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.

Hamed Abedtash

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

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

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

API	Application programming interface
CART	Classification and regression tree
CCD	Continuity of care
C-CDA	Consolidated clinical document architecture
CDA	Clinical document architecture
CDC	Centers for disease control and prevention
CDM	Common data model
CDS	Clinical decision support
CPT	Current procedural terminology
DDD	Defined daily dose
EHR	Electronic medical records
ETL	Extract transform load
HCPCS	Healthcare common procedure coding system
HDL	High density lipoprotein
HIE	Health information exchange
HL7	Health level seven
ICD-9-CM	The international classification of diseases, ninth revision, clinical modification
ICD-10-CM	The international classification of diseases, tenth revision, clinical modification
IT	Information technology
LOINC	Logical observation identifiers names and codes
NDC	National drug code
NDF-RT	the national drug file – reference terminology
OMOP	Observational medical outcomes partnership
O-PMML	OMOP-compliant predictive model markup language
PHRST	Personalized health risk scoring tool
PMML	Predictive model markup language
REST	Representational state transfer

RIM	Reference information model
ROC	Receiver operating characteristic
SBD	Semantic brand drug
SBP	Systolic blood pressure
SCD	Semantic clinical drug
SNOMED CT	Systematized nomenclature of medicine – clinical terms
SVM	Support vector machine
VA	Veterans affairs
WHO	World health organization
XML	Extensible markup language

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

(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 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.

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 of 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 venders 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.

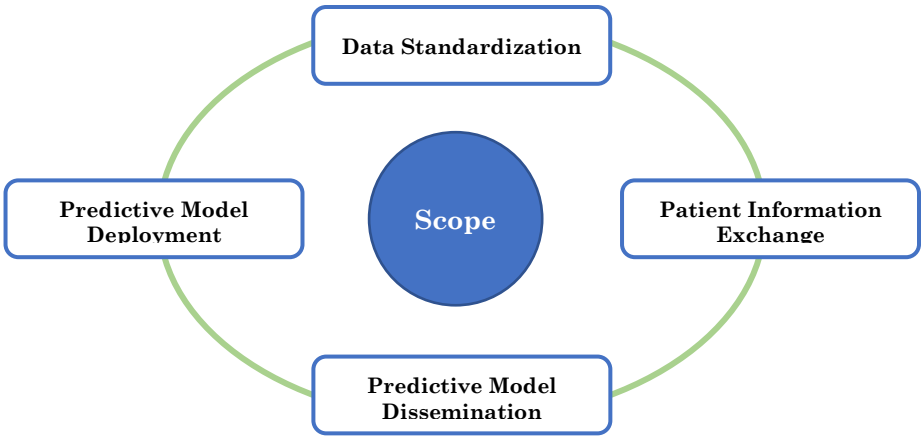

Figure 1. The scope of the project is to deploy predictive models on patient data exchanged via HL7 messaging standard. The project involves standardizing transferred patient information into OMOP CDM, introducing a new standard for disseminating OMOP-based predictive models, and building a pilot solution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

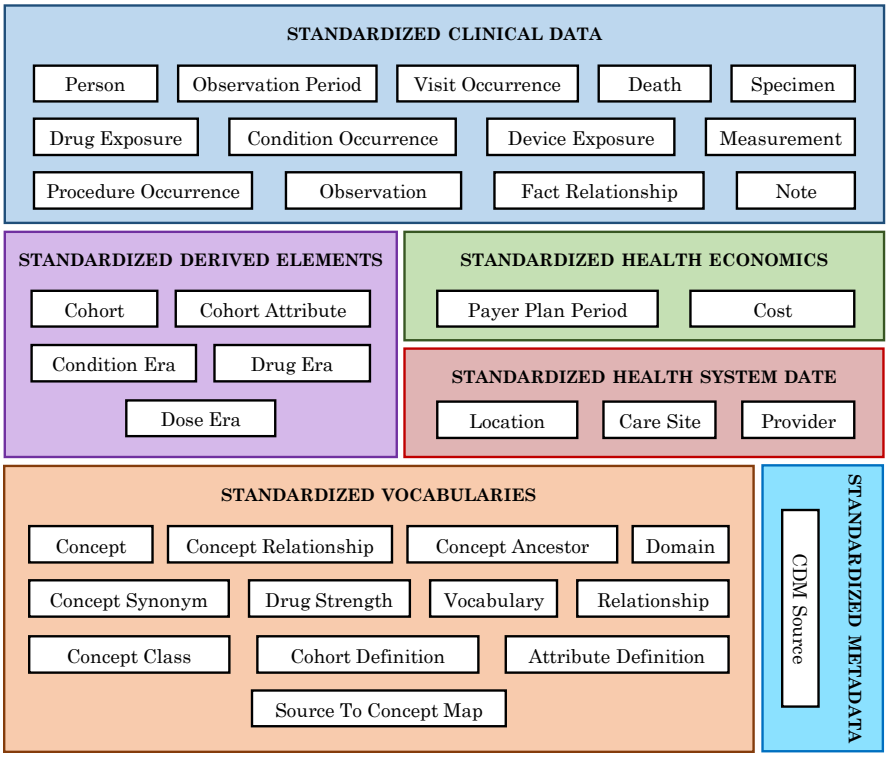

Figure 2. OMOP Common Data Model Version 5.1 conceptual model (83).

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).

Section Name	Object Identifier (OID)	Description
US Realm Header	2.16.840.1.113883.10.20.22.1.1	Describes patient's demographics (e.g., gender, race, marital status) and common administrative information (e.g., patient's name and address, author, provider).
Advance Directives Section (entries optional)	2.16.840.1.113883.10.20.22.2.21	Describes patient's directives of living wills, resuscitation status, CPR orders, and healthcare proxies.
Allergies Section (entries required)	2.16.840.1.113883.10.20.22.2.6.1	Describes allergies to food, medicines, and other substances (e.g., latex), and reported adverse reactions.
Encounters Section (entries optional)	2.16.840.1.113883.10.20.22.2.22	Lists the history of encounters between patient and healthcare providers for diagnosis, treatment, or evaluating medical condition purposes.
Family History Section (entries optional)	2.16.840.1.113883.10.20.22.2.15	Describes health risk factors of patient's biologic parents that may affect the patient's risk of medical condition occurrences.
Functional Status Section (entries optional)	2.16.840.1.113883.10.20.22.2.14	Describes findings and evaluation results of patient's physical function, including basic and instrumental activities of daily living (ADLs).
Immunizations Section (entries optional)	2.16.840.1.113883.10.20.22.2.2.1	Lists history patient's immunization history.
Medical Equipment Section (entries optional)	2.16.840.1.113883.10.20.22.2.23	Lists external, implanted, or durable medical devices used to treat patient's medical condition or uphold the health status.
Medications Section (entries required)	2.16.840.1.113883.10.20.22.2.1.1	Lists patient's history of prescribed and dispensed medications.

Section Name	Object Identifier (OID)	Description
Payers Section (entries optional)	2.16.840.1.113883.10.20.22.2.18	Lists all financial, payment, insurance, and health plan coverage records pertinent to the healthcare services provided to the patient.
Plan of Care Section (entries optional)	2.16.840.1.113883.10.20.22.2.10	Lists all patient's ongoing, incomplete, pending, or unfulfilled services, orders, encounters, and procedures to be carried out in future.
Problem Section (entries required)	2.16.840.1.113883.10.20.22.2.5.1	Lists patient's history of clinical conditions and problems.
Procedures Section (entries required)	2.16.840.1.113883.10.20.22.2.7.1	Lists patient's history of surgical, therapeutics, and diagnostic procedures.
Results Section (entries required)	2.16.840.1.113883.10.20.22.2.3.1	Lists all results of laboratory tests, diagnostic procedures, and imaging procedures.
Social History Section (entries optional)	2.16.840.1.113883.10.20.22.2.17	Describes patient's social observations, e.g., smoking status, pregnancy history, tobacco use, cultural and religious practices, etc.
Vital Signs Section (entries optional)	2.16.840.1.113883.10.20.22.2.4	Describes patient's vital signs, e.g., blood pressure, temperature, pulse rate, height, weight, etc.

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.

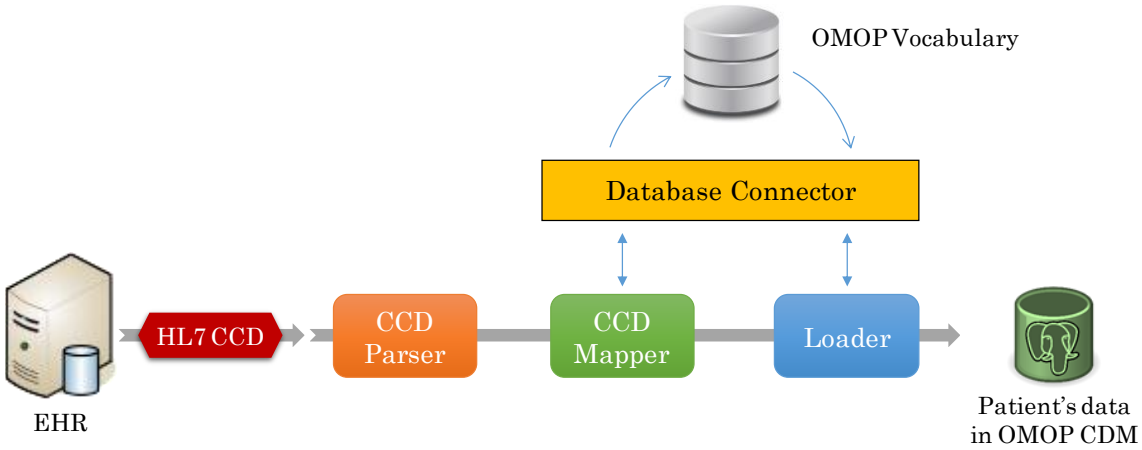

Figure 3. A schematic of CCD-TO-OMOP package

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

Destination Table Field	Matching C-CDA section	Applied Rule
person_id	–	Randomly generated unique identification per C-CDA document
gender_concept_id	–	OMOP standard Concept ID of *Gender Source Value*
year_of_birth	US Realm Header: *birthTime*	Extracted year part of the date of birth
month_of_birth	US Realm Header: *birthTime*	Extracted month part of the date of birth
day_of_birth	US Realm Header: *birthTime*	Extracted day part of the date of birth
race_concept_id	–	OMOP standard Concept ID of *Race Source Value*
ethnicity_concept_id	–	OMOP standard Concept ID of *Ethnicity Source Value*
gender_source_value	US Realm Header: *administrativeGenderCode*	
gender_source_concept_id	–	OMOP Concept ID of *Gender Source Value* If *Domain ID* = *"Gender"* and *Invalid Reason* = *Null*
race_source_value	US Realm Header: *raceCode*	
race_source_concept_id	–	OMOP Concept ID of *Race Source Value* If *Domain ID* = *"Race"* and *Invalid Reason* = *Null*
ethnicity_source_value	US Realm Header: *ethnicGroupCode*	
ethnicity_source_concept_id	–	OMOP concept id of *Ethnicity Source Value* If *Domain ID* = *"Gender"* and *Invalid Reason* = *Null*

Table 3. Matching OMOP concepts of HL7 administrative gender codes

HL7 V3 Administrative Gender		OMOP Standard Concept	
Code	**Description**	**Concept ID**	**Concept Name**
F	Female	8532	Female
M	Male	8507	Male
UN	Undifferentiated	8551	Unknown

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

CDC Race and Ethnicity Values		OMOP Standard Concept	
Code	**Description**	**Concept ID**	**Concept Name**
1002-5	American Indian or Alaska Native	8657	American Indian or Alaska Native
2028-9	Asian	8515	Asian
2054-5	Black or African American	8516	Black or African American
2076-8	Native Hawaiian or Other Pacific Islander	8557	Native Hawaiian or Other Pacific Islander
2106-3	White	8527	White
2135-2	Hispanic or Latino	38003563	Hispanic or Latino
2186-5	Not Hispanic or Latino	38003564	Not Hispanic or Latino

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

Destination Table Field	Applied Rule
observation_period_id	Auto-numbered unique identifier for each observation period
person_id	The patient's identifier from *Person* table
observation_period_start_date	The start date of consolidated observation period
observation_period_end_date	The end date of consolidated observation period
period_type_concept_id	44814725 (Period inferred by algorithm)

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.

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

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.

Destination Table Field	Matching C-CDA section	Applied Rule
condition_occurrence_id	–	Auto-numbered unique identifier for each condition occurrence
person_id	–	The patient's identifier from *Person* table
condition_concept_id	–	OMOP standard Concept ID of *Condition Source Value*
condition_start_date	Allergies Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: *effectiveTime/low* Problems Problem Section: Problem Concern Act: Problem Observation: *effectiveTime/low*	Allergies Only if it is an ongoing allergy, i.e., *Status Code* = *"active"* under *Allergy Problem Act*
condition_end_date	Problems Problem Section: Problem Concern Act: Problem Observation: *effectiveTime/high*	Allergies Null; an active allergy concern does not have end date. Problems If an ongoing concern, i.e., *Status Code* = *"Active"* under *Problem Concern Act*, then end date is Null. If resolved problem concern, i.e., *Status Code* = *"completed"* under *Problem Concern Act*, then the reported end date.
condition_type_concept_id	–	38000245 (EHR problem list entry)

Destination Table Field	Matching C-CDA section	Applied Rule
visit_occurrence_id	–	The identifier of corresponding visit occurrence from *Visit Occurrence* table. Null if *Condition Start Date* did not match with any visit occurrence.
condition_source_value	Allergies Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: *code*	If the source code is a condition concept (i.e., *Domain ID = "Condition"* and *Invalid Reason = Null*). Allergies
	Participating agent in allergies Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: Product: Product Detail: *playingEntity/code*	No space in CDM table to record the participating agent. The SNOMED code of the concept that represents the allergy or adverse event finding of the allergen is recorded (e.g., allergy to penicillin).
	Problems Problem Section: Problem Concern Act: Problem Observation: *code*	Problems The source code of reported condition or diagnosis
condition_source_concept_id	–	Allergies OMOP concept id of *Condition Source Value* which is the same as *Condition Concept ID* Problems OMOP concept id of *Condition Source Value*

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 8. Applied rules to build *Condition Era* table

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

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

Destination Table Field	Matching C-CDA section	Applied Rule
procedure_occurrence_id	–	Auto-numbered unique identifier for each procedure occurrence
person_id	–	The patient's identifier from *Person* table
procedure_concept_id	–	OMOP standard Concept ID of *Procedure Source Value*
procedure_date	<u>Procedures</u> Procedures Section: Procedure Activity Procedure: *effectiveTime* Procedures Section: Procedure Activity Observation: *effectiveTime* Procedures Section: Procedure Activity Act: *effectiveTime* <u>Problems</u> Problem Section: Problem Concern Act: Problem Observation: *effectiveTime*	The date when procedure occurred If *stausCode* = *"completed"*
procedure_type_concept_id	–	<u>Procedures</u> Procedure Activity Procedure: 38000275 (EHR order list entry) Procedure Activity Observation: 38000275 (EHR order list entry)

Destination Table Field	Matching C-CDA section	Applied Rule
		Procedure Activity Act: 38000275 (EHR order list entry) <u>Problems</u> Problem Observation: 38000245 (EHR problem list entry)
visit_occurrence_id	–	The identifier of corresponding visit occurrence from *Visit Occurrence* table. Null if *Procedure Date* did not match with any visit occurrence.
procedure_source_value	<u>Procedures</u> Procedures Section: Procedure Activity Procedure: *code* Procedures Section: Procedure Activity Observation: *code* Procedures Section: Procedure Activity Act: *code* <u>Problems</u> Problem Section: Problem Concern Act: Problem Observation: *code*	If the source code is a procedure concept (i.e., *Domain ID* = *"Procedure"* and *Invalid Reason* = *Null*), AND the service was completed (i.e., *stausCode* = *"completed"*).
procedure_source_concept_id	–	OMOP concept id of *Procedure Source Value*

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., *substanceAdministration[@moodCode] = "EVN"* and *substanceAdministration/statusCode = "completed"*.
2. *Immunization Activity* must have *substanceAdministration[@negationInd] = "false"*, indicating that immunization was truly administered.
3. Medication entries must have *Supply Activity* indicating that the medication was dispensed (*supply[@moodCode] = "EVN"* and *supply/statusCode = "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:

$$Days\ of\ supply\ = \frac{Quantity\ of\ supplied\ drug}{Daily\ dose\ interval \times Quantity\ of\ drug\ per\ dose} \quad (1)$$

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 it 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 10. Corresponding quantity units to CCD quantity values in *Drug Exposure* table

Dose form	**Quantity unit in CCD**	**Quantity unit in *Drug Exposure* table**
Aerosol	aero	Actuation
inhalation solution	nebu	Milliliter
Inhaler	aepb, inhaler, inhalatn	Actuation
Injectable solution	amp, soln, vial	Milliliter
Injectable suspension	ml, vial	Milliliter
Oral capsule	caps, capsule, bottle, cpdr	Capsule
Oral solution	bottle, ml, soln	Milliliter
Oral suspension	bottle, ml, powder	Milliliter
Oral tablet	tabs, tablet, bottle, chew tab, tbdp	Tablet
Patch	patch	Patch
Prefilled syringe	syringe	Milliliter
Suppository	supp	Suppository
Topical cream	g, cream, gm, jar, tube	Gram
Topical gel	g, gel, jar, tube	Gram
Topical ointment	g, oint, jar, tube	Gram
vaginal cream	tube	Gram

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

Destination Table Field	Matching C-CDA section	Applied Rule
drug_exposure_id	–	Auto-numbered unique identifier for each drug exposure occurrence
person_id	–	The patient's identifier from *Person* table
drug_concept_id	–	OMOP standard Concept ID of *Drug Source Value*
drug_exposure_start_date	Immunizations Immunizations Section: Immunization Activity: *effectiveTime[@value]* Medications Medications Section: Medication Activity: *effectiveTime/low*	Medication *effectiveTime[@xsi:type]* = *"IVL_TS"*
drug_exposure_end_date	–	Immunization: Null Medication The expiry date was calculated by adding days of drug supply to the start date.
drug_type_concept_id	–	38000177 (Prescription written)
refills	Medications Medications Section: Medication Activity: *repeatNumber*	Immunization: Null
quantity	Medications Medications Section: Medication Activity: Supply Activity: *quantity*	

Destination Table Field	Matching C-CDA section	Applied Rule
days_supply	–	Immunizations: Null Medications: Using Days of Supply equation (1)
sig	Medications Medications Section: Medication Activity: *text*	
route_concept_id	–	OMOP Concept ID of route of administration based on *Drug Concept ID* looked up in *Concept Relationship* table.
lot_number	Immunizations Immunizations Section: Immunization Activity: Immunization Medication Information: *manufacturedMaterial/lotNumberText* Medications No lot number	
visit_occurrence_id	–	The identifier of corresponding visit occurrence from *Visit Occurrence* table. Null if *Drug Exposure Start Date* did not match with any visit occurrence.
drug_source_value	Immunizations Immunizations Section: Immunization Activity: Immunization Medication Information: *manufacturedMaterial/code* Medications	Immunization 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*). Medication

Destination Table Field	Matching C-CDA section	Applied Rule
	Medications Section: Medication Activity: Medication Information: *manufacturedMaterial/code*	If *Medication Activity* was completed, AND *Medication Dispense* exists, AND the source code was a drug concept (i.e., *Domain Id = "Drug",* AND *Invalid Reason = Null*).
drug_source_concept_id	–	OMOP Concept ID of *Drug Source Value*
route_source_value	<u>Immunizations</u> Immunizations Section: Immunization Activity: *routeCode[@code]* <u>Medications</u> Medications Section: Medication Activity: *routeCode[@code]*	If the route code does not exist, enter Null.

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

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

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

Destination Table Field	Matching C-CDA section	Applied Rule
measurement_id	–	Auto-numbered unique identifier for each measurement
person_id	–	The patient's identifier from *Person* table
measurement_concept_id	–	OMOP standard Concept ID of *Measurement Source Value*
measurement_date	Vital signs Vital Signs Section: Vital Signs Organizer: Vital Sign Observation: *effectiveTime[@value]* Results Results Section: Result Organizer: Result Observation: *effectiveTime[@value]* Problems Problem Section: Problem Concern Act: *effectiveTime/low*	
measurement_type_concept_id	–	Vital signs: 44818701 (From physical examination) Results: 44818702 (Lab result) Problems: 38000245 (EHR problem list entry)
operator_concept_id	–	Separate operator (<, >, ≤, or ≥) from *Value Source Value* if represents a value as number. Then, map to OMOP standard concept of *Domain ID = 'Meas Value Operator'*.
value_as_number	Vital signs	Vital signs

Destination Table Field	Matching C-CDA section	Applied Rule
	Vital Signs Section: Vital Signs Organizer: Vital Sign Observation: *value[@value]*	If the value is a physical quantity or real number, i.e., *value[@xsi:type] = "PQ" or "REAL"*.
	<u>Results</u> Results Section: Result Organizer: Result Observation: *value[@value]*	<u>Results</u> If the value is a physical quantity, i.e., *value[@xsi:type] = "PQ"*.
value_as_concept_id	–	If *Value Source Value* is a concept code form a coding system, e.g., SNOMED CT
unit_concept_id	–	OMOP Concept ID of *Unit Source Value*
visit_occurrence_id	–	The identifier of corresponding visit occurrence from *Visit Occurrence* table. Null if *Measurement Date* did not match with any visit occurrence.
measurement_source_value	<u>Vital signs</u> Vital Signs Section: Vital Signs Organizer: Vital Sign Observation: *code*	<u>Vital signs</u> If *Vital Sign Observation* was completed, AND the source code was a measurement concept (i.e., *Domain ID = "Measurement"* AND *Invalid Reason = Null*).
	<u>Results</u> Results Section: Result Organizer: Result Observation: *code* <u>Problems</u> Problem Section: Problem Concern Act: *code*	<u>Results</u> If *Result Observation* was completed, AND the source code was a measurement concept (i.e., *Domain ID = "Measurement"* AND *Invalid Reason = Null*).
measurement_source_concept_id	–	OMOP Concept ID of *Measurement Source Value*
unit_source_value	<u>Vital signs</u> Vital Signs Section: Vital Signs Organizer: Vital Sign Observation: *value[@unit]*	<u>Vital signs</u> If the value is a physical quantity, i.e., *value[@xsi:type] = "PQ"*.

Destination Table Field	Matching C-CDA section	Applied Rule
	Results Results Section: Result Organizer: Result Observation: *value[@unit]*	Results If the value is a physical quantity, i.e., *value[@xsi:type] = "PQ"*.
value_source_value	Vital signs Vital Signs Section: Vital Signs Organizer: Vital Sign Observation: *value[@value]*	
	Results Results Section: Result Organizer: Result Observation: *value[@value]*	
	Problems Problem Section: Problem Concern Act: *code*	

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 14. Matching OMOP concepts of HL7 smoking status value set

HL7 smoking status value		OMOP Standard Concept	
Code	**Code Name**	**Concept ID**	**Concept Name**
8517006	Former smoker	4310250	Ex-smoker
77176002	Smoker, current status unknown	4298794	Smoker
266927001	Unknown if ever smoked	4141786	Tobacco smoking consumption unknown
449868002	Current every day smoker	42709996	Smokes tobacco daily
266919005	Never smoker (Never Smoked)	4144272	Never smoked tobacco
428041000124106	Current some day smoker	4298794	Smoker

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

Destination Table Field	Matching C-CDA section	Applied Rule
observation_id	–	Auto-numbered unique identifier for each observation
person_id	–	The patient's identifier from *Person* table
observation_concept_id	–	OMOP standard Concept ID of *Observation Source Value*
observation_date	Smoking status Social History Section: Smoking Status Observation: *effectiveTime/low* or *effectiveTime/high* Problems Problem Section: Problem Concern Act: Problem Observation: *effectiveTime/low* Allergies Allergies Section: Allergy Problem Act: *effectiveTime/low*	Smoking status If this is an active or unknow smoking behavior, the only existing time point is stored (i.e., *effectiveTime/low*). If this is a past smoking behavior (former smoker). Both start and end of smoking behavior exist, but the end point shows when smoking behavior changed (i.e., *effectiveTime/high*). Problems 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.
observation_type_concept_id	–	Smoking: 44814721 (Patient reported) Problems: 38000276 (Problem list from EHR) Allergies: 38000280 (Observation recorded from EHR)

Destination Table Field	Matching C-CDA section	Applied Rule
visit_occurrence_id	–	The identifier of corresponding visit occurrence from *Visit Occurrence* table. Null if *Condition Start Date* did not match with any visit occurrence.
observation_source_value	Smoking status: Social History Section: Smoking Status Observation: *value[@code]* Problems Problem Section: Problem Concern Act: Problem Observation: *code* Allergies Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: *code* Participating agent in allergies Allergies Section: Allergy Problem Act: Allergy-Intolerance Observation: Product: Product Detail: *playingEntity/code*	If the source code is an observation concept (i.e., *Domain ID = "Observation"* and *Invalid Reason = Null*). Problems The observation was truly observed (*negationInd = "false"* under *Problem Observation*). Allergies OMOP concept id of *Condition Source Value* considering the participating allergen under *Allergy-Intolerance Observation*
observation_source_concept_id	–	OMOP Concept ID of *Observation Source Value*

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. To test the

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 16. An excerpt of records in *Condition Era* table.

Condition Concept ID	Condition Name	Condition Era Start Date	Condition Era End Date
4008576	Diabetes mellitus without complication	5/12/2012	5/12/2012
4008576	Diabetes mellitus without complication	12/28/2012	12/29/2012
4009042	Acute abscess of nasal sinus	11/10/2012	11/11/2012
4010024	Localized abdominal pain	10/7/2012	10/8/2012
4023183	Gastric reflux	10/8/2012	10/8/2012
4023183	Gastric reflux	6/8/2015	6/8/2015
4028741	Benign hypertension	4/16/2012	4/16/2012
4028741	Benign hypertension	6/8/2015	6/8/2015
4029305	Hypercholesterolemia	10/8/2012	10/8/2012
4029305	Hypercholesterolemia	6/8/2015	6/8/2015

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

Drug Concept ID	Drug Name	Drug Era Start Date	Drug Era End Date	Drug Exposure Counts	Gap Days
1307046	Metoprolol	9/21/2013	3/13/2015	4	-2
1307046	Metoprolol	4/14/2015	7/13/2015	3	0
1307046	Metoprolol	9/9/2015	3/4/2016	6	-3
1308216	Lisinopril	9/10/2013	2/9/2016	8	-288
1332418	Amlodipine	9/10/2013	4/13/2015	5	10
1332418	Amlodipine	11/9/2015	2/7/2016	3	0
1539403	Simvastatin	9/12/2013	10/12/2013	1	0
1539403	Simvastatin	11/12/2013	2/10/2014	2	30
1539403	Simvastatin	3/19/2014	4/18/2014	1	0
1539403	Simvastatin	5/28/2014	3/5/2015	4	41
1539403	Simvastatin	4/6/2015	6/3/2015	2	-2

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).

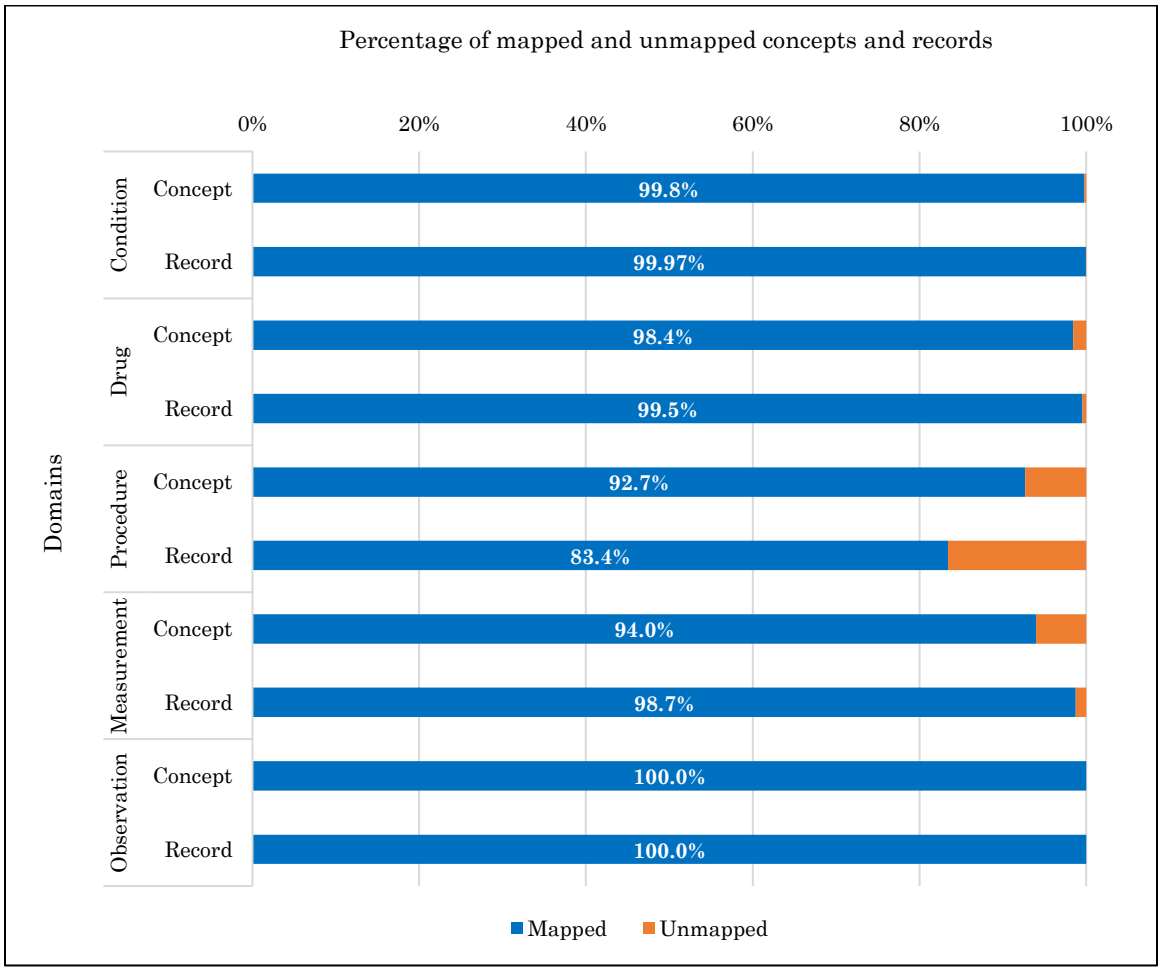

Figure 4. Overall mapping performance of concepts and records to OMOP CDM vocabulary by domain.

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

Domain and Code System	Concepts			Records		
	Mapped, *n* (%)	Unmapped, *n* (%)	Total	Mapped, *n* (%)	Unmapped, *n* (%)	Total
Condition			**1,459**			**12,648**
SNOMED CT	1,456 (99.8)	3 (0.2)	1,459	12,644 (99.97)	4 (0.03)	12,648
Drug	**758 (98.4)**	**12 (1.6)**	**770**	**7,230 (99.5)**	**33 (0.5)**	**7,263**
CVX		1 (100)	1		5 (100)	5
Local code		4 (100)	4		7 (100)	7
RxNorm	758 (99.1)	7 (0.9)	765	7,230 (99.7)	21 (0.3)	7,251
Procedure	**367 (92.7)**	**29 (7.3)**	**396**	**941 (83.4)**	**187 (16.6)**	**1,128**
CPT-4	80 (79.2)	21 (20.8)	101	195 (75.6)	63 (24.4)	258
ICD-9-CM	285 (99.3)	2 (0.7)	287	744 (99.1)	7 (0.9)	751
Local code		4 (100)	4		8 (100)	8
SNOMED CT	2 (50)	2 (50)	4	2 (0.8)	109 (98.2)	111
Measurement	**642 (94.0)**	**41 (6.0)**	**683**	**35,724 (98.7)**	**460 (1.3)**	**36,184**
Local code	3 (9.7)	28 (90.3)	31	663 (64.6)	363 (35.4)	1,026
LOINC	599 (99.0)	6 (1.0)	605	34,788 (99.9)	33 (0.1)	34,821
SNOMED CT	40 (85.1)	7 (14.9)	47	273 (81.0)	64 (19.0)	337
Observation			**71**			**280**
SNOMED CT	71 (100)		71	280 (100)		280
Race			**4**			**74**
CDC Race and Ethnicity	4 (4)		4	74 (100)		

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 vocabular 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

Concept's Domain	Mapped Code		Unmapped Code		Recall (%)	Precision (%)
	Correct (TP)	Incorrect (FP)	Correct (TN)	Incorrect (FN)		
Condition	1,456	0	0	3	99.8	100
Drug	758	0	4	8	99.0	100
Procedure	367	0	0	25	93.6	100
Measurement	642	0	0	13	98.0	100
Observation	71	0	0	0	100	100
Total	3,294	0	4	49	98.5	100

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.

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

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).

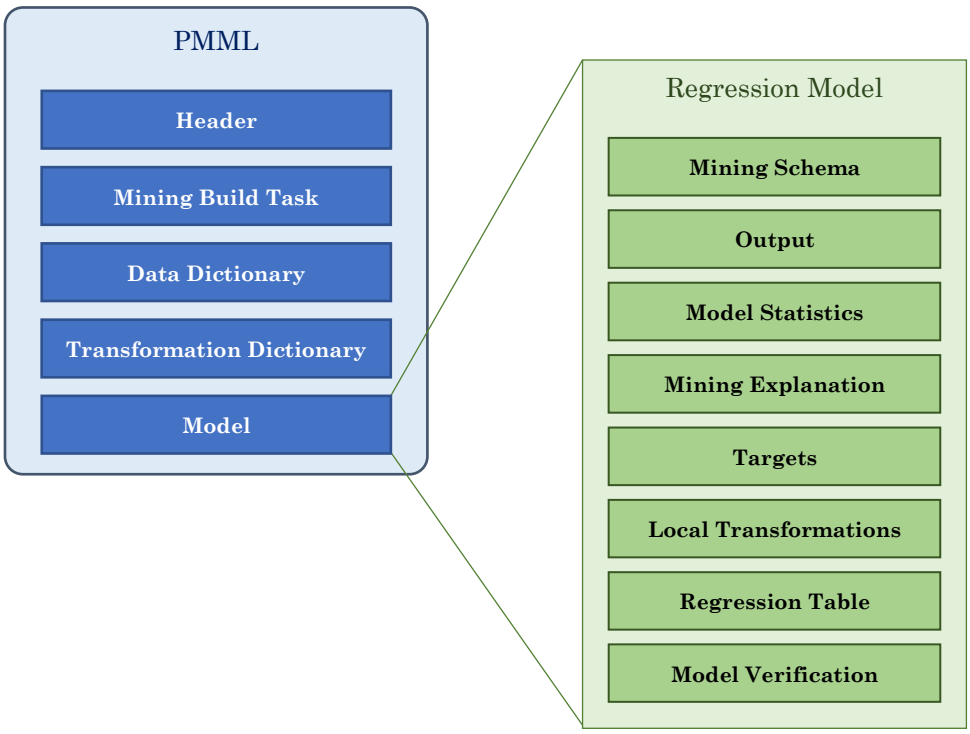

Figure 5. The general structure of a PMML document (left), and the schema of sections under Regression Model (right).

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 constrains 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).

```
<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 'I' 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).

```
<xs:element name="Header">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" ref="Extension"/>
      <xs:element ref="Application"/>
      <xs:element ref="Annotation"/>
      <xs:element ref="Timestamp"/>
    </xs:sequence>
    <xs:attribute name="copyright" type="xs:string"/>
    <xs:attribute name="description" type="xs:string"/>
    <xs:attribute name="modelVersion" type="xs:string"/>
  </xs:complexType>
</xs:element>

<xs:element name="Application">
  <xs:complexType>
    <xs:attribute name="name" type="xs:string" use="required"/>
    <xs:attribute name="version" type="xs:string"/>
  </xs:complexType>
</xs:element>

<xs:element name="Annotation" type="xs:string"/>

<xs:element name="Timestamp" type="xs:dateTime"/>

<xs:element name="Extension">
  <xs:complexType>
    <xs:choice>
      <xs:element minOccurs="0" ref="OmopCdm"/>
      <xs:element minOccurs="0" ref="author"/>

          etc.

    </xs:choice>
    <xs:attribute name="extender" type="xs:string"/>
  </xs:complexType>
</xs:element>

<xs:element name="OmopCdm">
  <xs:complexType>
    <xs:attribute name="version" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="author">
  <xs:complexType>
    <xs:attribute name="name" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
```

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

```
<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: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:element name="Value">
  <xs:complexType>
    <xs:attribute name="value" type="xs:string" use="required"/>
    <xs:attribute name="displayValue" type="xs:string"/>
    <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>
```

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

```
<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: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>
```

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.

```
<xs:element name="MiningBuildTask">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" maxOccurs="unbounded" ref="Extension"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="Extension">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="0"  ref="InputParameters"/>
      <xs:element minOccurs="0" ref="Statement"/>

         etc.

    </xs:sequence>
    <xs:attribute name="extender" type="xs:string" use="optional"
fixed="omop"/>
       <xs:attribute name="name" type="xs:string" use="optional"/>
  </xs:complexType>
</xs:element>

<xs:element name="InputParameters">
  <xs:complexType>
    <xs:sequence>
         <xs:element minOccurs="1" ref="InputParameter"/>
       </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="InputParameter">
  <xs:complexType>
       <xs:attribute name="name" type="xs:string" use="required"/>
       <xs:attribute name="displayName" type="xs:string" use="optional"/>
       <xs:attribute name="optype" type="OPTYPE" use="required"/>
    <xs:attribute name="dataType" type="DATATYPE" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="Statement">
  <xs:complexType>
       <xs:attribute name="dialect" type="STATEMENT-DIALECT" use="required"/>
  </xs:complexType>
</xs:element>

<xs:simpleType name="STATEMENT-DIALECT">
  <xs:restriction base="xs:string">
    <xs:enumeration value="postgresql"/>
    <xs:enumeration value="mssql"/>
    <xs:enumeration value="mysql"/>
    <xs:enumeration value="netezza"/>
    <xs:enumeration value="r"/>
    <xs:enumeration value="java"/>
    <xs:enumeration value="sql"/>
  </xs:restriction>
</xs:simpleType>
```

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.

```
<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.

```
<xs:element name="MiningSchema">
  <xs:complexType>
    <xs:sequence>
      <xs:element maxOccurs="unbounded" ref="MiningField"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="MiningField">
  <xs:complexType>
    <xs:sequence>
    </xs:sequence>
    <xs:attribute name="name" type="FIELD-NAME" use="required"/>
    <xs:attribute name="usageType" type="FIELD-USAGE-TYPE" default="active"/>
    <xs:attribute name="optype" type="OPTYPE"/>
    ...
  </xs:complexType>
</xs:element>
```

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).

```
<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>
```

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*.

```
<xs:element name="Output">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="OutputField" minOccurs="1" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</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: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:complexType>
</xs:element>
```

Figure 14. The XML schema of *Output* section and *Output Filed* 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.

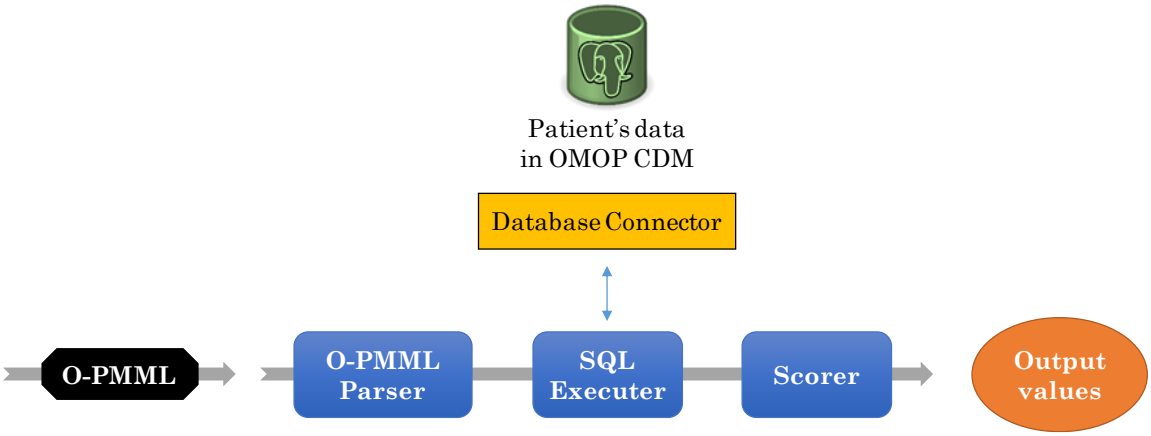

Figure 15. A schematic of O-PMML scoring engine

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.

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

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

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

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

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

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.

```
<MiningBuildTask>
  <Extension name="age" extender="omop">
    <InputParameters>
      <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"
        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>

  etc.

<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

Data Field Name	Data Field Display Name	Description
age	Age	The age of patient at index date
TCL	Total cholesterol level	Serum total cholesterol level (OMOP Concept ID = 3027114) within 30 days of index date
HDL	HDL cholesterol level	Serum HDL cholesterol level (OMOP Concept ID = 3007070) within 30 days of index date
HTNTRT	Antihypertensive medication use	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
SBP	Systolic blood pressure	Systolic blood pressure within 30 days of index date
smoker	Cigarette smoking status	Report of smoking within 90 days of index date
diabetic	Diabetes status	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

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:

```
case when HTNTRT=0 then logSBP else 0 end as logSBP_NOTTRT
case when HTNTRT=1 then logSBP else 0 end as logSBP_TRT
```

```
<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>
```

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

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,

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"*).

```
<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 fucntion="-">
        <Constant>1.0</Constant>
        <Apply fucntion="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

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

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.

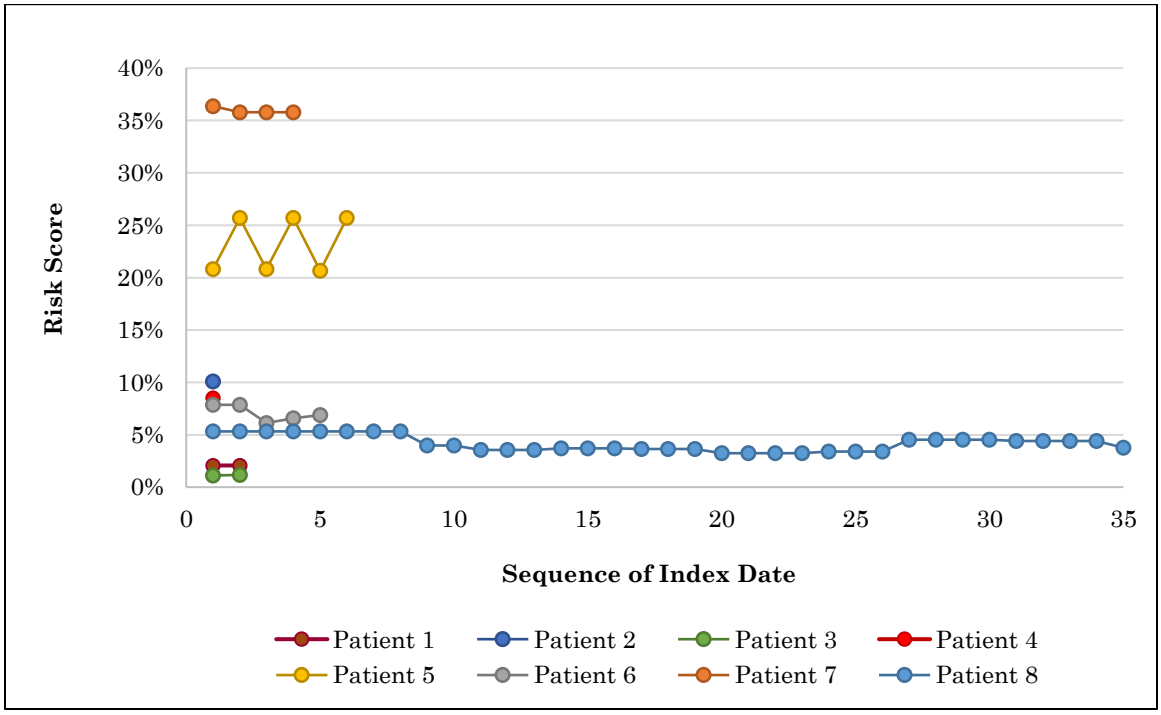

Figure 22. The timeline of estimated 10-year risk score of cardiovascular disease of patients that had full set of required values to generate scores.

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.

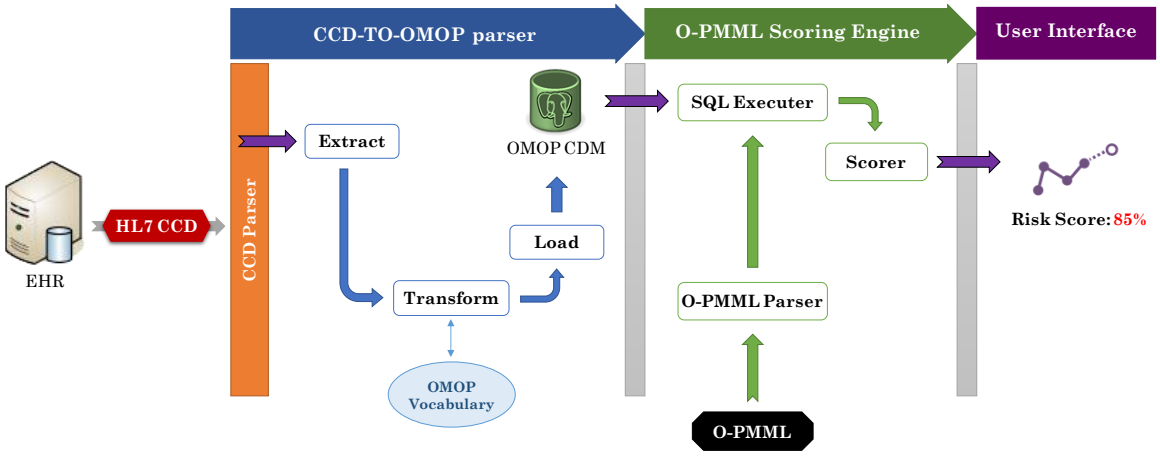

Figure 23. A schematic of the interoperable framework for delivering personalized health outcome predictions

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.

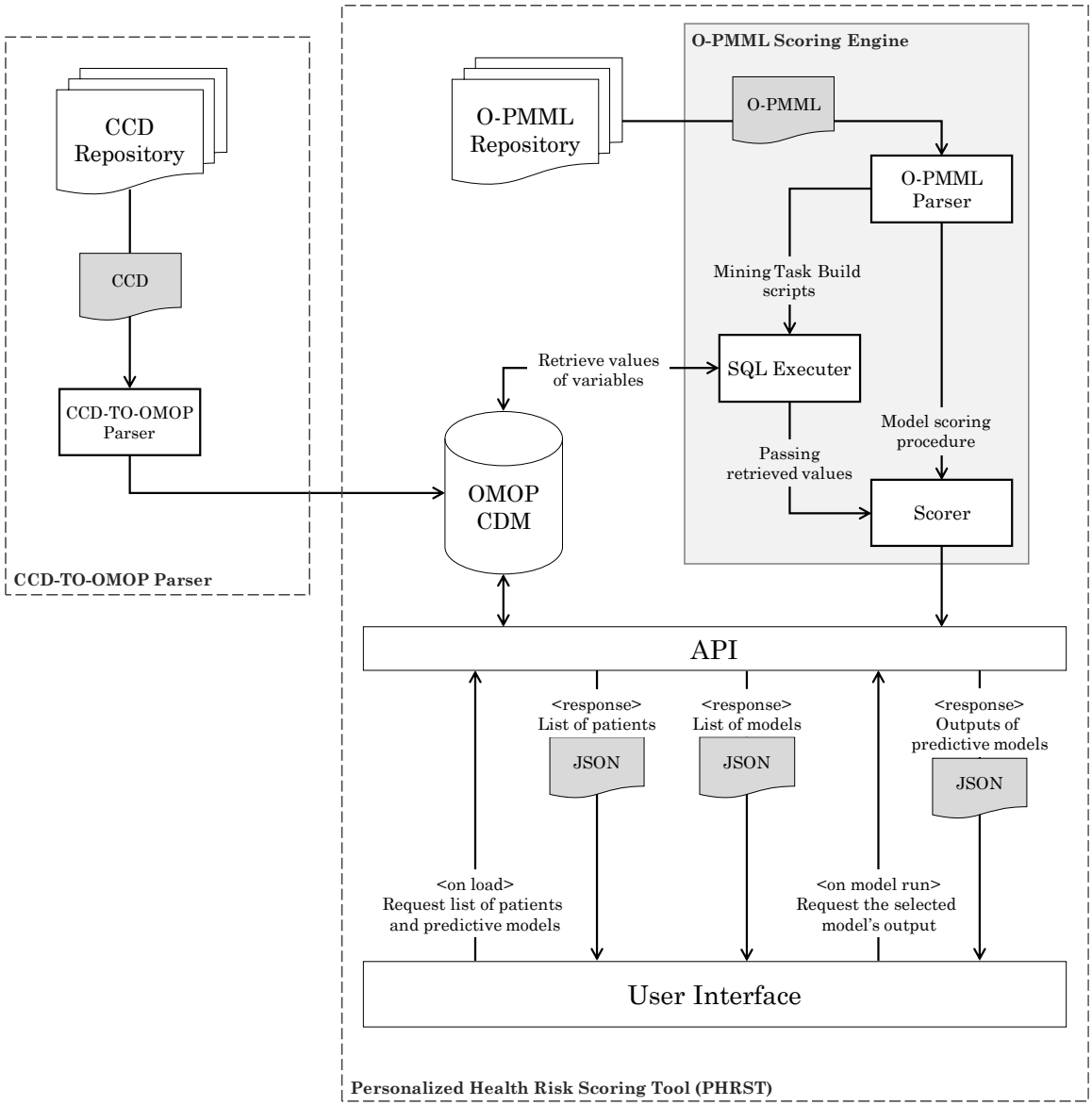

Figure 24. A diagram of the architecture and data flow of personalized health outcome prediction framework

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).

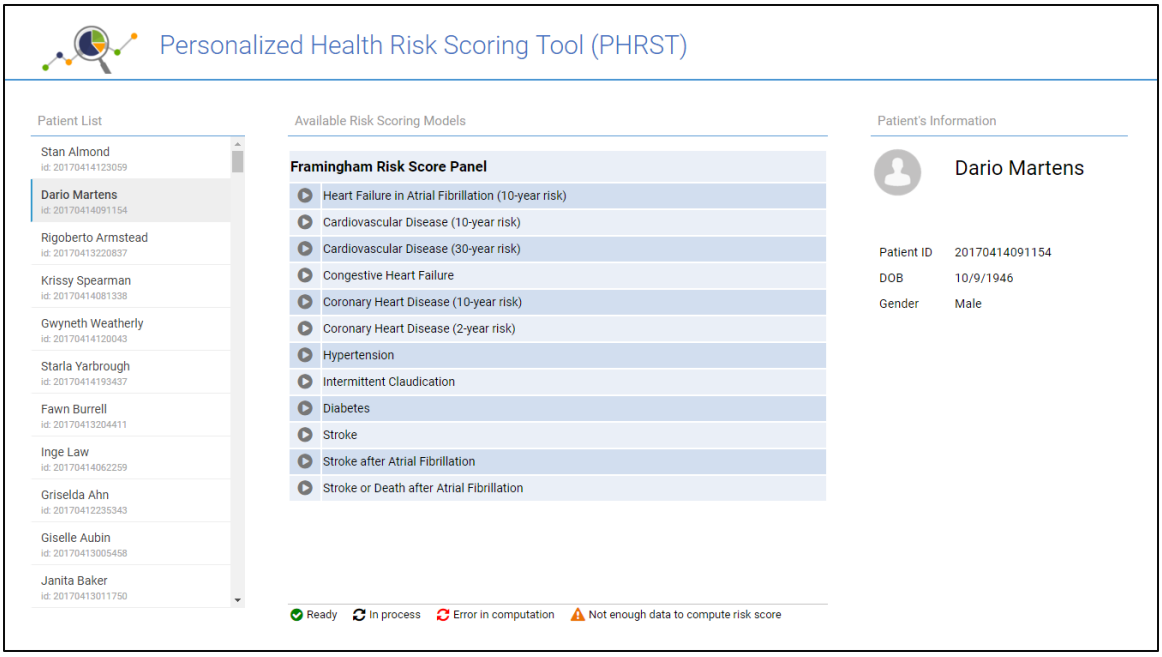

Figure 25. The view of PHRST application that allows users to select patients and order risk score estimates. The names of patients are fictitious.

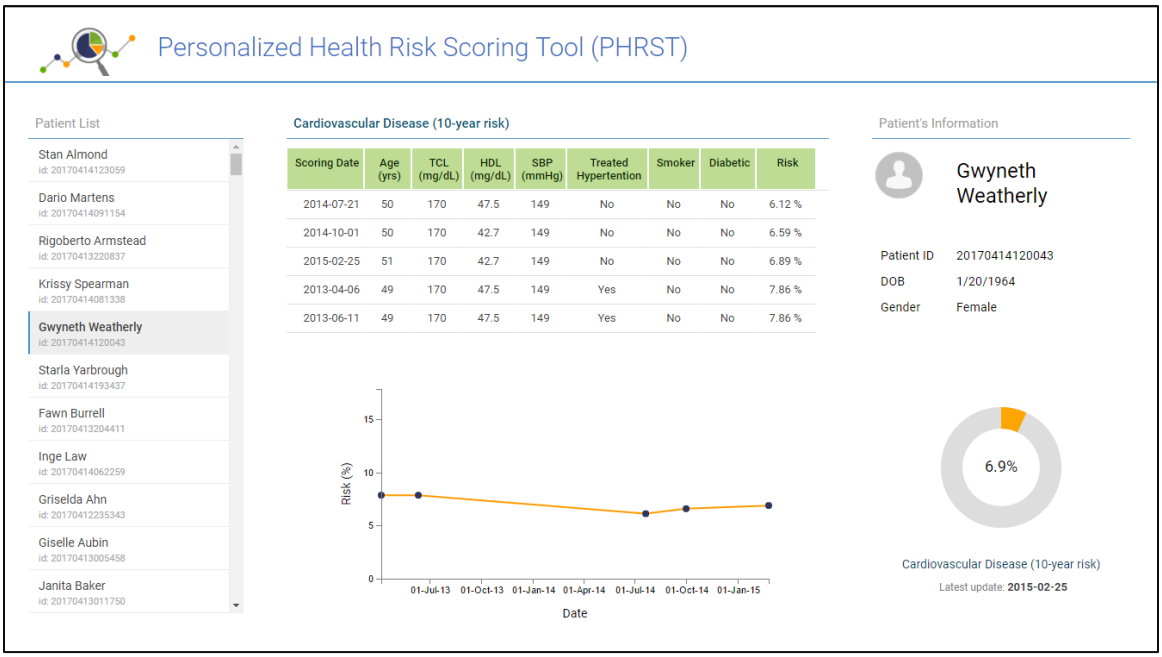

Figure 26. The view of PHRST application that displays more details of the selected patient's risk factor data and the estimated risk score. The names of patients are fictitious.

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 heterogenous 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 heterogenous 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 patent 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

Template Name	Template Type	Template ID
US Realm	Header	2.16.840.1.113883.10.20.22.1.1
Allergies Section (entries required)	Section	2.16.840.1.113883.10.20.22.2.6.1
Allergy Problem Act	Entry	2.16.840.1.113883.10.20.22.4.30
Allergy-Intolerance Observation	Entry	2.16.840.1.113883.10.20.22.4.7
Encounters Section (entries optional)	Section	2.16.840.1.113883.10.20.22.2.22
Encounter Activity	Entry	2.16.840.1.113883.10.20.1.21
Immunizations Section (entries optional)	Section	2.16.840.1.113883.10.20.22.2.2
Immunization Activity	Entry	2.16.840.1.113883.10.20.22.4.52
Immunization Medication Information	Entry	2.16.840.1.113883.10.20.22.4.54
Medications Section (entries required)	Section	2.16.840.1.113883.10.20.22.2.1.1
Medication Activity	Entry	2.16.840.1.113883.10.20.22.4.16
Medication Information	Entry	2.16.840.1.113883.10.20.22.4.23
Supply Activity	Entry	2.16.840.1.113883.10.20.1.34
Problem Section (entries required)	Section	2.16.840.1.113883.10.20.22.2.5.1
Problem Concern Act	Entry	2.16.840.1.113883.10.20.22.4.3
Problem Observation	Entry	2.16.840.1.113883.10.20.22.4.4
Procedures Section (entries required)	Section	2.16.840.1.113883.10.20.22.2.7.1
Procedure Activity Act		2.16.840.1.113883.10.20.22.4.12
Procedure Activity Observation		2.16.840.1.113883.10.20.22.4.13
Procedure Activity Procedure		2.16.840.1.113883.10.20.22.4.14

Template Name	Template Type	Template ID
Results Section (entries required)	Section	2.16.840.1.113883.10.20.22.2.3.1
Result Organizer	Entry	2.16.840.1.113883.10.20.22.4.1
Result Observation	Entry	2.16.840.1.113883.10.20.22.4.2
Social History Section (entries optional)	Section	2.16.840.1.113883.10.20.22.2.17
Smoking Status Observation	Entry	2.16.840.1.113883.10.20.22.4.78
Vital Signs Section (entries optional)	Section	2.16.840.1.113883.10.20.22.2.4
Vital Signs Organizer	Entry	2.16.840.1.113883.10.20.22.4.26
Vital Sign Observation	Entry	2.16.840.1.113883.10.20.1.31

Appendix 2. SQL query to map source codes to OMOP source concepts

```
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, C.CONCEPT_ID AS TARGET_CONCEPT_ID, C.VOCABULARY_ID AS
   TARGET_VOCABULARY_ID, C.DOMAIN_ID AS TARGET_DOMAIN_ID, C.CONCEPT_CLASS_ID AS
   TARGET_CONCEPT_CLASS_ID, C.INVALID_REASON AS TARGET_INVALID_REASON,
   C.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
FROM CONCEPT C
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 3. SQL query to map source codes to OMOP standard concepts

```
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_VOCABUALRY_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

CVX Code	CPT Code	OMOP Concept ID	CVX Short Description
01	90701	No Concept ID	DTP
02	90712	2213470	OPV
03	90707	2213466	MMR
04	90708	2213467	M/R
05	90705	2213464	measles
06	90706	2213465	rubella
07	90704	2213463	mumps
08	90744	2213491	Hep B, adolescent or pediatric
09	90714	2213472	Td (adult), adsorbed
09	90718	No Concept ID	Td (adult), adsorbed
10	90713	2213471	IPV
12	90296	2213401	diphtheria antitoxin
13	90389	2213410	TIG
14	90741	No Concept ID	IG, unspecified formulation
15	No Code	No Concept ID	influenza, split (incl. purified surface antigen)
16	90659	No Concept ID	influenza, whole
17	90737	No Concept ID	Hib, unspecified formulation
18	90675	2213449	rabies, intramuscular injection
19	90585	2213425	BCG
19	90728	No Concept ID	BCG
20	90700	2213459	DTaP
21	90716	2213474	varicella
22	90720	2213478	DTP-Hib
23	90727	2213482	plague
24	90581	2213424	anthrax
25	90690	2213453	typhoid, oral
26	90725	2213481	cholera, unspecified formulation
27	90287	2213398	botulinum antitoxin
28	90702	2213461	DT (pediatric)
29	90291	2213400	CMVIG
30	90371	2213402	HBIG
31	No Code	No Concept ID	HepA Pediatric (Unspecified, historical)

CVX Code	CPT Code	OMOP Concept ID	CVX Short Description
32	90733	2213484	meningococcal MPSV4
33	90732	2213483	pneumococcal polysaccharide PPV23
34	90375	2213403	RIG
34	90376	2213404	RIG
35	90703	2213462	tetanus toxoid, adsorbed
36	90396	2213412	VZIG
37	90717	2213475	yellow fever
39	90735	2213486	Japanese encephalitis SC
40	90676	2213450	rabies, intradermal injection
41	90692	2213455	typhoid, parenteral
42	90745	No Concept ID	Hep B, adolescent/high risk infant
43	90739	43527982	Hep B, adult
43	90743	2213490	Hep B, adult
43	90746	2213492	Hep B, adult
44	90740	2213489	Hep B, dialysis
44	90747	2213493	Hep B, dialysis
45	90731	No Concept ID	Hep B, unspecified formulation
46	90646	2213432	Hib (PRP-D)
47	90645	2213431	Hib (HbOC)
48	90648	2213434	Hib (PRP-T)
49	90647	2213433	Hib (PRP-OMP)
50	90721	2213479	DTaP-Hib
51	90748	2213494	Hib-Hep B
52	90632	2213427	Hep A, adult
53	90693	2213456	typhoid, parenteral, AKD (U.S. military)
54	90476	2213422	adenovirus, type 4
55	90477	2213423	adenovirus, type 7
62	90649	2213435	HPV, quadrivalent
66	90665	No Concept ID	Lyme disease
71	90379	No Concept ID	RSV-IGIV
79	90393	2213411	vaccinia immune globulin
83	90633	2213428	Hep A, ped/adol, 2 dose
84	90634	2213429	Hep A, ped/adol, 3 dose
85	90730	No Concept ID	Hep A, unspecified formulation

CVX Code	CPT Code	OMOP Concept ID	CVX Short Description
86	90281	2213395	IG
87	90283	2213396	IGIV
88	90724	No Concept ID	influenza, unspecified formulation
90	90726	No Concept ID	rabies, unspecified formulation
91	90714	2213472	typhoid, unspecified formulation
93	90378	2213405	RSV-MAb
94	90710	2213469	MMRV
100	90669	2213447	pneumococcal conjugate PCV 7
101	90691	2213454	typhoid, ViCPs
104	90636	2213430	Hep A-Hep B
106	90700	2213459	DTaP, 5 pertussis antigens
110	90723	2213480	DTaP-Hep B-IPV
111	90660	2213442	influenza, live, intranasal
113	90714	2213472	Td (adult) preservative free
114	90734	2213485	meningococcal MCV4P
115	90715	2213473	Tdap
116	90680	2213451	rotavirus, pentavalent
118	90650	2213436	HPV, bivalent
119	90681	2213452	rotavirus, monovalent
120	90698	2213458	DTaP-Hib-IPV
121	90736	2213487	zoster
125	90664	40756887	Novel Influenza-H1N1-09, nasal
126	90666	40756874	Novel influenza-H1N1-09, preservative-free
127	90668	40757097	Novel influenza-H1N1-09
128	90470	No Concept ID	Novel Influenza-H1N1-09, all formulations
128	90663	No Concept ID	Novel Influenza-H1N1-09, all formulations
130	90696	2213457	DTaP-IPV
133	90670	2213448	Pneumococcal conjugate PCV 13
134	90738	2213488	Japanese Encephalitis IM
135	90662	2213444	Influenza, high dose seasonal
136	90734	2213485	Meningococcal MCV4O
140	90655	2213437	Influenza, seasonal, injectable, preservative free

CVX Code	CPT Code	OMOP Concept ID	CVX Short Description
140	90656	2213438	Influenza, seasonal, injectable, preservative free
141	90657	2213439	Influenza, seasonal, injectable
141	90658	2213440	Influenza, seasonal, injectable
144	90654	42742499	influenza, seasonal, intradermal, preservative free
146	90697	No Concept ID	DTaP,IPV,Hib,HepB
148	90644	40757102	Meningococcal C/Y-HIB PRP
149	90672	43527981	influenza, live, intranasal, quadrivalent
150	90686	44816520	influenza, injectable, quadrivalent, preservative free
153	90661	2213443	Influenza, injectable, MDCK, preservative free
155	90673	44816443	influenza, recombinant, injectable, preservative free
158	90687	44816519	influenza, injectable, quadrivalent
158	90688	44816518	influenza, injectable, quadrivalent
161	90685	44816521	Influenza, injectable, quadrivalent, preservative free, pediatric
162	90621	No Concept ID	meningococcal B, recombinant
163	90620	No Concept ID	meningococcal B, OMV
165	90651	46257428	HPV9
166	90630	46257714	influenza, intradermal, quadrivalent, preservative free
168	90653	43527980	influenza, trivalent, adjuvanted
171	90674	No Concept ID	Influenza, injectable, MDCK, preservative free, quadrivalent
174	90625	No Concept ID	cholera, live attenuated
175	90675	2213449	Rabies - IM Diploid cell culture
176	90675	2213449	Rabies - IM fibroblast culture

Appendix 5. SQL query to find ingredients of OMOP standard drug concepts

```
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 INTERPRETATION')

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
(PERSON_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

```
<xs:schema xmlns:xs="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 -->
<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>
<!-- Header -->
<xs:element name="Header">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" ref="Extension"/>
      <xs:element ref="Application"/>
      <xs:element ref="Annotation"/>
      <xs:element ref="Timestamp"/>
    </xs:sequence>
    <xs:attribute name="copyright" type="xs:string"/>
    <xs:attribute name="description" type="xs:string"/>
    <xs:attribute name="modelVersion" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:element name="Application">
  <xs:complexType>
    <xs:attribute name="name" type="xs:string" use="required"/>
    <xs:attribute name="version" type="xs:string"/>
  </xs:complexType>
</xs:element>
<xs:element name="Annotation" type="xs:string"/>
<xs:element name="Timestamp" type="xs:dateTime"/>
<xs:element name="OmopCdm">
  <xs:complexType>
    <xs:attribute name="version" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
<xs:element name="author">
  <xs:complexType>
    <xs:attribute name="name" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
<!-- Extensions -->
<xs:element name="Extension">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="0" ref="OmopCdm"/>
      <xs:element minOccurs="0" ref="author"/>
```

```
        <xs:element minOccurs="0" ref="InputParameters"/>
        <xs:element minOccurs="0" ref="Statement"/>
    </xs:sequence>
    <xs:attribute name="extender" type="xs:string" use="optional"
fixed="omop"/>
      <xs:attribute name="name" type="xs:string" use="optional"/>
  </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: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: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="time"/>
    <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: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>
<!-- Transformation Dictionary -->
<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>
<!-- Local Transforamtions -->
<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>
<!-- Transformation Expressions -->
<xs:group name="EXPRESSION">
  <xs:choice>
    <xs:element ref="Constant"/>
    <xs:element ref="FieldRef"/>
    <xs:element ref="NormContinuous"/>
    <xs:element ref="NormDiscrete"/>
```

```
    <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:attribute name="field" type="FIELD-NAME" use="required"/>
    <xs:attribute name="mapMissingTo" type="xs:string"/>
    <xs:attribute name="defaultValue" type="xs:string"/>
    <xs:attribute name="dataType" type="DATATYPE"/>
  </xs:complexType>
</xs:element>
<xs:element name="DiscretizeBin">
  <xs:complexType>
    <xs:sequence>

      <xs:element ref="Interval"/>
```

```
    </xs:sequence>
    <xs:attribute name="binValue" type="xs:string" use="required"/>
  </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:sequence>
    <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:restriction>
      </xs:simpleType>
    </xs:attribute>
    <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:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="wordSeparatorCharacterRE" type="xs:string"
default="\s"/>
    <xs:attribute name="tokenize" type="xs:boolean" default="true"/>
  </xs:complexType>
</xs:element>
<xs:element name="TextIndexNormalization">
  <xs:complexType>
    <xs:sequence>
      <xs:choice minOccurs="0">
        <xs:element ref="TableLocator"/>
```

```
        <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:sequence>
      <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: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 name="Apply">
  <xs:complexType>
    <xs:sequence>
      <xs:group ref="EXPRESSION" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="function" type="xs:string" use="required"/>
    <xs:attribute name="mapMissingTo" type="xs:string"/>
    <xs:attribute name="defaultValue" type="xs:string"/>
    <xs:attribute name="invalidValueTreatment" type="INVALID-VALUE-TREATMENT-
METHOD" default="returnInvalid"/>
  </xs:complexType>
</xs:element>
<!-- Mining Build Task -->
<xs:element name="MiningBuildTask">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" maxOccurs="unbounded" ref="Extension"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="InputParameters">
  <xs:complexType>
    <xs:sequence>
        <xs:element minOccurs="1" ref="InputParameter"/>
      </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="InputParameter">
  <xs:complexType>
      <xs:attribute name="name" type="xs:string" use="required"/>
      <xs:attribute name="displayName" type="xs:string" use="optional"/>
      <xs:attribute name="optype" type="OPTYPE" use="required"/>
    <xs:attribute name="dataType" type="DATATYPE" use="required"/>
  </xs:complexType>
</xs:element>
<xs:element name="Statement">
  <xs:complexType>
      <xs:attribute name="dialect" type="STATEMENT-DIALECT" use="required"/>
  </xs:complexType>
</xs:element>
<xs:simpleType name="STATEMENT-DIALECT">
  <xs:restriction base="xs:string">
    <xs:enumeration value="postgresql"/>
    <xs:enumeration value="mssql"/>
    <xs:enumeration value="mysql"/>
    <xs:enumeration value="netezza"/>
    <xs:enumeration value="r"/>
    <xs:enumeration value="java"/>
    <xs:enumeration value="sql"/>
  </xs:restriction>
</xs:simpleType>
<!-- Regression Model -->
<xs:element name="RegressionModel">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="MiningSchema"/>
      <xs:element ref="RegressionTable" maxOccurs="unbounded"/><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="LocalTransformations" 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>
<!-- Mining Schema -->
<xs:element name="MiningSchema">
  <xs:complexType>
    <xs:sequence>
      <xs:element maxOccurs="unbounded" ref="MiningField"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="MiningField">
  <xs:complexType>
    <xs:sequence>
    </xs:sequence>
    <xs:attribute name="name" type="FIELD-NAME" use="required"/>
    <xs:attribute name="usageType" type="FIELD-USAGE-TYPE" default="active"/>
    <xs:attribute name="optype" type="OPTYPE"/>
    <xs:attribute name="importance" type="PROB-NUMBER"/>
    <xs:attribute name="outliers" type="OUTLIER-TREATMENT-METHOD"
default="asIs"/>
    <xs:attribute name="lowValue" type="NUMBER"/>
    <xs:attribute name="highValue" type="NUMBER"/>
    <xs:attribute name="missingValueReplacement" type="xs:string"/>
    <xs:attribute name="missingValueTreatment" type="MISSING-VALUE-TREATMENT-
METHOD"/>
    <xs:attribute name="invalidValueTreatment" type="INVALID-VALUE-TREATMENT-
METHOD" default="returnInvalid"/>
  </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:simpleType name="REGRESSIONNORMALIZATIONMETHOD">
  <xs:restriction base="xs:string">
    <xs:enumeration value="none"/>
```

```
    <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 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 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 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>
<!-- Output -->
<xs:element name="Output">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="OutputField" minOccurs="1" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</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: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:complexType>
</xs:element>
<xs:simpleType 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:choice>
        <xs:element ref="PredictiveModelQuality" minOccurs="0"
maxOccurs="unbounded"/>
      </xs:choice>
      <xs:element ref="Correlations" minOccurs="0"/>
    </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>
    <xs:attribute name="meanError" type="NUMBER" use="optional"/>
    <xs:attribute name="meanAbsoluteError" type="NUMBER" use="optional"/>
    <xs:attribute name="meanSquaredError" type="NUMBER" use="optional"/>
    <xs:attribute name="rootMeanSquaredError" type="NUMBER" use="optional"/>
    <xs:attribute name="r-squared" type="NUMBER" use="optional"/>
    <xs:attribute name="adj-r-squared" type="NUMBER" use="optional"/>
    <xs:attribute name="sumSquaredError" 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>
<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: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 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:attribute>
  </xs:complexType>
</xs:element>
<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:group ref="FrequenciesType" minOccurs="0"/>
    </xs:sequence>
    <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:element ref="MultivariateStat" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="targetCategory" type="xs:string" use="optional"/>
  </xs:complexType>
</xs:element>
<xs:element name="MultivariateStat">
  <xs:complexType>
    <xs:attribute name="name" type="xs:string"/>
    <xs:attribute name="category" type="xs:string"/>
    <xs:attribute name="exponent" type="INT-NUMBER" default="1"/>
    <xs:attribute name="isIntercept" type="xs:boolean" default="false"/>
    <xs:attribute name="importance" type="PROB-NUMBER"/>
    <xs:attribute name="stdError" type="NUMBER"/>
    <xs:attribute name="tValue" type="NUMBER"/>
    <xs:attribute name="chiSquareValue" type="NUMBER"/>
    <xs:attribute name="fStatistic" type="NUMBER"/>
    <xs:attribute name="dF" type="NUMBER"/>
    <xs:attribute name="pValueAlpha" type="PROB-NUMBER"/>
    <xs:attribute name="pValueInitial" type="PROB-NUMBER"/>
    <xs:attribute name="pValueFinal" type="PROB-NUMBER"/>
    <xs:attribute name="confidenceLevel" type="PROB-NUMBER" default="0.95"/>
    <xs:attribute name="confidenceLowerBound" type="NUMBER"/>
    <xs:attribute name="confidenceUpperBound" type="NUMBER"/>
  </xs:complexType>
</xs:element>
<xs:element name="Anova">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="AnovaRow" minOccurs="3" maxOccurs="3"/>
    </xs:sequence>
    <xs:attribute name="target" type="FIELD-NAME"/>
  </xs:complexType>
</xs:element>
<xs:element name="AnovaRow">
  <xs:complexType>
    <xs:attribute name="type" use="required">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="Model"/>
          <xs:enumeration value="Error"/>
          <xs:enumeration value="Total"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="sumOfSquares" type="NUMBER" use="required"/>
    <xs:attribute name="degreesOfFreedom" type="NUMBER" use="required"/>
    <xs:attribute name="meanOfSquares" type="NUMBER"/>
    <xs:attribute name="fValue" type="NUMBER"/>
    <xs:attribute name="pValue" type="PROB-NUMBER"/>
  </xs:complexType>
</xs:element>
<xs:element name="Partition">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="PartitionFieldStats" minOccurs="0"
maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="name" type="xs:string" use="required"/>
    <xs:attribute name="size" type="NUMBER"/>
  </xs:complexType>
</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: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:attribute>
  </xs:complexType>
</xs:element>
<!-- Targets -->
<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:sequence>
    <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:complexType>
</xs:element>
<xs:element name="InlineTable">
  <xs:complexType>
    <xs:sequence>
```

```
      <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>
```

Appendix 9. The O-PMML containing predictive model for estimating Framingham 10-year risk of cardiovascular disease for men.

```
<?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"
   xsi:schemaLocation="http://www.dmg.org/PMML-4_3 http://www.dmg.org/v4-3/pmml-
   4-3.xsd">
  <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"/>
      <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="continous" 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="continous" 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>
<Extension name="HDL" extender="omop">
  <InputParameters>
    <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"
    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
    &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 (3007070) 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 hdl
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=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="continous" 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>
```

```
    <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"
    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 &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
    (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
          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)
    select value_as_number from sbp
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=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="continous" 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
    (4218741,4246415,4276526,4052947,4052029,4052030,4218917,4298794,427099
    96,4043053,4144273,4144273,4043056)
             and person_id=@PERSON_ID and observation_date
    &lt;=to_date('@INDEXT_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
    &lt;=90;
  </Statement>
</Extension>
<Extension name="diabetic" extender="omop">
  <InputParameters>
```

```
  <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"
 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,
 46221581,42899447,19090244,19090229,19090247,19090249,1513876,19090180,19
 013926,19091621,19090187,19013951,1590165,1586346,1513849,1562586,1909020
 4,1513843,40239216,40170911,44506754,1503297,1510202,1502826,19033909,152
 5215,1517998,1596977,1516766,1547504,40166035,1580747,1502809,1502855,190
 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>=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
 (201820,45757674,45757474,4096666,4008576,201254,4152858,201531,4099214,4
 43412,201826,4196141,201530,4151282,4198296,4200875,4099651,4193704)
            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
          )
        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>
        <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>
  <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 fucntion="-">
          <Constant>1.0</Constant>
          <Apply fucntion="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>
</PMML>
```

Appendix 10. The O-PMML containing predictive model for estimating Framingham 10-year risk of cardiovascular disease for women.

```
<?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"
   xsi:schemaLocation="http://www.dmg.org/PMML-4_3 http://www.dmg.org/v4-3/pmml-
   4-3.xsd">
  <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="continous" 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="continous" 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>
<Extension name="HDL" extender="omop">
  <InputParameters>
    <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"
    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
    &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 (3007070) 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 hdl
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=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="continous" 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>
```

```
    <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"
    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 &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
    (3028737,3004249,3018586,3035856,3018822,21490779,21492239,4161413)
          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)
    select value_as_number from sbp
    where to_date('@INDEX_DATE','YYYY-MM-DD')-MEAS_DATE &lt;=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="continous" 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
    (4218741,4246415,4276526,4052947,4052029,4052030,4218917,4298794,427099
    96,4043053,4144273,4144273,4043056)
             and person_id=@PERSON_ID and observation_date
    &lt;=to_date('@INDEXT_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
    &lt;=90;
  </Statement>
</Extension>
<Extension name="diabetic" extender="omop">
  <InputParameters>
```

```
  <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"
  dataType="string"/>
</InputParameters>
<Statement dialect="postgresql">
  with medrec as (select * from omop.drug_era
          where drug_concept_id in
  (1529331,1530014,44816332,43013884,19035533,43526465,19033498,1594973,4
  4785829,45774435,45774751,1583722,19001409,19059796,1597756,1560171,190
  97821,1559684,19001441,1000979,19122121,35604829,35605670,35602717,1516
  976,40056629,1502905,1588986,1550023,1531601,1567198,1544838,40051349,4
  0051350,46221581,42899447,19090244,19090229,19090247,19090249,1513876,1
  9090180,19013926,19091621,19090187,19013951,1590165,1586346,1513849,156
  2586,19090204,1513843,40239216,40170911,44506754,1503297,1510202,150282
  6,19033909,1525215,1517998,1596977,1516766,1547504,40166035,1580747,150
  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>=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
          )
        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>
        <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>
  <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 fucntion="-">
          <Constant>1.0</Constant>
          <Apply fucntion="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.

Patient ID	Gender	Age	TCL	HDL	Antihypertensive Use (1: Yes, 0: No)	SBP	Smoker (1: Yes, 0: No)	Diabetic (1: Yes, 0: No)	Risk Score
Patient 1	F	31	161	42	1	133	0	0	2.07%
Patient 1	F	31	161	42	1	133	0	0	2.07%
Patient 2	M	44	219	47	0	162	0	0	10.10%
Patient 3	F	30	184	31	0	106	0	0	1.11%
Patient 3	F	30	184	31	0	108	0	0	1.17%
Patient 4	F	52	200	35	0	140.5	0	0	8.51%
Patient 5	M	67	143	36	1	134.9	0	0	25.69%
Patient 5	M	67	143	36	1	119.6	0	0	20.81%
Patient 5	M	67	143	36	1	134.9	0	0	25.69%
Patient 5	M	67	143	36	1	119.6	0	0	20.81%
Patient 5	M	67	143	36	1	134.9	0	0	25.69%
Patient 5	M	67	143	36	1	119.1	0	0	20.66%
Patient 6	F	49	170	47.5	1	149	0	0	7.86%
Patient 6	F	49	170	47.5	1	149	0	0	7.86%
Patient 6	F	50	170	47.5	0	149	0	0	6.12%
Patient 6	F	50	170	42.7	0	149	0	0	6.59%
Patient 6	F	51	170	42.7	0	149	0	0	6.89%
Patient 7	M	64	87	25	1	149.5	0	1	36.37%
Patient 7	M	64	87	25	1	148	0	1	35.79%
Patient 7	M	64	87	25	1	148	0	1	35.79%
Patient 7	M	64	87	25	1	148	0	1	35.79%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%

Patient ID	Gender	Age	TCL	HDL	Antihypertensive Use (1: Yes, 0: No)	SBP	Smoker (1: Yes, 0: No)	Diabetic (1: Yes, 0: No)	Risk Score
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	1	130	0	0	5.32%
Patient 8	F	49	140	35	0	130	0	0	3.98%
Patient 8	F	49	140	35	0	130	0	0	3.98%
Patient 8	F	49	140	35	0	124.5	0	0	3.54%
Patient 8	F	49	140	35	0	124.5	0	0	3.54%
Patient 8	F	49	140	35	0	124.5	0	0	3.54%
Patient 8	F	50	140	35	0	124.5	0	0	3.71%
Patient 8	F	50	140	35	0	124.5	0	0	3.71%
Patient 8	F	50	140	35	0	124.5	0	0	3.71%
Patient 8	F	50	149	40	0	124.5	0	0	3.64%
Patient 8	F	50	149	40	0	124.5	0	0	3.64%
Patient 8	F	50	149	40	0	124.5	0	0	3.64%
Patient 8	F	50	149	40	0	119.3	0	0	3.24%
Patient 8	F	50	149	40	0	119.3	0	0	3.24%
Patient 8	F	50	149	40	0	119.3	0	0	3.24%
Patient 8	F	50	149	40	0	119.3	0	0	3.24%
Patient 8	F	51	149	40	0	119.3	0	0	3.39%
Patient 8	F	51	149	40	0	119.3	0	0	3.39%
Patient 8	F	51	149	40	0	119.3	0	0	3.39%
Patient 8	F	51	149	40	1	119.3	0	0	4.51%
Patient 8	F	51	149	40	1	119.3	0	0	4.51%
Patient 8	F	51	149	40	1	119.3	0	0	4.51%
Patient 8	F	51	149	40	1	119.3	0	0	4.51%

Patient ID	Gender	Age	TCL	HDL	Antihypertensive Use (1: Yes, 0: No)	SBP	Smoker (1: Yes, 0: No)	Diabetic (1: Yes, 0: No)	Risk Score
Patient 8	F	52	149	40	1	116.3	0	0	4.40%
Patient 8	F	52	149	40	1	116.3	0	0	4.40%
Patient 8	F	52	149	40	1	116.3	0	0	4.40%
Patient 8	F	52	149	40	1	116.3	0	0	4.40%
Patient 8	F	52	143	43	1	114	0	0	3.77%

REFERENCES

1. National Research Council, Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington (DC): National Academies Press; 2011.
2. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. Jama. 2011;306(15):1688-98.
3. Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, et al. Statistical models and patient predictors of readmission for heart failure: a systematic review. Archives of internal medicine. 2008;168(13):1371-86.
4. Hu Z, Hao S, Jin B, Shin AY, Zhu C, Huang M, et al. Online Prediction of Health Care Utilization in the Next Six Months Based on Electronic Health Record Information: A Cohort and Validation Study. Journal of medical Internet research. 2015;17(9):e219.
5. Khazaei H, McGregor C, Eklund JM, El-Khatib K. Real-Time and Retrospective Health-Analytics-as-a-Service: A Novel Framework. JMIR medical informatics. 2015;3(4):e36.
6. Toerper MF, Flanagan E, Siddiqui S, Appelbaum J, Kasper EK, Levin S. Cardiac catheterization laboratory inpatient forecast tool: a prospective evaluation. Journal of the American Medical Informatics Association : JAMIA. 2015.
7. Mitchell C, Meredith P, Richardson M, Greengross P, Smith GB. Reducing the number and impact of outbreaks of nosocomial viral gastroenteritis: time-series analysis of a multidimensional quality improvement initiative. BMJ quality & safety. 2015.
8. Brady TM, Neu AM, Miller ER, 3rd, Appel LJ, Siberry GK, Solomon BS. Real-time electronic medical record alerts increase high blood pressure recognition in children. Clinical pediatrics. 2015;54(7):667-75.
9. Bates DW, Cohen M, Leape LL, Overhage JM, Shabot MM, Sheridan T. Reducing the frequency of errors in medicine using information technology.

Journal of the American Medical Informatics Association : JAMIA. 2001;8(4):299-308.

10. Charles K, Cannon M, Hall R, Coustasse A. Can utilizing a computerized provider order entry (CPOE) system prevent hospital medical errors and adverse drug events? Perspectives in health information management / AHIMA, American Health Information Management Association. 2014;11:1b.
11. Barrett JS, Mondick JT, Narayan M, Vijayakumar K, Vijayakumar S. Integration of modeling and simulation into hospital-based decision support systems guiding pediatric pharmacotherapy. BMC medical informatics and decision making. 2008;8:6.
12. Nuckols TK, Asch SM, Patel V, Keeler E, Anderson L, Buntin MB, et al. Implementing Computerized Provider Order Entry in Acute Care Hospitals in the United States Could Generate Substantial Savings to Society. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2015;41(8):341-50.
13. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. Journal of the American Medical Informatics Association : JAMIA. 2003;10(6):523-30.
14. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. Journal of the American Medical Informatics Association : JAMIA. 2001;8(6):527-34.
15. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. Journal of the American Medical Informatics Association : JAMIA. 2006;13(2):138-47.
16. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ (Clinical research ed). 2013;346:f657.
17. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in

outpatients. Journal of the American Medical Informatics Association : JAMIA. 2014;21(3):487-91.

18. Hatton RC, Rosenberg AF, Morris CT, McKelvey RP, Lewis JR. Evaluation of contraindicated drug-drug interaction alerts in a hospital setting. The Annals of pharmacotherapy. 2011;45(3):297-308.
19. FitzHenry F, Doran J, Lobo B, Sullivan TM, Potts A, Feldott CC, et al. Medication-error alerts for warfarin orders detected by a bar-code-assisted medication administration system. American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists. 2011;68(5):434-41.
20. Magid SK, Pancoast PE, Fields T, Bradley DG, Williams RB. Employing clinical decision support to attain our strategic goal: the safe care of the surgical patient. Journal of healthcare information management : JHIM. 2007;21(2):18-25.
21. Topaz M, Seger DL, Slight SP, Goss F, Lai K, Wickner PG, et al. Rising drug allergy alert overrides in electronic health records: an observational retrospective study of a decade of experience. Journal of the American Medical Informatics Association : JAMIA. 2015.
22. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annual Symposium proceedings AMIA Symposium. 2007:26-30.
23. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health affairs (Project Hope). 2011;30(12):2310-7.
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.
25. American Recovery and Reinvestment Act of 2009, Public Law 111–5 (2009).

26. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. Health Affairs. 2014;33(7):1123-31.
27. Barrett TW, Martin AR, Storrow AB, Jenkins CA, Harrell FE, Jr., Russ S, et al. A clinical prediction model to estimate risk for 30-day adverse events in emergency department patients with symptomatic atrial fibrillation. Annals of emergency medicine. 2011;57(1):1-12.
28. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e2):e243-52.
29. Echouffo-Tcheugui JB, Greene SJ, Papadimitriou L, Zannad F, Yancy CW, Gheorghiade M, et al. Population risk prediction models for incident heart failure: a systematic review. Circulation Heart failure. 2015;8(3):438-47.
30. Ustun B, Westover MB, Rudin C, Bianchi MT. Clinical Prediction Models for Sleep Apnea: The Importance of Medical History over Symptoms. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine. 2015.
31. Lee BJ, Kim JY. Identification of Type 2 Diabetes Risk Factor using Phenotypes consisting of Anthropometry and Triglycerides based on Machine Learning. IEEE journal of biomedical and health informatics. 2015.
32. Russo J, Katon W, Zatzick D. The development of a population-based automated screening procedure for PTSD in acutely injured hospitalized trauma survivors. General hospital psychiatry. 2013;35(5):485-91.
33. Cichosz SL, Johansen MD, Hejlesen O. Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. Journal of diabetes science and technology. 2015.
34. Menendez ME, Janssen SJ, Ring D. Electronic health record-based triggers to detect adverse events after outpatient orthopaedic surgery. BMJ quality & safety. 2015.
35. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive

mining of electronic health record data. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e1):e118-24.

36. Inacio MC, Paxton EW, Chen Y, Harris J, Eck E, Barnes S, et al. Leveraging electronic medical records for surveillance of surgical site infection in a total joint replacement population. Infection control and hospital epidemiology. 2011;32(4):351-9.

37. Wojtusiak J, Levy CR, Williams AE, Alemi F. Predicting Functional Decline and Recovery for Residents in Veterans Affairs Nursing Homes. The Gerontologist. 2015.

38. Wasfy JH, Singal G, O'Brien C, Blumenthal DM, Kennedy KF, Strom JB, et al. Enhancing the Prediction of 30-Day Readmission After Percutaneous Coronary Intervention Using Data Extracted by Querying of the Electronic Health Record. Circulation Cardiovascular quality and outcomes. 2015;8(5):477-85.

39. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ. Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. Neurosurgical focus. 2015;39(6):E13.

40. Lee JS, Kim CK, Kang J, Park JM, Park TH, Lee KB, et al. A Novel Computerized Clinical Decision Support System for Treating Thrombolysis in Patients with Acute Ischemic Stroke. Journal of stroke. 2015;17(2):199-209.

41. Croskerry P. From mindless to mindful practice--cognitive bias and clinical decision making. The New England journal of medicine. 2013;368(26):2445-8.

42. Witteveen A, Vliegen IM, Sonke GS, Klaase JM, MJ IJ, Siesling S. Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients. Breast cancer research and treatment. 2015;152(3):627-36.

43. Buckeridge D, Huang A, Hanley J, Kelome A, Reidel K, Verma A, et al. Risk of injury associated with opioid use in older adults. Journal of the American Geriatrics Society. 2010;58(9):1664-70.

44. Ando M, Okamoto I, Yamamoto N, Takeda K, Tamura K, Seto T, et al. Predictive factors for interstitial lung disease, antitumor response, and

survival in non-small-cell lung cancer patients treated with gefitinib. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24(16):2549-56.

45. James ML, Andersen ND, Swaminathan M, Phillips-Bute B, Hanna JM, Smigla GR, et al. Predictors of electrocerebral inactivity with deep hypothermia. The Journal of thoracic and cardiovascular surgery. 2014;147(3):1002-7.

46. Bondareva IB, Jelliffe RW, Andreeva OV, Bondareva KI. Predictability of individualized dosage regimens of carbamazepine and valproate mono- and combination therapy. Journal of clinical pharmacy and therapeutics. 2011;36(5):625-36.

47. Kawazoe Y, Miyo K, Kurahashi I, Sakurai R, Ohe K. Prediction-based threshold for medication alert. Studies in health technology and informatics. 2013;192:229-33.

48. Wang G, Lam KM, Deng Z, Choi KS. Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. Computers in biology and medicine. 2015;63:124-32.

49. Tsoukalas A, Albertson T, Tagkopoulos I. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. JMIR medical informatics. 2015;3(1):e11.

50. Hu YH, Wu F, Lo CL, Tai CT. Predicting warfarin dosage from clinical data: a supervised learning approach. Artificial intelligence in medicine. 2012;56(1):27-34.

51. Wolfson J, Bandyopadhyay S, Elidrisi M, Vazquez-Benitez G, Vock DM, Musgrove D, et al. A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. Statistics in medicine. 2015;34(21):2941-57.

52. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care. 2010;48(11):981-8.

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

comparison between clinician assessment and proprietary database severity rankings. The Annals of pharmacotherapy. 2010;44(11):1718-24.

54. Parke C, Santiago E, Zussy B, Klipa D. Reduction of clinical support warnings through recategorization of severity levels. American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists. 2015;72(2):144-8.

55. Beccaro MA, Villanueva R, Knudson KM, Harvey EM, Langle JM, Paul W. Decision Support Alerts for Medication Ordering in a Computerized Provider Order Entry (CPOE) System: A systematic approach to decrease alerts. Applied clinical informatics. 2010;1(3):346-62.

56. Cornu P, Steurbaut S, Gentens K, Van de Velde R, Dupont AG. Pilot evaluation of an optimized context-specific drug-drug interaction alerting system: A controlled pre-post study. International journal of medical informatics. 2015;84(9):617-29.

57. van der Sijs H, Baboe I, Phansalkar S. Human factors considerations for contraindication alerts. Studies in health technology and informatics. 2013;192:132-6.

58. Lee EK, Wu TL, Senior T, Jose J. Medical alert management: a real-time adaptive decision support tool to reduce alert fatigue. AMIA Annual Symposium proceedings AMIA Symposium. 2014;2014:845-54.

59. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2013;20(3):489-93.

60. Levin SR, Harley ET, Fackler JC, Lehmann CU, Custer JW, France D, et al. Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. Critical care medicine. 2012;40(11):3058-64.

61. Tamblyn R, Eguale T, Buckeridge DL, Huang A, Hanley J, Reidel K, et al. The effectiveness of a new generation of computerized drug alerts in reducing the risk of injury from drug side effects: a cluster randomized trial. Journal of the American Medical Informatics Association : JAMIA. 2012;19(4):635-43.

62. Shamir RR, Dolber T, Noecker AM, Walter BL, McIntyre CC. Machine Learning Approach to Optimizing Combined Stimulation and Medication Therapies for Parkinson's Disease. Brain stimulation. 2015;8(6):1025-32.
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.
64. Montgomery VL, Strotman JM, Ross MP. Impact of multiple organ system dysfunction and nosocomial infections on survival of children treated with extracorporeal membrane oxygenation after heart surgery. Critical care medicine. 2000;28(2):526-31.
65. Beeler GW, Jr. Taking HL7 to the next level. MD computing : computers in medical practice. 1999;16(2):21-4.
66. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. Journal of biomedical informatics. 2012;45(4):689-96.
67. Field TS, Rochon P, Lee M, Gavendo L, Subramanian S, Hoover S, et al. Costs Associated with Developing and Implementing a Computerized Clinical Decision Support System for Medication Dosing for Patients with Renal Insufficiency in the Long-term Care Setting. Journal of the American Medical Informatics Association : JAMIA. 2008;15(4):466-72.
68. Schweitzer EJ. Reconciliation of the cloud computing model with US federal electronic health record regulations. Journal of the American Medical Informatics Association : JAMIA. 2012;19(2):161-5.
69. Kuo AM-H. Opportunities and Challenges of Cloud Computing to Improve Health Care Services. Journal of medical Internet research. 2011;13(3):e67.
70. Kudtarkar P, DeLuca TF, Fusaro VA, Tonellato PJ, Wall DP. Cost-Effective Cloud Computing: A Case Study Using the Comparative Genomics Tool, Roundup. Evolutionary Bioinformatics Online. 2010;6:197-203.
71. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. Genome medicine. 2010;2(8):51.

72. Bildosola I, Río-Belver R, Cilleruelo E, Garechana G. Design and Implementation of a Cloud Computing Adoption Decision Tool: Generating a Cloud Road. PLoS ONE. 2015;10(7):e0134563.
73. Mell P, Grance T. The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology. Gaithersburg, MD: 2011.
74. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. Pharmacoepidemiology and drug safety. 2012;21 Suppl 1:18-22.
75. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiology and drug safety. 2012;21 Suppl 1:1-8.
76. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: Turning a dream into reality. Journal of the American Medical Informatics Association. 2014;21(4):576-7.
77. Kindig D, Stoddart G. What is population health? American journal of public health. 2003;93(3):380-3.
78. Health Level Seven International. HL7 Implementation Guide for CDA Release 2: IHE Health Story Consolidation, DSTU Release 1.1 (US Realm) Draft Standard for Trial Use: Health Level Seven, Inc.; 2012.
79. Kahn MG, Batson D, Schilling LM. Data Model Considerations for Clinical Effectiveness Researchers. Medical care. 2012;50(0):10.1097/MLR.0b013e318259bff4.
80. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association : JAMIA. 2012;19(1):54-60.
81. Guazzelli A, Zeller M, Lin W-C, Williams G. PMML: An open standard for sharing models. The R Journal. 2009;1(1):60-5.
82. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Annals of internal medicine. 2010;153(9):600-6.

83. Observational Health Data Sciences and Informatics. OMOP Common Data Model V5.1 [Internet]. 2017 [March 11, 2017]. Available from: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:single-page.
84. The Observational Health Data Sciences and Informatics. ATHENA Download Page Standardized Vocabularies for OMOP CDM. 2016 [October 23, 2016]. Available from: http://athena.ohdsi.org/.
85. Observational Health Data Sciences and Informatics. CPT4 [Internet]. 2017 [April 27, 2017]. Available from: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:cpt4.
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.
87. WHO International Working Group for Drug Statistics Methodology, WHO Collaborating Centre for Drug Statistics Methodology, WHO Collaborating Centre for Drug Utilization Research and Clinical Pharmacological Services. Drug utilization metrics and their applications. 2003 [cited April 4, 2017]. In: Introduction to Drug Utilization Research [Internet]. Available from: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/Drug%20utilization%20research.pdf.
88. Hall WD, Mant A, Mitchell PB, Rendle VA, Hickie IB, McManus P. Association between antidepressant prescribing and suicide in Australia, 1991-2000: trend analysis. BMJ (Clinical research ed). 2003;326(7397):1008.
89. Dormuth CR, Glynn RJ, Neumann P, Maclure M, Brookhart AM, Schneeweiss S. Impact of two sequential drug cost-sharing policies on the use of inhaled medications in older patients with chronic obstructive pulmonary disease or asthma. Clinical therapeutics. 2006;28(6):964-78; discussion 2-3.
90. Gagne JJ, Maio V, Rabinowitz C. Prevalence and predictors of potential drug-drug interactions in Regione Emilia-Romagna, Italy. Journal of clinical pharmacy and therapeutics. 2008;33(2):141-51.
91. Maxwell M, Heaney D, Howie JG, Noble S. General practice fundholding: observations on prescribing patterns and costs using the defined daily dose method. British Medical Journal. 1993;307(6913):1190-4.

92. Cars O, Mölstad S, Melander A. Variation in antibiotic use in the European Union. The Lancet.357(9271):1851-3.
93. Wertheimer AI. The defined daily dose system (DDD) for drug utilization review. Hosp Pharm. 1986;21(3):233-4, 9-41, 58.
94. Sinnott SJ, Polinski JM, Byrne S, Gagne JJ. Measuring drug exposure: concordance between defined daily dose and days' supply depended on drug class. Journal of clinical epidemiology. 2016;69:107-13.
95. Polk RE, Fox C, Mahoney A, Letcavage J, MacDougall C. Measurement of Adult Antibacterial Drug Use in 130 US Hospitals: Comparison of Defined Daily Dose and Days of Therapy. Clinical Infectious Diseases. 2007;44(5):664-70.
96. Rønning M, Salvesen Blix H, Tange Harbø B, Strøm H. Different versions of the anatomical therapeutic chemical classification system and the defined daily dose – are drug utilisation data comparable? European Journal of Clinical Pharmacology. 2000;56(9):723-7.
97. Graham JW. Missing data analysis: making it work in the real world. Annual review of psychology. 2009;60:549-76.
98. Twisk J, de Vente W. Attrition in longitudinal studies: How to deal with missing data. Journal of clinical epidemiology. 2002;55(4):329-37.
99. Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. Journal of clinical epidemiology. 2002;55(2):184-91.
100. Groenwold RHH, Donders ART, Roes KCB, Harrell JFE, Moons KGM. Dealing With Missing Outcome Data in Randomized Trials and Observational Studies. American Journal of Epidemiology. 2012;175(3):210-7.
101. Grossman R, Bailey S, Ramu A, Malhi B, Hallstrom P, Pulleyn I, et al. Management and mining of multiple predictive models using the predictive modeling markup language. Information and Software Technology. 1999;41(9):589-95.
102. The Data Mining Group. PMML 4.3 - PMML Conformance. 2016 [April 29, 2017]. Available from: http://dmg.org/pmml/v4-3/Conformance.html.

103. The Data Mining Group. PMML 4.3 - General Structure. 2016 [April 29, 2017]. Available from: http://dmg.org/pmml/v4-3/GeneralStructure.html.
104. The Data Mining Group. PMML Version 4.3. 2016 [April 29, 2017]. Available from: http://dmg.org/pmml/pmml-v4-3.html.
105. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.
106. Marquardt A, Newman S, Hattarki D, Srinivasan R, Sushmita S, Ram P, et al., editors. HealthSCOPE: An Interactive Distributed Data Mining Framework for Scalable Prediction of Healthcare Costs. 2014 IEEE International Conference on Data Mining Workshop; 2014 14-14 Dec. 2014.
107. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet (London, England). 2009;373(9665):739-45.
108. Schnabel RB, Rienstra M, Sullivan LM, Sun JX, Moser CB, Levy D, et al. Risk assessment for incident heart failure in individuals with atrial fibrillation. European journal of heart failure. 2013;15(8):843-9.
109. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119(24):3078-84.
110. Kannel WB, D'Agostino RB, Silbershatz H, Belanger AJ, Wilson PW, Levy D. Profile for estimating risk of heart failure. Archives of internal medicine. 1999;159(11):1197-204.
111. Expert Panel on D, Evaluation, and Treatment of High Blood Cholesterol in A. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). JAMA. 2001;285(19):2486-97.
112. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.

113. D'Agostino RB, Russell MW, Huse DM, Ellison RC, Silbershatz H, Wilson PW, et al. Primary and subsequent coronary risk appraisal: new results from the Framingham study. Am Heart J. 2000;139(2 Pt 1):272-81.
114. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Archives of internal medicine. 2007;167(10):1068-74.
115. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Annals of internal medicine. 2008;148(2):102-10.
116. Murabito JM, D'Agostino RB, Silbershatz H, Wilson WF. Intermittent claudication. A risk profile from The Framingham Heart Study. Circulation. 1997;96(1):44-9.
117. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. Stroke. 1994;25(1):40-3.
118. Wang TJ, Massaro JM, Levy D, Vasan RS, Wolf PA, D'Agostino RB, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. Jama. 2003;290(8):1049-56.
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 Guildelines. 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06); 2006 0-0 0.
121. Sartipi K, Yarmand MH, Down DG, editors. Mined-Knowledge and Decision Support Services in Electronic Health. Systems Development in SOA Environments, 2007 SDSOA '07: ICSE Workshops 2007 International Workshop on; 2007 20-26 May 2007.
122. Jeffery AD. Methodological Challenges in Examining the Impact of Healthcare Predictive Analytics on Nursing-Sensitive Patient Outcomes. Computers, informatics, nursing : CIN. 2015;33(6):258-64.

123. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearbook of medical informatics. 2014;9:42-7.
124. The Health Insurance Portability and Accountability Act of 1996 (HIPAA), 45 CFR 160 (2007).

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 2006-2009
Hospital Pharmacy Intern

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

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

Publications

Peer-reviewed Journal Articles and Conference Proceedings

Abedtash H, Holden RJ (2017). *Systematic Review of the Effectiveness of Health-Related Behavioral Interventions Using Portable Activity Sensing Devices (PASDs)*. Journal of the American Medical Informatics Association (JAMIA), 1–14, doi: 10.1093/jamia/ocx006

Abedtash, H., Duke, J. D. (2015). An Interactive User Interface for Drug Labeling to Improve Readability and Decision-Making. *AMIA Annual Symposium Proceedings*, *2015*, 278–286.

Ghasemi Y, **Abedtash H**, Morowvat MH, Mohagheghzadeh A, Ardeshir-Rouhani-Fard S (2015). *Essential oil composition and bioinformatic analysis of Spanish broom (Spartium junceum L.)*. Trends in Pharmaceutical Sciences 1(2):97-104.

Abedtash H, Finnell JT (2013). *A Pilot Study: Integrating an Emergency Department with Indiana's Prescription Drug Monitoring Program*. In Universal Access in Human-Computer Interaction. Applications and Services for Quality of Life (pp. 419-425). Springer Berlin Heidelberg.

Namazi S, Ardeshir-Rouhani-Fard S, **Abedtash H** (2011). *The effect of renin angiotensin system on tamoxifen resistance*. Medical Hypotheses, 77:152-155.

Namazi S, Ardeshir-Rouhani-Fard S, **Abedtash H** (2008). Role of endothelin-1 in tamoxifen resistance: mechanism for a new possible treatment strategy in breast cancer. Medical Hypotheses, 70:109-111.

Karimzadeh I, **Abedtash H**, Mohagheghzadeh A (2005). *A review on molecular modeling and inhibition effect of flavonoids on cytochrome P450 enzyme*. Razi 5:11-25. In Persian.

Peer-reviewed Conference Abstract Podiums

Abedtash H, Finnell JT (2013). *Integrating an Emergency Department with a Prescription Drug Monitoring Program*. AMIA 2013 Annual Symposium, Washington, DC, USA.

*Peer-reviewed Conference Poster Presentation*s

Abedtash H, Duke JD (2016). *CCD2OMOP: An Interoperable Extract-Transform-Load Package to Support the Implementation of OHDSI Software Tools Across*

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

Abedtash H, Duke JD (2015). *An Interoperable Electronic Medical Record-Based Platform for Personalized Predictive Analytics*. OHDSI 2015, Washington, DC, USA.

Sligh J, **Abedtash H**, yang M, Zhang E, Jones J (2016). *A Novel Pipeline for Targeting Breast Cancer Patients on Twitter for Clinical Trial Recruitment*. IUPUI Research Day, Indianapolis, IN, USA.

Yang M, **Abedtash H**, Jones J (2015). *Medical Dictionary for Twitter Pharmacovigilance (MedTP): A Possible Solution to the Challenge of Twitter Mining for Drug-Related Adverse Events*. 2015 Indy Big Data, Indianapolis, IN, USA.

Abedtash H, Finnell JT (2014). *Emergency Physicians Assessment of Opiate Risk from Prescription Drug Monitoring Program Data*. Society for Academic Emergency Medicine (SAEM) 2014 Annual Meeting, Dallas, TX, USA.

Abedtash H, Mohagheghzadeh A, Daneshamouz S (2009). *Release of rutin from controlled release proniosomes/proliposomes, a mechanistically study*. 4th Iranian Controlled Release Conference, Zanjan, Iran.

Abedtash H, Mohagheghzadeh A, Daneshamouz S (2008). *Niosome as a carrier for oral delivery of rutin: preparation and physicochemical characterization*. 11th Iranian Pharmaceutical Sciences Conference, Kerman, Iran.

Ardeshir-Rouhani-Fard S, **Abedtash H**, Morowvat MH, Mohagheghzadeh A, Ghasemi Y (2008). *Analysis of the essential oil and 18S rRNA gene of a new chemotype of Spartium junceum L*. 11th Iranian Pharmaceutical Sciences Conference, Kerman, Iran.

Ardeshir-Rouhani-Fard S, Namazi S, **Abedtash H** (2008). *Tamoxifen resistance mechanism in breast cancer, an approach to a new possible mechanism*. 11th Iranian Pharmaceutical Sciences Conference, Kerman, Iran.

Ardeshir-Rouhani-Fard S, **Abedtash H**, Namazi S (2008). *Role of endothelin-1 in tamoxifen resistance: mechanism for a new possible treatment strategy in breast cancer*. 8th Annual Research Congress of Iranian Medical Sciences Students, Shiraz, Iran.

Daneshamouz S, Mohagheghzadeh A, **Abedtash H** (2007). *Proniosome as a novel carrier for oral delivery of rutin: Preparation and physicochemical characterization*. 67th Congress of FIP, Beijing, China.

Ghasemi Y, **Abedtash H**, Faridi P, Gholami A, Mehregan I, Shams-Ardekani MR, Mohagheghzadeh A (2006). *Antimicrobial essential oil from Smyrniopsis aucheri Boiss*. 10th Iranian Pharmaceutical Sciences Conference, Tehran, Iran.

Rezaei Z, Khabnadideh S, Khalafi-Nezhad A, Roosta A, Heiran M, **Abedtash H**, Faridi P (2005). *Synthesis of clotrimazole derivatives as antifungal agents*. Medicine and Health in the Tropics, Marseille, France.

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

Ghasemi Y, Morowvat MH, **Abedtash H**, Mohagheghzadeh A, Ardeshir-Rouhani-Fard S (2008). *PCR amplification of 18S rRNA gene of Spartium junceum HPSUMS 280*. Pubmed Nucleotide, Accession code: EU605787.