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

Project Webinar

September 15, 2021

The Office of the National Coordinator for
Health Information Technology

Meeting Agenda

1. Introductions
2. Webinar Objective
3. Introduction to ONC
4. Project Overview
    - Enhancing PCOR Infrastructure
    - Goal and Objectives
    - Project Approach
5. Kidney Disease & Use Case
6. Project Activities – Methodology & Results
    - USRDS Data Mapping to Use Case
    - Training Dataset
    - Machine Learning (ML) Models
7. Recommendations for Future Applications of AI/ML to PCOR and Health Care
8. Project Resources – Overview & Locations
9. Q&A

Introductions

Matt Rahn

Deputy Director, Standards Division
The Office of the National Coordinator for Health IT

Susan Tenney, Ph.D.

Project Director
Booz Allen Hamilton

Michelle Estrella, M.D.

Executive Director
Kidney Health Research Collaborative/University of California San Francisco

Summer Rankin, Ph.D.

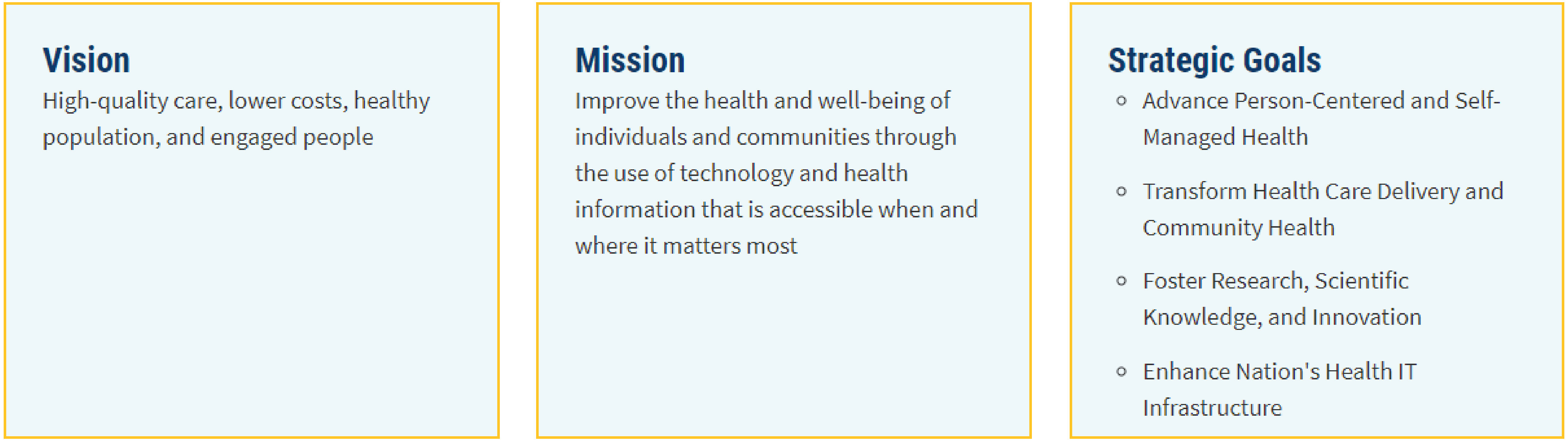
Senior Data Scientist
Booz Allen Hamilton

Webinar Objective

- Disseminate Project activities and resources to external audiences, including Federal agencies, academic institutions, health IT organizations, patient and professional health organizations
  - Describe the background of the project
  - Share high-level project activities and findings
  - Inform the location of resources generated in the Project (Final Report, Implementation Guide, etc.) for other researchers to use

Introduction to ONC

- The Office of the National Coordinator for Health Information Technology (ONC) is located within the Office of the Secretary of the Department of Health and Human Services (HHS)
  - Principal federal entity charged with coordination of nationwide efforts to implement and use the most advanced health information technology and the electronic exchange of health information.
  - Focused on two strategic objectives
    - Advance the development and use of health IT capabilities, and
    - Establish expectations for data sharing



Vision

High-quality care, lower costs, healthy population, and engaged people

Mission

Improve the health and well-being of individuals and communities through the use of technology and health information that is accessible when and where it matters most

Strategic Goals

- Advance Person-Centered and Self-Managed Health
- Transform Health Care Delivery and Community Health
- Foster Research, Scientific Knowledge, and Innovation
- Enhance Nation's Health IT Infrastructure

Project Overview – Enhancing PCOR Infrastructure

- ONC in coordination with NIH's National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) conducted foundational work to support future applications of artificial intelligence (AI)/ machine learning (ML) for PCOR, and in turn enhance the adoption and implementation of a PCOR data infrastructure
- PCOR aims to produce new scientific evidence that informs and supports the health care decisions of patients, families, and their health care providers
- Project is funded through the PCOR Trust Fund (PCORTF), established under the Patient Protection and Affordable Care Act of 2010, and managed by DHHS Assistant Secretary for Planning and Evaluation (ASPE)
- PCORTF supports intradepartmental projects to build data capacity for PCOR infrastructure
- Project period: September 2019 – September 2021

ASPE
OFFICE OF THE ASSISTANT SECRETARY
FOR PLANNING AND EVALUATION

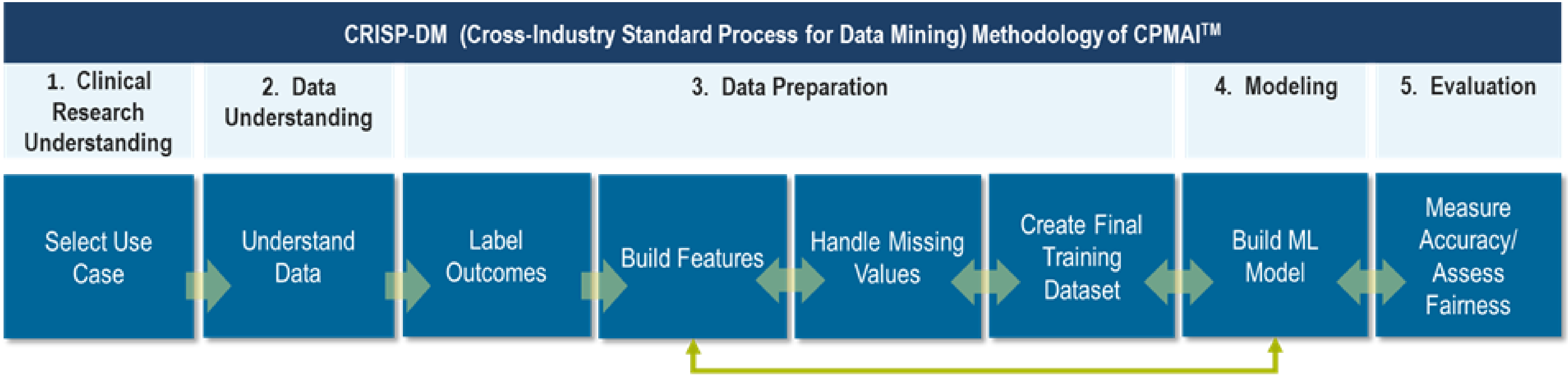
OS PCOR TF
OFFICE OF THE SECRETARY
PATIENT-CENTERED OUTCOMES
RESEARCH TRUST FUND

Project Overview – Project Goal & Objectives

- Goal: Conduct foundational work to advance the future application of AI/ML for PCOR by generating high-quality training datasets that can be used in ML models for a chronic 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 – predicting mortality within the first 90 days of dialysis
  - Develop ML models based on three algorithms – eXtreme gradient boosting (XGBoost), logistic regression, and multilayer perceptron (MLP) – 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 resources generated in the project that future researchers can refer to when preparing training datasets and ML models for new kidney disease use cases

Project Overview – Approach

- A Technical Expert Panel (TEP) composed of experts from AI/ML and health information technology and a patient advocate provided feedback and insight on the Project approach, the criteria for high-quality training datasets, and the methods and results from building the training datasets and ML models
- Used the Cognitive Project Management for Artificial Intelligence methodology (CPMAI^TM) – a detailed implementation of the widely used Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which defines a robust and proven approach for applying analytics to practical challenges (*Note*: the sixth phase of deployment is not applicable to this Project)

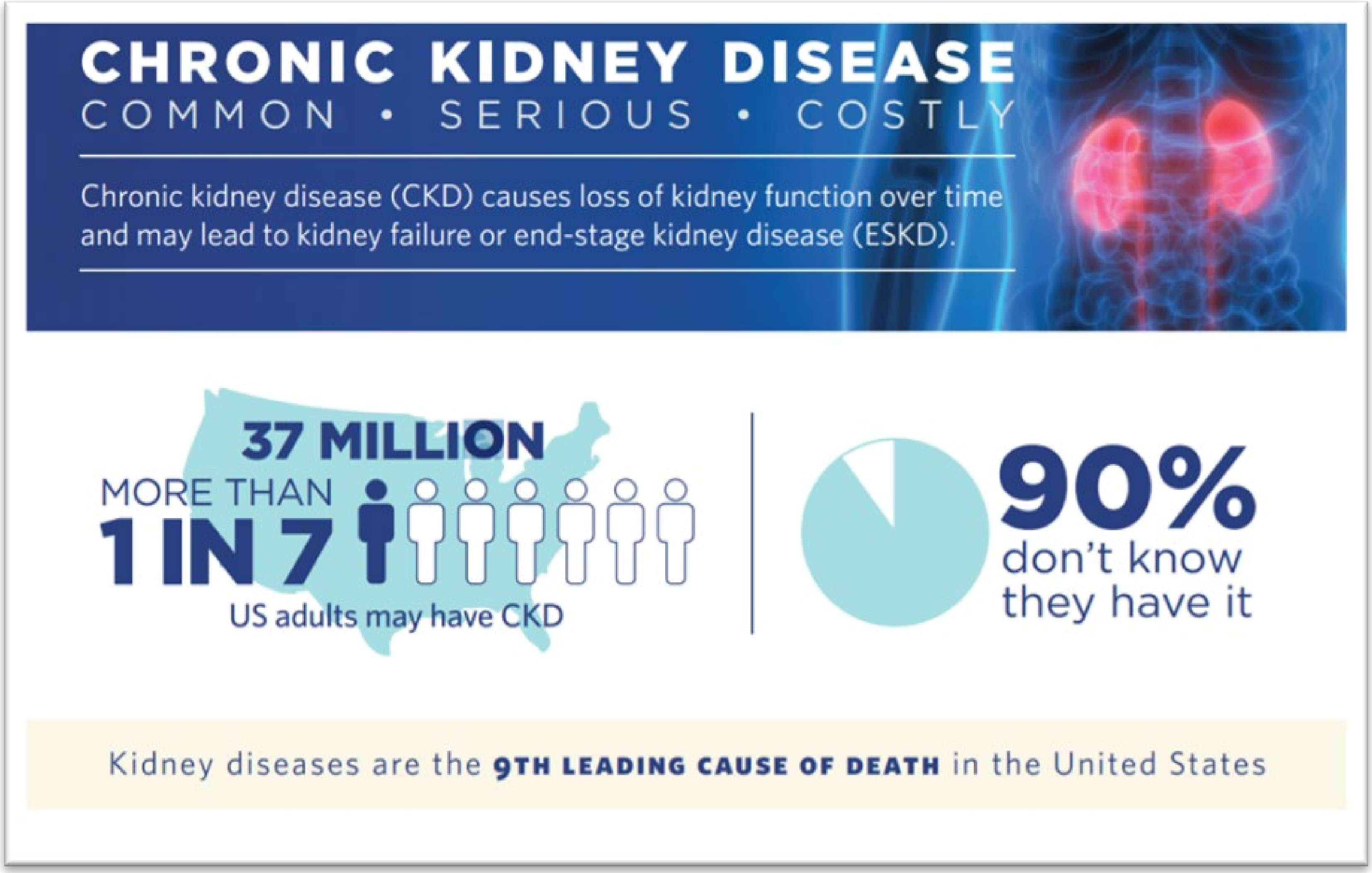


CRISP-DM (Cross-Industry Standard Process for Data Mining) Methodology of CPMAI^TM

1. Clinical Research Understanding	2. Data Understanding	3. Data Preparation				4. Modeling	5. Evaluation
Select Use Case	Understand Data	Label Outcomes	Build Features	Handle Missing Values	Create Final Training Dataset	Build ML Model	Measure Accuracy/ Assess Fairness

Kidney Disease – Overview*

- Kidney disease is common and under-recognized
- Most patients see a nephrologist when they're near ESKD, in need of dialysis or worse, crash into dialysis
- Kidney disease disproportionately affects African Americans
- Dialysis or kidney transplant are the only two treatment options for ESKD patients



CHRONIC KIDNEY DISEASE

COMMON • SERIOUS • COSTLY

Chronic kidney disease (CKD) causes loss of kidney function over time and may lead to kidney failure or end-stage kidney disease (ESKD).

37 MILLION

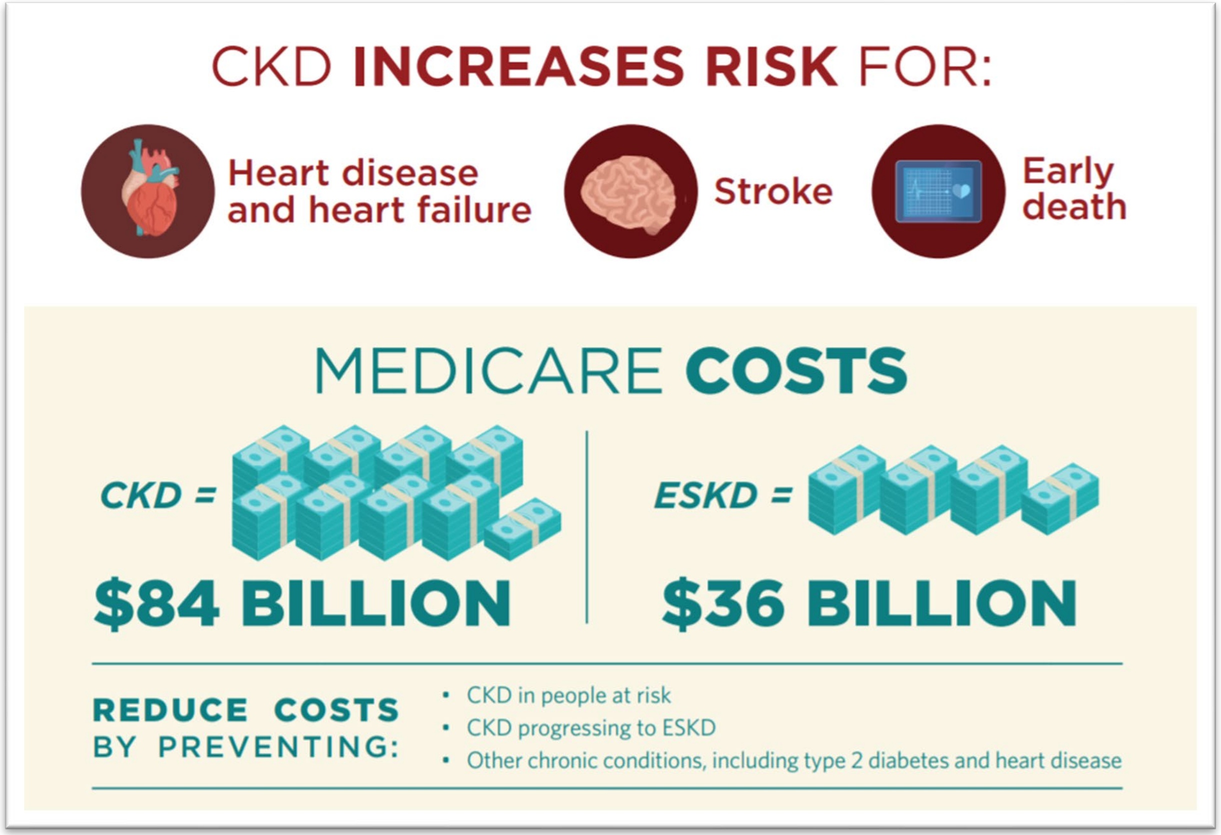
MORE THAN 1 IN 7 US adults may have CKD

90% don't know they have it

Kidney diseases are the **9TH LEADING CAUSE OF DEATH** in the United States

* https://www.cdc.gov/kidneydisease/pdf/CKD-common-serious-costly-h.pdf

Kidney Disease – Overview* (contd.)



CKD **INCREASES RISK** FOR:

- Heart disease and heart failure
- Stroke
- Early death

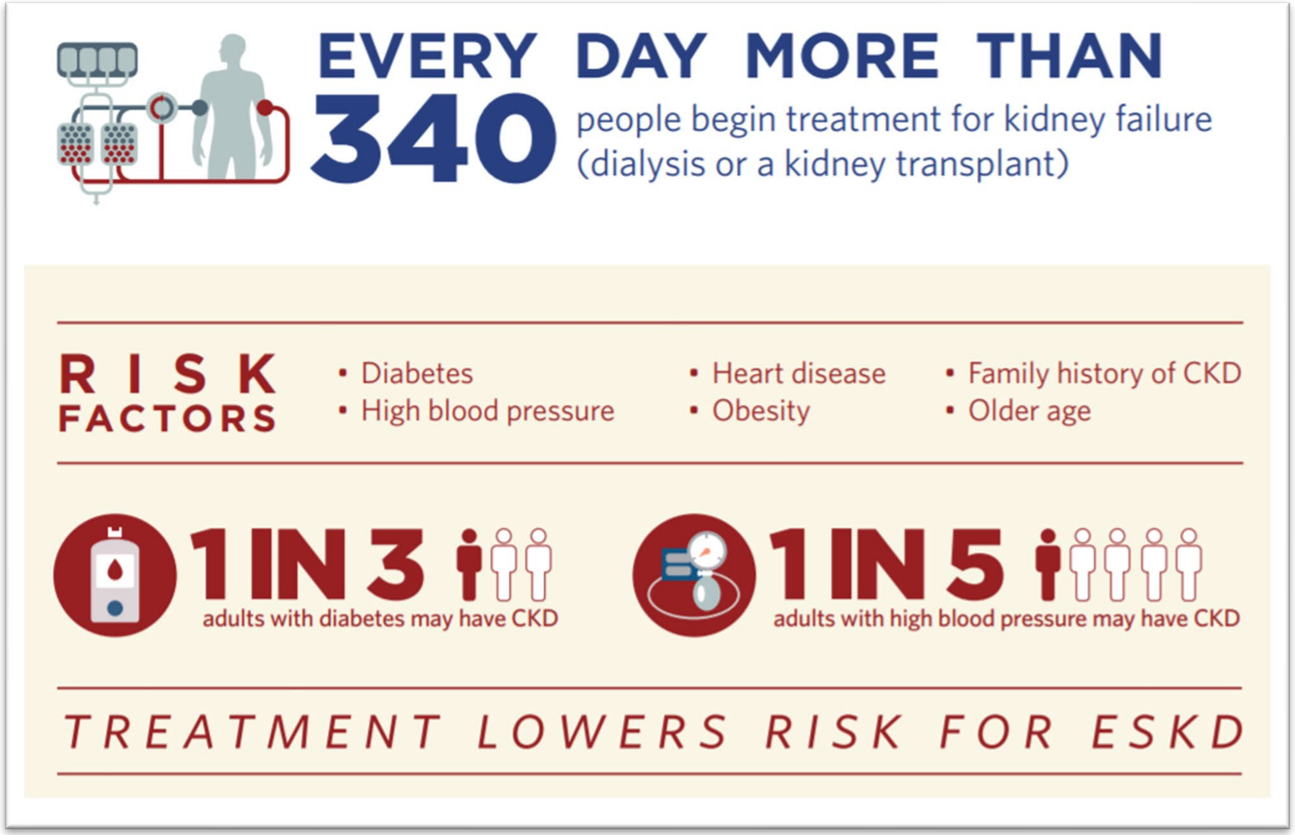
MEDICARE **COSTS**

CKD = **$84 BILLION**

ESKD = **$36 BILLION**

REDUCE COSTS BY PREVENTING:

- CKD in people at risk
- CKD progressing to ESKD
- Other chronic conditions, including type 2 diabetes and heart disease



EVERY DAY MORE THAN 340 people begin treatment for kidney failure (dialysis or a kidney transplant)

RISK FACTORS

- Diabetes
- High blood pressure
- Heart disease
- Obesity
- Family history of CKD
- Older age

1 IN 3 adults with diabetes may have CKD

1 IN 5 adults with high blood pressure may have CKD

TREATMENT LOWERS RISK FOR ESKD

* https://www.cdc.gov/kidneydisease/pdf/CKD-common-serious-costly-h.pdf

Kidney Disease – Use Case

Predicting mortality in the first 90 days of dialysis

The first 90 days following initiation of chronic dialysis represent a high-risk period for adverse outcomes, including mortality

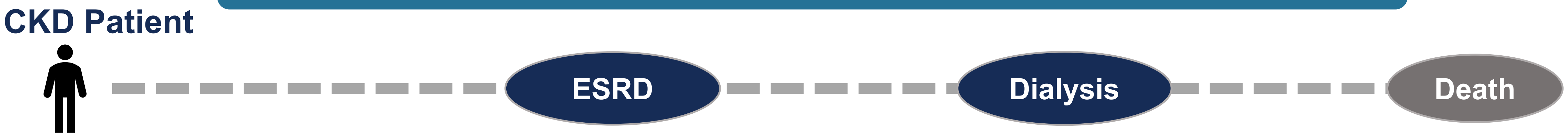
While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated

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

Project Activities – USRDS Data Mapping to Use Case

Selected use case: *Predicting mortality in the first 90 days of dialysis*



CKD Patient

ESRD

Dialysis

Death

1. **CMS Pre-ESRD Claims Datasets**
   - Parts A and B claims prior to ESRD diagnosis
   - Used to build features, such as prior nephrology care

2. **ESRD Medical Evidence Report (MEDEVID) (CMS 2728)/ PATIENTS Dataset**
   - Form is completed when a patient is diagnosed as ESRD and receives their first chronic dialysis treatment(s) or transplant
   - Used to build features such as patient demographics, comorbid conditions, primary cause of renal failure, and laboratory values

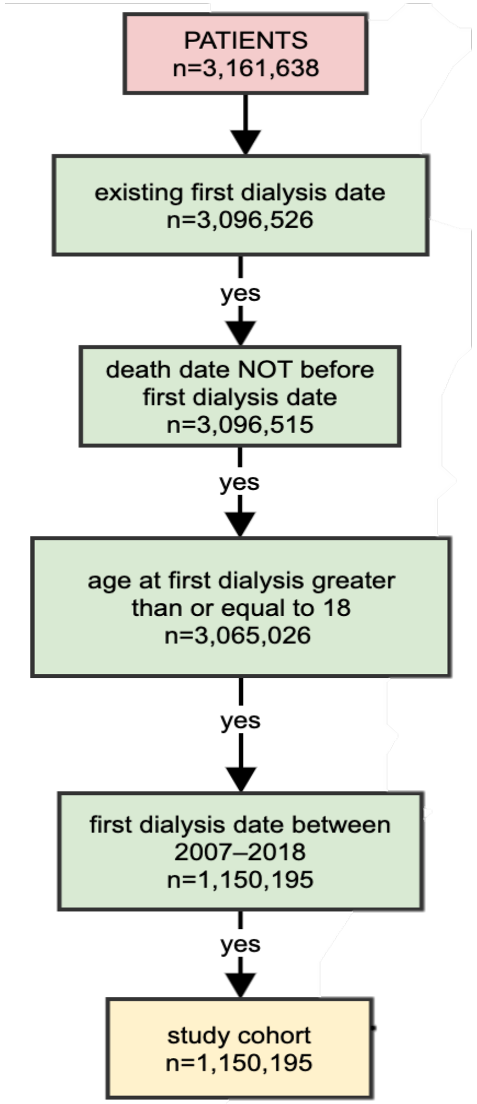
2A. PATIENTS Dataset
- Provides basic demographic and ESRD-related data
- Used to obtain dialysis start date and modality
- Used in conjunction with MEDEVID to build demographic features such as age, sex, race, etc.

2B. Transplant Dataset (TX)
- Provides information on kidney transplants such as list date/data on eligibility pre-dialysis
- Used to build features such as transplant waitlist status

3. **PATIENTS Dataset/ DEATH Dataset (CMS ESRD Death Notification Form 2726)**
   - Used to determine if a patient died in the first 90 days after dialysis start

Training Datasets Development – Methodology

Inclusion/Exclusion Criteria



PATIENTS
n=3,161,638

existing first dialysis date
n=3,096,526

yes

death date NOT before first dialysis date
n=3,096,515

yes

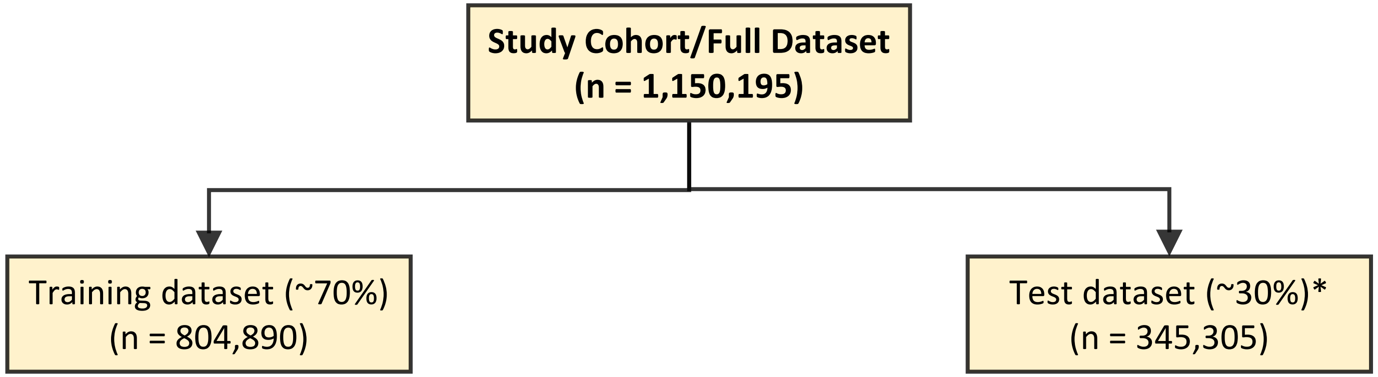
age at first dialysis greater than or equal to 18
n=3,065,026

yes

first dialysis date between 2007–2018
n=1,150,195

yes

study cohort
n=1,150,195



Study Cohort/Full Dataset
(n = 1,150,195)

Training dataset (~70%)
(n = 804,890)

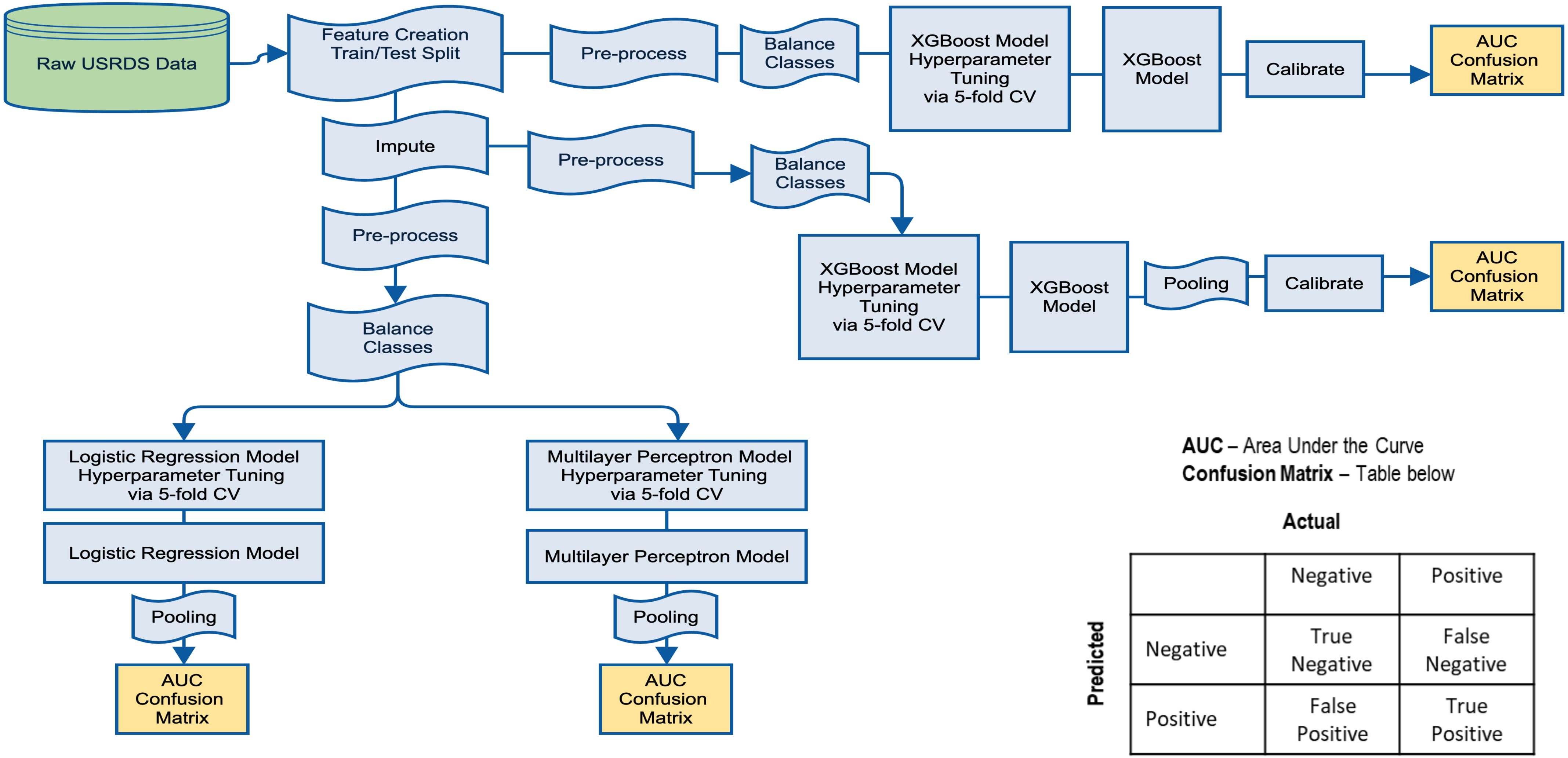
Test dataset (~30%)*
(n = 345,305)

*Test dataset not used until evaluating the final ML model

Criteria for a high-quality training dataset

- Well-labeled (removed or flagged outliers and erroneous/suspicious/duplicate/ inconsistent values, documented how outliers/inconsistencies were addressed, etc.)
- Well-structured (inclusion/exclusion criteria correctly applied, merging/joining done correctly, missing data patterns documented and addressed, centering/scaling etc.)
- Use common data elements (common data elements used for constructed variables)

ML Model Development – Methodology



Raw USRDS Data

Feature Creation Train/Test Split

Pre-process

Balance Classes

XGBoost Model Hyperparameter Tuning via 5-fold CV

XGBoost Model

Calibrate

AUC Confusion Matrix

Impute

Pre-process

Balance Classes

XGBoost Model Hyperparameter Tuning via 5-fold CV

XGBoost Model

Pooling

Calibrate

AUC Confusion Matrix

Pre-process

Balance Classes

Logistic Regression Model Hyperparameter Tuning via 5-fold CV

Logistic Regression Model

Pooling

AUC Confusion Matrix

Multilayer Perceptron Model Hyperparameter Tuning via 5-fold CV

Multilayer Perceptron Model

Pooling

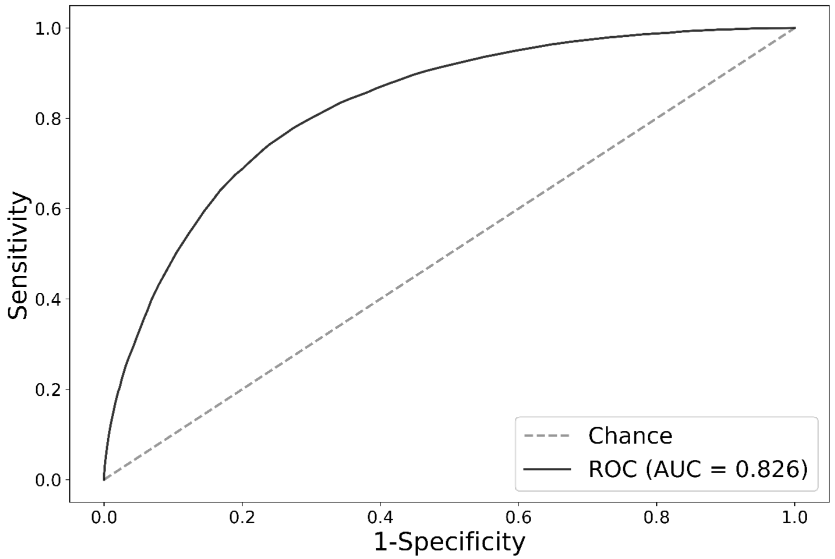
AUC Confusion Matrix

AUC – Area Under the Curve
Confusion Matrix – Table below

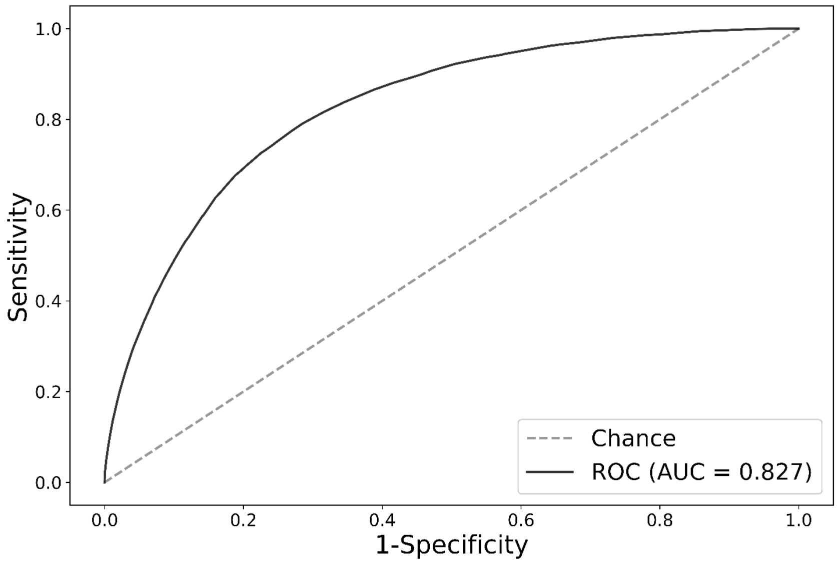
		Actual	
		Negative	Positive
Predicted	Negative	True Negative	False Negative
	Positive	False Positive	True Positive

ML Model Results – Area under the curve

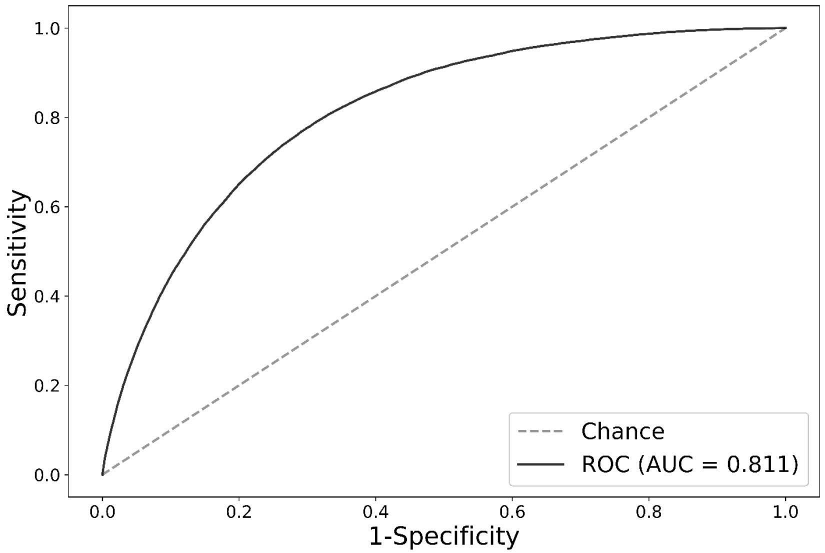
XGBoost Non-imputed



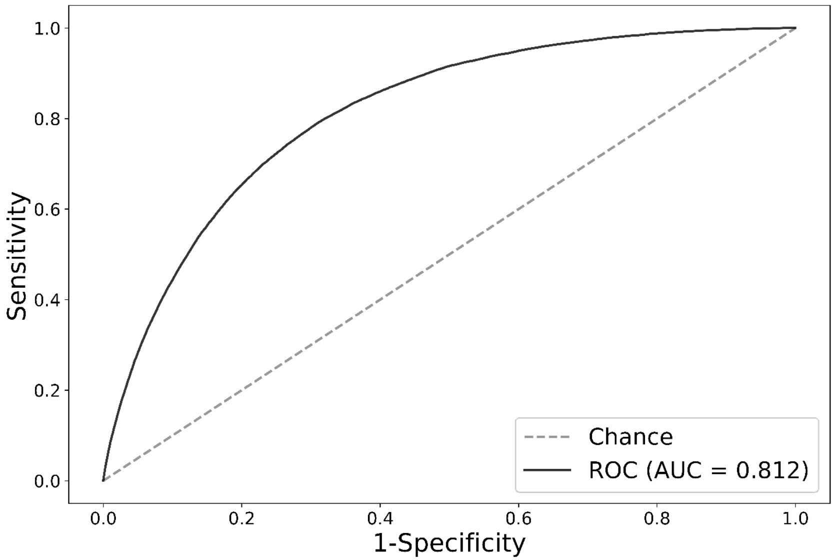
XGBoost Imputed



Logistic Regression



Multilayer Perceptron



ML Model Results – Top Features*

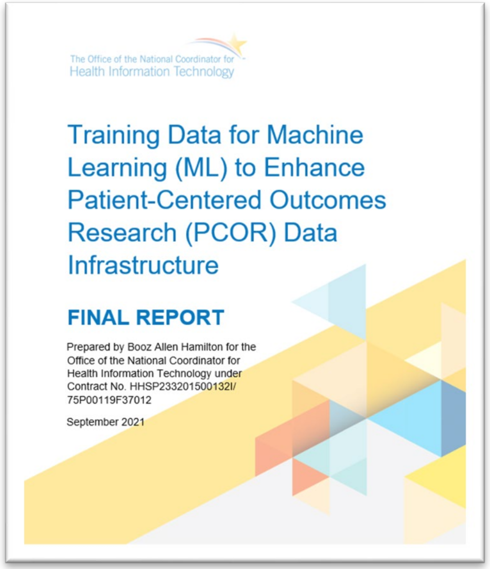
	Feature	Explanation
1.	**Age**	• Older age is associated with worse survival
2.	**Inpatient stays**	• Longer inpatient stays is more common in older and sicker patients and has been associated with early mortality
3.	**Received erythropoietin (EPO)**	• EPO hormone is produced by kidneys when it senses low oxygen levels in the blood; EPO triggers bone marrow to produce more red blood cells which raises blood oxygen • Patients on EPO typically have advanced CKD at the time of dialysis and are under the care of a nephrologist • Patients with kidney failure produces less EPO; therefore, are given EPO
4.	**Albumin**	• Albumin reflects the patient’s overall health status (including nutrition and inflammation) • Risk of death is increased by poor serum albumin levels reflecting inadequate nutrition
5.	**Arteriovenous Fistula (AVF)**	• The presence of a maturing AVF indicates prior nephrology care • Hemodialysis through AVF access is associated with reduced mortality

*Top feature rankings are from XGBoost (non-imputed and imputed) and logistic regression models

Strategic Recommendations for Future Applications of AI/ML for PCOR and Health Care

- Develop an industry-wide strategy to address the ongoing challenge of accessing data in a timely manner (specifically EHR data) for applying AI/ML to important clinical use cases that can significantly impact patient-provider decisions and advance PCOR
- Ensure that a base set of widely used data elements such as demographics, clinical conditions and history, basic laboratory measures and values, clinical outcomes, etc. are captured comprehensively, completely, and accurately in national registries that serve as data sources for AI/ML applications
- Emphasize engagement and close collaboration between AI/ML practitioners and clinical domain experts with AI/ML understanding when developing prediction models that could potentially be deployed to support provider-patient decisions
- Utilize a standardized framework with checklists and best practices for prediction modeling and standardized metrics for evaluating ML models for addressing clinical use cases that have the potential to be deployed in the clinic will further expand the applications of AI/ML in health care more broadly

Project Resources – Overview & Locations



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

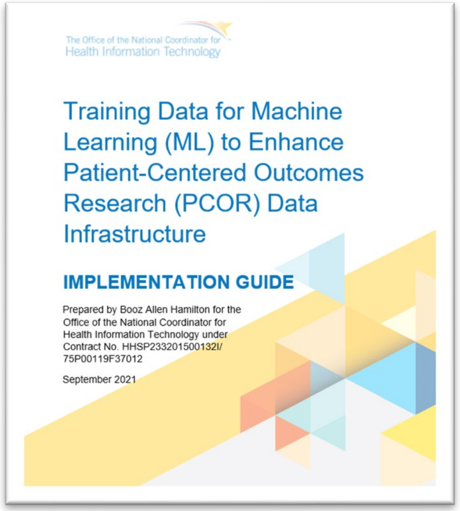
FINAL REPORT

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

September 2021

- Purpose: Comprehensive description of project activities undertaken to build the training datasets and ML models and develop a set of recommendations for future work in applying ML to PCOR and health care
- Content: Executive summary, introduction/background, high-level description of the methodology used to create the training dataset and test ML models, considerations for future researchers, and recommendations for future work
- Location: A section 508-compliant Final Report is posted on the Project page of the HealthIT.gov website

- Purpose: Detailed methodology and points to consider/lessons learned from building high-quality training datasets and machine learning models that can be utilized by other PCOR/health care researchers
- Content: A high-level project overview and a step-by-step guide (Implementation Guidance) of the methodologies used to develop the training datasets and ML models with snippets of code and explanations of the methodologies chosen
- Location: The Implementation Guide is posted on the Project page of the HealthIT.gov website



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

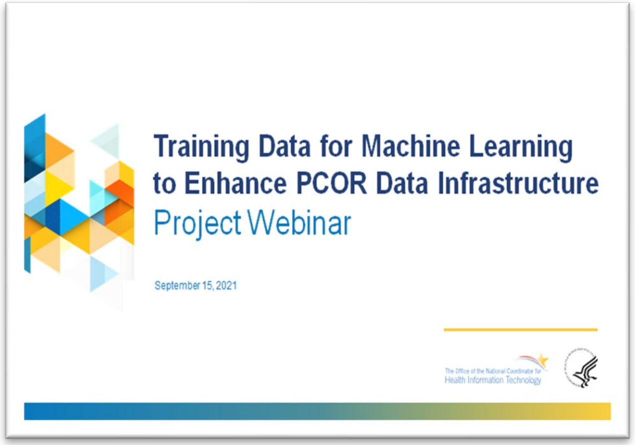
Prepared by Booz Allen Hamilton for the Office of the National Coordinator for Health Information Technology under Contract No. HHSP233201500132I/ 75P00119F37012

September 2021

Project Resources – Overview & Locations (contd.)

- Purpose: Allows future PCOR researchers to access the code used to create the training dataset and ML models and adapt the code base for other projects
- Content: Python and R code used to create the training datasets and ML models (non-imputed XGBoost, imputed XGBoost, logistic regression, and multilayer perceptron)
- Location: The code is hosted in the ONC GitHub repository

- Purpose: Disseminate project activities and resources for the Training Data for ML to Enhance PCOR Data Infrastructure
- Content: Introduction to ONC, background for the Project, kidney disease use case selected for the project, high level project activities undertaken, and location of project resources
- Location: The webinar slides is posted on the Project page of the HealthIT.gov website shortly



Training Data for Machine Learning to Enhance PCOR Data Infrastructure

Project Webinar

September 15, 2021

The Office of the National Coordinator for Health Information Technology

Acknowledgements

- ONC partner for the Project: Dr. Ken Wilkins, The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
- Technical Expert Panel members:
  - 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
- Interagency Assembly members (60+ from over 10 government agencies, including NIH, FDA, CDC, VA, CMS, AHRQ, etc.)

Q&A

The Office of the National Coordinator for Health Information Technology

The Office of the National Coordinator for
Health Information Technology

Contact ONC

Phone: 202-690-7151

Health IT Feedback Form: https://www.healthit.gov/form/healthit-feedback-form

Twitter: @onc_healthIT

LinkedIn: Search "Office of the National Coordinator for Health Information Technology"

HealthIT.gov

Subscribe to our weekly eblast at healthit.gov for the latest updates!