The Office of the National Coordinator for Health Information Technology

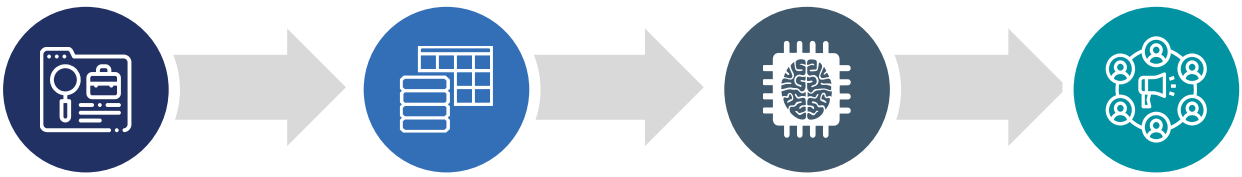
Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research Data Infrastructure

Project Goal

Conduct foundational work to advance the future application of artificial intelligence (AI)/machine learning (ML) for patient-centered outcomes research (PCOR) by generating high-quality training datasets that can be used in ML models for a kidney disease use case

Objectives

- Prepare high-quality training datasets from the United States Renal Data System (USRDS) data to address a kidney disease use case
- Develop ML models based on three algorithms – eXtreme gradient boosting (XGBoost), logistic regression, and multilayer perceptron – to provisionally test the training datasets
- Validate the approaches for building the ML models by evaluating their performance using conventional metrics such as area under the curve (AUC)
- Disseminate project outputs that future researchers can refer to when preparing training datasets and ML models for new kidney disease use cases



DATA SOURCE & USE CASE SELECTION

HIGH-QUALITY TRAINING DATASETS DEVELOPMENT

MACHINE LEARNING MODELS DEVELOPMENT

PROJECT OUTPUTS – DISSEMINATION

Prepared for the Office of the National Coordinator for Health Information Technology under contract number HHSP233201500132I

Data Source & Use Case Selection

Data Source: **United States Renal Data System (USRDS)**

Use Case: **Predicting mortality in the first 90 days of dialysis**

The first 90 days following initiation of chronic dialysis in end-stage kidney disease patients represent a high-risk period for adverse outcomes, including mortality.

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.

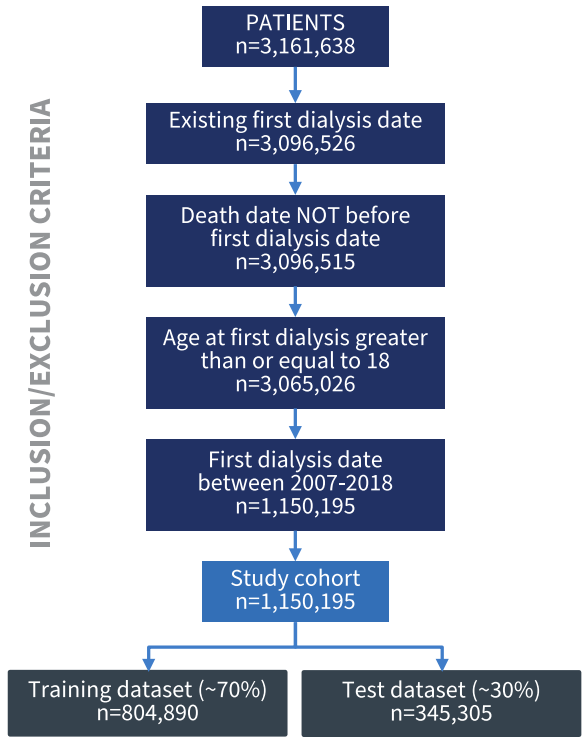
Studies of the end-stage kidney population have conventionally excluded the first 90 days from analyses.

Tools to identify patients at highest-risk for poor outcomes during this early period are lacking.

High Quality Training Datasets Development

USRDS DATA → COHORT SELECTION → FEATURE ENGINEERING → TRAINING DATASET

METHODOLOGY



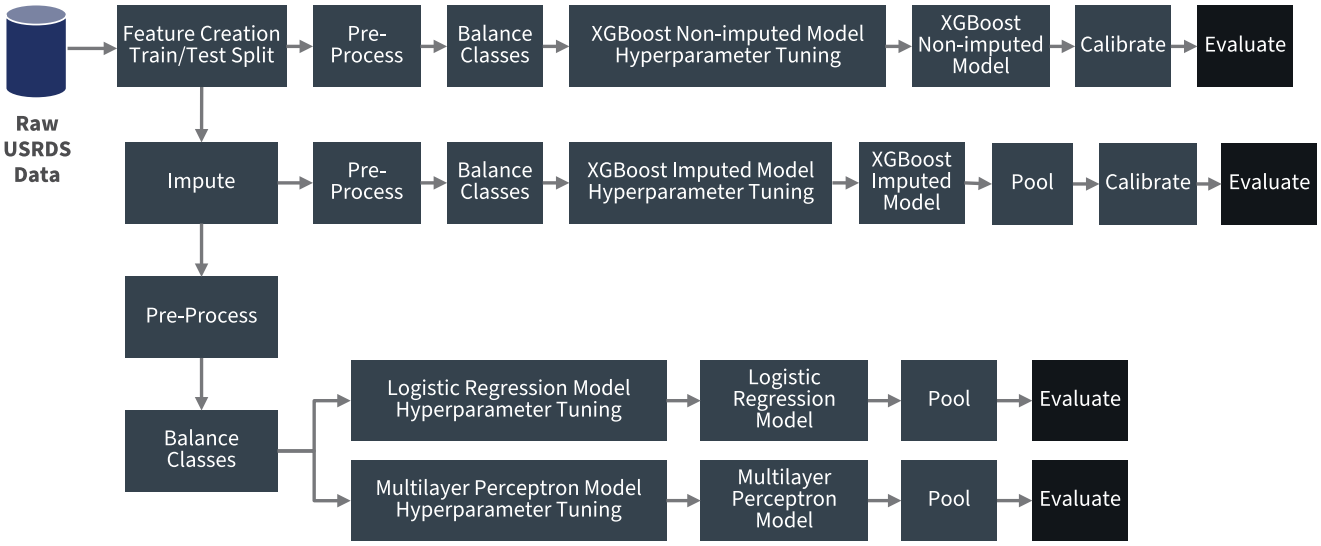
RESULTS

- High-quality training dataset criteria:
  - *Features cleaned and correctly labeled (well-labeled)*
  - *Dataset reliable and well curated (well-structured)*
  - *Features use common data elements*
- 7.5% of patients died in the first 90 days of dialysis in the study cohort
- Training dataset includes 188 features, including demographics, prior care, clinical variables, comorbidities, patient education
- Two versions of the dataset prepared: imputed (using multiple imputations by chained equations) and non-imputed
- Full dataset divided into a training and a test dataset using a 70%-30% split

Machine Learning Models Development

TRAINING DATASET → PRE-PROCESS DATA → TUNE HYPERPARAMETERS → ML MODELS

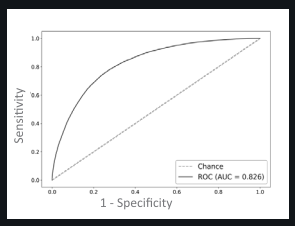
METHODOLOGY



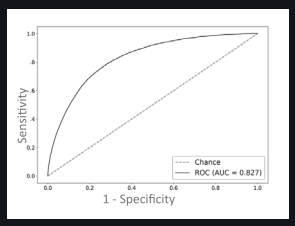
RESULTS

- Area under the receiver operating characteristic curve (ROC AUC) ranged from 0.811 to 0.827
- Top features ranked in XGBoost and logistic regression models include patient age, whether the patient had inpatient stay claims, had received exogenous erythropoietin (anti-anemic treatment), serum albumin value, and presence of arteriovenous fistula (for hemodialysis)

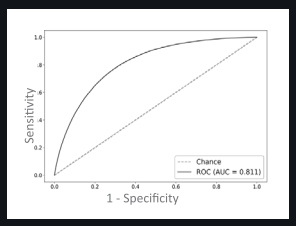
Area Under the Curve (AUC)



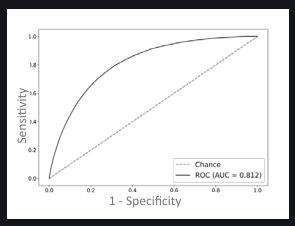
XGBoost Non-imputed
AUC = 0.826



XGBoost Imputed
AUC = 0.827



Logistic Regression
AUC = 0.811



Multilayer Perceptron
AUC = 0.812

Project Outputs - Dissemination

Click Below to Access

Final Report | **Implementation Guide** | **Code** | **Webinar Slides**