

The O-PMML standard in conjunction with the developed scoring engine offer the capability to “plug-and-play” predictive models on OMOP-formatted medical data repositories. There are numerous reports of individualized predictive models using the EHR in the literature that are supported by PMML structure, such as logistic regression to predict prognosis of health outcomes (38-40, 42), Cox proportional hazards model for estimating opioid dose-related risk of injuries in older adults (43) and survival analysis (44), linear regression for exploring outcome predictors (45), and personalizing medicine dosage (46) and risk estimations (39), and random forest for individualized medicine doses (37, 47). Researchers have also examined several machine learning methods to improve the accuracy and generalizability of the developed predictive models, such as support vector machine (SVM) for early detection of myocardial infarction (47) and mortality risk of radical
cystectomy (48), Markov Decision Process for predicting mortality and length of hospitalization in septic patients (49), k-nearest neighbor for warfarin dosage estimation (50), naïve Bayes network for cardiovascular disease risk (51), classification and regression tree (CART) for heart failure patients’ readmission risk using EHR data (52). Ultimately, this standard will help develop interoperable personalized scoring systems offering real-time, personalized predictions to clinicians about patients at the point-of-care. It can also help test the performance of predictive models across databases through prediction-as-a-service (106) with minimum implementation efforts needed.

In addition, the new structure of Mining Build Task section within O-PMML grants more flexibility to the O-PMML to score predictive models at certain index dates. Therefore, the O-PMMLs can not only share the specifications of predictive model, but also allow running the scoring algorithms multiple times on patient data to deliver the trend of measures throughout time. One example use case is to monitor drug adherence among patients through episodes of care, and explore the effectiveness of interventions on improving medication use. In this case, applying an O-PMML containing an algorithm to estimate drug adherence on patient medical records will deliver series of scores at intervention time points.

3.8.3. Limitations

This study only focused on sharing a regression model using the O-PMML to deliver predictions about health-related events. The standard needs to be tested for other types of predictive models such as logistic regression, decision tree, and random forest that are widely used in clinical research. The performance assessment of O-PMML was also limited to small sample of patients. The standard needs to be tested on larger, multi-center sample of patients to ensure the accuracy of processing pipelines, and the feasibility of implementing O-PMML across systems.
CHAPTER 4. THE INTEROPERABLE SYSTEM FOR DELIVERING PERSONALIZED HEALTH OUTCOME PREDICTIONS

4.1. Background

The literature has many examples of predictive modeling developed to support clinical decision-making through offering estimations about the health status of patients. Typical examples of predictive models include predicting prognostic risk of health events (27-29), disease screening (30-32), and managing short- and long-term complications (33-36).

Although numerous EHR-based predictive models have been reported in the literature, there are only few reports of deployment of health outcome predictive models, and no report exist that applies or evaluates the models across multiple data warehouses for clinical effectiveness and cost efficiency. For example, Hu et al., 2015 (4) reported an online application for predicting the next 6-month healthcare resource utilization by chronic disease patients. The prediction system could make real-time risk assessment using a health information exchange (HIE) electronic health records (EHR) warehouse; however, it was not evaluated in other HIE networks.

In another effort in Canada, Khazaei et al., 2015 (5) proposed a cloud-based Analytics-as-a-Service framework for real-time patient monitoring in the clinical edition and retrospective health analytics in the research edition across multiple EHR systems. They deployed an algorithm to identify septic neonates in an intensive care unit; however, the report did not provide any information about data exchange standards or data transformation processes (e.g., inter-vocabulary mapping of concept code). Although no predictive model was deployed, this study sheds light on novel approaches of integrating advanced analytics in healthcare system. Toerper et al., 2015 (6) also developed a web-based application that predicts daily admission bed needs based on EHR data to improve patient flow management. Despite providing a real-time tool for monitoring and forecasting patient flow in a hospital setting, the proposed system is not interoperable to work across other centers to retrieve new incoming patients’ data for better prediction performance.
The existing healthcare IT solutions have not fully leveraged forecasting capabilities of predictive models in medical practice to provide clinicians with personalized recommendations about patients’ health status. Testing the quality of predictions is also challenging as deployment of predictive models required extensive efforts and teams of analysts, statisticians, IT professionals, and computer programmer. This project intends to fill the gap via designing, evaluating, implementing an interoperable solution that receives patient electronic health records via Health Level Seven (HL7) messaging standard, transforms the records to a common data model for population health research, and applies risk prediction models received in PMML format on patient data to make predictions about health outcomes, and delivers the results to healthcare professionals.

4.2. Objectives

This paper describes an interoperable framework for delivering real-time, personalized predictions about patient health status by applying predictive models on exchanged patient information. To proof the concept, Personalized Health Risk Scoring Tool (PHRST), a web-based, interoperable OMOP-based model scoring tool was developed that receives patient medical records from OMOP CDM repository, applies risk score models obtained in a standard PMML-based format on patient information, and delivers personalized risk scores to the end user who can be a healthcare professional.

4.3. Methods

4.3.1. The personalized health outcome prediction framework

Figure 23 shows the proposed interoperable framework that delivers personalized health outcome predictions through applying predictive models on patient medical records. The framework is designed to be interoperable in two directions, patient’s data input and predictive model input. It receives the medical records in HL7 C-CDA-based CCDs, and accepts trained predictive models in OMOP-compliant PMML. The system is capable of performing predictions through a wide range of OMOP-compliant machine learning models, such as regression (general linear, multinomial logistic, ordinal multinomial, generalized linear, Cox regression), decision tree, support vector machine, and random forest.

Figure 24 presents the data flow in the personalized health outcome prediction framework. Once the system receives patient health information enclosed
in HL7 CCD, a parser developed in the previous study called CCD-TO-OMOP parser (Chapter 2) extracts data from the document, transform the data to OMOP CDM, and stores in a PostgreSQL database. The module is equipped with an Extract-Load-Transform (ETL) processor that maps diagnoses, laboratory tests, drugs, demographics, observations, allergies, and other data element requirements into OMOP CDM version 5.1 (Figure 2). The resulting transformed data are passed to an PMML scoring engine developed in a previous study (Chapter 3) to estimate outcome predictions. Both PMMLs and the scoring engine are compliant with OMOP CDM.

The engine receives O-PMML containing predictive models from an internal or externa repository, and applies the model on retrieved patient information from the CDM. The framework allows to run multiple predictive models received in O-PMML documents from an O-PMML Repository on many patients’ medical records at once. As a final step, the prediction output can be send back to the client in a message or displayed on a user interface.

Figure 23. A schematic of the interoperable framework for delivering personalized health outcome predictions
4.3.2. Architecture and dataflow of PHRST

To show case the concept of interoperable personalized health outcome prediction framework, a web-based application was developed based on the described framework to assess compatibility of components to deliver outputs of predictive models. The Personalized Health Risk Scoring Tool (PHRST) solution (pronounced First) composes of four components (Figure 24): A data warehouse that contains OMOP-transformed patient medical records (Chapter 2); the O-PMML scoring engine (Chapter 3) that generates predictions by applying predictive models from OMOP-compliant PMMLs on patient information from the database; a user interface that displays the prediction outputs to the user; and a representational state transfer (REST) application programming interface (API) that connects the user interface to the database and scoring engine. All responses from the server are JavaScript Object Notation (JSON) documents.

On load, the PHRST shows a list of patients stored in the OMOP CDM to the user. It also pulls the list of available predictive models on the server. Once a patient is selected, the system proposes the list of models to user for selection. The user may choose to run one or multiple predictive models on the selected patients. Upon clicking on the start button, a request is sent to the scoring engine on server through the API to parse the corresponding O-PMML document and compute outputs of the selected models. Then, prediction outputs are transferred through the API in JSON documents back to user interface to display in tables. The solution can deliver both cross-sectional and longitudinal estimations of the health outcome, depending on the design of O-PMMLs.
Figure 24. A diagram of the architecture and data flow of personalized health outcome prediction framework
4.3.3. Case study: Framingham risk functions in action

This case study intended to deploy Framingham risk functions in plug-and-play manner on patient medical records using PHRST solution, and deliver contemporaneous risk predictions based on their medical history.

4.3.3.1. Data source

The proof-of-concept study involved medical records of 250 patients obtained from Regenstrief Institute in HL7 consolidated CCDs. The content of CCDs was transformed to OMOP CDM using the CCD-TO-OMOP as described in Chapter 2.

4.3.3.2. Predictive models

Framingham Heart Study has generated series of robust risk predictive models estimating various health outcomes to help clinicians better assess the risk of diseases and assist them in evidence-based clinical decision-making process. The algorithms include risk estimation of atrial fibrillation (107, 108), cardiovascular disease(105, 109), congestive heart failure (110), coronary heart disease (111-113), diabetes (114), hypertension (115), intermittent claudication (116), and stroke (117, 118). This case study tested deployment of 10-year risk of cardiovascular disease model on PHRST using the designed O-PMML in Chapter 3 (Appendix 9 and Appendix 10).

4.4. Results

4.4.1. PHRST development

The single-page user interface of PHRST web application was developed in HTML5 and CSS3, and APIs were Java web services. The solution offers risk predictions of diseases personalized to each patient, allowing users to select one patient at a time and order one or multiple risk predictions. The application alerts the user when the risk estimates are available, an error has occurred, or not sufficient data exist to compute risk scores (Figure 25). The user can also see more details of granular data values when selects a generated risk score (Figure 26).
**Figure 25.** The view of PHRST application that allows users to select patients and order risk score estimates. The names of patients are fictitious.

**Figure 26.** The view of PHRST application that displays more details of the selected patient’s risk factor data and the estimated risk score. The names of patients are fictitious.
4.4.2. Deployment of Framingham risk functions on PHRST

The solution successfully called O-PMMLs on the server and generated risk estimates. Depending on the size of available patients’ medical records in the database, the processing time took from seconds to under one minute. The specifications of O-PMMLs were cached upon first call, then used in the subsequent requests; thus, the processing reduced after the first model was processed.

The solution is designed and programmed in a way that plugs new algorithms into play once O-PMMLs reach the server. Upon arrival in the server via a message or simply copied to the disc, PHRST lists the models on the panel for user selection. Each model is assigned a unique ID which is called upon a request to compute the model’s output.

4.5. Discussion

4.5.1. An overview

The PHRST solution was developed based on the described conceptual framework of personalized health outcome predictions to allow delivering risk estimates of patient health outcome to clinicians, and ease deployment of new predictive models in clinical setting. New predictive algorithms can easily plug into the solution when exchanged in O-PMML format to the server, and will be ready-to-use on OMOP CDM medical records with no further implementation efforts. The user interface of PHRST not only displays the estimated risk score at a certain point of time, but also present a timeline of estimations in the past. This can be very useful when monitoring patients for treatment efficacy and medication adherence.

4.5.2. Advantages of the framework

Interoperability is the main benefit of the framework. The system not only can work with other electronic medical record systems to receive patient information, but also can “plug-and-play” predictive models from third parties immediately. The solution can be implemented as a cloud-based system external to the EHR and the recipient of prediction outputs. This feature promotes the framework as a suitable solution for prediction-as-a-service (106).

Another advantage of this framework is that the system is database-structure agnostic, meaning that accepts patient medical records from other EHR systems independent of the architecture and coding systems. The CCD-TO-OMOP parser transforms medical records to OMOP CDM that unifies most of commonly
used coding systems to standard concepts. This feature is crucial when the solution is hosted on a cloud server and receives heterogeneous medical records from diverse, disparate databases to deliver health outcome predictions.

4.5.3. Similar studies

There are very few reports of similar frameworks in the literature, and no study was found on dual-interoperable solution that supports OMOP CDM to deliver personalized health outcome estimations. There is a sequel of reports from a team on an interoperable system designed to exchange predictive models in PMML, and transfer patient data in HL7 standard in clinical decision support systems (119-121). Although the reported frameworks were similarly designed to deliver output of model scoring operations on EHR data, they still needed custom-coding to adopt diversity of data structure across data repositories (120). This is a critical challenge for a cloud-based solution intended to process heterogeneous EHR data that our study addressed successfully.

4.5.4. Limitations

The framework was tested on a small set of patient data and limited number of regression models. To ensure high performance and accurate prediction delivery, the framework needs to be assessed on larger multi-center patient datasets with diverse formats, and other commonly used machine learning algorithms in healthcare research and practice.
CHAPTER 5. CONCLUSION

5.1. OMOP CDM Accommodates HL7 Consolidated CCD Data

The OMOP CDM demonstrated the capability to accommodate concepts and data elements of HL7 CCD documents with high accuracy. The CDM allows researchers to analyze patients’ medical information similarly across discrete EHR systems when transferred via HL7 CCD messaging standard. The performance of the developed data transformation pipeline needs to be validated on a larger pool of CCD files from diverse providers.

5.2. A New Standard Enables Sharing Health Risk Prediction Models

The O-PMML, a customized version of PMML could disseminate predictive models to operate on OMOP CDM and generate predictions about health outcome. This is very important to not only clinical research, but also evidence-based clinical practice. Using this standard, we can share and deploy newly developed predictive models across databases to evaluate the performance of algorithms in order to improve the quality of predictions. It also allows to deliver the predictions about patient health outcomes at the-point-care to support clinical decision-making.

5.3. An Interoperable System Delivers Personalized Health Outcome Predictions

The PHRST solution proved the conceptual interoperable framework to deliver personalized health outcome predictions by applying predictive models on patient data on-the-fly. The proposed solution can handle heterogenous EHR data coded in diverse terminologies since patient information is transferred to OMOP CDM before undergoing computation process. This is one of the central benefits of using this framework in a multi-center setting.

5.4. Future Work

The O-PMML standard was designed to facilitate delivering estimations about health-related events, such as risk score and drug adherence indexes to support clinical decision-making, improving quality of healthcare service, and ultimately and patient’s quality of life. In future work, we need to evaluate the impact of O-PMML standard on these topics.
The user interface of PHRST is now in the proof-of-concept stage, and there are high potentials for further developments. It is imperative to test other types of machine learning algorithms and risk prediction models across multiple center databases to ensure integrity of the system and quality of predictions.

Finally, privacy and security of the developed solutions was out of scope of this project; however, we need to address confidentiality, privacy, technical, implementation cost challenges (122, 123) in order to achieve a real-time predictive analytics system that delivers patient-specific predictions. The HIPAA privacy and security rules require covered entities including providers, hospitals, insurers, and their business associates to protect the privacy and confidentiality of individually identifiable health information (124). The entities are always reluctant to send out patient data to other parties unless they ensure that privacy and security safeguards are prepared; thus, these topics deserve high attention in future investigations to make sure existing security and privacy protocol for exchanging and maintaining IT solutions properly apply to the solutions.
### APPENDICES

**Appendix 1.** Template IDs of C-CDA Continuity of Care Document (CCD) Release 1.1 templates used by CCD parser to locate entries

<table>
<thead>
<tr>
<th>Template Name</th>
<th>Template Type</th>
<th>Template ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Realm</td>
<td>Header</td>
<td>2.16.840.1.113883.10.20.22.1.1</td>
</tr>
<tr>
<td>Allergies Section (entries required)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.6.1</td>
</tr>
<tr>
<td>Allergy Problem Act</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.30</td>
</tr>
<tr>
<td>Allergy-Intolerance Observation</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.7</td>
</tr>
<tr>
<td>Encounters Section (entries optional)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.22</td>
</tr>
<tr>
<td>Encounter Activity</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.1.21</td>
</tr>
<tr>
<td>Immunizations Section (entries optional)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.2</td>
</tr>
<tr>
<td>Immunization Activity</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.52</td>
</tr>
<tr>
<td>Immunization Medication Information</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.54</td>
</tr>
<tr>
<td>Medications Section (entries required)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.1.1</td>
</tr>
<tr>
<td>Medication Activity</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.16</td>
</tr>
<tr>
<td>Medication Information</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.23</td>
</tr>
<tr>
<td>Supply Activity</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.1.34</td>
</tr>
<tr>
<td>Problem Section (entries required)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.5.1</td>
</tr>
<tr>
<td>Problem Concern Act</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.3</td>
</tr>
<tr>
<td>Problem Observation</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.4</td>
</tr>
<tr>
<td>Procedures Section (entries required)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.7.1</td>
</tr>
<tr>
<td>Procedure Activity Act</td>
<td></td>
<td>2.16.840.1.113883.10.20.22.4.12</td>
</tr>
<tr>
<td>Procedure Activity Observation</td>
<td></td>
<td>2.16.840.1.113883.10.20.22.4.13</td>
</tr>
<tr>
<td>Procedure Activity Procedure</td>
<td></td>
<td>2.16.840.1.113883.10.20.22.4.14</td>
</tr>
<tr>
<td>Template Name</td>
<td>Template Type</td>
<td>Template ID</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>---------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Results Section (entries required)</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.3.1</td>
</tr>
<tr>
<td>Result Organizer</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.1</td>
</tr>
<tr>
<td>Result Observation</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.2</td>
</tr>
<tr>
<td>Social History Section (entries</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.17</td>
</tr>
<tr>
<td>optional)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking Status Observation</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.78</td>
</tr>
<tr>
<td>Vital Signs Section (entries</td>
<td>Section</td>
<td>2.16.840.1.113883.10.20.22.2.4</td>
</tr>
<tr>
<td>optional)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vital Signs Organizer</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.22.4.26</td>
</tr>
<tr>
<td>Vital Sign Observation</td>
<td>Entry</td>
<td>2.16.840.1.113883.10.20.1.31</td>
</tr>
</tbody>
</table>
Appendix 2. SQL query to map source codes to OMOP source concepts

```sql
SELECT C.CONCEPT_CODE AS SOURCE_CODE, C.CONCEPT_ID AS SOURCE_CONCEPT_ID,
       C.VOCABULARY_ID AS SOURCE_VOCABULARY_ID, C.DOMAIN_ID AS SOURCE_DOMAIN_ID,
       C.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID, C.INVALID_REASON AS SOURCE_INVALID_REASON,
       TARGET_CONCEPT_ID AS SOURCE_CONCEPT_ID
FROM CONCEPT C
UNION
SELECT SOURCE_CODE, SOURCE_CONCEPT_ID, SOURCE_VOCABULARY_ID, C1.DOMAIN_ID AS
       SOURCE_DOMAIN_ID, C1.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
       C1.INVALID_REASON AS SOURCE_INVALID_REASON, TARGET_CONCEPT_ID
FROM SOURCE_TO_CONCEPT_MAP STCM
LEFT OUTER JOIN CONCEPT C1
ON C1.CONCEPT_ID = STCM.SOURCE_CONCEPT_ID
LEFT OUTER JOIN CONCEPT C2
ON C2.CONCEPT_ID = STCM.TARGET_CONCEPT_ID
WHERE STCM.INVALID_REASON IS NULL
```
Appendix 3. SQL query to map source codes to OMOP standard concepts

```sql
SELECT C.CONCEPT_CODE AS SOURCE_CODE, C.CONCEPT_ID AS SOURCE_CONCEPT_ID,
C.VOCABULARY_ID AS SOURCE_VOCABULARY_ID, C.DOMAIN_ID AS SOURCE_DOMAIN_ID,
C.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID, C.INVALID_REASON AS
SOURCE_INVALID_REASON, C1.CONCEPT_ID AS TARGET_CONCEPT_ID, C1.VOCABULARY_ID
AS TARGET_VOCABULARY_ID, C1.DOMAIN_ID AS TARGET_DOMAIN_ID,
C1.CONCEPT_CLASS_ID AS TARGET_CONCEPT_CLASS_ID, C1.INVALID_REASON AS
TARGET_INVALID_REASON, C1.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
FROM CONCEPT C
JOIN CONCEPT_RELATIONSHIP CR
    ON C.CONCEPT_ID = CR.CONCEPT_ID_1
       AND CR.INVALID_REASON IS NULL
       AND CR.RELATIONSHIP_ID = 'Maps to'
JOIN CONCEPT C1
    ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID
       AND C1.INVALID_REASON IS NULL
UNION
SELECT SOURCE_CODE, SOURCE_CONCEPT_ID, SOURCE_VOCABULARY_ID, C1.DOMAIN_ID AS
SOURCE_DOMAIN_ID, C2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
STCM.INVALID_REASON AS SOURCE_INVALID_REASON, TARGET_CONCEPT_ID,
TARGET_VOCABULARY_ID, C2.DOMAIN_ID AS TARGET_DOMAIN_ID, C2.CONCEPT_CLASS_ID
AS TARGET_CONCEPT_CLASS_ID, C2.INVALID_REASON AS TARGET_INVALID_REASON,
C2.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
FROM SOURCE_TO_CONCEPT_MAP STCM
LEFT OUTER JOIN CONCEPT C1
    ON C1.CONCEPT_ID = STCM.SOURCE_CONCEPT_ID
LEFT OUTER JOIN CONCEPT C2
    ON C2.CONCEPT_ID = STCM.TARGET_CONCEPT_ID
WHERE STCM.INVALID_REASON IS NULL
```
### Appendix 4. Mapped CVX codes to OMOP standard Concept IDs

<table>
<thead>
<tr>
<th>CVX Code</th>
<th>CPT Code</th>
<th>OMOP Concept ID</th>
<th>CVX Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>90701</td>
<td>No Concept ID</td>
<td>DTP</td>
</tr>
<tr>
<td>02</td>
<td>90712</td>
<td>2213470</td>
<td>OPV</td>
</tr>
<tr>
<td>03</td>
<td>90707</td>
<td>2213466</td>
<td>MMR</td>
</tr>
<tr>
<td>04</td>
<td>90708</td>
<td>2213467</td>
<td>M/R</td>
</tr>
<tr>
<td>05</td>
<td>90705</td>
<td>2213464</td>
<td>measles</td>
</tr>
<tr>
<td>06</td>
<td>90706</td>
<td>2213465</td>
<td>rubella</td>
</tr>
<tr>
<td>07</td>
<td>90704</td>
<td>2213463</td>
<td>mumps</td>
</tr>
<tr>
<td>08</td>
<td>90744</td>
<td>2213491</td>
<td>Hep B, adolescent or pediatric</td>
</tr>
<tr>
<td>09</td>
<td>90714</td>
<td>2213472</td>
<td>Td (adult), adsorbed</td>
</tr>
<tr>
<td>09</td>
<td>90718</td>
<td>No Concept ID</td>
<td>Td (adult), adsorbed</td>
</tr>
<tr>
<td>10</td>
<td>90713</td>
<td>2213471</td>
<td>IPV</td>
</tr>
<tr>
<td>12</td>
<td>90296</td>
<td>2213401</td>
<td>diphtheria antitoxin</td>
</tr>
<tr>
<td>13</td>
<td>90389</td>
<td>2213410</td>
<td>TIG</td>
</tr>
<tr>
<td>14</td>
<td>90741</td>
<td>No Concept ID</td>
<td>IG, unspecified formulation</td>
</tr>
<tr>
<td>15</td>
<td>No Code</td>
<td>No Concept ID</td>
<td>influenza, split (incl. purified surface antigen)</td>
</tr>
<tr>
<td>16</td>
<td>90659</td>
<td>No Concept ID</td>
<td>influenza, whole</td>
</tr>
<tr>
<td>17</td>
<td>90737</td>
<td>No Concept ID</td>
<td>Hib, unspecified formulation</td>
</tr>
<tr>
<td>18</td>
<td>90675</td>
<td>2213449</td>
<td>rabies, intramuscular injection</td>
</tr>
<tr>
<td>19</td>
<td>90585</td>
<td>2213425</td>
<td>BCG</td>
</tr>
<tr>
<td>19</td>
<td>90728</td>
<td>No Concept ID</td>
<td>BCG</td>
</tr>
<tr>
<td>20</td>
<td>90700</td>
<td>2213459</td>
<td>DTaP</td>
</tr>
<tr>
<td>21</td>
<td>90716</td>
<td>2213474</td>
<td>varicella</td>
</tr>
<tr>
<td>22</td>
<td>90720</td>
<td>2213478</td>
<td>DTP-Hib</td>
</tr>
<tr>
<td>23</td>
<td>90727</td>
<td>2213482</td>
<td>plague</td>
</tr>
<tr>
<td>24</td>
<td>90581</td>
<td>2213424</td>
<td>anthrax</td>
</tr>
<tr>
<td>25</td>
<td>90690</td>
<td>2213453</td>
<td>typhoid, oral</td>
</tr>
<tr>
<td>26</td>
<td>90725</td>
<td>2213481</td>
<td>cholera, unspecified formulation</td>
</tr>
<tr>
<td>27</td>
<td>90287</td>
<td>2213398</td>
<td>botulinum antitoxin</td>
</tr>
<tr>
<td>28</td>
<td>90702</td>
<td>2213461</td>
<td>DT (pediatric)</td>
</tr>
<tr>
<td>29</td>
<td>90291</td>
<td>2213400</td>
<td>CMVIG</td>
</tr>
<tr>
<td>30</td>
<td>90371</td>
<td>2213402</td>
<td>HBIG</td>
</tr>
<tr>
<td>31</td>
<td>No Code</td>
<td>No Concept ID</td>
<td>HepA Pediatric (Unspecified, historical)</td>
</tr>
<tr>
<td>CVX Code</td>
<td>CPT Code</td>
<td>OMOP Concept ID</td>
<td>CVX Short Description</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>-----------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>32</td>
<td>90733</td>
<td>2213484</td>
<td>meningococcal MPSV4</td>
</tr>
<tr>
<td>33</td>
<td>90732</td>
<td>2213483</td>
<td>pneumococcal polysaccharide PPV23</td>
</tr>
<tr>
<td>34</td>
<td>90375</td>
<td>2213403</td>
<td>RIG</td>
</tr>
<tr>
<td>34</td>
<td>90376</td>
<td>2213404</td>
<td>RIG</td>
</tr>
<tr>
<td>35</td>
<td>90703</td>
<td>2213462</td>
<td>tetanus toxoid, adsorbed</td>
</tr>
<tr>
<td>36</td>
<td>90396</td>
<td>2213412</td>
<td>VZIG</td>
</tr>
<tr>
<td>37</td>
<td>90717</td>
<td>2213475</td>
<td>yellow fever</td>
</tr>
<tr>
<td>39</td>
<td>90735</td>
<td>2213486</td>
<td>Japanese encephalitis SC</td>
</tr>
<tr>
<td>40</td>
<td>90676</td>
<td>2213450</td>
<td>rabies, intradermal injection</td>
</tr>
<tr>
<td>41</td>
<td>90692</td>
<td>2213455</td>
<td>typhoid, parenteral</td>
</tr>
<tr>
<td>42</td>
<td>90745</td>
<td>No Concept ID</td>
<td>Hep B, adolescent/high risk infant</td>
</tr>
<tr>
<td>43</td>
<td>90739</td>
<td>43527982</td>
<td>Hep B, adult</td>
</tr>
<tr>
<td>43</td>
<td>90743</td>
<td>2213490</td>
<td>Hep B, adult</td>
</tr>
<tr>
<td>43</td>
<td>90746</td>
<td>2213492</td>
<td>Hep B, adult</td>
</tr>
<tr>
<td>44</td>
<td>90740</td>
<td>2213489</td>
<td>Hep B, dialysis</td>
</tr>
<tr>
<td>44</td>
<td>90747</td>
<td>2213493</td>
<td>Hep B, dialysis</td>
</tr>
<tr>
<td>45</td>
<td>90731</td>
<td>No Concept ID</td>
<td>Hep B, unspecified formulation</td>
</tr>
<tr>
<td>46</td>
<td>90646</td>
<td>2213432</td>
<td>Hib (PRP-D)</td>
</tr>
<tr>
<td>47</td>
<td>90645</td>
<td>2213431</td>
<td>Hib (HbOC)</td>
</tr>
<tr>
<td>48</td>
<td>90648</td>
<td>2213434</td>
<td>Hib (PRP-T)</td>
</tr>
<tr>
<td>49</td>
<td>90647</td>
<td>2213433</td>
<td>Hib (PRP-OMP)</td>
</tr>
<tr>
<td>50</td>
<td>90721</td>
<td>2213479</td>
<td>DTaP-Hib</td>
</tr>
<tr>
<td>51</td>
<td>90748</td>
<td>2213494</td>
<td>Hib-Hep B</td>
</tr>
<tr>
<td>52</td>
<td>90632</td>
<td>2213427</td>
<td>Hep A, adult</td>
</tr>
<tr>
<td>53</td>
<td>90693</td>
<td>2213456</td>
<td>typhoid, parenteral, AKD (U.S. military)</td>
</tr>
<tr>
<td>54</td>
<td>90476</td>
<td>2213422</td>
<td>adenovirus, type 4</td>
</tr>
<tr>
<td>55</td>
<td>90477</td>
<td>2213423</td>
<td>adenovirus, type 7</td>
</tr>
<tr>
<td>62</td>
<td>90649</td>
<td>2213435</td>
<td>HPV, quadrivalent</td>
</tr>
<tr>
<td>66</td>
<td>90665</td>
<td>No Concept ID</td>
<td>Lyme disease</td>
</tr>
<tr>
<td>71</td>
<td>90379</td>
<td>No Concept ID</td>
<td>RSV-IGIV</td>
</tr>
<tr>
<td>79</td>
<td>90393</td>
<td>2213411</td>
<td>vaccinia immune globulin</td>
</tr>
<tr>
<td>83</td>
<td>90633</td>
<td>2213428</td>
<td>Hep A, ped/adol, 2 dose</td>
</tr>
<tr>
<td>84</td>
<td>90634</td>
<td>2213429</td>
<td>Hep A, ped/adol, 3 dose</td>
</tr>
<tr>
<td>85</td>
<td>90730</td>
<td>No Concept ID</td>
<td>Hep A, unspecified formulation</td>
</tr>
<tr>
<td>CVX Code</td>
<td>CVX Short Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>IG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>IGIV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>influenza, unspecified formulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>rabies, unspecified formulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>typhoid, unspecified formulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>RSV-MAb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>MMRV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>pneumococcal conjugate PCV 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>typhoid, ViCPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>Hep A-Hep B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>DTaP, 5 pertussis antigens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>DTaP-Hep B-IPV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>influenza, live, intranasal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>Td (adult) preservative free</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>meningococcal MCV4P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>Tdap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>rotavirus, pentavalent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>HPV, bivalent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>rotavirus, monovalent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>DTaP-Hib-IPV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>zoster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>Novel Influenza-H1N1-09, nasal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>Novel influenza-H1N1-09, preservative-free</td>
<td></td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>Novel influenza-H1N1-09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Novel Influenza-H1N1-09, all formulations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>DTaP-IPV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>133</td>
<td>Pneumococcal conjugate PCV 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>134</td>
<td>Japanese Encephalitis IM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>Influenza, high dose seasonal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>136</td>
<td>Meningococcal MCV4O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>Influenza, seasonal, injectable, preservative free</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVX Code</td>
<td>CPT Code</td>
<td>OMOP Concept ID</td>
<td>CVX Short Description</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>-----------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>140</td>
<td>90656</td>
<td>2213438</td>
<td>Influenza, seasonal, injectable, preservative free</td>
</tr>
<tr>
<td>141</td>
<td>90657</td>
<td>2213439</td>
<td>Influenza, seasonal, injectable</td>
</tr>
<tr>
<td>141</td>
<td>90658</td>
<td>2213440</td>
<td>Influenza, seasonal, injectable</td>
</tr>
<tr>
<td>144</td>
<td>90654</td>
<td>42742499</td>
<td>influenza, seasonal, intradermal, preservative free</td>
</tr>
<tr>
<td>146</td>
<td>90697</td>
<td>No Concept ID</td>
<td>DTaP,IPV,Hib,HepB</td>
</tr>
<tr>
<td>148</td>
<td>90644</td>
<td>40757102</td>
<td>Meningococcal C/Y-HIB PRP</td>
</tr>
<tr>
<td>149</td>
<td>90672</td>
<td>43527981</td>
<td>influenza, live, intranasal, quadrivalent</td>
</tr>
<tr>
<td>150</td>
<td>90686</td>
<td>44816520</td>
<td>influenza, injectable, quadrivalent, preservative free</td>
</tr>
<tr>
<td>153</td>
<td>90661</td>
<td>2213443</td>
<td>Influenza, injectable, MDCK, preservative free</td>
</tr>
<tr>
<td>155</td>
<td>90673</td>
<td>44816443</td>
<td>influenza, recombinant, injectable, preservative free</td>
</tr>
<tr>
<td>158</td>
<td>90687</td>
<td>44816519</td>
<td>influenza, injectable, quadrivalent</td>
</tr>
<tr>
<td>158</td>
<td>90688</td>
<td>44816518</td>
<td>influenza, injectable, quadrivalent</td>
</tr>
<tr>
<td>161</td>
<td>90685</td>
<td>44816521</td>
<td>Influenza, injectable, quadrivalent, preservative free, pediatric</td>
</tr>
<tr>
<td>162</td>
<td>90621</td>
<td>No Concept ID</td>
<td>meningococcal B, recombinant</td>
</tr>
<tr>
<td>163</td>
<td>90620</td>
<td>No Concept ID</td>
<td>meningococcal B, OMV</td>
</tr>
<tr>
<td>165</td>
<td>90651</td>
<td>46257428</td>
<td>HPV9</td>
</tr>
<tr>
<td>166</td>
<td>90630</td>
<td>46257714</td>
<td>influenza, intradermal, quadrivalent, preservative free</td>
</tr>
<tr>
<td>168</td>
<td>90653</td>
<td>43527980</td>
<td>influenza, trivalent, adjuvanted</td>
</tr>
<tr>
<td>171</td>
<td>90674</td>
<td>No Concept ID</td>
<td>Influenza, injectable, MDCK, preservative free, quadrivalent</td>
</tr>
<tr>
<td>174</td>
<td>90625</td>
<td>No Concept ID</td>
<td>cholera, live attenuated</td>
</tr>
<tr>
<td>175</td>
<td>90675</td>
<td>2213449</td>
<td>Rabies - IM Diploid cell culture</td>
</tr>
<tr>
<td>176</td>
<td>90675</td>
<td>2213449</td>
<td>Rabies - IM fibroblast culture</td>
</tr>
</tbody>
</table>
Appendix 5. SQL query to find ingredients of OMOP standard drug concepts

```sql
SELECT DISTINCT A.CONCEPT_ID AS DRUG_EXPOSURE_CONCEPT_ID,
C.CONCEPT_ID AS INGREDIENT_CONCEPT_ID
FROM CONCEPT C
JOIN CONCEPT_ANCESTOR CA
    ON CA.ANCESTOR_CONCEPT_ID = C.CONCEPT_ID
    AND C.VOCABULARY_ID = 'RxNorm'
    AND C.CONCEPT_CLASS_ID = 'Ingredient'
    AND INVALID_REASON IS NULL
JOIN CONCEPT A
    ON CA.DESCENDANT_CONCEPT_ID = A.CONCEPT_ID
```
Appendix 6. SQL query to map measurement observations to the corresponding clinical evaluation and measurement value concepts

WITH SNOMED_TO_MEAS AS (  
    SELECT DISTINCT C1.CONCEPT_ID AS STANDARD_CONCEPT_ID, C1.CONCEPT_NAME AS  
        STANDARD_CONCEPT_NAME, R.RELATIONSHIP_ID,  
        C2.CONCEPT_ID AS MEASUREMENT_CONCEPT_ID, C2.CONCEPT_NAME AS  
        MEASUREMENT_NAME, C2.DOMAIN_ID AS MEASUREMENT_DOMAIN_ID, C2.CONCEPT_CLASS_ID AS  
        MEASUREMENT_CLASS_ID  
        FROM OMOP.CONCEPT C1, OMOP.CONCEPT_RELATIONSHIP R, OMOP.CONCEPT C2  
        WHERE C1.CONCEPT_ID = R.CONCEPT_ID_1 AND C2.CONCEPT_ID = R.CONCEPT_ID_2  
        AND R.RELATIONSHIP_ID IN ('HAS INTERPRETATION', 'HAS INTERPRETS')  
        AND R.INVALID_REASON IS NULL AND C1.INVALID_REASON IS NULL AND  
        C2.INVALID_REASON IS NULL  
        AND C1.DOMAIN_ID = 'MEASUREMENT'  
        AND C1.VOCABULARY_ID IN ('SNOMED') AND C2.DOMAIN_ID IN  
        ('MEASUREMENT', 'PROCEDURE', 'MEAS VALUE')  
        ORDER BY C1.CONCEPT_ID),  
    MEASUREMENT AS (  
        SELECT  
            DISTINCT STM.STANDARD_CONCEPT_ID, STM.STANDARD_CONCEPT_NAME, STM.MEASUREMENT_CONCEPT_ID, STM.MEASUREMENT_NAME  
            FROM SNOMED_TO_MEAS STM  
            WHERE STM.RELATIONSHIP_ID = 'HAS INTERPRETS')  
    ,  
    MEASVALUE AS (  
        SELECT DISTINCT STM.STANDARD_CONCEPT_ID, STM.STANDARD_CONCEPT_NAME, STM.MEASUREMENT_CONCEPT_ID AS MEASVALUE_CONCEPT_ID, STM.MEASUREMENT_NAME AS MEASVALUE_NAME  
            FROM SNOMED_TO_MEAS STM  
            WHERE STM.RELATIONSHIP_ID = 'HAS INTERPRETS')  
    
    SELECT M.*, MV.MEASVALUE_CONCEPT_ID, MV.MEASVALUE_NAME  
    FROM MEASUREMENT M  
    LEFT JOIN MEASVALUE MV  
    ON M.STANDARD_CONCEPT_ID = MV.STANDARD_CONCEPT_ID;
Appendix 7. SQL script to add table constrains to OMOP CDM

Primary key constraints

ALTER TABLE OMOP.PERSON ADD CONSTRAINT XPK_PERSON PRIMARY KEY ( PERSON_ID ) ;
ALTER TABLE OMOP.OBSERVATION_PERIOD ADD CONSTRAINT XPK_OBSERVATION_PERIOD
PRIMARY KEY ( OBSERVATION_PERIOD_ID ) ;
ALTER TABLE OMOP.VISIT_OCCURRENCE ADD CONSTRAINT XPK_VISIT_OCCURRENCE PRIMARY
KEY ( VISIT_OCCURRENCE_ID ) ;
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT XPK_PROCEDURE_OCCURRENCE
PRIMARY KEY ( PROCEDURE_OCCURRENCE_ID ) ;
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT XPK_DRUG_EXPOSURE PRIMARY KEY ( DRUG_EXPOSURE_ID ) ;
ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT XPK_CONDITION_OCCURRENCE
PRIMARY KEY ( CONDITION_OCCURRENCE_ID ) ;
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT XPK_MEASUREMENT PRIMARY KEY ( MEASUREMENT_ID ) ;
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT XPK_OBSERVATION PRIMARY KEY ( OBSERVATION_ID ) ;
ALTER TABLE OMOP.DRUG_ERA ADD CONSTRAINT XPK_DRUG_ERA PRIMARY KEY ( DRUG_ERA_ID ) ;
ALTER TABLE OMOP.CONDITION_ERA ADD CONSTRAINT XPK_CONDITION_ERA PRIMARY KEY ( CONDITION_ERA_ID ) ;

Foreign key constraints

ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_GENDER_CONCEPT FOREIGN KEY
(GENDER_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_RACE_CONCEPT FOREIGN KEY
(RACE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_ETHNICITY_CONCEPT FOREIGN KEY
(ETHNICITY_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_GENDER_CONCEPT_S FOREIGN KEY
(GENDER_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_RACE_CONCEPT_S FOREIGN KEY
(RACE_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PERSON ADD CONSTRAINT FPK_PERSON_ETHNICITY_CONCEPT_S FOREIGN KEY
(ETHNICITY_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION_PERIOD ADD CONSTRAINT
FPK_OBSERVATION_PERIOD_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON
(PERSON_ID);
ALTER TABLE OMOP.OBSERVATION_PERIOD ADD CONSTRAINT
FPK_OBSERVATION_PERIOD_CONCEPT FOREIGN KEY (PERIOD_TYPE_CONCEPT_ID) REFERENCES
OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.VISIT_OCCURRENCE ADD CONSTRAINT FPK_VISIT_PERSON FOREIGN KEY
(PER SON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.VISIT_OCCURRENCE ADD CONSTRAINT FPK_VISIT_CONCEPT FOREIGN KEY
(VISIT_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.VISIT_OCCURRENCE ADD CONSTRAINT FPK_VISIT_TYPE_CONCEPT FOREIGN KEY
(VISIT_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.VISIT_OCCURRENCE ADD CONSTRAINT FPK_VISIT_CONCEPT_S FOREIGN KEY
(VISIT_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_CONCEPT FOREIGN KEY (PROCEDURE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_TYPE_CONCEPT FOREIGN KEY (PROCEDURE_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_MODIFIER FOREIGN KEY (MODIFIER_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_VISIT FOREIGN KEY (VISIT_OCCURRENCE_ID) REFERENCES OMOP.VISIT_OCCURRENCE (VISIT_OCCURRENCE_ID);
ALTER TABLE OMOP.PROCEDURE_OCCURRENCE ADD CONSTRAINT FPK_PROCEDURE_CONCEPT_S FOREIGN KEY (PROCEDURE_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_CONCEPT FOREIGN KEY (DRUG_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_TYPE_CONCEPT FOREIGN KEY (DRUG_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_ROUTE_CONCEPT FOREIGN KEY (ROUTE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_DOSE_UNIT_CONCEPT FOREIGN KEY (DOSE_UNIT_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_VISIT FOREIGN KEY (VISIT_OCCURRENCE_ID) REFERENCES OMOP.VISIT_OCCURRENCE (VISIT_OCCURRENCE_ID);
ALTER TABLE OMOP.DRUG_EXPOSURE ADD CONSTRAINT FPK_DRUG_CONCEPT_S FOREIGN KEY (DRUG_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT FPK_CONDITION_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT FPK_CONDITION_CONCEPT FOREIGN KEY (CONDITION_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT FPK_CONDITION_TYPE_CONCEPT FOREIGN KEY (CONDITION_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT FPK_CONDITION_VISIT FOREIGN KEY (VISIT_OCCURRENCE_ID) REFERENCES OMOP.VISIT_OCCURRENCE (VISIT_OCCURRENCE_ID);
ALTER TABLE OMOP.CONDITION_OCCURRENCE ADD CONSTRAINT FPK_CONDITION_CONCEPT_S FOREIGN KEY (CONDITION_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_CONCEPT FOREIGN KEY (MEASUREMENT_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_TYPE_CONCEPT FOREIGN KEY (MEASUREMENT_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_OPERATOR FOREIGN KEY (OPERATOR_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_VALUE FOREIGN KEY (VALUE_AS_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_UNIT FOREIGN KEY (UNIT_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_VISIT FOREIGN KEY (VISIT_OCCURRENCE_ID) REFERENCES OMOP.VISIT_OCCURRENCE (VISIT_OCCURRENCE_ID);
ALTER TABLE OMOP.MEASUREMENT ADD CONSTRAINT FPK_MEASUREMENT_CONCEPT_S FOREIGN KEY (MEASUREMENT_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_CONCEPT FOREIGN KEY (OBSERVATION_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_TYPE_CONCEPT FOREIGN KEY (OBSERVATION_TYPE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_VALUE FOREIGN KEY (VALUE_AS_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_QUALIFIER FOREIGN KEY (QUALIFIER_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_UNIT FOREIGN KEY (UNIT_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_VISIT FOREIGN KEY (VISIT_OCCURRENCE_ID) REFERENCES OMOP.VISIT_OCCURRENCE (VISIT_OCCURRENCE_ID);
ALTER TABLE OMOP.OBSERVATION ADD CONSTRAINT FPK_OBSERVATION_CONCEPT_S FOREIGN KEY (OBSERVATION_SOURCE_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.DRUG_ERA ADD CONSTRAINT FPK_DRUG_ERA_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.DRUG_ERA ADD CONSTRAINT FPK_DRUG_ERA_CONCEPT FOREIGN KEY (DRUG_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);

ALTER TABLE OMOP.CONDITION_ERA ADD CONSTRAINT FPK_CONDITION_ERA_PERSON FOREIGN KEY (PERSON_ID) REFERENCES OMOP.PERSON (PERSON_ID);
ALTER TABLE OMOP.CONDITION_ERA ADD CONSTRAINT FPK_CONDITION_ERA_CONCEPT FOREIGN KEY (CONDITION_CONCEPT_ID) REFERENCES OMOP.CONCEPT (CONCEPT_ID);
Appendix 8. The XML schema of O-PMML standard

<x:schema xmlns:x="http://www.w3.org/2001/XMLSchema"
xmlns="http://www.dmg.org/PMML-4_3"
elementFormDefault="unqualified"
targetNamespace="http://www.dmg.org/PMML-4_3">
<!-- PMML -->
<x:element name="PMML">
  <x:complexType>
    <x:sequence>
      <x:element ref="Header"/>
      <x:element ref="DataDictionary"/>
      <x:element ref="TransformationDictionary" minOccurs="0"/>
      <x:element ref="MiningBuildTask"/>
      <x:sequence minOccurs="1">
        <x:group ref="MODEL-ELEMENT"/>
      </x:sequence>
    </x:sequence>
    <x:attribute name="version" type="xs:string" use="required"/>
  </x:complexType>
</x:element>
<x:group name="MODEL-ELEMENT">
  <x:choice>
    <x:element ref="RegressionModel"/>
  </x:choice>
</x:group>
<!-- Header -->
<x:element name="Header">
  <x:complexType>
    <x:sequence>
      <x:element minOccurs="1" ref="Extension"/>
      <x:element ref="Application"/>
      <x:element ref="Annotation"/>
      <x:element ref="Timestamp"/>
    </x:sequence>
    <x:attribute name="copyright" type="xs:string"/>
    <x:attribute name="description" type="xs:string"/>
    <x:attribute name="modelVersion" type="xs:string"/>
  </x:complexType>
</x:element>
<x:element name="Application">
  <x:complexType>
    <x:attribute name="name" type="xs:string" use="required"/>
    <x:attribute name="version" type="xs:string"/>
  </x:complexType>
</x:element>
<x:element name="Annotation" type="xs:string"/>
<x:element name="Timestamp" type="xs:dateTime"/>
<x:element name="OmopCdm">
  <x:complexType>
    <x:attribute name="version" type="xs:string" use="required"/>
  </x:complexType>
</x:element>
<x:element name="author">
  <x:complexType>
    <x:attribute name="name" type="xs:string" use="required"/>
  </x:complexType>
</x:element>
<!-- Extensions -->
<x:element name="Extension">
  <x:complexType>
    <x:sequence>
      <x:element minOccurs="0" ref="OmopCdm"/>
      <x:element minOccurs="0" ref="author"/>
    </x:sequence>
  </x:complexType>
</x:element>
</x:schema>
<xs:element minOccurs="0" ref="InputParameters"/>
<xs:element minOccurs="0" ref="Statement"/>
</xs:sequence>
</xs:complexType>
</xs:element>

<!-- Data Dictionary -->
<xs:element name="DataDictionary">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="DataField" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="numberOfFields" type="xs:nonNegativeInteger"/>
  </xs:complexType>
</xs:element>

<xs:element name="DataField">
  <xs:complexType>
    <xs:sequence>
      <xs:sequence>
        <xs:element ref="Interval" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="Value" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="name" type="FIELD-NAME" use="required"/>
      <xs:attribute name="displayName" type="xs:string"/>
      <xs:attribute name="optype" type="OPTYPE" use="required"/>
      <xs:attribute name="dataType" type="DATATYPE" use="required"/>
    </xs:complexType>
  </xs:element>
</xs:complexType>

<xs:simpleType name="OPTYPE">
  <xs:restriction base="xs:string">
    <xs:enumeration value="categorical"/>
    <xs:enumeration value="ordinal"/>
    <xs:enumeration value="continuous"/>
  </xs:restriction>
</xs:simpleType>

<xs:simpleType name="DATATYPE">
  <xs:restriction base="xs:string">
    <xs:enumeration value="string"/>
    <xs:enumeration value="integer"/>
    <xs:enumeration value="float"/>
    <xs:enumeration value="double"/>
    <xs:enumeration value="boolean"/>
    <xs:enumeration value="date"/>
    <xs:enumeration value="dateTime"/>
    <xs:enumeration value="dateDaysSince[0]"/>
    <xs:enumeration value="dateDaysSince[1960]"/>
    <xs:enumeration value="dateDaysSince[1970]"/>
    <xs:enumeration value="dateDaysSince[1980]"/>
    <xs:enumeration value="timeSeconds"/>
    <xs:enumeration value="dateTimeSecondsSince[0]"/>
    <xs:enumeration value="dateTimeSecondsSince[1960]"/>
    <xs:enumeration value="dateTimeSecondsSince[1970]"/>
    <xs:enumeration value="dateTimeSecondsSince[1980]"/>
  </xs:restriction>
</xs:simpleType>

<xs:element name="Value">
  <xs:complexType>
    <xs:attribute name="value" type="xs:string" use="required"/>
    <xs:attribute name="displayValue" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:attribute name="property" default="valid">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="valid"/>
      <xs:enumeration value="invalid"/>
      <xs:enumeration value="missing"/>
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
</xs:complexType>
</xs:element>

<xs:element name="Interval">
  <xs:complexType>
    <xs:attribute name="closure" use="required">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="openClosed"/>
          <xs:enumeration value="openOpen"/>
          <xs:enumeration value="closedOpen"/>
          <xs:enumeration value="closedClosed"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="leftMargin" type="NUMBER"/>
    <xs:attribute name="rightMargin" type="NUMBER"/>
  </xs:complexType>
</xs:element>

<xs:element name="TransformationDictionary">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="DefineFunction" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="DerivedField" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="LocalTransformations">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="DerivedField" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="DerivedField">
  <xs:complexType>
    <xs:sequence>
      <xs:group ref="EXPRESSION"/>
      <xs:element ref="Value" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="name" type="FIELD-NAME"/>
    <xs:attribute name="displayName" type="xs:string"/>
    <xs:attribute name="optype" type="OPTYPE" use="required"/>
    <xs:attribute name="dataType" type="DATATYPE" use="required"/>
  </xs:complexType>
</xs:element>

<xs:group name="EXPRESSION">
  <xs:choice>
    <xs:element ref="Constant"/>
    <xs:element ref="FieldRef"/>
    <xs:element ref="NormContinuous"/>
    <xs:element ref="NormDiscrete"/>
  </xs:choice>
</xs:group>
<xs:element ref="Discretize"/>
<xs:element ref="MapValues"/>
<xs:element ref="TextIndex"/>
<xs:element ref="Apply"/>
<xs:element ref="Aggregate"/>
<xs:element ref="Lag"/>
</xs:choice>
</xs:group>
<xs:element name="Constant">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="dataType" type="DATATYPE"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="FieldRef">
  <xs:complexType>
    <xs:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:attribute name="mapMissingTo" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:element name="NormContinuous">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="LinearNorm" minOccurs="2" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="mapMissingTo" type="NUMBER"/>
    <xs:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:attribute name="outliers" type="OUTLIER-TREATMENT-METHOD" default="asIs"/>
  </xs:complexType>
</xs:element>
<xs:element name="LinearNorm">
  <xs:complexType>
    <xs:attribute name="orig" type="NUMBER" use="required"/>
    <xs:attribute name="norm" type="NUMBER" use="required"/>
  </xs:complexType>
</xs:element>
<xs:element name="NormDiscrete">
  <xs:complexType>
    <xs:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:attribute name="value" type="xs:string" use="required"/>
    <xs:attribute name="mapMissingTo" type="NUMBER"/>
  </xs:complexType>
</xs:element>
<xs:element name="Discretize">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="DiscretizeBin" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="DiscretizeBin">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Interval"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="MapValues">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="FieldColumnPair"/>
      <xs:choice minOccurs="0">
        <xs:element ref="TableLocator"/>
        <xs:element ref="InlineTable"/>
      </xs:choice>
    </xs:sequence>
    <xs:attribute name="mapMissingTo" type="xs:string"/>
    <xs:attribute name="defaultValue" type="xs:string"/>
    <xs:attribute name="outputColumn" type="xs:string" use="required"/>
    <xs:attribute name="dataType" type="DATATYPE"/>
  </xs:complexType>
</xs:element>

<xs:element name="FieldColumnPair">
  <xs:complexType>
    <xs:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:attribute name="column" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="TextIndex">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="TextIndexNormalization" minOccurs="0" maxOccurs="unbounded"/>
      <xs:group ref="EXPRESSION"/>
      <xs:element ref="TextIndexNormalization" minOccurs="0" maxOccurs="unbounded"/>
      <xs:attribute name="textField" type="FIELD-NAME" use="required"/>
      <xs:attribute name="localTermWeights" default="termFrequency"/>
      <xs:simpleType>
        <xs:restriction base="xs:string"/>
        <xs:enumeration value="termFrequency"/>
        <xs:enumeration value="binary"/>
        <xs:enumeration value="logarithmic"/>
        <xs:enumeration value="augmentedNormalizedTermFrequency"/>
      </xs:simpleType>
      <xs:attribute name="isCaseSensitive" type="xs:boolean" default="false"/>
      <xs:attribute name="maxLevenshteinDistance" type="xs:integer" default="0"/>
      <xs:attribute name="countHits" default="allHits"/>
      <xs:simpleType>
        <xs:restriction base="xs:string"/>
        <xs:enumeration value="allHits"/>
        <xs:enumeration value="bestHits"/>
      </xs:simpleType>
      <xs:attribute name="wordSeparatorCharacterRE" type="xs:string" default="\s"/>
      <xs:attribute name="tokenize" type="xs:boolean" default="true"/>
    </xs:attribute>
  </xs:complexType>
</xs:element>

<xs:element name="TextIndexNormalization">
  <xs:complexType>
    <xs:sequence>
      <xs:choice minOccurs="0">
        <xs:element ref="TableLocator"/>
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element ref="InlineTable"/>
</xs:choice>
</xs:sequence>
<xs:attribute name="inField" type="xs:string" default="string"/>
<xs:attribute name="outField" type="xs:string" default="stem"/>
<xs:attribute name="regexField" type="xs:string" default="regex"/>
<xs:attribute name="recursive" type="xs:boolean" default="false"/>
<xs:attribute name="isCaseSensitive" type="xs:boolean"/>
<xs:attribute name="maxLevenshteinDistance" type="xs:integer"/>
<xs:attribute name="wordSeparatorCharacterRE" type="xs:string"/>
<xs:attribute name="tokenize" type="xs:boolean"/>
</xs:complexType>
</xs:element>
<xs:element name="Aggregate">
<xs:complexType>
<xs:attribute name="field" type="FIELD-NAME" use="required"/>
<xs:attribute name="function" use="required">
<xs:simpleType>
<xs:restriction base="xs:string">
<xs:enumeration value="count"/>
<xs:enumeration value="sum"/>
<xs:enumeration value="average"/>
<xs:enumeration value="min"/>
<xs:enumeration value="max"/>
<xs:enumeration value="multiset"/>
</xs:restriction>
</xs:simpleType>
</xs:attribute>
<xs:attribute name="groupField" type="FIELD-NAME"/>
<xs:attribute name="sqlWhere" type="xs:string"/>
</xs:complexType>
</xs:element>
<xs:element name="Lag">
<xs:complexType>
<xs:sequence>
<xs:element ref="BlockIndicator" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
<xs:attribute name="field" type="FIELD-NAME" use="required"/>
<xs:attribute name="n" type="xs:positiveInteger" default="1"/>
</xs:complexType>
</xs:element>
<xs:element name="BlockIndicator">
<xs:complexType>
<xs:attribute name="field" type="FIELD-NAME" use="required"/>
</xs:complexType>
</xs:element>
<xs:element name="DefineFunction">
<xs:complexType>
<xs:element ref="ParameterField" minOccurs="1" maxOccurs="unbounded"/>
<xs:group ref="EXPRESSION"/>
<xs:sequence>
<xs:attribute name="name" type="xs:string" use="required"/>
<xs:attribute name="optype" type="OPTYPE" use="required"/>
<xs:attribute name="dataType" type="DATATYPE"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="ParameterField">
<xs:complexType>
<xs:attribute name="name" type="xs:string" use="required"/>
<xs:attribute name="optype" type="OPTYPE"/>
<xs:attribute name="dataType" type="DATATYPE"/>
</xs:complexType>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
</xs:complexType>
</xs:element>
<xs:element ref="ModelVerification" minOccurs="0"/>
</xs:sequence>
<xs:attribute name="modelName" type="xs:string"/>
<xs:attribute name="functionName" type="MINING-FUNCTION" use="required"/>
<xs:attribute name="algorithmName" type="xs:string"/>
<xs:attribute name="modelType" use="optional" type="xs:string">
<xs:restriction base="xs:string">
<xs:enumeration value="linearRegression"/>
<xs:enumeration value="stepwisePolynomialRegression"/>
<xs:enumeration value="logisticRegression"/>
</xs:restriction>
</xs:attribute>
<xs:attribute name="targetFieldName" type="FIELD-NAME" use="optional"/>
<xs:attribute name="normalizationMethod" type="REGRESSIONNORMALIZATIONMETHOD" default="none"/>
<xs:attribute name="isScorable" type="xs:boolean" default="true"/>
</xs:complexType>
</xs:element>
<!-- Mining Schema -->
<xs:element name="MiningSchema">
<xs:complexType>
<xs:sequence>
<xs:element maxOccurs="unbounded" ref="MiningField"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<!-- Regression Table -->
<xs:element name="RegressionTable">
<xs:complexType>
<xs:sequence>
<xs:element ref="NumericPredictor" minOccurs="0" maxOccurs="unbounded"/>
<xs:element ref="CategoricalPredictor" minOccurs="0" maxOccurs="unbounded"/>
<xs:element ref="PredictorTerm" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
<xs:attribute name="intercept" type="REAL-NUMBER" use="required"/>
<xs:attribute name="targetCategory" type="xs:string"/>
</xs:complexType>
</xs:element>
<xs:restriction base="REGRESSIONNORMALIZATIONMETHOD">
<xs:enumeration value="none"/>
</xs:restriction>

123
<xs:enumeration value="simplemax"/>
<xs:enumeration value="softmax"/>
<xs:enumeration value="logit"/>
<xs:enumeration value="probit"/>
<xs:enumeration value="cloglog"/>
<xs:enumeration value="exp"/>
<xs:enumeration value="loglog"/>
<xs:enumeration value="cauchit"/>
</xs:restriction>
</xs:simpleType>
</xs:element>
<xs:element name="NumericPredictor">
  <xs:complexType>
    <xs:attribute name="name" type="FIELD-NAME" use="required"/>
    <xs:attribute name="exponent" type="INT-NUMBER" default="1"/>
    <xs:attribute name="coefficient" type="REAL-NUMBER" use="required"/>
  </xs:complexType>
</xs:element>
</xs:element>
<xs:element name="CategoricalPredictor">
  <xs:complexType>
    <xs:attribute name="name" type="FIELD-NAME" use="required"/>
    <xs:attribute name="value" type="xs:string" use="required"/>
    <xs:attribute name="coefficient" type="REAL-NUMBER" use="required"/>
  </xs:complexType>
</xs:element>
</xs:element>
<xs:element name="PredictorTerm">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="FieldRef" minOccurs="1" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="name" type="FIELD-NAME"/>
    <xs:attribute name="coefficient" type="REAL-NUMBER" use="required"/>
  </xs:complexType>
</xs:element>
</xs:element>
<xs:element name="Output">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="OutputField" minOccurs="1" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:element>
<xs:element name="OutputField">
  <xs:complexType>
    <xs:sequence>
      <xs:sequence minOccurs="0" maxOccurs="1">
        <xs:element ref="Decisions" minOccurs="0" maxOccurs="1"/>
        <xs:group ref="EXPRESSION" minOccurs="1" maxOccurs="1"/>
      </xs:sequence>
      <xs:attribute name="name" type="FIELD-NAME" use="required"/>
      <xs:attribute name="displayName" type="xs:string"/>
      <xs:attribute name="optype" type="OPTYPE"/>
      <xs:attribute name="dataType" type="DATATYPE" use="required"/>
      <xs:attribute name="targetField" type="FIELD-NAME"/>
      <xs:attribute name="feature" type="RESULT-FEATURE" default="predictedValue"/>
      <xs:attribute name="value" type="xs:string"/>
      <xs:attribute name="isFinalResult" type="xs:boolean" default="true"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:complexType name="MINING-FUNCTION">
<xs:restriction base="xs:string">
  <xs:enumeration value="associationRules"/>
<xs:enumeration value="sequences"/>
<xs:enumeration value="classification"/>
<xs:enumeration value="regression"/>
<xs:enumeration value="clustering"/>
<xs:enumeration value="timeSeries"/>
<xs:enumeration value="mixed"/>
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="FIELD-USAGE-TYPE">
<xs:restriction base="xs:string">
<xs:enumeration value="active"/>
<xs:enumeration value="predicted"/>
<xs:enumeration value="target"/>
<xs:enumeration value="supplementary"/>
<xs:enumeration value="group"/>
<xs:enumeration value="order"/>
<xs:enumeration value="frequencyWeight"/>
<xs:enumeration value="analysisWeight"/>
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="OUTLIER-TREATMENT-METHOD">
<xs:restriction base="xs:string">
<xs:enumeration value="asIs"/>
<xs:enumeration value="asMissingValues"/>
<xs:enumeration value="asExtremeValues"/>
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="MISSING-VALUE-TREATMENT-METHOD">
<xs:restriction base="xs:string">
<xs:enumeration value="asIs"/>  
<xs:enumeration value="asMean"/>  
<xs:enumeration value="asMode"/>  
<xs:enumeration value="asMedian"/>  
<xs:enumeration value="asValue"/>  
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="INVALID-VALUE-TREATMENT-METHOD">
<xs:restriction base="xs:string">
<xs:enumeration value="returnInvalid"/>
<xs:enumeration value="asIs"/>
<xs:enumeration value="asMissing"/>
</xs:restriction>
</xs:simpleType>
<!-- Model Explanation -->
<xs:element name="ModelExplanation">
<xs:complexType>
<xs:sequence>
<xs:element ref="PredictiveModelQuality" minOccurs="0" maxOccurs="unbounded"/>
<xs:choice>
<xs:element ref="Correlations" minOccurs="0"/>
</xs:choice>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="PredictiveModelQuality">
<xs:complexType>
<xs:sequence>
<xs:element ref="ROC" minOccurs="0"/>
</xs:sequence>
<xs:attribute name="targetField" type="xs:string" use="required"/>
<xs:attribute name="dataName" type="xs:string" use="optional"/>
<xs:attribute name="dataUsage" default="training"/>
<xs:simpleType>
    <xs:restriction base="xs:string">
        <xs:enumeration value="training"/>
        <xs:enumeration value="test"/>
        <xs:enumeration value="validation"/>
    </xs:restriction>
</xs:simpleType>

<xs:attribute name="meanError" type="NUMBER" use="optional"/>
<xs:attribute name="meanAbsoluteError" type="NUMBER" use="optional"/>
<xs:attribute name="rootMeanSquaredError" type="NUMBER" use="optional"/>
<xs:attribute name="adj-r-squared" type="NUMBER" use="optional"/>
<xs:attribute name="sumSquaredRegression" type="NUMBER" use="optional"/>
<xs:attribute name="numOfRecords" type="NUMBER" use="optional"/>
<xs:attribute name="numOfRecordsWeighted" type="NUMBER" use="optional"/>
<xs:attribute name="numOfPredictors" type="NUMBER" use="optional"/>
<xs:attribute name="degreesOfFreedom" type="NUMBER" use="optional"/>
<xs:attribute name="fStatistic" type="NUMBER" use="optional"/>
<xs:attribute name="AIC" type="NUMBER" use="optional"/>
<xs:attribute name="BIC" type="NUMBER" use="optional"/>
<xs:attribute name="AICC" type="NUMBER" use="optional"/>
</xs:complexType>

<xs:element name="ROC">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="ROCGraph"/>
        </xs:sequence>
        <xs:attribute name="positiveTargetFieldValue" type="xs:string" use="required"/>
        <xs:attribute name="positiveTargetFieldDisplayValue" type="xs:string"/>
        <xs:attribute name="negativeTargetFieldValue" type="xs:string"/>
        <xs:attribute name="negativeTargetFieldDisplayValue" type="xs:string"/>
    </xs:complexType>
</xs:element>

<xs:element name="ROCGraph">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="XCoordinates"/>
            <xs:element ref="YCoordinates"/>
            <xs:element ref="BoundaryValues" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<xs:element name="Correlations">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="CorrelationFields"/>
            <xs:element ref="CorrelationValues"/>
            <xs:element ref="CorrelationMethods" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<xs:element name="CorrelationFields">
    <xs:complexType>
        <xs:sequence>
            <xs:group ref="STRING-ARRAY"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="CorrelationValues">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="Matrix"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="CorrelationMethods">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="Matrix"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<!-- Model Verification -->
<xs:element name="ModelVerification">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="VerificationFields"/>
            <xs:element ref="InlineTable"/>
        </xs:sequence>
        <xs:attribute name="recordCount" type="INT-NUMBER" use="optional"/>
        <xs:attribute name="fieldCount" type="INT-NUMBER" use="optional"/>
    </xs:complexType>
</xs:element>
<xs:element name="VerificationFields">
    <xs:complexType>
        <xs:sequence>
            <xs:element maxOccurs="unbounded" ref="VerificationField"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="VerificationField">
    <xs:complexType>
        <xs:attribute name="field" type="xs:string" use="required"/>
        <xs:attribute name="column" type="xs:string" use="optional"/>
        <xs:attribute name="precision" type="xs:double" default="1E-6"/>
        <xs:attribute name="zeroThreshold" type="xs:double" default="1E-16"/>
    </xs:complexType>
</xs:element>
<xs:element name="ResultField">
    <xs:complexType>
        <xs:attribute name="name" type="FIELD-NAME" use="required"/>
        <xs:attribute name="displayName" type="xs:string"/>
        <xs:attribute name="optype" type="OPTYPE"/>
        <xs:attribute name="dataType" type="DATATYPE"/>
        <xs:attribute name="feature" type="RESULT-FEATURE"/>
        <xs:attribute name="value" type="xs:string"/>
    </xs:complexType>
</xs:element>
<xs:simpleType name="RESULT-FEATURE">
    <xs:restriction base="xs:string">
        <xs:enumeration value="predictedValue"/>
        <xs:enumeration value="predictedDisplayValue"/>
        <xs:enumeration value="transformedValue"/>
        <xs:enumeration value="decision"/>
        <xs:enumeration value="probability"/>
        <xs:enumeration value="affinity"/>
        <xs:enumeration value="residual"/>
        <xs:enumeration value="standardError"/>
        <xs:enumeration value="clusterId"/>
        <xs:enumeration value="clusterAffinity"/>
        <xs:enumeration value="entityId"/>
    </xs:restriction>
</xs:simpleType>
<xs:enumeration value="entityAffinity"/>
<xs:enumeration value="warning"/>
<xs:enumeration value="ruleValue"/>
<xs:enumeration value="reasonCode"/>
<xs:enumeration value="antecedent"/>
<xs:enumeration value="consequent"/>
<xs:enumeration value="rule"/>
<xs:enumeration value="ruleId"/>
<xs:enumeration value="confidence"/>
<xs:enumeration value="support"/>
<xs:enumeration value="lift"/>
<xs:enumeration value="leverage"/>
</xs:restriction>
</xs:simpleType>
</xs:element>
<xs:element name="Decisions">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Decision" minOccurs="1" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="businessProblem" type="xs:string"/>
    <xs:attribute name="description" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:element name="Decision">
  <xs:complexType>
    <xs:attribute name="value" type="xs:string" use="required"/>
    <xs:attribute name="displayValue" type="xs:string"/>
    <xs:attribute name="description" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:simpleType name="RULE-FEATURE">
  <xs:restriction base="xs:string">
    <xs:enumeration value="antecedent"/>
    <xs:enumeration value="consequent"/>
    <xs:enumeration value="rule"/>
    <xs:enumeration value="ruleId"/>
    <xs:enumeration value="confidence"/>
    <xs:enumeration value="support"/>
    <xs:enumeration value="lift"/>
    <xs:enumeration value="leverage"/>
    <xs:enumeration value="affinity"/>
  </xs:restriction>
</xs:simpleType>
<!-- Model Statistics -->
<xs:element name="ModelStats">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="UnivariateStats" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="MultivariateStats" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="UnivariateStats">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Counts" minOccurs="0"/>
      <xs:element ref="NumericInfo" minOccurs="0"/>
      <xs:element ref="DiscrStats" minOccurs="0"/>
      <xs:element ref="ContStats" minOccurs="0"/>
      <xs:element ref="Anova" minOccurs="0"/>
    </xs:sequence>
    <xs:attribute name="field" type="FIELD-NAME"/>
    <xs:attribute name="weighted" default="0"/>
<xs:simpleType>
  <xs:restriction base="xs:string">
    <xs:enumeration value="0"/>
    <xs:enumeration value="1"/>
  </xs:restriction>
</xs:simpleType>

<xs:element name="Counts">
  <xs:complexType>
    <xs:attribute name="totalFreq" type="NUMBER" use="required"/>
    <xs:attribute name="missingFreq" type="NUMBER"/>
    <xs:attribute name="invalidFreq" type="NUMBER"/>
    <xs:attribute name="cardinality" type="xs:nonNegativeInteger"/>
  </xs:complexType>
</xs:element>

<xs:element name="NumericInfo">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Quantile" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="minimum" type="NUMBER"/>
    <xs:attribute name="maximum" type="NUMBER"/>
    <xs:attribute name="mean" type="NUMBER"/>
    <xs:attribute name="standardDeviation" type="NUMBER"/>
    <xs:attribute name="median" type="NUMBER"/>
    <xs:attribute name="interQuartileRange" type="NUMBER"/>
  </xs:complexType>
</xs:element>

<xs:element name="Quantile">
  <xs:complexType>
    <xs:attribute name="quantileLimit" type="PERCENTAGE-NUMBER" use="required"/>
    <xs:attribute name="quantileValue" type="NUMBER" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="DiscrStats">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Array" minOccurs="0" maxOccurs="2"/>
    </xs:sequence>
    <xs:attribute name="modalValue" type="xs:string"/>
  </xs:complexType>
</xs:element>

<xs:element name="ContStats">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Interval" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute ref="FrequenciesType" minOccurs="0"/>
    <xs:attribute name="totalValuesSum" type="NUMBER"/>
    <xs:attribute name="totalSquaresSum" type="NUMBER"/>
  </xs:complexType>
</xs:element>

<xs:group name="FrequenciesType">
  <xs:sequence>
    <xs:group ref="NUM-ARRAY" minOccurs="1" maxOccurs="3"/>
  </xs:sequence>
</xs:group>

<xs:element name="MultivariateStats">
  <xs:complexType>
    <xs:sequence>
    </xs:sequence>
</xs:element>
<xs:element name="PartitionFieldStats">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Counts" minOccurs="0"/>
      <xs:element ref="NumericInfo" minOccurs="0"/>
      <xs:group ref="FrequenciesType" minOccurs="0"/>
    </xs:sequence>
    <xs:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="0"/>
        <xs:enumeration value="1"/>
      </xs:restriction>
    </xs:simpleType>
    <xs:attribute name="weighted" default="0"/>
  </xs:complexType>
</xs:element>

<xs:element name="Targets">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Target" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="Target">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="TargetValue" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="field" type="FIELD-NAME"/>
    <xs:attribute name="optype" type="OPTYPE"/>
    <xs:attribute name="castInteger">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="round"/>
          <xs:enumeration value="ceiling"/>
          <xs:enumeration value="floor"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="min" type="xs:double"/>
    <xs:attribute name="max" type="xs:double"/>
    <xs:attribute name="rescaleConstant" type="xs:double" default="0"/>
    <xs:attribute name="rescaleFactor" type="xs:double" default="1"/>
  </xs:complexType>
</xs:element>

<xs:element name="TargetValue">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Partition" minOccurs="0"/>
      <xs:attribute name="value" type="xs:string"/>
      <xs:attribute name="displayValue" type="xs:string"/>
      <xs:attribute name="priorProbability" type="PROB-NUMBER"/>
      <xs:attribute name="defaultValue" type="NUMBER"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="InlineTable">
  <xs:complexType>
    <xs:sequence>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element ref="row" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="row">
<xs:complexType>
<xs:complexContent mixed="true">
<xs:restriction base="xs:anyType">
<xs:sequence>
<xs:any processContents="skip" minOccurs="2" maxOccurs="unbounded"/>
</xs:sequence>
</xs:restriction>
</xs:complexContent>
</xs:complexType>
</xs:element>
<xs:simpleType name="NUMBER">
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="INT-NUMBER">
<xs:restriction base="xs:integer">
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="REAL-NUMBER">
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="PROB-NUMBER">
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="PERCENTAGE-NUMBER">
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
<xs:simpleType name="FIELD-NAME">
<xs:restriction base="xs:string">
</xs:restriction>
</xs:simpleType>
<xs:element name="Matrix">
<xs:complexType>
<xs:choice minOccurs="0">
<xs:group ref="NUM-ARRAY" maxOccurs="unbounded"/>
<xs:element ref="MatCell" maxOccurs="unbounded"/>
</xs:choice>
<xs:attribute name="kind" use="optional" default="any">
<xs:simpleType>
<xs:restriction base="xs:string">
<xs:enumeration value="diagonal"/>
<xs:enumeration value="symmetric"/>
<xs:enumeration value="any"/>
</xs:restriction>
</xs:simpleType>
</xs:attribute>
<xs:attribute name="nbRows" type="INT-NUMBER" use="optional"/>
<xs:attribute name="nbCols" type="INT-NUMBER" use="optional"/>
<xs:attribute name="diagDefault" type="REAL-NUMBER" use="optional"/>
<xs:attribute name="offDiagDefault" type="REAL-NUMBER" use="optional"/>
</xs:complexType>
</xs:element>
<xs:element name="MatCell">
<xs:complexType>
<xs:simpleContent>
<xs:extension base="xs:string">
  <xs:attribute name="row" type="INT-NUMBER" use="required"/>
  <xs:attribute name="col" type="INT-NUMBER" use="required"/>
</xs:extension>
</xs:simpleContent>
</xs:complexType>
</xs:element>
</xs:schema>

```xml
<?xml version="1.0"?>
<PMML version="4.3" xmlns="http://www.dmg.org/PMML-4_3"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <Header copyright="Copyright (c) 2017 Hamed Abedtash
  [hamed.abedtash@gmail.com] description="Framingham 10-year risk of cardiovascular disease for men">
    <Extension extender="omop">
      <OmopCdm version="5.1"/>
    </Extension>
    <Application name="OMOP-PMML Writer" version="0.1"/>
    <Timestamp>2017-04-30 16:52:26</Timestamp>
  </Header>
  <MiningBuildTask>
    <Extension name="age" extender="omop">
      <InputParameters>
        <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
            optype="continuous" dataType="date"/>
        <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
            optype="continuous" dataType="bigint"/>
        <InputParameter name="SCHEMA" displayName="Database schema"
            dataType="string"/>
      </InputParameters>
      <Statement dialect="postgresql">
        select distinct extract(year from date '@INDEX_DATE')
           -year_of_birth as AGE from @SCHEMA.person where person_id=@PERSON_ID;
      </Statement>
    </Extension>
    <Extension name="TCL" extender="omop">
      <InputParameters>
        <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
            optype="continuous" dataType="date"/>
        <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
            optype="continuous" dataType="bigint"/>
        <InputParameter name="SCHEMA" displayName="Database schema"
            dataType="string"/>
      </InputParameters>
      <Statement dialect="postgresql">
        with tcl as (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
           FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from @SCHEMA.measurement
               where measurement_concept_id in (3027114) and measurement_date
               &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')
               and person_id=@PERSON_ID) meas
           join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE
               from @SCHEMA.measurement
               where measurement_concept_id in (3027114) and measurement_date
               &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')
               and person_id=@PERSON_ID) measmax
           on meas.person_id=measmax.person_id and meas.measurement_date=measmax.MEAS_DATE)
        select value_as_number from tcl
        where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=30
        order by meas_date desc limit 1;
      </Statement>
    </Extension>
  </MiningBuildTask>
</PMML>
```
with hdl as (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from @SCHEMA.measurement
        where measurement_concept_id in (3007070) and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD')
        and person_id=@PERSON_ID) meas
      join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE
            from @SCHEMA.measurement
            where measurement_concept_id in (3007070) and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD')
            and person_id=@PERSON_ID) measmax
        on meas.person_id=measmax.person_id and meas.measurement_date=measmax.MEAS_DATE)
select value_as_number from hdl
where to_date('@INDEX_DATE','YYYY-MM-DD') - MEAS_DATE < 30
order by meas_date desc limit 1;

with htnmed as (select distinct c2.concept_id
              from omop.concept c1,omop.concept_ancestor a,omop.concept c2
              where c1.concept_code in ('C02A', 'C02B', 'C02C', 'C02D', 'C02K', 'C09AA', 'C09CA', 'C07AA', 'C07AB', 'C07AG', 'C08CA', 'C08D', 'C03AA', 'C03BA', 'C03BX', 'C03CA', 'C03DA', 'C03DB')
              and c1.vocabulary_id='ATC'
              and a.ancestor_concept_id=c1.concept_id
              and a.descendant_concept_id=c2.concept_id
              and c2.concept_class_id in ('Ingredient')
              ),
rec as (select * from omop.drug_era
        where drug_concept_id in (select distinct concept_id from htnmed)
        and (to_date('@INDEX_DATE','YYYY-MM-DD') between drug_era_start_date and drug_era_end_date)
        and person_id=@PERSON_ID)
select case when count(1)>0 then 1 else 0 end as HTNTRT from rec;
<InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
  optype="continuous" dataType="date"/>
<InputParameter name="PERSON_ID" displayName="OMOP Person ID"
  optype="continuous" dataType="bigint"/>
<InputParameter name="SCHEMA" displayName="Database schema"
  dataType="string"/>
</InputParameters>

<Statement dialect="postgresql">
  with sbp as (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
    FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from @SCHEMA.measurement
      where measurement_concept_id in
        (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
      and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD') and
      person_id=@PERSON_ID) meas
    join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE
      from @SCHEMA.measurement
      where measurement_concept_id in
        (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
      and person_id=@PERSON_ID
      and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD')) measmax
    on meas.person_id=measmax.person_id and
    meas.measurement_date=measmax.MEAS_DATE
  select value_as_number from sbp
  where to_date('@INDEX_DATE','YYYY-MM-DD') - MEAS_DATE < 30
  order by meas_date desc limit 1;
</Statement>
</Extension>

<Extension name="smoker" extender="omop">
  <InputParameters>
    <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
  optype="continuous" dataType="date"/>
    <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
  optype="continuous" dataType="bigint"/>
    <InputParameter name="SCHEMA" displayName="Database schema"
  dataType="string"/>
  </InputParameters>
  <Statement dialect="postgresql">
    with smoker as (SELECT DISTINCT person_id,observation_concept_id, observation_date
      FROM omop.observation
      where observation_concept_id IN
        (4218741,4246415,4276526,4052947,4052029,4052030,4218917,4298794,4270996,4043053,4144273,4144273,4043056)
      and person_id=@PERSON_ID and observation_date < to_date('@INDEX_DATE','YYYY-MM-DD')
    )
    ,
    last_smoke as (select max(observation_date) over(partition by person_id) as latest
      from smoker
    )
    select case when count(1)>0 then 1 else 0 end as SMOKER
    from last_smoke
    where to_date('@INDEX_DATE','YYYY-MM-DD') - latest < 90;
  </Statement>
</Extension>

<Extension name="diabetic" extender="omop">
  <InputParameters>
    <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
  optype="continuous" dataType="date"/>
    <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
  optype="continuous" dataType="bigint"/>
    <InputParameter name="SCHEMA" displayName="Database schema"
  dataType="string"/>
  </InputParameters>
  <Statement dialect="postgresql">
    with smoker as (SELECT DISTINCT person_id,observation_concept_id, observation_date
      FROM omop.observation
      where observation_concept_id IN
        (4218741,4246415,4276526,4052947,4052029,4052030,4218917,4298794,4270996,4043053,4144273,4144273,4043056)
      and person_id=@PERSON_ID and observation_date < to_date('@INDEX_DATE','YYYY-MM-DD')
    )
    ,
    last_smoke as (select max(observation_date) over(partition by person_id) as latest
      from smoker
    )
    select case when count(1)>0 then 1 else 0 end as SMOKER
    from last_smoke
    where to_date('@INDEX_DATE','YYYY-MM-DD') - latest < 90;
  </Statement>
</Extension>
<InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
    optype="continuous" dataType="date"/>

<InputParameter name="PERSON_ID" displayName="OMOP Person ID"
    optype="continuous" dataType="bigint"/>

<InputParameter name="SCHEMA" displayName="Database schema"
    dataType="string"/>

</InputParameters>

<Statement dialect="postgresql">
    with medrec as (select * from omop.drug_era
    where drug_concept_id in
    (1529331,1530014,44816332,43013884,19035533,43526465,19033498,1594973,447
    85829,45774435,45774751,1583722,19001409,19059796,1597756,1560171,1909782
    1,1559684,19001441,1000979,19122121,35604829,35605670,35602717,1516976,40
    056629,1502905,1588986,1550023,1531601,1567198,1544838,40051349,40051350,
    46221581,42899447,1909024,19090229,19090247,19090249,1513876,19090180,19
    013926,19090187,19013951,1590165,1586346,1513849,1562586,1909020
    4,1513843,40239216,40170911,44506754,1503297,1510202,1502826,19033909,152
    5215,1517998,1596777,1516766,1547504,40166035,1580747,1502809,1502855,190
    42191,1515249,19090226,19090221,1586369,19122137)
    and (to_date('@INDEX_DATE','YYYY-MM-DD') between
    drug_era_start_date and drug_era_end_date
    OR to_date('@INDEX_DATE','YYYY-MM-DD') between
    drug_era_end_date and drug_era_end_date+30)
    and person_id=@PERSON_ID
    )

    , hasdiabmed as (select case when count(1)>0 then 1 else 0 end as
    DIABETIC
    from medrec 
    )

    , lab AS (select distinct meas.person_id,
    meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
    FROM (select distinct person_id, measurement_concept_id,
    value_as_number, measurement_date from omop.measurement
    where measurement_concept_id in (3037110)
    and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')
    and value_as_number&gt;=126 and person_id=@PERSON_ID) meas
    join (select distinct person_id, measurement_concept_id,
    max(measurement_date) over(partition by
    person_id,measurement_concept_id) as MEAS_DATE
    from omop.measurement
    where measurement_concept_id in (3037110) and
    person_id=@PERSON_ID
    and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')) measmax
    on meas.person_id=measmax.person_id and
    meas.measurement_date=measmax.MEAS_DATE
    )

    , hasdiablalab as (select case when count(1)>0 then 1 else 0 end as
    DIABETIC
    from lab 
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=60
    )

    , diabdx as (select max(condition_era_end_date) as latest
    from (SELECT DISTINCT *
    FROM omop.condition_era
    where condition_concept_id IN
    (201820,45757674,45757474,4096666,4008576,201254,4152858,201531,4099214,4
    43412,201826,4196141,201530,4151282,4198296,4202875,4099651,4193704)
    and person_id=@PERSON_ID and condition_era_end_date
    &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')) as dx
    )

    , hasdx as (select case when count(1)&gt;0 then 1 else 0 end as DIABETIC
    from diabdx 
    where latest &lt;='@INDEX_DATE'"
from diabdx
)
select case when sum(diabetic) > 0 then 1 else 0 end as DIABETIC from
(select diabetic from hasdiabmed
union all
select diabetic from hasdiablab
union all
select diabetic from hasdx) u
;
</Statement>
</Extension>
</MiningBuildTask>
<DataDictionary numberOfFields="8">
<DataField name="hazard" displayName="Cumulative Hazard" optype="continuous" dataType="double"/>
<DataField name="age" displayName="Age" optype="continuous" dataType="double">
  <Interval closure="closedClosed" leftMargin="30" rightMargin="74"/>
</DataField>
<DataField name="TCL" displayName="Total Cholesterol" optype="continuous" dataType="double"/>
<DataField name="HDL" displayName="HDL" optype="continuous" dataType="double"/>
<DataField name="HTNTRT" displayName="Antihypertensive medication use (y/n)" optype="categorical" dataType="boolean">
  <Value value="1"/>
  <Value value="0"/>
</DataField>
<DataField name="SBP" displayName="Systolic Blood Pressure" optype="continuous" dataType="double"/>
<DataField name="smoker" displayName="Smoker(y/n)" optype="categorical" dataType="integer">
  <Value value="1"/>
  <Value value="0"/>
</DataField>
<DataField name="diabetic" displayName="Diabetic(y/n)" optype="categorical" dataType="integer">
  <Value value="1"/>
  <Value value="0"/>
</DataField>
</DataDictionary>
<TransformationDictionary>
<DerivedField name="logAge" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="age"/>
  </Apply>
</DerivedField>
<DerivedField name="logTCL" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="TCL"/>
  </Apply>
</DerivedField>
<DerivedField name="logHDL" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="HDL"/>
  </Apply>
</DerivedField>
<DerivedField name="logSBP" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="SBP"/>
  </Apply>
</DerivedField>
<DerivedField name="logSBP_NOTTRT" dataType="double" optype="continuous"/>
<Apply function="if">
  <Apply function="equal" dataType="boolean">
    <FieldRef field="HTNTRT"/>
    < Constant dataType="integer">1</Constant>
  </Apply>
  <Constant dataType="integer">0</Constant>
</Apply>
</DerivedField>

<DerivedField name="logSBP_TRT" dataType="double" optype="continuous">
  <Apply function="if">
    <Apply function="equal" dataType="boolean">
      <FieldRef field="HTNTRT"/>
      <Constant dataType="integer">1</Constant>
    </Apply>
    <FieldRef field="logSBP"/>
    <Constant dataType="integer">0</Constant>
  </Apply>
</DerivedField>

<RegressionModel modelName="framingham10ycvdmen" functionName="regression"
  algorithmName="Cox proportional-hazards regression" isScorable="true">
  <MiningSchema>
    <MiningField name="hazard" usageType="predicted"/>
    <MiningField name="logAge" usageType="active"/>
    <MiningField name="logTCL" usageType="active"/>
    <MiningField name="logHDL" usageType="active"/>
    <MiningField name="logSBP_TRT" usageType="active"/>
    <MiningField name="logSBP_NOTTRT" usageType="active"/>
    <MiningField name="smoker" usageType="active"/>
    <MiningField name="diabetic" usageType="active"/>
  </MiningSchema>

  <Output>
    <OutputField name="hazard" optype="continuous" dataType="double"
      feature="predictedValue" isFinalResult="false"/>
    <OutputField name="hazard_ratio" optype="continuous" dataType="double"
      feature="transformedValue" isFinalResult="false">
      <Apply function="exp">
        <FieldRef field="hazard"/>
      </Apply>
    </OutputField>
    <OutputField name="risk" optype="continuous" dataType="double"
      feature="transformedValue" isFinalResult="true">
      <Apply function="-">
        <Constant>1.0</Constant>
        <Apply function="pow">
          <Constant>0.88936</Constant>
          <FieldRef field="hazard_ratio"/>
        </Apply>
      </Apply>
    </OutputField>
  </Output>

  <RegressionTable intercept="-23.9802">
    <NumericPredictor name="logAge" coefficient="3.06117"/>
    <NumericPredictor name="logTCL" coefficient="1.12370"/>
    <NumericPredictor name="logHDL" coefficient="-0.93263"/>
    <NumericPredictor name="logSBP_TRT" coefficient="1.99881"/>
    <NumericPredictor name="logSBP_NOTTRT" coefficient="1.93303"/>
    <CategoricalPredictor name="smoker" value="1" coefficient="0.65451"/>
    <CategoricalPredictor name="diabetic" value="1" coefficient="0.57367"/>
  </RegressionTable>
</RegressionModel>

```xml
<?xml version="1.0"?>
<PMML version="4.3" xmlns="http://www.dmg.org/PMML-4_3"
  <Header copyright="Copyright (c) 2017 Hamed Abedtash
  (hamed.abedtash@gmail.com)" description="Framingham 10-year risk of
cardiovascular disease for women">
    <Extension extender="omop">
      <OmopCdm version="5.1"/>
      <author name="Hamed Abedtash"/>
    </Extension>
    <Application name="OMOP-PMML Writer" version="0.1"/>
    <Timestamp>2017-04-30 16:52:26</Timestamp>
  </Header>
  <MiningBuildTask>
    <Extension name="age" extender="omop">
      <InputParameters>
        <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
optype="continuous" dataType="date"/>
        <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
optype="continuous" dataType="bigint"/>
        <InputParameter name="SCHEMA" displayName="Database schema"
dataType="string"/>
      </InputParameters>
      <Statement dialect="postgresql">
        select distinct extract(year from date '@INDEX_DATE') - year_of_birth as AGE from @SCHEMA.person where person_id=@PERSON_ID;
      </Statement>
    </Extension>
    <Extension name="TCL" extender="omop">
      <InputParameters>
        <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
optype="continuous" dataType="date"/>
        <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
optype="continuous" dataType="bigint"/>
        <InputParameter name="SCHEMA" displayName="Database schema"
dataType="string"/>
      </InputParameters>
      <Statement dialect="postgresql">
        with tcl as (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from @SCHEMA.measurement where measurement_concept_id in (3027114) and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD') and person_id=@PERSON_ID) meas
        join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE from @SCHEMA.measurement where measurement_concept_id in (3027114) and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD') and person_id=@PERSON_ID) measmax
        on meas.person_id=measmax.person_id and meas.measurement_date=measmax.MEAS_DATE)
        select value_as_number from tcl
        where to_date('@INDEX_DATE','YYYY-MM-DD') - MEAS_DATE &lt;=30
        order by meas_date desc limit 1;
      </Statement>
    </Extension>
  </MiningBuildTask>
</PMML>
```
<Extension name="HDL" extender="omop">
<InputParameters>
  <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
    optype="continuous" dataType="date"/>
  <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
    optype="continuous" dataType="bigint"/>
  <InputParameter name="SCHEMA" displayName="Database schema"
    dataType="string"/>
</InputParameters>
<Statement dialect="postgresql">
  with hdl as (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
    FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from @SCHEMA.measurement
      where measurement_concept_id in (3007070) and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD') and person_id=@PERSON_ID) meas
    join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE
      from @SCHEMA.measurement
      where measurement_concept_id in (3007070) and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD') and person_id=@PERSON_ID) measmax
    on meas.person_id=measmax.person_id and meas.measurement_date=measmax.MEAS_DATE)
  select value_as_number from hdl
  where to_date('@INDEX_DATE','YYYY-MM-DD') - MEAS_DATE < 30
  order by meas_date desc limit 1;
</Statement>
</Extension>

<Extension name="HTNTRT" extender="omop">
<InputParameters>
  <InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
    optype="continuous" dataType="date"/>
  <InputParameter name="PERSON_ID" displayName="OMOP Person ID"
    optype="continuous" dataType="bigint"/>
  <InputParameter name="SCHEMA" displayName="Database schema"
    dataType="string"/>
</InputParameters>
<Statement dialect="postgresql">
  with htnmed as (select distinct c2.concept_id
    from omop.concept c1,omop.concept_ancestor a,omop.concept c2
    where c1.concept_code in ('C02A', 'C02B', 'C02C', 'C02D', 'C02K', 'C09AA', 'C09CA', 'C07AA', 'C07AB', 'C07AG', 'C08CA', 'C08D', 'C03AA', 'C03BA', 'C03Bx', 'C03CA', 'C03DA', 'C03DB')
    and c1.vocabulary_id='ATC'
    and a.ancestor_concept_id=c1.concept_id
    and a.descendant_concept_id=c2.concept_id
    and c2.concept_class_id in ('Ingredient')
  ),
  rec as (select * from omop.drug_era
    where drug_concept_id in (select distinct concept_id from htnmed)
    and (to_date('@INDEX_DATE','YYYY-MM-DD') between drug_era_start_date and drug_era_end_date)
    and person_id=@PERSON_ID
  )
  select case when count(1)>0 then 1 else 0 end as HTNTRT from rec;
</Statement>
</Extension>

<Extension name="SBP" extender="omop">
<InputParameters>
</InputParameters>
</Extension>
<Statement dialect="postgresql">
  with sbp as (select distinct meas.person_id,
  meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
  FROM (select distinct person_id, measurement_concept_id,
  value_as_number, measurement_date from @SCHEMA.measurement
  where measurement_concept_id in
  (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
  and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD')
  and person_id=@PERSON_ID) meas
  join (select distinct person_id, measurement_concept_id,
  max(measurement_date) over(partition by
  person_id,measurement_concept_id) as MEAS_DATE
  from @SCHEMA.measurement
  where measurement_concept_id in
  (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
  and person_id=@PERSON_ID
  and measurement_date < to_date('@INDEX_DATE','YYYY-MM-DD'))
  measmax
  on meas.person_id=measmax.person_id and
  meas.measurement_date=measmax.MEAS_DATE
  select value_as_number from sbp
  where to_date('@INDEX_DATE','YYYY-MM-DD') - MEAS_DATE < 30
  order by meas_date desc limit 1;
</Statement>
</Extension>

<Extension name="smoker" extender="omop">
</Extension>

<Statement dialect="postgresql">
  with smoker as (SELECT DISTINCT person_id,observation_concept_id,
  observation_date
  FROM omop.observation
  where observation_concept_id IN
  (4218741,4246415,4276526,4052947,4052029,4052030,4218917,4298794,4270996,4043053,4144273,4144273,4043056)
  and person_id=@PERSON_ID and observation_date
  < to_date('@INDEX_DATE','YYYY-MM-DD'))
  )
  ,
  last_smoke as (select max(observation_date)over(partition by
  person_id) as latest
  from smoker
  )
  select case when count(1)>0 then 1 else 0 end as SMOKER
  from last_smoke where to_date('@INDEX_DATE','YYYY-MM-DD') - latest < 90;
</Statement>
</Extension>

<Extension name="diabetic" extender="omop">
</Extension>
<InputParameter name="INDEX_DATE" displayName="Index date (YYYY-MM-DD)"
    optype="continous" dataType="date"/>
<InputParameter name="PERSON_ID" displayName="OMOP Person ID"
    optype="continuous" dataType="bigint"/>
<InputParameter name="SCHEMA" displayName="Database schema"
    data="string"/>
</InputParameters>

<Statement dialect="postgresql">

with medrec as (select * from omop.drug_era
    where drug_concept_id in
        (1529331, 1530014, 44816332, 43013884, 19035533, 43526465, 19033498, 1594973, 4
        785829, 45774435, 45774751, 1587322, 19001409, 19059796, 1597756, 1560171, 190
        97821, 1559684, 19001441, 1000979, 19122121, 35604829, 35605670, 35602717, 1516
        976, 40056629, 1502905, 1588986, 1550023, 1531601, 1567198, 1544838, 40051349, 4
        0051350, 46221581, 42899447, 19090244, 19090229, 19090247, 19090249, 1513876, 19
        090180, 19013926, 1909187, 19013951, 1590165, 1586346, 1513849, 1562586, 190
        90204, 1513843, 40239216, 40170911, 44506754, 1503297, 1510202, 150282
6, 19033909, 1525215, 1517998, 1596977, 1547504, 40166035, 1580747, 150
2809, 1502855, 19042191, 1515249, 19090226, 19090221, 1586369, 19122137)
    and (to_date('@INDEX_DATE','YYYY-MM-DD') between
        drug_era_start_date and drug_era_end_date
        OR to_date('@INDEX_DATE','YYYY-MM-DD') between
        drug_era_end_date and drug_era_end_date+30)
    and person_id=@PERSON_ID)

, hasdiabmed as (select case when count(1)>0 then 1 else 0 end as DIABETIC
    from medrec
)

, lab AS (select distinct meas.person_id, meas.measurement_concept_id, meas.value_as_number, measmax.MEAS_DATE
    FROM (select distinct person_id, measurement_concept_id, value_as_number, measurement_date from omop.measurement
        where measurement_concept_id in (3037110)
        and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')
        and value_as_number&gt;=126 and person_id=@PERSON_ID) meas
    join (select distinct person_id, measurement_concept_id, max(measurement_date) over(partition by person_id,measurement_concept_id) as MEAS_DATE
        from omop.measurement
        where measurement_concept_id in (3037110) and
        person_id=@PERSON_ID
        and measurement_date &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')) measmax
    on meas.person_id=measmax.person_id and
    meas.measurement_date=measmax.MEAS_DATE
)

, hasdiablab as (select case when count(1)>0 then 1 else 0 end as DIABETIC
    from lab
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=60
)

, diabdx as (select max(condition_era_end_date) as latest
    from (SELECT DISTINCT *
        FROM omop.condition_era
        where condition_concept_id IN
            (201820, 45757674, 45757474, 4096666, 4008576, 201254, 4152858, 201531, 4099214
            , 443412, 201826, 4196141, 201530, 4151282, 4198296, 4200875, 4099651, 4193704)
        and person_id=@PERSON_ID and condition_era_end_date
        &lt;=to_date('@INDEX_DATE','YYYY-MM-DD')) as dx
)

, hasdx as (select case when count(1)>0 then 1 else 0 end as DIABETIC
    from diabdx
    where to_date('@INDEX_DATE','YYYY-MM-DD')-condition_era_end_date &lt;=30
)}
from diabdx
)
select case when sum(diabetic) > 0 then 1 else 0 end as DIABETIC from
(select diabetic from hasdiabmed
union all
select diabetic from hasdiablabs
union all
select diabetic from hasdx) u
;
</Statement>
</Extension>
</MiningBuildTask>
<DataDictionary numberOfFields="8">
<DataField name="hazard" displayName="Cumulative Hazard"
  optype="continuous" dataType="double"/>
<DataField name="age" displayName="Age" optype="continuous"
  dataType="double"/>
  <Interval closure="closedClosed" leftMargin="30" rightMargin="74"/>
</DataField>
<DataField name="TCL" displayName="Total Cholesterol" optype="continuous"
  dataType="double"/>
<DataField name="HDL" displayName="HDL" optype="continuous"
  dataType="double"/>
<DataField name="HTNTRT" displayName="Antihypertensive medication use (y/n)"
  optype="categorical" dataType="boolean"/>
  <Value value="1"/>
  <Value value="0"/>
</DataField>
<DataField name="SBP" displayName="Systolic Blood Pressure"
  optype="continuous" dataType="double"/>
<DataField name="smoker" displayName="Smoker(y/n)" optype="categorical"
  dataType="integer"/>
  <Value value="1"/>
  <Value value="0"/>
</DataField>
<DataField name="diabetic" displayName="Diabetic(y/n)" optype="categorical"
  dataType="integer"/>
  <Value value="1"/>
  <Value value="0"/>
</DataField>
</DataDictionary>
<TransformationDictionary>
<DerivedField name="logAge" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="age"/>
  </Apply>
</DerivedField>
<DerivedField name="logTCL" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="TCL"/>
  </Apply>
</DerivedField>
<DerivedField name="logHDL" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="HDL"/>
  </Apply>
</DerivedField>
<DerivedField name="logSBP" dataType="double" optype="continuous">
  <Apply function="ln">
    <FieldRef field="SBP"/>
  </Apply>
</DerivedField>
<DerivedField name="logSBP_NOTTRT" dataType="double" optype="continuous">
<Apply function="if">
  <Apply function="equal" dataType="boolean">
    <FieldRef field="HTNTRT"/>
    <Constant dataType="integer">1</Constant>
  </Apply>
  <FieldRef field="logSBP"/>
</Apply>
</DerivedField>

<DerivedField name="logSBP_TRT" dataType="double" optype="continuous">
  <Apply function="if">
    <Apply function="equal" dataType="boolean">
      <FieldRef field="HTNTRT"/>
      <Constant dataType="integer">1</Constant>
    </Apply>
    <FieldRef field="logSBP"/>
    <Constant dataType="integer">0</Constant>
  </Apply>
</DerivedField>

<DerivedField>
</TransformationDictionary>
<RegressionModel modelName="framingham10ycvdmen" functionName="regression" algorithmName="Cox proportional-hazards regression" isScorable="true">
  <MiningSchema>
    <MiningField name="hazard" usageType="predicted"/>
    <MiningField name="logAge" usageType="active"/>
    <MiningField name="logTCL" usageType="active"/>
    <MiningField name="logHDL" usageType="active"/>
    <MiningField name="logSBP_TRT" usageType="active"/>
    <MiningField name="logSBP_NOTTRT" usageType="active"/>
    <MiningField name="smoker" usageType="active"/>
    <MiningField name="diabetic" usageType="active"/>
  </MiningSchema>
  <Output>
    <OutputField name="hazard" optype="continuous" dataType="double" feature="predictedValue" isFinalResult="false"/>
    <OutputField name="hazard_ratio" optype="continuous" dataType="double" feature="transformedValue" isFinalResult="false">
      <Apply function="exp">
        <FieldRef field="hazard"/>
      </Apply>
    </OutputField>
    <OutputField name="risk" optype="continuous" dataType="double" feature="transformedValue" isFinalResult="true">
      <Apply function="-">
        <Constant>1.0</Constant>
        <Apply function="pow">
          <Constant>0.95012</Constant>
          <FieldRef field="hazard_ratio"/>
        </Apply>
      </Apply>
    </OutputField>
  </Output>
  <RegressionTable intercept="-26.1931">
    <NumericPredictor name="logAge" coefficient="2.32888"/>
    <NumericPredictor name="logTCL" coefficient="1.20904"/>
    <NumericPredictor name="logHDL" coefficient="-0.70833"/>
    <NumericPredictor name="logSBP_TRT" coefficient="2.82263"/>
    <NumericPredictor name="logSBP_NOTTRT" coefficient="2.76157"/>
    <CategoricalPredictor name="smoker" value="1" coefficient="0.52873"/>
    <CategoricalPredictor name="diabetic" value="1" coefficient="0.69154"/>
  </RegressionTable>
</RegressionModel>
</PMML>
### Appendix 11

Estimated 10-year risk score of cardiovascular disease for 56 records of 8 unique patients. F: Female, M: Male, TCL: Total cholesterol level, HDL: High-density lipoprotein cholesterol level, SBP: Systolic blood pressure.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Gender</th>
<th>Age</th>
<th>TCL</th>
<th>HDL</th>
<th>Antihypertensive Use (1: Yes, 0: No)</th>
<th>SBP</th>
<th>Smoker (1: Yes, 0: No)</th>
<th>Diabetic (1: Yes, 0: No)</th>
<th>Risk Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>F</td>
<td>31</td>
<td>161</td>
<td>42</td>
<td>1</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>2.07%</td>
</tr>
<tr>
<td>Patient 1</td>
<td>F</td>
<td>31</td>
<td>161</td>
<td>42</td>
<td>1</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>2.07%</td>
</tr>
<tr>
<td>Patient 2</td>
<td>M</td>
<td>44</td>
<td>219</td>
<td>47</td>
<td>0</td>
<td>162</td>
<td>0</td>
<td>0</td>
<td>10.10%</td>
</tr>
<tr>
<td>Patient 3</td>
<td>F</td>
<td>30</td>
<td>184</td>
<td>31</td>
<td>0</td>
<td>106</td>
<td>0</td>
<td>0</td>
<td>1.11%</td>
</tr>
<tr>
<td>Patient 3</td>
<td>F</td>
<td>30</td>
<td>184</td>
<td>31</td>
<td>0</td>
<td>108</td>
<td>0</td>
<td>0</td>
<td>1.17%</td>
</tr>
<tr>
<td>Patient 4</td>
<td>F</td>
<td>52</td>
<td>200</td>
<td>35</td>
<td>0</td>
<td>140.5</td>
<td>0</td>
<td>0</td>
<td>8.51%</td>
</tr>
<tr>
<td>Patient 5</td>
<td>M</td>
<td>67</td>
<td>143</td>
<td>36</td>
<td>1</td>
<td>134.9</td>
<td>0</td>
<td>0</td>
<td>25.69%</td>
</tr>
<tr>
<td>Patient 5</td>
<td>M</td>
<td>67</td>
<td>143</td>
<td>36</td>
<td>1</td>
<td>134.9</td>
<td>0</td>
<td>0</td>
<td>25.69%</td>
</tr>
<tr>
<td>Patient 6</td>
<td>F</td>
<td>49</td>
<td>170</td>
<td>47.5</td>
<td>1</td>
<td>149</td>
<td>0</td>
<td>1</td>
<td>36.37%</td>
</tr>
<tr>
<td>Patient 6</td>
<td>F</td>
<td>49</td>
<td>170</td>
<td>47.5</td>
<td>1</td>
<td>149</td>
<td>0</td>
<td>1</td>
<td>36.37%</td>
</tr>
<tr>
<td>Patient 6</td>
<td>F</td>
<td>51</td>
<td>170</td>
<td>42.7</td>
<td>0</td>
<td>149</td>
<td>0</td>
<td>1</td>
<td>35.79%</td>
</tr>
<tr>
<td>Patient 7</td>
<td>M</td>
<td>64</td>
<td>87</td>
<td>25</td>
<td>1</td>
<td>149.5</td>
<td>0</td>
<td>1</td>
<td>35.79%</td>
</tr>
<tr>
<td>Patient 7</td>
<td>M</td>
<td>64</td>
<td>87</td>
<td>25</td>
<td>1</td>
<td>148</td>
<td>0</td>
<td>1</td>
<td>35.79%</td>
</tr>
<tr>
<td>Patient 7</td>
<td>M</td>
<td>64</td>
<td>87</td>
<td>25</td>
<td>1</td>
<td>148</td>
<td>0</td>
<td>1</td>
<td>35.79%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient ID</td>
<td>Gender</td>
<td>Age</td>
<td>TCL</td>
<td>HDL</td>
<td>Antihypertensive Use (1: Yes, 0: No)</td>
<td>SBP</td>
<td>Smoker (1: Yes, 0: No)</td>
<td>Diabetic (1: Yes, 0: No)</td>
<td>Risk Score</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------------------------------------</td>
<td>-----</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>1</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>5.32%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>3.98%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>124.5</td>
<td>0</td>
<td>0</td>
<td>3.54%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>49</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>124.5</td>
<td>0</td>
<td>0</td>
<td>3.54%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>124.5</td>
<td>0</td>
<td>0</td>
<td>3.71%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>124.5</td>
<td>0</td>
<td>0</td>
<td>3.71%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>124.5</td>
<td>0</td>
<td>0</td>
<td>3.64%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.24%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>35</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.24%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.24%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.24%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.39%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>3.39%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>4.51%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>4.51%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>50</td>
<td>140</td>
<td>40</td>
<td>0</td>
<td>119.3</td>
<td>0</td>
<td>0</td>
<td>4.51%</td>
</tr>
<tr>
<td>Patient ID</td>
<td>Gender</td>
<td>Age</td>
<td>TCL</td>
<td>HDL</td>
<td>Antihypertensive Use (1: Yes, 0: No)</td>
<td>SBP</td>
<td>Smoker (1: Yes, 0: No)</td>
<td>Diabetic (1: Yes, 0: No)</td>
<td>Risk Score</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-------------------------------------</td>
<td>-----</td>
<td>------------------------</td>
<td>-------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>52</td>
<td>149</td>
<td>40</td>
<td>1</td>
<td>116.3</td>
<td>0</td>
<td>0</td>
<td>4.40%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>52</td>
<td>149</td>
<td>40</td>
<td>1</td>
<td>116.3</td>
<td>0</td>
<td>0</td>
<td>4.40%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>52</td>
<td>149</td>
<td>40</td>
<td>1</td>
<td>116.3</td>
<td>0</td>
<td>0</td>
<td>4.40%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>52</td>
<td>149</td>
<td>40</td>
<td>1</td>
<td>116.3</td>
<td>0</td>
<td>0</td>
<td>4.40%</td>
</tr>
<tr>
<td>Patient 8</td>
<td>F</td>
<td>52</td>
<td>143</td>
<td>43</td>
<td>1</td>
<td>114</td>
<td>0</td>
<td>0</td>
<td>3.77%</td>
</tr>
</tbody>
</table>
REFERENCES


24. The Office of the National Coordinator for Health Information Technology (ONC) Office of the Secretary, United States Department of Health and Human Services. Report to congress: Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information. 2014.


53. Smithburger PL, Kane-Gill SL, Benedict NJ, Falcione BA, Seybert AL. Grading the severity of drug-drug interactions in the intensive care unit: a


63. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current cardiology reports. 2014;16(1):441.


86. CDC National Center for Immunization and Respiratory Diseases. CPT Codes Mapped to CVX Codes [Internet]. 2017 [April 1, 2017]. Available from: https://www2a.cdc.gov/vaccines/IIS/IISStandards/vaccines.asp?rpt=cvx.


119. Kazemzadeh RS, Sartipi K, editors. Interoperability of Data and Knowledge in Distributed Health Care Systems. 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05); 2005 0-0 0.

120. Kazemzadeh RS, Sartipi K, editors. Incoporating Data Mining Applications into Clinical Guidelines. 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06); 2006 0-0 0.


CURRICULUM VITAE

Hamed Abedtash

Education

Doctor of Philosophy (PhD) of Informatics August 2012-July 2017
Indiana University (IUPUI), Indianapolis, IN, USA

Doctor of Pharmacy (PharmD) September 2001-June 2009
Shiraz University of Medical Sciences (SUMS), Shiraz, Iran

Honors, Awards

Travel Fellowship November 2015
Indiana University (IUPUI) Graduate Office

IUPUI Elite 50 Award April 2015
Indiana University (IUPUI)

Indiana HIMSS 2013 Scholarship November 2013
Health Information Management Systems Society (HIMSS), Indiana Chapter

IUPUI Educational Enhancement Grant November 2013
Indiana University (IUPUI) Graduate Office

Award in Recognition of Excellence in Research 2008
SUMS

Conference Travel Grant 2008
SUMS Gifted and Talented Office

Professional Experience

Eli Lilly and Company, Indianapolis, IN September 2016-June 2017
Real-World Evidence (RWE) IT Developer at Global Patient Outcome and RWE (GPORWE)

IUPUI School of Informatics and Computing, Indianapolis, IN Aug. 2012-May 2017
Research Assistant, Teaching Assistant, and Instructor in Health Informatics Program

Eli Lilly and Company, Indianapolis, IN Jan. 2016-May 2016
Academic Researcher, Research IT

Eli Lilly and Company, Indianapolis, IN May 2015-August 2015
Intern at Data Sciences and Solutions

Regenstrief Institute, Inc., Indianapolis, IN May 2014-August 2014
Intern, Information Technology and Application Developer

Community Pharmacist, Tehran, Iran 2009-2012
Community Pharmacist, Shiraz, Iran 2007-2009
Shiraz University of Medical Sciences, Shiraz, Iran  
Hospital Pharmacy Intern  
2006-2009

Shiraz University of Medical Sciences, Shiraz, Iran  
Clinical Pharmacy Intern  
2006-2009

Shiraz University of Medical Sciences, Shiraz, Iran  
Pharmacy Technician  
2004-2006

Publications

Peer-reviewed Journal Articles and Conference Proceedings


Peer-reviewed Conference Abstract Podiums


Peer-reviewed Conference Poster Presentations

Non-OMOP-based Electronic Health Records. OHDSI 2016, Washington, DC, USA.


**Abedtash H**, Faridi P, Montaseri H (2004). *Calculating kinetic mechanism via MS Excel®*. 5th Research Conference of SUMS School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.

Karimzadeh I, Abedtash H, Mohagheghzadeh A, Ghasemi Y (2004). *A review on molecular modeling and inhibition effect of flavonoids on cytochrome P450 enzyme*. 5th Research Conference of SUMS School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.

*Registered Gene in PubMed Nucleotide Database*