The Office of the National Coordinator for Health Information Technology

Training Data for Machine Learning (ML) to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure

IMPLEMENTATION GUIDE

Prepared by Booz Allen Hamilton for the Office of the National Coordinator for Health Information Technology under Contract No. HHSP233201500132I/ 75P00119F37012

September 2021

Acknowledgements

The authors would like to recognize the important contributions made by the members of the Technical Expert Panel who shared their expertise and provided guidance in the development of this Project:

- Peter Chang, M.D., Co-Director, Center for AI in Diagnostic Medicine, UC Irvine School of Medicine
- Mark DePristo, Ph.D., Founder & Chief Executive Officer, BigHat Biosciences
- Kevin Fowler, President, The Voice of the Patient
- James Hickman, Product Lead, Epic Systems
- Eileen Koski, Director for Health and Data Insights, International Business Machines Corporation (IBM)
- Jarcy Zee, Ph.D., Assistant Professor of Biostatistics, University of Pennsylvania

Table of Contents

1. Introduction

The *Training Data for Machine Learning to Enhance PCOR Data Infrastructure* project (hereafter the Project) led by the Office of the National Coordinator for Health Information Technology (ONC) conducted foundational work to support future applications of artificial intelligence (AI), specifically focused on machine learning (ML) to further health, health care, and patient-centered outcomes research (PCOR), and in turn enhance the adoption and implementation of a PCOR data infrastructure[i]. PCOR is "designed to produce scientific evidence to inform and support health care decisions of patients, families, and providers. PCOR focuses on studying the effectiveness of prevention and treatment options with consideration of the preferences, values, and questions patients face when making health care choices"[ii]. This Project is funded through the PCOR Trust Fund (PCORTF) [iii], created under the Patient Protection and Affordable Care Act of 2010, and managed by the Department of Health and Human Services (HHS) Assistant Secretary for Planning and Evaluation (ASPE). ASPE partners with 12 HHS agencies to lead intradepartmental projects that build data capacity and infrastructure for conducting PCOR.

AI/ML applications have the power to utilize large amounts of real-world clinical data in varied and complex formats to rapidly identify effective treatments, potentially accelerating clinical innovation and supporting evidence-based decisions in clinical settings[iv,v,vi]. However, the wide-spread application and adoption of AI/ML in health care and PCOR is wrought with challenges, including the lack of high-quality training data from which to build and maintain AI applications in health[vii]. This Project was undertaken to address the challenge of the lack of availability of high-quality training datasets. This Project informs future work that aims to leverage AI/ML to develop scientific approaches to support personalized medicine so that providers can eventually match patients to the best treatments based on their specific health conditions, life-experiences, and genetic/phenotypic profiles.

To support the goal of conducting foundational work that will facilitate future applications of AI/ML and enhance PCOR data infrastructure, ONC partnered with the National Institutes of Health (NIH) National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Through this Project, ONC and NIDDK have advanced the application of AI and ML algorithms in PCOR by defining requirements for high-quality training datasets. The Project used data from the United States Renal Data System (USRDS)[viii] to prepare high-quality training datasets and to apply ML techniques for a chronic kidney disease use case of predicting mortality within the first 90 days of dialysis.

A technical expert panel (TEP) assembled for the Project, composed of AI/ML and health IT experts and a patient advocate, was instrumental in vetting the methodology, interpreting the findings, and helping to address the challenges encountered during the training dataset and ML development process. The TEP offered directional guidance and recommendations for other PCOR investigators to build upon the results of this Project and future opportunities related to the development and application of AI/ML to health, healthcare, and PCOR.

This Project facilitates the broader application of AI/ML by PCOR researchers through the resources generated from this Project including the methodology used and lessons learned in building the training dataset and ML models, and recommendations for future projects gathered from the technical experts assembled for this Project (these resources include this Implementation Guide and Final Report).

Foundational knowledge gathered from this Project aligns with the goals of other PCORTF and ASPE funded projects aimed at enhancing the PCOR data infrastructure, including the Patient Matching, Aggregation, and Linking project that developed a framework to address data quality and data sharing, the privacy-preserving record linkage project that facilitates the linking of data from diverse data sources, and the more recent projects such as the building infrastructure and evidence for COVID-19 related research by developing synthetic linked data files or using split-learning ML techniques to enable health information exchange. Evidence generated from this Project also supports multiple federal and HHS investments, including the Precision Medicine Initiative (PMI), the Transitions in Care program conducted in coordination with the Department of Veterans Affairs, and agency-specific, and related NIDDK-funded kidney research programs such as the Kidney Precision Medicine Project.

2. Purpose of the Implementation Guide

This Implementation Guide (IG) prepared for this Project will provide PCOR and other researchers with the detailed methodology for, and lessons learned from, building high quality training datasets and ML models. The methodology described in this IG was vetted with the TEP. The detailed methodology and points to consider described in this document aims to facilitate the application of ML for other kidney disease use cases and thereby enhance adoption and implementation of a PCOR data infrastructure.

3. Guidance to the Readers

The following table provides a list of publicly available project documentation that can help the reader more fully understand the context and content of the Implementation Guide.

Content	What it Contains and its relationship to the Training Data IG
Project Overview	The artifact provides an overview of the Project, the use case selected for the Project, and the overall approach and methodology for building the training datasets and ML models
Implementation Guidance	This artifact provides details on the methodology and snippets of the code used for building the training datasets and ML models in this Project and points to consider for other researchers as they undertake similar work
Data Dictionary	This artifact describes the list of features (variables) in the training dataset that were either taken directly or constructed from the USRDS datasets along with the construction method

4. Glossary and Acronymns

GLOSSARY

Term	Definition
Algorithm	A procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation. Current term of choice for a problem-solving procedure, *algorithm*, is commonly used nowadays for the set of rules a machine (and especially a computer) follows to achieve a particular goal.[ix]
Artificial Intelligence (AI)	Artificial intelligence (AI) is a branch of computer science dealing with the simulation of intelligent behavior in computers; the capability of a machine to imitate intelligent human behavior.[x]
Artificial neural network (ANN)	An artificial neural network consists of a collection of simulated neurons. Each neuron is a node which is connected to other nodes via links that correspond to biological axon-synapse-dendrite connections. Each link has a weight, which determines the strength of one node's influence on another.[xi]
Area under the curve (AUC)	An evaluation metric that considers all possible classification thresholds along a curve that characterizes tradeoffs in classification operating characteristics (such as sensitivity, specificity, precision, recall, etc.). The area under the receiver operating characteristic curve (AUC ROC) is the probability that a classifier will be more confident that a randomly chosen positive example is actually positive than that a randomly chosen negative example is positive.[xii] Similarly, some researchers prefer, for the sake of its distinct interpretation of the resulting evaluation metric, area under the Precision-Recall curve.
Common data element (CDE)	A common data element (CDE) refers to a data element that is common to multiple data sets across different studies, surveys, or registries. The intentional use of CDEs improves data quality and promotes data sharing.[xiii]
Cross validation	A mechanism for estimating how well a model will generalize to new data by testing the model against one or more non-overlapping data subsets withheld from the training set.[xii]
Confusion Matrix	A table with two rows and two columns that reports the number of true positives, true negatives, false positives, and false negatives. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa.[xiv]
Electronic health record (EHR)	An EHR is a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does

Term	Definition
	contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care. One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers across more than one healthcare organization. EHRs are built to share information with other healthcare providers and organizations—such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics—so they contain information from *all clinicians involved in a patient's care.*[xv]
End Stage Kidney (Renal) Disease (ESKD/ESRD)	A medical condition in which a person's kidneys cease functioning on a permanent basis leading to the need for a regular course of long-term dialysis or a kidney transplant to maintain life.[xvi]
Feature	An individual measurable property or characteristic of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.[xvii]
Hyperparameter	The "knobs" that you tweak during successive runs of training a model. For example, learning rate is a hyperparameter.[xii]
Machine Learning (ML)	The process by which a computer is able to improve its own performance (as in analyzing image files) by continuously incorporating new data into an existing statistical model.[xviii]
Model	The representation of what a machine learning system has learned from the training data.[xii]
Patient-Centered Outcomes Research (PCOR)	PCOR helps people and their caregivers communicate and make informed healthcare decisions, allowing their voices to be heard in assessing the value of healthcare options.[xix]
Protected health information (PHI)	The Privacy Rule protects all "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral. The Privacy Rule calls this information "protected health information".[xx]
Personally identifiable information (PII)	As defined in OMB Memorandum M-07-1616, PII refers to information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual.[xxi]

ACRONYMNS

Acronym	Explanation
AI	Artificial intelligence
ANN	Artificial neural network
ASPE	Assistant Secretary for Planning and Evaluation
AUC	Area under the curve
AVF	Arteriovenous fistula
AVG	Arteriovenous graft
AWS	Amazon Web Services
BMI	Body Mass Index
CDC	Centers for Disease Control and Prevention
CDE	Common data element
CKD	Chronic kidney disease
CMS	Centers for Medicare & Medicaid Services
CPMAI	Cognitive Project Management for Artificial Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
CV	Cross validation
EHR	Electronic health record
EPO	Erythropoietin
ESKD	End-stage kidney disease
ESRD	End stage renal disease
GFR-EPI	(Estimated) Glomerular Filtration Rate – CKD Epidemiology Collaboration
HH	Home health
HHS	The Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HS	Hospice
IA	Interagency Assembly
ICD	International Classification of Diseases
IP	Inpatient
IRB	Institutional Review Board
IT	Information Technology
LR	Logistic regression
MEDEVID	Medical Evidence
MICE	Multiple imputations by chained equations
ML	Machine learning
MLP	Multilayer perceptron
MNAR	Missing not at random
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
NIH	National Institutes of Health
ONC	Office of the National Coordinator for Health Information Technology
OP	Outpatient
PCOR	Patient-centered outcomes research
PCORTF	PCOR Trust Fund

Acronym	Explanation
PDIS	Primary disease causing renal failure
PHI	Protected health information
PII	Personally identifiable information
PMI	Precision Medicine Initiative
PMM	Predictive mean matching
Postgres	PostgreSQL
PR	Precision recall
ROC	Receiver operating characteristic
SDOH	Social determinants of health
SGD	Stochastic gradient descent
SN	Skilled nursing unit
TEP	Technical Expert Panel
TX	Transplant
UCSF	University of California San Francisco
UNOS	United Network for Organ Sharing
USRDS	The United States Renal Data System
VA	Veterans Affairs
XGBoost	eXtreme gradient boosting

5. Project Overview

BACKGROUND

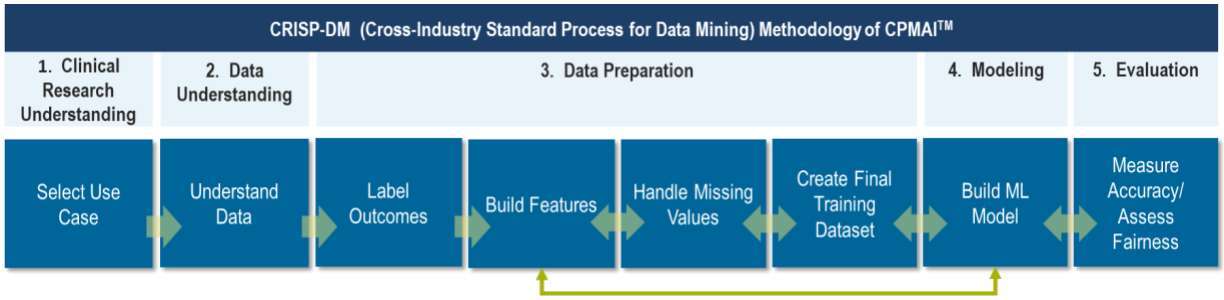
The Project was undertaken to help address the lack of availability of high-quality training datasets for PCOR research. High-quality training datasets that are well-labeled, well-structured, and use common data elements are essential to train prediction models that use ML algorithms, extract features most relevant to specified research goals, and reveal meaningful associations. To start, there is no standard definition of what constitutes a high-quality training dataset and, since ML models are custom tailored to the dataset on which it is trained, many ML practitioners define quality as a function of the ML model that will be developed (for example: some algorithms can inherently handle non-informatively missing values and others cannot). Nevertheless, there are baseline characteristics that all training datasets must have for successful use in developing ML applications. Towards identifying these baseline characteristics, and to develop a high-quality training dataset that can be employed for addressing a kidney disease use case, this Project was implemented based on the following principles:

- Engaging clinical domain experts throughout the course of the Project to ensure that the training datasets and ML models are clinically relevant and patient-centered
- Pre-defining the quality criteria for, and validating its quality of, the prepared training dataset (e.g., by testing the goodness of the imputations performed for missing values)
- Vetting the approaches and methodology used to build the training dataset and ML models, and reviewing the results and findings with a TEP consisting of AI/ML domain experts with broad experience in advanced ML techniques such as deep learning, health information technology (IT) solutions, and patient advocacy
- Capturing and incorporating recommendations and points to consider when building training datasets and ML models provided by various stakeholders throughout the course of the Project
- Disseminating Project progress and obtaining feedback from an Interagency Assembly (IA) with clinical and AI experts from across the federal agencies, including the NIH, FDA, the Centers for Medicare & Medicaid Services (CMS), Veterans Affairs (VA), Centers for Disease Control and Prevention (CDC), Census Bureau, etc.

OVERALL APPROACH FOR BUILDING THE TRAINING DATASET AND ML MODELS

The overall approach for building the training dataset and the ML models is based on the Cognitive Project Management for Artificial Intelligence[xxii] (CPMAI™) methodology, a detailed implementation of the widely used Cross-Industry Standard Process for Data Mining[xxiii] (CRISP-DM) methodology, which defines a robust and proven approach for applying analytics to practical challenges. The CRISP-DM methodology has six phases, five of which are shown in Figure 1 below. The last phase of 'Deployment'—the step of making the model available to end users of the model, such as in a clinic or a hospital or dialysis center—is beyond the scope of this Project.

Figure 1: CRISP-DM Methodology Adapted for Clinical Research Applications



The detailed methodology for each of the steps used in the Project as described in this Implementation Guide aligns to the Patient-Centered Outcomes Research Institute (PCORI) Methodology Standards Checklist to ensure that the overall study design addresses patient centeredness appropriately.

USE CASE AND DATA SOURCE SELECTED FOR THE PROJECT

Use Case

For applying ML in PCOR and health care, clinically compelling patient centric use cases should be identified first rather than tailoring a use case to an existing, easily accessible (open) dataset. From a patient centered perspective, ML is particularly useful to predict *potential outcomes* prior to decisions that patients, in coordination with their providers, must make regarding whether to undergo treatment, which treatment to choose, and how to address potential adverse events once a treatment choice is made. Key to implementing ML for such prediction use cases is access to EHR and clinical research data that has been already collected and stored in various data repositories, such as the federally sponsored one employed in this project.

At the initiation of this Project, upstream kidney disease use cases were considered based on discussions with the TEP, which included a patient advocate, who emphasized the need to move PCOR to focus on research prior to being diagnosed with kidney disease or earlier in kidney disease progression. Such use cases require access to EHR data, which offer high granular information on relevant features at the system-, provider- and patient-level. It is to be noted that EHR data are particularly useful for a broad range of use cases focused on kidney disease. However, the Project Team faced the following challenges in trying to access EHR data stored in multiple federal and private repositories in a timely manner to address an upstream kidney disease use case within the two year project period (for reasons that may impact others pursuing similar applications of machine learning, thus are listed here for others' benefit):

- Data security concerns surrounding patient privacy and confidentiality
- Contractual agreements with health systems that incur additional costs or (as in case of this Project) raise concerns about data-sharing among partnering organizations (e.g., when not all parties are HIPAA-covered entities)
- Requirement for approval by ethical and other regulatory bodies, including the Institutional Review Boards (IRBs), and the differing processes for such approvals across health systems and repositories

Therefore, for this Project, the data source (USRDS) was selected before identifying the use case – *predicting mortality in the first 90 days of dialysis.* This use case focused on patients who had already progressed to ESKD/ESRD to build the training dataset and ML models. USRDS is the national data registry that stores and distributes data on the outcomes and treatments of Chronic Kidney Disease (CKD) and End Stage Kidney Disease (ESKD or End Stage Renal Disease, ESRD[xxiv]) population in the U.S. The kidney disease use case defined for the Project therefore was focused on ESKD/ESRD. Studies focused on ESKD/ESRD are particularly important as it is the only chronic kidney disease stage that is covered through CMS Medicare regardless of the age of the patient (that is, all ESKD/ESRD patients under or over 65 years of age are covered).

ESKD/ESRD is associated with exceedingly high morbidity and mortality. Unfortunately, mortality in the first 90 days of dialysis initiation also remains notably high[xxv,xxvi]. Although risk models do exist for predicting ESKD/ESRD, mortality in the first 90 days of dialysis is not well studied[xxvii,xxviii]. From a patient-centered perspective, a model that predicts mortality in the first 90 days could inform patient-provider joint clinical decisions on whether to initiate dialysis and if so, which type of dialysis to initiate. Therefore, the specific use case—*predicting mortality in the first 90 days of dialysis*—was selected for the following reasons:

- The first 90 days following initiation of chronic dialysis represent a high-risk period for adverse outcomes, including mortality
- Studies of the end-stage kidney population have conventionally excluded this time period from analyses
- While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated[xxix,xxx]
- Tools to identify patients at highest risk for poor outcomes during this early period are lacking; however, such tools may inform discussions between clinicians and patients and their shared decision-making regarding dialysis initiation

The purpose of this use case is to predict mortality in the first 90 days of dialysis initiation to potentially inform shared decision-making between patient and provider.

Data Source

Datasets were obtained from USRDS – the national data registry developed from resources initiated by CMS and its funded ESKD/ESRD networks and maintained by NIDDK – stores and distributes data on the outcomes and treatments of CKD and ESKD/ESRD population in the U.S. While USRDS data does not include complete EHRs for patients suffering from ESKD/ESRD, the data offers multiple advantages for preparing training datasets for this Project:

- It provides the most comprehensive capture of ESKD/ESRD patients who initiated or are currently on dialysis
- It links to several databases, including those related to organ transplantation and mortality
- It incorporates the CMS Form 2728 (the “medical evidence” form) which covers all Americans suffering from ESKD/ESRD, so it is a relevant dataset on which to apply ML to predict ESKD/ESRD-specific outcomes.

- As of 2006, CMS Form 2728 (MEDEVID dataset in USRDS) includes some information on how well prepared the patient was for dialysis – for example: whether the patient was under a nephrologist's care prior to ESKD/ESRD and for how long.
- It incorporates CMS claims data for patients before diagnosis with ESKD/ESRD, which contains information (such as claims for nephrology care) on how well prepared the patient was for dialysis.

However, there are certain limitations with using the USRDS data for the use case—these include:

- CMS claims data are only available for the Medicare population (65 and older, or younger patients diagnosed with ESKD/ESRD).
- CMS Form 2728 is manually completed by clinical providers; therefore, it is prone to data entry errors.
- CMS Form 2728 does not contain the full range of data relevant to kidney risk. For example, Form 2728 has serum creatinine and serum albumin readings but not urine creatinine or urine albumin (other biomarkers diagnostic of kidney disease and prognostic of its progression for some patients).
- Sudden changes in serum creatinine levels contain important information about kidney function; the data on Form 2728 may not be collected frequently enough to detect these changes.
- USRDS data lack continuous validation of its data collection/curation methods, lack complete comorbidity and laboratory data at registration, an initial survival bias in the data due to not including patients who die soon after ESKD/ESRD diagnosis (yet lack Medicare claims or data from CMS Forms), and a lack of accuracy of attributed cause-of-death reporting.

Notwithstanding the limitations, based on the advantages listed above, a robust training dataset of approximately 1.15 million sample size was prepared from the USRDS datasets for applying ML to predict mortality in the first 90 days of dialysis.

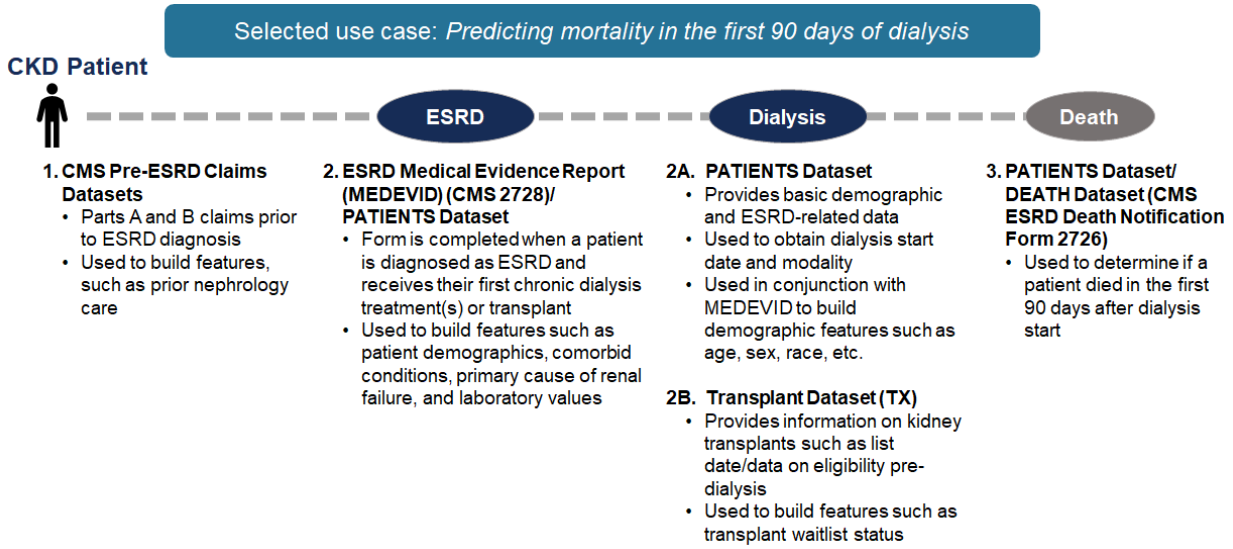
USRDS Dataset Mapping to the Selected Use Case

The overall training dataset was prepared using variables in the USRDS data with clinical relevance and prognostic value for mortality in the first 90 days after dialysis initiation as determined by kidney disease experts from UCSF. Variables selected for the training dataset only include those known on or prior to the first day of dialysis. To ensure the training dataset and ML models are broadly applicable, the training dataset was prepared from routinely collected data available in the following USRDS datasets:

- USRDS core tables: MEDEVID (Medical Evidence), PATIENTS, kidney transplant waitlist tables (WAITSEQ_KI, WAITSEQ_KP, and TX), from 2008 through 2017
- Medicare pre-ESKD claims data (for assessing the degree to which a patient has been prepared for dialysis) from 2008 through 2017

Figure 2 below shows the various USRDS datasets that were used to prepare the training dataset to address the selected use case of *predicting mortality in the first 90 days of dialysis.*

Figure 2: USRDS Dataset[xxiv] Mapping to the Use Case



DEFINING HIGH QUALITY FOR THE TRAINING DATASET

High quality training data sets are essential to train prediction models that use ML algorithms, extract features most relevant to specified research goals, and reveal meaningful associations. Challenges surrounding the availability of high-quality training datasets include:

- Real world data collected via electronic health record (EHR) systems or from clinical research studies, registry-based data, and other data collection systems are complex, diverse, and often noisy, error-prone, have incorrect, outlier or missing values, and have inconsistent measures and values across multiple facilities, even within the same health care setting
- Variables, even those often considered to be core features in a training dataset (e.g., dates, sex, race, ethnicity), are often not collected in a standardized format and can lack proper annotations Duplicate datasets for patients within the same EHR or data collection systems due to lack of provenance or audit trail of the data
- Representativeness of observations/patients captured within an EHR system
- Insufficient quantity of data with desired features for a specific ML use case
- Regulatory and proprietary obstacles to accessing EHR data

Health care providers and patients alike need to have high confidence the clinical decision supporting predictive or classifier AI tools developed are accurate and reliable. The availability of high-quality training datasets is therefore a fundamental requirement for developing and deploying ML tools in clinical settings.

Building a high-quality training dataset and capturing the details of the methodology used and the lessons learned in the process was a primary objective of the Project. Towards that objective, the criteria for high quality were defined with input from various stakeholders, including the TEP. The criteria[xxxi] and how they were applied to the training dataset are shown in Table 1 below.

Table 1: Criteria for a High-Quality Training Dataset

Quality Criteria	How addressed in the Training Dataset
Features cleaned and correctly labeled (well-labeled)	• Removed or flagged outliers, erroneous, suspicious, duplicate, and inconsistent values • Documented how outliers/inconsistencies were addressed across USRDS datasets (e.g., inconsistent coding practices, units, definitions) • Documented and validated any constructed or derived features, to ensure that methods/ equations were selected and applied correctly
Dataset reliable and well curated (well-structured)	• Merging and joining done correctly • Inclusion and exclusion criteria applied correctly (such as only including patients with valid dialysis start date, excluding patients <18, etc.) • Missing data patterns documented and addressed (Medicare pre-ESKD/ESRD claims are missing for those who do not qualify for Medicare prior to ESKD/ESRD diagnosis) • Centering/scaling/standardizing some variables for analysis or balancing the data based on the algorithm that was used • Excluded operational factors such as location, provider, and masked dates when building features • Train/test/validation split done such that the training data is representative of the rest of the data • Data dictionary created
Use common data elements (CDEs)	• Used CDEs for constructed features • For features pulled directly from USRDS dataset, CDEs were based on what was used by USRDS

METHDOLOGY FOR BUILDING THE TRAINING DATASET AND ML MODELS – OVERVIEW

Data De-identification

USRDS provides 'limited datasets' with most of the personally identifiable information (PII) removed but retaining certain limited PII such as dates and geographic (location) variables under a 'controlled-access' model requiring oversight (per Federal Human Subjects Protection regulation) by an Institutional Review Board (IRB; often referred to as Ethics Committees in other nations). To comply with requirements from the Project Team's study IRB (from UCSF), these two variables were de-identified before use in this Project. USRDS data received in sas7bdat format were de-identified as per the Safe Harbor method of the Health Insurance Portability and Accountability Act (HIPAA)[xxxii] using a SAS script. All date variables in USRDS—other than variables which contain only the year (with no month or day information)—were de-identified by offsetting all date fields by a randomly-chosen number specific to each patient included in the USRDS data. For location variables, the zip code and county Federal Information Processing Standard Publication (FIPS)

codes variables were deleted. The accuracy of the date de-identification was validated by comparing a sample of the relative date ranges in the de-identified data to the relative date ranges in the source data (for additional details, refer to section 6.2.0 in the Implementation Guidance).

Building the Cohort and Outcome Variable

The following criteria was applied to the dataset for selecting the cohort for the Project:

- An existing date of first dialysis treatment (n=3,096,526)
- Death date not before first dialysis treatment (n=3,096,515)
- Adults (age >=18 years old) (n=3,065,026)
- Incident year from 2008-2017 (n=1,150,195)

The outcome variable for the selected use case is whether a patient died within the first 90 days of dialysis initiation. Additional details for preparing the study cohort are available in sections 6.2.1 and 6.2.4 in the Implementation Guidance.

Feature Selection

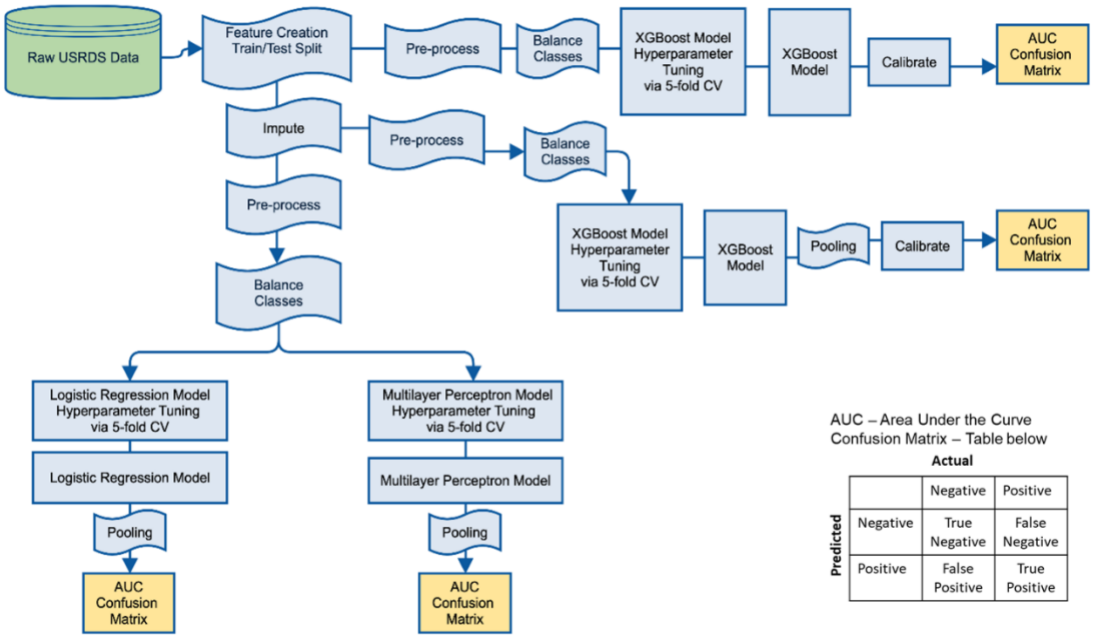
The full training dataset of 1,150,195 derived from the raw USRDS datasets was developed by building features that had clinical relevance and prognostic value to the use case – *predicting mortality in the first 90 days of dialysis*. Each feature captures information known about a patient on or prior to the date of dialysis initiation. The final structure of the training dataset, which was used to train and test the ML models, consisted of 188 features, and has one record per patient. Two sets of features were included in the training dataset – features taken directly from the USRDS datasets (e.g., age, race, and hemoglobin) and features that were constructed (e.g., time on kidney transplant waitlist, number of pre-ESKD/ESRD claims). A detailed list of both sets of features, including the construction methods are provided in the Data Dictionary. Kidney disease experts from UCSF (part of the Project Team) defined the upper and lower bounds of the clinical and laboratory features (e.g., height, weight, serum creatinine) that were used in the training dataset, such that any values outside these bounds were considered clinically impossible; these outlier values were set as missing (for additional details, refer to section 6.2.18 in the Implementation Guidance).

Building the Training Datasets and the ML Models

Figure 3 below shows the overall methodology undertaken to build the training datasets, the data flow through the ML models, and the output of those models. The training dataset with the full set of features was partitioned randomly into 10 stratified non-intersecting subsets to handle the large data size more effectively for modeling. These 10 partitions were further split into training and testing datasets at approximately a 70/30 ratio to allow sufficient data to both train and robustly evaluate the models. Multiple imputation was done using the 'mice' (multiple imputations by chained equations[xxxiii]) library (version 3.13.0) in R and using five imputations to target at least 95% relative efficiency under commonly-adopted assumptions about the fraction of missing information (not explicitly derived, however, in the case of models and associated estimands/predictions entertained here)[xxxiv]. Two datasets—a non-imputed dataset and an imputed dataset—were prepared and utilized for ML modeling. More information on the MICE imputation and predictive mean matching (PMM) method used in this Project can be found in Section 6.2.19 of the Implementation Guidance.

Clinical and laboratory variables with fewer than 40% missing values were included as features in the training dataset because multiple imputations are not advised for any application beyond hypothesis generation when features contain more than 40% missing values, without very restrictive assumptions about how the missing values occurred[xxxv,xxxvi]. For both datasets, clinical and laboratory variables that had missing values for more than 40% of patients were excluded. Variables that were missing in less than 40% of patients were imputed to prepare the imputed dataset – these included height, weight, BMI[xxxvii], serum creatinine, serum albumin, hemoglobin, and GFR-EPI[xxxviii].

Figure 3: Overview of the Methodology for Building the Training Dataset and the ML Models



Three ML algorithms were selected with input from the TEP to provisionally test the training dataset: eXtreme gradient boosting (XGBoost), logistic regression (LR), and multilayer perceptron (MLP), an artificial neural network implementation. These algorithms are a mixture of non-parametric (XGBoost) and parametric (LR and MLP) models.

- XGBoost is a popular implementation of gradient boosted decision trees because it performs especially well for tabular data, can be applied to a wide array of use cases, data types, and desired prediction outcomes (regression vs classification), and can handle non-informative randomly-missing values by default[xxxix]. Such tree-based algorithms learn branch directions for missing values during training, which allows for a comparison between models run on non-imputed data versus models run on imputed data.
- Logistic regression is a classic categorization model that can be used to examine the association of (categorical or continuous) independent variable(s) with one binary dependent variable. However, it requires that the input dataset have no missing values.

- Multilayer perceptron is a class of hierarchical artificial neural network (ANN) that consists of at least three layers of nodes—an input layer, a hidden layer and an output layer—to carry out the process of ML. They are used for tabular datasets and classification prediction problems.

The XGBoost models were prepared using both the non-imputed dataset containing missing lab values and the imputed dataset, whereas the LR and MLP models were prepared using only the imputed datasets as these cannot handle missing values.

ML algorithms have differing requirements for the input training dataset; hence, to prepare the training dataset for XGBoost, logistic regression, and multilayer perceptron models, several additional data processing steps were performed (for additional details, refer to sections 6.3.1.1, 6.3.2.1, 6.3.3.1, and 6.3.4.3 in the Implementation Guidance).

- The input of all three models must be numeric so all categorical features were one-hot encoded into numeric indicators of each factor in the categorical features (e.g., the sex feature was converted into 3 columns: sex_1 (male), sex_2 (female), sex_3 (unknown) through one-hot encoding). Since XGBoost models take numeric values as input and can handle missing values and class imbalance, the XGBoost model can use the training dataset after one-hot encoding the categorical features.
- Logistic regression and multilayer perceptron models[xl] cannot inherently handle missing values as opposed to a tree-based model like XGBoost which learns to handle missing values during training; therefore, the specific numeric pre-ESKD/ESRD claims features with a large percentage of missing data (~40%) were removed from the training dataset[xli]. Only the binary pre-ESKD/ESRD features, which were converted to categorical (i.e., 0=not present, 1=present, 2=missing), were retained in the training dataset for these two models. This effectively allowed retaining the meaning of whether the data was present or missing for the claims features.
- The numeric features for logistic regression and multilayer perceptron models were scaled and normalized as follows:
  - Removed features that had zero variance (variables that have only a single value) from the training dataset because the presence of these variables does not add information to the model[xlii,xliii]
  - Numeric variables constructed from the pre-ESKD/ESRD Medicare claims with missing values (such as claims counts, diagnosis groupings, etc.) were removed and only the binary features (such as indicators for claims in each care setting, indicators for each diagnosis group, and indicators for pre-ESKD/ESRD claims) were kept.

  - Standardized each numeric feature to have a mean of zero and a standard deviation of one—the mean of each numeric feature was subtracted from each value and then divided by the standard deviation. Standardization allows for comparison of multiple features in different units and the penalty (e.g., L1) will be applied more equally across the features. Both logistic regression and multilayer perceptron models will learn the importance of features better and faster when they aren't overwhelmed by a feature with a much larger range than the others.
- Neither logistic regression nor multilayer perceptron models perform well without some additional tailoring when the outcome variable is imbalanced (or heavily skewed towards one outcome). The outcome variable (died in 90 days) in the training dataset for these two models was rebalanced through weighting (the weight parameter in the model to give more weight to the minority class and less to the majority class). Balancing the data ensures that the models have sufficient data from both outcome classes (died vs. survived) on which to train. This results in a better trade-off between sensitivity and specificity, which is important for this dataset where mortality is predicted.

Hyperparameter tuning varied between the non-imputed and imputed datasets. Hyperparameters were tuned for the *non-imputed* dataset with a Bayesian optimization approach, and then a 5-fold cross validation was used to identify the optimal hyperparameters for the model. The best performing model was evaluated by the selecting the hyperparameter combination with the highest AUC. Hyperparameters were tuned for the *imputed* datasets using a two-tiered approach: first, Bayesian optimization and 5-fold cross validation were used for each imputed dataset to narrow the ranges for the hyperparameter space. The highest and lowest values for each hyperparameter over the five imputed datasets[xliv] were set as the new ranges to use in a random grid search. From the new hyperparameter space, 25 hyperparameter combinations were randomly generated and tested. For each hyperparameter combination, the prediction scores for each imputed dataset were pooled via averaging per Rubin's rules[xlv] (performing analysis on each imputed dataset and averaging the parameter estimates to obtain a single estimate so that the variance estimates would reflect the appropriate uncertainty surrounding parameter estimates). These averaged predictions were used to calculate the AUC for each hyperparameter combination. The best performing model was evaluated by the selecting the hyperparameter combination with the highest AUC (for additional details, refer to sections 6.3.1.2, 6.3.2.2, 6.3.3.1, and 6.3.4.4 in the Implementation Guidance).

To explore how closely the predicted events rates align to the observed rates (which is more informative to clinicians) across the full range of predicted risk scores, both XGBoost models were calibrated using a non-parametric isotonic regressor trained on 66% of the testing dataset and evaluated on the remaining 33% of the testing dataset. Calibration (reliability) curves were plotted to reveal each prediction score decile, the number of patients that fall into each decile, and the proportion of patients in each decile who actually died in the first 90 days following dialysis initiation.

All models were evaluated using conventional metrics – receiver operating characteristic (ROC) area under the curve (AUC) and a confusion matrix (used to calculate metrics such as sensitivity, specificity, positive predictive value, likelihood ratio, F1 score, etc.). Additional details are provided for the models in sections 6.3.1 – 6.3.4 in the Implementation Guidance.

6. Implementation Guidance

Implementation Guidance provides details on the methodology along with snippets of the code used for building the training datasets and ML models in this Project, and points to consider for other researchers as they undertake similar work.

7. Data Dictionary

The final structure of the training dataset, which was used to train and test the ML models, consists of 188 features, and has one observation per patient. Two sets of features were included in the training dataset: features taken directly from the USRDS datasets and features that were constructed. The full list of features and the detailed method for the features that were constructed from PATIENTS, MEDEVID and Medicare pre-ESKD/ESRD claims data are provided in the Data Dictionary.

8. References

[i] https://www.healthit.gov/topic/scientific-initiatives/building-data-infrastructure-support-patient-centered-outcomes-research
[ii] https://aspe.hhs.gov/patient-centered-outcomes-research-trust-fund
[iii] https://aspe.hhs.gov/training-data-machine-learning-enhance-patient-centered-outcomes-research-pcor-data-infrastructure
[iv] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018 Oct;2(10):719-731. doi: 10.1038/s41551-018-0305-z. Epub 2018 Oct 10. PMID: 31015651.
[v] Girosi F, Mann S, Kareddy V. Narrative Review and Evidence Mapping: Artificial Intelligence in Clinical Care. Patient-Centered Outcomes Research Institute; February 2021. Prepared by RAND under Contract No. IDIQ-TO#22-RAND-ENG-AOSEPP-04-01-2020. https://www.pcori.org/research-results/evidence-synthesis/evidence-maps-and-evidence-visualizations/artificial
[vi] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial Intelligence Transforms the Future of Health Care. Am J Med. 2019 Jul;132(7):795-801. doi: 10.1016/j.amjmed.2019.01.017. Epub 2019 Jan 31. PMID: 30710543; PMCID: PMC6669105.
[vii]Artificial Intelligence for Health and Health Care, JASON Report 2017: https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealth care12122017.pdf
[viii] https://www.usrds.org/
[ix] https://www.merriam-webster.com/dictionary/algorithm#note-1
[x] https://www.merriam-webster.com/dictionary/artificial%20intelligence
[xi] Artificial intelligence (3rd ed.). Addison-Wesley Pub. Co. 1992. ISBN 0-201-53377-4.
[xii] https://developers.google.com/machine-learning/glossary#a
[xiii]https://registries.ncats.nih.gov/glossary/common-data-element-cde/#:~:text=A%20common%20data%20element%20(CDE,quality%20and%20promotes%20data%20sharing
[xiv] Tharwat A. (August 2018). "Classification assessment methods". Applied Computing and Informatics. doi:10.1016/j.aci.2018.08.003
[xv] https://www.healthit.gov/faq/what-electronic-health-record-ehr
[xvi] https://www.cms.gov/Medicare/Coordination-of-Benefits-and-Recovery/Coordination-of-Benefits-and-Recovery-Overview/End-Stage-Renal-Disease-ESRD/ESRD#:~:text=End%2DStage%20Renal%20Disease%20(ESRD)%20is%20a%20medical%20condition,kidney%20transplant%20to%20maintain%20life.
[xvii] https://en.wikipedia.org/wiki/Feature_(machine_learning)
[xviii] https://www.merriam-webster.com/dictionary/machine%20learning
[xix] https://www.pcori.org/research-results/about-our-research/patient-centered-outcomes-research
[xx] https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html#:~:text=The%20Privacy%20Rule%20protects%20all,health%20information%20(PHI).%22
[xxi] https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act#:~:text=The%20term%20%E2%80%9CPII%2C%E2%80%9D%20as,linkable%20to%20a%20specific%20individual.
[xxii] https://www.cognilytica.com/cpmai-methodology/
[xxiii] https://www.sv-europe.com/crisp-dm-methodology/
[xxiv] USRDS site and dataset refers to ESKD as ESRD; however, ESKD is used for the purposes of this document.
[xxv] JM Souci, et al. Early death in dialysis patients: risk factors and impact on incidence and mortality rates. Journal of the American Society of Nephrology. 1996; 7 (10): 2169-2175. DOI: https://[doi.org/10.1681/ASN.V7102169
[xxvi] Chan KE, Maddux FW, Tolkoff-Rubin N et al. Early outcomes among those initiating chronic dialysis in the United States. Clin J Am Soc Nephrol 2011; 6: 2642–2649. DOI: https://[doi.org/10.2215/CJN.03680411
[xxvii] Kevin E. Chan, et al. Early Outcomes among Those Initiating Chronic Dialysis in the United States. Clin J Am Soc Nephrol. 2011 Nov; 6(11): 2642–2649. DOI: https://doi.org/10.2215/CJN.03680411 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3359565/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3359565/
[xxviii] Marlies Noordjiz, Kitty J. Jager, Increased mortality early after dialysis initiation: a universal phenomenon. Official Journal of the International Society of Nephrology, January 2014. Volume 85, Issue 1, Pages 12–14. https://www.kidney-international.org/article/S0085-2538(15)56151-6/fulltext DOI: https://doi.org/10.1038/ki.2013.316

[xxix] Amber O. Molnar et al., Risk factors for unplanned and crash dialysis starts: a protocol for a systematic review and meta-analysis. Systematic Reviews 2016; 5: 117. Published online 2016 Jul 19. doi: 10.1186/s13643-016-0297-2. PMCID: PMC4950106. PMID: 27431915
[xxx] Kevin E. Chan et al., Early Outcomes among Those Initiating Chronic Dialysis in the United States. Clin J Am Soc Nephrol. 2011 Nov; 6(11): 2642–2649. doi: 10.2215/CJN.03680411. PMCID: PMC3359565. PMID: 21959599
[xxxi] List of criteria for defining a high quality training dataset is not available in the published literature; therefore, the specific criteria were compiled based on experiences of the Project Team and ML community websites and vetted with the Technical Expert Panel (TEP) to finalize. Some of the websites reviewed include: http://www.cs.ust.hk/~qyang/Docs/2003/Data_Preparation_for_Data_Mining_ZZY.pdf; https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/; https://towardsdatascience.com/5-steps-to-correctly-prep-your-data-for-your-machine-learning-model-c06c24762b73; https://www.kdnuggets.com/2018/12/six-steps-master-machine-learning-data-preparation.html; https://www.cloudfactory.com/training-data-guide; https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better
[xxxii] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
[xxxiii] van Buuren S, Groothuis-Oudshoorn K (2011). “mice: Multivariate Imputation by Chained Equations in R.” Journal of Statistical Software, 45(3), 1-67. https://www.jstatsoft.org/v45/i03/.https://www.jstatsoft.org/v45/i03/ DOI: http://dx.[doi.org/10.18637/jss.v045.i03
[xxxiv] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. In D. B. Rubin (Ed.), Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. (pp. 114). DOI:10.1002/9780470316696
[xxxv] Janus Christian Jakobsen et. al., When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowchart. BMC Med Res Methodol. 2017 Dec 6;17(1):162. doi: 10.1186/s12874-017-0442-1
[xxxvi] Missing values will be imputed in the ‘Missing Data Imputation’ section, therefore only features with fewer than 40% missing values are included as features in the training dataset.
[xxxvii] BMI is calculated as part of the imputation process from the imputed values of height and weight. GFR-EPI is calculated as part of the imputation process from the imputed values of serum creatinine, using the CKD-EPI equation. These variables are not imputed.
[xxxviii] https://www.kidney.org/content/ckd-epi-creatinine-equation-2009
[xxxix] https://xgboost.readthedocs.io/en/latest/faq.html
[xl] García-Laencina, P.J., Sancho-Gómez, J., Figueiras-Vidal, A.R. (2009). Pattern classification with missing data: a review. Neural Comput & Applic. DOI 10.1007/s00521-009-0295-6. https://sci2s.ugr.es/keel/pdf/specific/articulo/pattern-classification-with-missin-data-a-review-2009.pdf
[xli] Since the pre-ESKD/ESRD claims features have a reason behind the missingness (not all patients in the study cohort have Medicare pre-ESKD/ESRD claims), and do so in a way that does not meet an implicit assumption of missing at random (i.e., that patients with Medicare pre-ESKD/ESRD claimn are NOT a random sample of all study cohort members), imputation should not be used to fill in these missing values. In other words, it is suspected pre-ESKD/ESRD claims were informatively missing across the study cohort, and applying any method without acknowledging this assumption would lead to biased estimates of 90-day mortality risk.
[xlii] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer. ISBN: 9781461468493 DOI: 10.1007/978-1-4614-6849-3
[xliii] Kuhn, M. (2008). Building predictive models in R using the caret package. *J Stat Softw*, *28*(5), 1-26. DOI: http://dx.[doi.org/10.18637/jss.v028.i05
[xliv] 5 imputed datasets were generated as part of the multiple imputations to handle missing values
[xlv] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons. DOI:10.1002/9780470316696