Supplemental Material

Table of Contents

Study Dataset

Study Dataset Preparation Methodology

The source data for building the study dataset was obtained from the United States Renal Data System (USRDS), the national data registry maintained by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that stores and distributes data on the outcomes and treatments of Chronic Kidney Disease (CKD) and ESKD (or End Stage Renal Disease, ESRD[1]) population in the U.S. The study 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

Further details on the source USRDS datasets and the predictors for which they were used are provided in **Supplemental Table 1** below.

Supplemental Table 1: USRDS Datasets utilized in the project for predicting mortality

	USRDS Dataset	Description and Use in the Study Dataset
1	Centers for Medicare & Medicaid Services (CMS) Pre-ESKD Claims Datasets	• Parts A and B claims prior to ESKD diagnosis • Used to build predictors, such as prior nephrology care
2	ESKD Medical Evidence Report (MEDEVID) (CMS 2728)/ PATIENTS Dataset	• Form is completed when a patient is diagnosed with ESKD and receives their first chronic dialysis treatment(s) or transplant • Used to build predictors, such as patient demographics, comorbid conditions, primary cause of renal failure, and laboratory values
3	PATIENTS Dataset	• Provides basic demographic and ESKD-related data • Used to obtain dialysis start date and modality • Used in conjunction with MEDEVID to build demographic predictors such as age, sex, race, etc.
4	Transplant Dataset (TX)	• Used to obtain information on transplant list date/ data on eligibility pre-dialysis
5	PATIENTS Dataset / DEATH Dataset (CMS ESKD Death Notification Form 2726)	• Used to determine if a patient died in the first 90 days after dialysis start

To ensure that the study dataset was of high quality for training a model, the criteria shown in **Supplemental Table 2** below were applied.

Supplemental Table 2: Criteria for preparing a high-quality study dataset

Quality Criteria	Methodology Employed in the Study 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 derived predictors, 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 claims are missing for those who do not qualify for Medicare prior to ESKD diagnosis) • Centering/scaling/standardizing some variables for analysis or balancing the data based on the algorithm selected • Excluded operational factors such as location, provider, and masked dates when building predictors • Training/testing/validation split done such that the training data is representative of the rest of the data • Created data dictionary
Use common data elements (CDEs)	• For derived predictors, used CDEs, where possible • For predictors pulled directly from USRDS dataset, CDEs were based on what was used by USRDS

Data Dictionary

The study dataset consisted of 188 predictors that were limited to information that was known on or prior to the first day of dialysis. Two types of predictors were included in the study dataset: some predictors were taken directly from the USRDS datasets (e.g., age, race, hemoglobin), whereas other predictors were derived from variables in USRDS (e.g., time on kidney transplant waitlist, number of pre-ESKD claims). The full list of predictors and the methods for deriving predictors are shown in the Data Dictionary in **Supplemental Table 3** and **Supplemental Table 4**.

Predictors taken directly from the USRDS data

These included predictors from PATIENTS table—specifically, *demographic variables*: age, race, sex, and Hispanic ethnicity. Additionally, co-author kidney disease experts identified variables of clinical relevance from the MEDEVID table for inclusion in the study dataset. Out of fifteen *clinical and laboratory values* in the MEDEVID table, only seven were included in the study dataset—the rest had a high percentage of missing values (more than 40 percent) or contained duplicate clinical information, such as methods of estimating glomerular filtration rate (eGFR). Masked date variables from the MEDEVID table, such as patient signature date and clinician signature dates, were also excluded from the training dataset as they were considered operational variables that have little to no clinical relevance. The full list of predictors taken directly from the PATIENTS and the MEDEVID tables are shown in **Supplemental Table 3**.

Supplemental Table 3: Predictors selected directly from the USRDS datasets

	USRDS Table	Category	Predictor Description	Variable Type
1.	Patients	Demographics	Age	Measure (Years)
2.	Patients	Demographics	Race	Factor with 7 levels: 1=White, 2=Black/African American, 3=American Indian or Alaska Native, 4=Asian, 5=Native Hawaiian or Pacific Islander, 6=Other or Multiracial, 9=Unknown
3.	Patients	Demographics	Sex	Factor with 3 levels: 1=M, 2=F, 3=Unknown
4.	Patients	Demographics	Ethnicity	Factor with 5 levels: 1=Hispanic-Mexican, 2=Hispanic Other, 3=Non-Hispanic, 5=Hispanic Non-Specified, 9=Unknown
5.	Medical Evidence	Clinical Variables	BMI	Measure
6.	Medical Evidence	Clinical Variables	Weight	Measure (kg)
7.	Medical Evidence	Clinical Variables	Height	Measure (cm)
8.	Medical Evidence	Clinical Variables	Albumin	Measure (g/dl)
9.	Medical Evidence	Clinical Variables	Serum Creatinine	Measure (mg/dl)
10.	Medical Evidence	Clinical Variables	Hemoglobin	Measure (g/dl)
11.	Medical Evidence	Clinical Variables	Estimated glomerular filtration rate (eGFR)	Measure (mL/min)
12.	Medical Evidence	Comorbidities	Congestive heart failure	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
13.	Medical Evidence	Comorbidities	Atherosclerotic heart disease (ASHD)	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
14.	Medical Evidence	Comorbidities	Other cardiac disease	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
15.	Medical Evidence	Comorbidities	Cerebrovascular disease, Cerebrovascular accident (CVA), Transient ischemic attack (TIA)	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
16.	Medical Evidence	Comorbidities	Peripheral vascular disease	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
17.	Medical Evidence	Comorbidities	History of hypertension	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
18.	Medical Evidence	Comorbidities	Amputation	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
19.	Medical Evidence	Comorbidities	Diabetes, currently on insulin	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
20.	Medical Evidence	Comorbidities	Diabetes, on oral medications	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
21.	Medical Evidence	Comorbidities	Diabetes, without medications	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown

22.	Medical Evidence	Comorbidities	Diabetic retinopathy	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
23.	Medical Evidence	Comorbidities	Chronic obstructive pulmonary disease	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
24.	Medical Evidence	Comorbidities	Tobacco use (current smoker)	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
25.	Medical Evidence	Comorbidities	Malignant neoplasm, Cancer	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
26.	Medical Evidence	Comorbidities	Toxic nephropathy	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
27.	Medical Evidence	Comorbidities	Alcohol dependence	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
28.	Medical Evidence	Comorbidities	Drug dependence	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
29.	Medical Evidence	Comorbidities	Inability to ambulate	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
30.	Medical Evidence	Comorbidities	Inability to transfer	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
31.	Medical Evidence	Comorbidities	Needs assistance with daily activities	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
32.	Medical Evidence	Comorbidities	Institutionalized	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
33.	Medical Evidence	Comorbidities	Non-renal congenital abnormality	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
34.	Medical Evidence	Comorbidities	None	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
35.	Medical Evidence	Comorbidities	Institutionalized -- Assisted Living	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
36.	Medical Evidence	Comorbidities	Institutionalized -- Nursing Home	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
37.	Medical Evidence	Comorbidities	Institutionalized -- Other Institution	Factor with 3 levels: 1=Yes, 2=No, 3=Unknown
38.	Medical Evidence	Renal Failure	Primary disease causing end stage kidney disease (ESKD): detailed group	Factor with 8` levels: 1=Diabetes, 2=Hypertension, 3=Glomerulonephritis, 4=Cystic kidney, 5=Other urologic, 6=Other cause, 7=Unknown cause, 8=Missing cause
39.	Medical Evidence	Prior Care	Prior nephrology care	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
40.	Medical Evidence	Prior Care	Range of nephrology care	Factor with 3 levels: 0= < 6 months,

				1= 6-12 months, 2= > 12 months
41.	Medical Evidence	Prior Care	Access type (first outpatient dialysis access type)	Factor with 5 levels: 1=Arteriovenous fistula (AVF), 2=Graft, 3=Cath, 4=Other, 5=Unknown
42.	Medical Evidence	Prior Care	Is maturing arteriovenous fistula (AVF) present	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
43.	Medical Evidence	Prior Care	Is maturing arteriovenous graft (AVG) present	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
44.	Medical Evidence	Prior Care	Received exogenous erythropoietin (EPO)	Factor with 3 levels: Y=Yes, N=No, U=Unknown
45.	Medical Evidence	Prior Care	EPO range	Factor with 3 levels: 0= < 6 months, 1= 6-12 months, 2= > 12 months
46.	Medical Evidence	Prior Care	Under care kidney dietician	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
47.	Medical Evidence	Prior Care	Range of diet care	Factor with 3 levels: 0= < 6 months, 1= 6-12 months, 2= > 12 months
48.	Medical Evidence	Patient Education	Informed of transplant options	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
49.	Medical Evidence	Patient Education	Reason not informed of transplant options: medically unfit	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
50.	Medical Evidence	Patient Education	Reason not informed of transplant options: unsuitable due to age	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
51.	Medical Evidence	Patient Education	Reason not informed of transplant options: psychologically unfit	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
52.	Medical Evidence	Patient Education	Reason not informed of transplant options: patient declines information	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
53.	Medical Evidence	Patient Education	Reason not informed of transplant options: patient has not been assessed	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
54.	Medical Evidence	Patient Education	Reason not informed of transplant options: other	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
55.	Medical Evidence	Patient Education	Patient has/will complete training	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
56.	Medical Evidence	Patient Education	Self dialysis training type	Factor with 7 levels: 0=No Training, 1=Hemodialysis, 2=Intermittent peritoneal dialysis

				(IPD), 3=Continuous ambulatory peritoneal dialysis (CAPD), 4=Continuous cycling peritoneal dialysis (CCPD), 5=Other, 6=Unknown
57.	Medical Evidence	Other	Prior employment status	Factor with 9 levels: 1=Unemployed, 2=Employed full-time, 3=Employed part-time, 4=Homemaker, 5=Retired-age, 6=Retired-disabled, 7=Medical Leave of absence, 8=Student, 9=Other
58.	Medical Evidence	Other	Current employment status	Factor with 9 levels: 1=Unemployed, 2=Employed full-time, 3=Employed part-time, 4=Homemaker, 5=Retired-age, 6=Retired-disabled, 7=Medical Leave of absence, 8=Student, 9=Other
59.	Medical Evidence	Other	Insurance type: Medicaid	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
60.	Medical Evidence	Other	Insurance type: Medicare	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
61.	Medical Evidence	Other	Insurance type: Medicare Advantage	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
62.	Medical Evidence	Other	Insurance type: Employer Group Health	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
63.	Medical Evidence	Other	Insurance type: Veteran's Affairs (VA)	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
64.	Medical Evidence	Other	Insurance type: Other	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
65.	Medical Evidence	Other	Insurance type: None	Factor with 3 levels: 1=Yes, 2=No, 9=Unknown
66.	Medical Evidence	Other	Primary dialysis type	Factor with 4 levels: Hemodialysis, Continuous Ambulatory Peritoneal Dialysis (CAPD), Continuous Cycling Peritoneal Dialysis (CCPD), Other
67.	Medical Evidence	Other	Primary dialysis setting	Factor with 5 levels: Hospital Inpatient, Dialysis Facility, Home, Unknown, Skilled nursing facility

Predictors derived from USRDS datasets

Detailed method for the predictors derived from PATIENTS, MEDEVID and Medicare pre-ESKD claims data are provided in **Supplemental Table 4**. A summary description of the derivation method is provided below.

The *transplant waitlist status* predictor was derived from the dialysis start date from the PATIENTS tables and the start and end dates from the kidney transplant waitlist tables (WAITSEQ_KI, WAITSEQ_KP, and TX tables) to determine whether a patient was actively on the kidney transplant waitlist, removed from the waitlist, received a kidney transplant, or never on the waitlist prior to dialysis initiation. The *time on transplant waitlist* variable was derived for the patients who are on the transplant waitlist by subtracting the start date from the end date.

The *primary cause of renal failure (PDIS)* predictor was derived by taking the PDIS variable from the PATIENTS table and replacing the missing values with the PDIS values from the MEDEVID table to reduce the number of overall missing values. Next, the PDIS predictor was recoded from ICD-9 and ICD-10 codes in text format to numeric categories.

Four predictors (*number of comorbidities marked as: yes, no, unknown, or missing*) were built from the comorbidity variables in the MEDEVID tables by counting the number of comorbidities—out of 6—for each category (yes, no, unknown, or missing). These aggregate variables allow for better interpretation of the outputs of the XGBoost models, such as when assessing feature importance of the comorbidity variables. Binary variables were created for each clinical/laboratory predictor included in the study dataset to indicate whether the original values were missing and whether the original values were out of bounds. The *time in dialysis training* was derived by subtracting the training end date from the training start date variables in the MEDEVID table.

For the Medicare pre-ESKD claims datasets, predictors with clinical relevance were also identified by co-author kidney disease experts. The *total number of claims* and *total lengths of stay* predictors for each type of claim setting (inpatient—IP, outpatient—OP, skilled nursing unit—SN, home health—HH, and hospice—HS) were derived by counting the number of claims per patient and summing the total lengths of stays per type of claim. Binary variables were also created to indicate the presence or absence of a claim in each claim setting (IP, OP, HH, HS, SN) as well as the presence or absence of any pre-ESKD Medicare claim per patient in the study cohort. Features that indicate the time elapsed between first and last pre-ESKD Medicare claim were derived for each patient across all claims settings and also for each setting (IP, OP, HH, HS, SN) by subtracting the date of the first claim from the date of the last claim.

Diagnosis code groupings were created based on 12 major disease groups that were defined by co-author kidney disease experts: diabetes, hypertension, heart failure, cardiovascular arterial disease, cerebrovascular disease, peripheral arterial disease, kidney failure, pneumonia, malignant neoplasm, alcohol dependence, smoking, and opioid dependence. These major disease groups have clinical relevance to ESKD and are likely to have prognostic value. Through matching the primary diagnosis code[2] for each claim with the ranges of the ICD 9/10 codes associated with each major disease, variables for inpatient, outpatient, and skilled nursing unit settings were created for each primary diagnosis code, total number of claims/total length of stay, and type of claim combination (e.g., total number of claims for a hypertension primary diagnosis code for outpatient claim, total length of stays for a heart failure diagnosis code for an inpatient claim). A binary indicator for whether a patient has any claim in each disease group was also derived for all claim settings.

Supplemental Table 4: Predictors derived from USRDS datasets

	USRDS Dataset	Category	Predictor Description	Variable Type	Derivation Method
1.	Patients/Kidney Transplant Waitlist	Prior care	Transplant Waitlist Status	Categorical	"never" -- never on the waitlist if no entry for a patient in the transplant dataset or if the first list date is after the initial dialysis date; "active" -- currently on the waitlist if list date is before the initial dialysis start date and end date is after the initial dialysis start date; "transplanted" -- already transplanted if the list date and the end date are before the initial dialysis start date

					and the patient has a transplant event "removed" -- removed from the waitlist if the list date and the end date are before the initial dialysis start date and the patient had no transplant event
2.	Patients/Kidney Transplant Waitlist	Prior care	Number of days on transplant waitlist	Numeric	Number of days on the waitlist calculated by subtracting dialysis date from the list date
3.	Medical Evidence	Comorbidities	Number of missing comorbidities	Numeric	Count the number of comorbidities marked as "missing"
4.	Medical Evidence	Comorbidities	Number of comorbidities coded as No (N)	Numeric	Count the number of comorbidities marked as "N"
5.	Medical Evidence	Comorbidities	Number of comorbidities coded as Yes (Y)	Numeric	Count the number of comorbidities marked as "Y"
6.	Medical Evidence	Comorbidities	Number of comorbidities coded as Unknown (U)	Numeric	Count the number of comorbidities marked as "U"
7.	Patients	Renal failure	Recoding of PDIS. PDIS is the ICD-9 or ICD-10 encoding for primary cause of renal failure	Unordered factor	Step 1: Find the ICD-10 encoding for PDIS, which requires mapping ICD-9 to ICD-10 for incident years prior to 2014 by using the 2017_I9gem.txt within the zip file on the CMS website https://www.cms.gov/Medicare/Coding/ICD10/Downloads/2017-GEM-DC.zip. Step 2: Convert to a factor
8.	Medical Evidence	Clinical Variables	Indicator of whether original albumin value was an outlier	Binary	1 if original albumin value was an outlier, 0 otherwise
9.	Medical Evidence	Clinical Variables	Indicator of whether original serum creatinine value was an outlier	Binary	1 if original serum creatinine value was an outlier, 0 otherwise
10.	Medical Evidence	Clinical Variables	Indicator of whether original hemoglobin value was an outlier	Binary	1 if original hemoglobin value was an outlier, 0 otherwise
11.	Medical Evidence	Clinical Variables	Indicator of whether original height value was an outlier	Binary	1 if original height value was an outlier, 0 otherwise

12.	Medical Evidence	Clinical Variables	Indicator of whether original weight value was an outlier	Binary	1 if original weight value was an outlier, 0 otherwise
13.	Medical Evidence	Clinical Variables	Indicator of whether original body mass index (BMI) value was an outlier	Binary	1 if original BMI value was an outlier, 0 otherwise
14.	Medical Evidence	Clinical Variables	Indicator of whether original eGFR value was an outlier	Binary	1 if original gfr_epi value was an outlier, 0 otherwise
15.	Medical Evidence	Prior care	Time on training	Measure (days)	Masked dialysis training end date (masked_trend) minus masked dialysis training begin date (masked_trstdat)
16.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for malignant neoplasm	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
17.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for smoking	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "smoking" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
18.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for alcohol	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "alcohol" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
19.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for drug abuse	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "opioid" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
20.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for kidney failure	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date

21.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for pneumonia	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
22.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for diabetes	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "diabetes" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
23.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for hypertension	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "hypertension" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
24.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for heart failure	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "heart failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
25.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for cardiovascular arterial disease	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
26.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for cerebrovascular disease	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
27.	Pre-ESKD Claims	Prior care	Total length of inpatient stay for peripheral arterial disease	Measure (days)	Total length of inpatient stay for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date

28.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for malignant neoplasm	Numeric	Total inpatient claims for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes that a patient has
29.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for smoking	Numeric	Total inpatient claims for instances of primary diagnosis code in the "smoking" group of diagnosis codes that a patient has
30.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for alcohol	Numeric	Total inpatient claims for instances of primary diagnosis code in the "alcohol" group of diagnosis codes that a patient has
31.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for drug abuse	Numeric	Total inpatient claims for instances of primary diagnosis code in the "opioid" group of diagnosis codes that a patient has
32.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for kidney failure	Numeric	Total inpatient claims for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes that a patient has
33.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for pneumonia	Numeric	Total inpatient claims for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes that a patient has
34.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for diabetes	Numeric	Total inpatient claims for instances of primary diagnosis code in the "diabetes" group of diagnosis codes that a patient has
35.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for hypertension	Numeric	Total inpatient claims for instances of primary diagnosis code in the "hypertension" group of diagnosis codes that a patient has
36.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for heart failure	Numeric	Total inpatient claims for instances of primary diagnosis code in the "heart failure" group of diagnosis codes that a patient has
37.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for cardiovascular arterial disease	Numeric	Total inpatient claims for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes that a patient has
38.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for cerebrovascular disease	Numeric	Total inpatient claims for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes that a patient has

39.	Pre-ESKD Claims	Prior care	Total number of inpatient claims for peripheral arterial disease	Numeric	Total inpatient claims for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes that a patient has
40.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for malignant neoplasm	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
41.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for smoking	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "smoking" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
42.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for alcohol	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "alcohol" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
43.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for drug abuse	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "opioid" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
44.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for kidney failure	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
45.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for pneumonia	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
46.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for diabetes	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "diabetes" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date

47.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for hypertension	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "hypertension" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
48.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for heart failure	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "heart failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
49.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for cardiovascular arterial disease	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
50.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for cerebrovascular disease	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
51.	Pre-ESKD Claims	Prior care	Total length of outpatient claims for peripheral arterial disease	Measure (days)	Total length of outpatient stays for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
52.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for malignant neoplasm	Numeric	Total outpatient claims for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes that a patient has
53.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for smoking	Numeric	Total outpatient claims for instances of primary diagnosis code in the "smoking" group of diagnosis codes that a patient has
54.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for alcohol	Numeric	Total outpatient claims for instances of primary diagnosis code in the "alcohol" group of diagnosis codes that a patient has

55.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for drug abuse	Numeric	Total outpatient claims for instances of primary diagnosis code in the "opioid" group of diagnosis codes that a patient has
56.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for kidney failure	Numeric	Total outpatient claims for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes that a patient has
57.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for pneumonia	Numeric	Total outpatient claims for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes that a patient has
58.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for diabetes	Numeric	Total outpatient claims for instances of primary diagnosis code in the "diabetes" group of diagnosis codes that a patient has
59.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for hypertension	Numeric	Total outpatient claims for instances of primary diagnosis code in the "hypertension" group of diagnosis codes that a patient has
60.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for heart failure	Numeric	Total outpatient claims for instances of primary diagnosis code in the "heart failure" group of diagnosis codes that a patient has
61.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for cardiovascular arterial disease	Numeric	Total outpatient claims for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes that a patient has
62.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for cerebrovascular disease	Numeric	Total outpatient claims for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes that a patient has
63.	Pre-ESKD Claims	Prior care	Total number of outpatient claims for peripheral arterial disease	Numeric	Total outpatient claims for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes that a patient has
64.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for malignant neoplasm	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
65.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for smoking	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "smoking" group of diagnosis

					codes where days was calculated by subtracting the claim start date from the claim end date
66.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for alcohol	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "alcohol" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
67.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for drug abuse	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "opioid" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
68.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for kidney failure	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
69.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for pneumonia	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
70.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for diabetes	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "diabetes" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
71.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for hypertension	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "hypertension" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
72.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for heart failure	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "heart failure" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date

73.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for cardiovascular arterial disease	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
74.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for cerebrovascular arterial disease	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
75.	Pre-ESKD Claims	Prior care	Total length of skilled nursing unit claims for peripheral arterial disease	Measure (days)	Total length of skilled nursing stay for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes where days was calculated by subtracting the claim start date from the claim end date
76.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for malignant neoplasm	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes that a patient has
77.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for smoking	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "smoking" group of diagnosis codes that a patient has
78.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for alcohol	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "alcohol" group of diagnosis codes that a patient has
79.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for drug abuse	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "opioid" group of diagnosis codes that a patient has
80.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for kidney failure	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes that a patient has
81.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for pneumonia	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes that a patient has

82.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for diabetes	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "diabetes" group of diagnosis codes that a patient has
83.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for hypertension	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "hypertension" group of diagnosis codes that a patient has
84.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for heart failure	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "heart failure" group of diagnosis codes that a patient has
85.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for cardiovascular arterial disease	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "cardiovascular arterial disease " group of diagnosis codes that a patient has
86.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for cerebrovascular disease	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "cerebrovascular disease" group of diagnosis codes that a patient has
87.	Pre-ESKD Claims	Prior care	Total number of skilled nursing unit claims for peripheral arterial disease	Numeric	Total skilled nursing claims for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes that a patient has
88.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "malignant neoplasm" group of diagnosis codes; 0 if not
89.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "smoking" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "smoking" group of diagnosis codes; 0 if not
90.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "alcohol" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "alcohol" group of diagnosis codes; 0 if not

91.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "opioid" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "opioid" group of diagnosis codes; 0 if not
92.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "kidney failure" group of diagnosis codes; 0 if not
93.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "pneumonia" group of diagnosis codes ; 0 if not
94.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "diabetes" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "diabetes" group of diagnosis codes; 0 if not
95.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "hypertension" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "hypertension" group of diagnosis codes ; 0 if not
96.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "heart failure" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "heart failure" group of diagnosis codes ; 0 if not
97.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "cardiovascular arterial" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "cardiovascular arterial" group of diagnosis codes ; 0 if not
98.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary	Binary	1 if a patient has any claims for instances of primary diagnosis code

			diagnosis code in the "cerebrovascular" group of diagnosis codes		in the "cerebrovascular" group of diagnosis codes ; 0 if not
99.	Pre-ESKD Claims	Prior care	Whether a patient has any claims for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes	Binary	1 if a patient has any claims for instances of primary diagnosis code in the "peripheral arterial disease" group of diagnosis codes ; 0 if not
100.	Pre-ESKD Claims	Prior care	Number of Inpatient Stays	Measure	Number of pre-ESKD Medicare Inpatient hospitalization claims a patient has
101.	Pre-ESKD Claims	Prior care	Number of Outpatient Visits	Measure	Number of pre-ESKD Medicare Outpatient visit claims a patient has
102.	Pre-ESKD Claims	Prior care	Number of Home Health Claims	Measure	Number of pre-ESKD Medicare Home Health care claims a patient has
103.	Pre-ESKD Claims	Prior care	Number of Hospice Claims	Measure	Number of pre-ESKD Medicare Hospice care claims a patient has
104.	Pre-ESKD Claims	Prior care	Number of Skilled Nursing Unit Stays	Measure	Number of pre-ESKD Medicare Skilled Nursing Unit stay claims a patient has
105.	Pre-ESKD Claims	Prior care	Total Length of Inpatient Stays	Measure (days)	Total length of each inpatient stay in days was calculated by subtracting the claim start date from the claim end date and specifying days as the unit of measure
106.	Pre-ESKD Claims	Prior care	Total Length of Skilled Nursing Unit Stays	Measure (days)	Total length of each skilled nursing unit stay in days was calculated by subtracting the claim start date from the claim end date and specifying days as the unit of measure
107.	Pre-ESKD Claims	Prior care	Total Length of Outpatient Claims	Measure (days)	Total length of each outpatient claims in days was calculated by subtracting the claim start date from the claim end date and specifying days as the unit of measure. The total number of outpatient claims was summed to create this predictor
108.	Pre-ESKD Claims	Prior care	Total Length of Hospice Claims	Measure (days)	Total length of each hospice stay in days was calculated by subtracting the claim start date from the claim end date and specifying days as the unit of measure

109.	Pre-ESKD Claims	Prior care	Total Length of Home Health Claims	Measure (days)	Total length of each home health claim in days was calculated by subtracting the claim start date from the claim end date and specifying days as the unit of measure
110.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Inpatient Claim	Measure (days)	Elapse time between first and last inpatient claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date for all inpatient claims
111.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Outpatient Claim	Measure (days)	Elapse time between first and last outpatient claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date for all outpatient claims
112.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Home Health Claim	Measure (days)	Elapse time between first and last home health claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date for all home health claims
113.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Hospice Claim	Measure (days)	Elapse time between first and last hospice claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date for all hospice claims
114.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Skilled Nursing Unit Claim	Measure (days)	Elapse time between first and last skilled nursing unit claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date for all skilled nursing unit claims
115.	Pre-ESKD Claims	Prior care	Elapsed Time Between First and Last Claim	Measure (days)	Elapse time between first and last claim was calculated by subtracting the minimum pre-ESKD claim date from the maximum pre-ESKD claim date across all 5 settings (IP, OP, HH, HS, SN)
116.	Pre-ESKD Claims	Prior care	Has inpatient claim	Binary	1 if a patient has an inpatient claim; 0 if not
117.	Pre-ESKD Claims	Prior care	Has outpatient claim	Binary	1 if a patient has an outpatient claim; 0 if not

118.	Pre-ESKD Claims	Prior care	Has home health claim	Binary	1 if a patient has a home health claim; 0 if not
119.	Pre-ESKD Claims	Prior care	Has hospice claim	Binary	1 if a patient has a hospice claim; 0 if not
120.	Pre-ESKD Claims	Prior care	Has skilled nursing unit claim	Binary	1 if a patient has a skilled nursing unit claim; 0 if not
121.	Pre-ESKD Claims	Prior care	Has any pre-ESKD Medicare claim	Binary	1 if a patient has any kind of pre-ESKD claim; 0 if not

Outliers

Supplemental Table 5 shows the upper and lower bounds for the clinical and laboratory variables included in the study dataset that were defined by co-authors M.E and M.S. Values that fell outside these bounds were considered clinically impossible. Around 0.5-2.3 percent of values were determined to be outliers across the clinical variables. Binary variables were created for each clinical/laboratory predictor to indicate (1) whether the original values were missing and (2) whether the original values were out of bounds. The outlier values for each predictor were set as missing and a numerical value was imputed using the 'mice' (version 3.13.0) in R (multiple imputations by chained equations[3]).

Supplemental Table 5: Upper and lower bounds for clinical and laboratory values

Variable	Lower bound	Upper bound
Height (cm)	76	243
Weight (kg)	20	250
BMI (kg/m2)	13	75
Serum Creatinine (mg/dL)	0.5	50
Serum Albumin (g/dL)	0.5	8
Estimated GFR (eGFR)*	1	30
Hemoglobin (g/dL)	2	18

*Estimated GFR was calculated using the CKD – EPI (Chronic Kidney Disease Epidemiology Collaboration) equation

Machine Learning

Hyperparameter Tuning of XGBoost Models

The R package XGBoost (version 1.3.2.1[4]) was used for this project. **Supplemental Table 6** shows the model settings (hyperparameters) that were selected for tuning. Additional documentation for the parameters can be found in the XGBoost documentation. Three parameters were set for the XGBoost models outside of parameter tuning: (1) setting scale_pos_weight (the parameter that handles the class imbalance by weighting the minority class) to 3.5 (the square root of the ratio of the negative class to the positive class), which per XGBoost documentation is a typical value to consider; (2) setting the number of iterations as 100, which is a standard number used for XGBoost[5]; and (3) setting early stopping rounds (the parameter that ends model training if the area under the receiver operating characteristic curve (AUC ROC or c-statistic) had not increased in X number of iterations) to 15 to prevent model overfitting.

Hyperparameters were tuned for the non-imputed dataset with a Bayesian optimization approach, and 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 5 imputed datasets were set as the new ranges for 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 (between 0 and 1) for each imputed dataset were combined by averaging the model prediction scores per patient across the 5 imputations to result in one prediction per patient. 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.

Supplemental Table 6: Hyperparameter Tuning

Hyperparameter (model settings)	Definition	Range of values	Optimal value (non-imputed)	Optimal value (imputed)
NRounds	Number of learning iterations for each model	10 - 500	497	493
Eta	Learning rate	0.001 - .8	0.057	0.050
Depth	Maximum tree depth in generating splits	2 - 10	6	7
Alpha	Regularization parameters for L1-norm	0 - 9	6.230	7.273
Lambda	Regularization parameters for L2-norm	1 - 9	8.318	8.207
Gamma	Minimum loss for generating a split	0 - 9	5.474	2.937
Subsample	Percent of observations sampled	0.2 - 1.0	0.751	0.751
Colsample_by_tree	Percent of predictors used	0.3 - 1.0	0.621	0.661
Min_child_ weight	The minimum number of observations subtending from a node in the tree	1 - 5	2	2
Max bin	Controls the maximum number of times the algorithm can split (the greater, the more the splits)	255 - 1023	354	935

XGBoost Imputed Model Results

Supplemental Table 7 shows the predictor rankings for imputed XGBoost model fit of the multiply imputed data as measured through gain (the relative contribution of the predictor to the model—a higher gain implies a feature is more important for generating a prediction).

Supplemental Table 7: Top 20 predictors of mortality within 90 days of dialysis initiation and their ranking of importance for the XGBoost model fit of the multiply imputed data as measured through gain (the relative contribution of the predictor to the model)

Rank	Feature	Gain
1	Age	0.1575
2	Total inpatient hospital days	0.0788
3	No maturing AVF	0.0441
4	Missing information on EPO receipt	0.0357
5	Patient documented to be unsuitable for kidney transplant due to age	0.0298
6	Missing information on whether a patient was under the care of kidney dietician	0.0295
7	Duration of time between first and last claim	0.0239
8	Serum Creatinine	0.0221
9	Albumin	0.0220
10	Estimated GFR (eGFR)	0.0222
11	Duration of time between first and last inpatient claim	0.0211

12	Underlying cause of ESKD categorized as other	0.0211
13	Cause of ESKD	0.0201
14	Not a nursing home occupant	0.0175
15	Number of inpatient claims	0.0154
16	Duration of time between first and last outpatient claim	0.0152
17	Not institutionalized	0.0156
18	Able to ambulate	0.0156
19	Number of comorbidities	0.0127
20	BMI	0.0134

EPO – exogenous erythropoietin; AVF – arteriovenous fistula; ESKD – end stage kidney disease; eGFR – glomerular filtration rate calculated using the Chronic Kidney Disease Epidemiology Collaboration equation; AVG – arteriovenous graft, BMI – Body Mass Index

The tradeoffs between sensitivity and specificity at the predicted mortality risk thresholds of 10%, 20%, 30%, 40%, and 50% were assessed for the XGBoost model fit of the multiply imputed data, as shown in **Supplemental Table 8**.

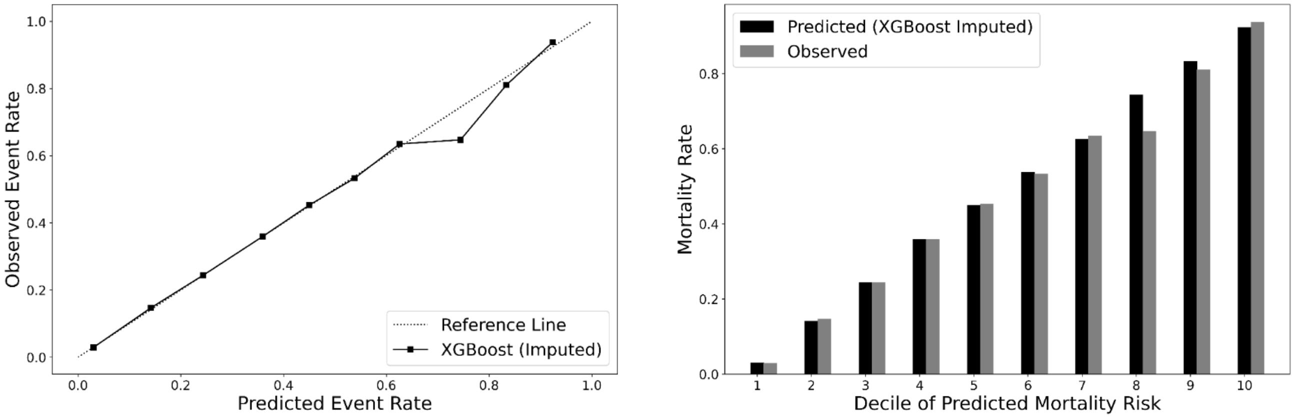
Supplemental Table 8: Predicted risk at 10, 20, 30, 40, 50% threshold for the XGBoost model fit of the multiply imputed data

Model Threshold	Sensitivity	Specificity	Likelihood Ratio (+)	Likelihood Ratio (–)	True Positive	False Positive	True Negative	False Negative
0.10	0.703	0.791	3.376	0.374	6,024	22,134	84,124	2,541
0.20	0.423	0.922	5.484	0.624	3,625	8,200	98,058	4,940
0.30	0.202	0.977	9.147	0.815	1,738	2,357	103,901	6,827
0.40	0.100	0.992	13.488	0.906	860	791	105,467	7,705
0.50	0.045	0.997	21.923	0.956	387	219	106,039	8,178

True positives = number of patients the model correctly predicted died in 90 days; False positives = number of patients the model incorrectly predicted died in 90 days; True negatives = number of patients the model correctly predicted survived in 90 days; False negatives = number of patients the model incorrectly predicted survived the first 90 days, Sensitivity = true positives/(true positives + false negatives) , Specificity = True negatives/(true negatives + false positives), likelihood ratio positive class = sensitivity/(1-specificity), likelihood ratio negative class = (1-sensitivity)/specificity.

Supplemental Figure 1 shows the calibration plot for the XGBoost imputed model predicted risks.

Supplemental Figure 1: Calibration plot for XGBoost imputed model predicted risks (a) Predicted risk by 10% intervals; (b) Predicted risk by decile



Project Resources as a Foundation for Future Work

The following resources generated from this project are available for broader use by investigators who are interested in pursuing future application of ML for kidney disease or other clinical use cases:

- Machine learning code developed for building the study dataset and ML models in this project are available at ONC GitHub.
- Implementation Guide with detailed methodology for building the study dataset and the ML models and considerations for future researchers based on the experiences of the project team, are available at HealthIT.gov project site.
- Final Report providing an overview of the project with methodology for building the study dataset and the ML models, considerations for future researchers based on the experiences of the project team, and recommendations for advancing ML in health care are available at HealthIT.gov project site.

[1] USRDS site refers to ESKD as ESRD; however, in this paper, references to ESRD have been changed to ESKD.

[2] Secondary diagnosis codes are not included in the USRDS data.

[3] 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/.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

[5] Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2020). Fair bayesian optimization. arXiv preprint arXiv:2006.05109.