**INTRODUCTION: comobid conditions that are predictive of ESRD**

eGFR alone was used as ESRD assessor, the the KFRE was built as a ESRD risk prediction model, but requires Albumin to calculate Albumin to creatnine ration, hence are there other comobidities can can predict ESRD, apart from lab values. Although there are proven therapies to improve outcomes in patients with progressive kidney disease, these therapies may also cause harm and add cost (Tangari 2011). Clinical decision making for CKD is challenging due to the heterogeneity of kidney dis- eases, variability in rates of disease pro- gression, and the competing risk of car- diovascular mortality (Tangari 2011). Accurate prediction of risk could facilitate indi- vidualized decision making, enabling early and appropriate patient care (Tangari 2011).

What comobid conditions can be used to predict ESRD other than lab values, to be used for early intervention in low resourse setting, for early intervention. CKD is exarbated by socio economic factors, hence in low resource settings,the diagnosis of comobid conditions can be used to calculate the risk of ESRD. The Models used today are lab based. We were especially interested in models that rely solely on information available to a primary care physician, enabling reporting the risk of kidney failure without laboratory test results. The ideal model to predict progression would be accurate, easytoimplement,andhighlygeneral- izable across a spectrum of patients with CKD in independent populations (Tangari 2011).

**METHODS**

**Study population and data source**

This retrospective cohort study utilized data from the "All of Us Research Program Controlled Tier Dataset v7," a comprehensive national database designed to capture health information from diverse populations across the United States

to advance precision medicine. Data extraction for this study was conducted up to July 1, 2022 (version C2022Q4R11,released on December 5, 2023). The study cohort consisted of 127,783 participants with recorded serum creatinine values. Within this cohort, 1,655 participants were identified as cases of End-Stage Renal Disease (ESRD), with 803 cases among Black participants, 794 among White participants, and 58 among Asian participants. The control group included 126,128 participants without

ESRD, comprising 29,199 Black participants, 92,844 White participants, and 4,085 Asian participants. In our analysis, we included 125,006 observations who had positive time to event.

**Cohort Description**

Table 1: Characteristics Developmental Cohort, Validation Cohort

**Prediction model construction**

The outcome of interest was kidney failure defined by SNOMED code 46177005. The time horizons for risk prediction were 2 and 5 years. The dataset was developed using SNOMED standardized variables from the All of Us Research Program. A comprehensive manual compilation was conducted to identify candidate predictors, guided by the ONCE framework and existing literature. Predictors were selected based on individual item counts, ensuring that the analysis captured the frequency of specific events or measurements. Systolic blood pressure was recorded as the closest value with measurement date time closest to index date. The index date for each participant was defined as the date of the first recorded serum creatinine

measurement. A 1-year retrospective feature window prior to the index date was used to generate predictors, with only data available during this period included. This feature window allowed us to capture count-based data for each predictor, reflecting the occurrence rates of specific conditions or exposures. Participants with end-stage renal disease (ESRD), defined by SNOMED code 46177005, were excluded if the diagnosis occurred before or on the the index date to get only positive time to event. The disease date is defined as the date of first diagnosis of ESRD. For the controls, the censored date is defined as the latest date of condition end date, device exposure end date, drug exposure end date, lab measurement date, observation date or procedure date.

**Statistical Analysis**

**Model Development**

The data preprocessing involved several critical steps. The First, a 70-30 split was performed, randomly assigning 70 percentof dataset on white and black participants to training and the remaining 30 percent to testing. 100 percent of the data on Asians was assigned to testing as there was not enough data on this group. The dataset was split to maintain a balanced distribution across training and testing sets based on race and case-control status. For the training set, Black or African American individuals had 19,953 controls and 489 cases, while White individuals had 63,798 controls and 507 cases. In the testing set, Black or African American individuals included 8,538 controls and 223 cases, White individuals had 27,322 controls and 238 cases, and Asian individuals comprised 3,894 controls and 44 cases. This stratification was implemented to ensure adequate representation across racial groups in both the training and testing phases, supporting robust model evaluation across diverse demographics. To address missing data, all NA values in the count data were replaced with zero ensuring data consistency and treating missing observations as zero counts. Multicollinearity was handled by setting a correlation threshold of 0.7, where only the variable with the higher case count was retained among correlated variables, while the other was excluded. Predictors were then examined for perfect separation to ensure clear differentiation between cases and controls. Variables with NA variance were excluded to avoid computational instability, ensuring that only stable and informative features were retained to enhance model robustness. Right truncation was applied, limiting data to two years for 2-year risk prediction and five years for 5-year risk prediction. Lastly, a log(x + 1) transformation followed by standard scaling was applied to normalize numeric variables, address skewness, and enhance interpretability. The addition

of 1 in the transformation prevented zero-value issues and ensured consistency across the data set.

To focus on a high-risk population, we restricted the dataset to individuals with an estimated glomerular filtration rate (eGFR) below 60, that is Patients with CKD stages 3 to 5 (estimated GFR, 60 mL/min/1.73 m2) calculated using the eGFR CKD-EPI 2021 equation, thereby concentrating the analysis on those with reduced kidney function. Each variable was then assessed using univariable Cox proportional hazards models, with non-significant variables (p > 0.10) (tangari 2011) excluded to refine the model. We used cox proportional hazards model to develop a sequencial series of models for 2 year and for 5 year predictions, comparing simple models to more complicated models. All cox proportional hazard model assumptions were assessed and the proportional hazards assumption was tested and on variables that it was violated we stratified on them. All final models satisfied the proportional hazards assumption on each variable and on the global test. We also compared the cox models to XGBoost, Random Forest, Lasso, Ridge and Elastic net models.

**Prediction Model Assessment**

We used a series of methods to evaluate the performance of the models in the de- velopment and the validation data sets. The baseline hazard function and coefficients from the de- veloped model were fixed and applied to the validation data set (tangari 2011).

***Discrimination***: Discrimination refers to the ability of a model to correctly distinguish between 2 classes of outcomes (kidney failure vs no kidney failure). Concordance statistics (C statistics) and integrated discrimina- tion improvement were computed as measures of discrimination.

***Calibration*:** Calibration describes how closely the predicted probabilities agree numerically with the observed out- comes. We compared the observed vs predicted risk of kidney failure for each quintile of predicted risk and determined the magnitude of the deviation using the Nam and D’Agostino chi-sq statistic.

***Goodness of Fit*:** Overall model fit for sequential models was compared using the Akaike Information Criterion (AIC), which takes into account both the sta- tistical goodness of fit and the number of parameters required to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters.

“***Shaply***”: Check prof Duan notes from meeting

All statistical analyses were performed using R on the ALL of us platform. Two- sided *P*  .05 was considered statistically significant. (Note: try shaply value, scale, + impact and – impact)

**RESULTS**

Figure 1: Observed vs Predicted Probability of Kidney Failure for 2 yr( Lasso and stepwise) and 2 yr( Lasso and stepwise) on the full test data

**Table 2 ( Rac Specifice vs Race Free Models):**

**Development Dataset**

**2 years**

**Model= All**

**1 2 3 4 5**

**Metric**

C Statistics (CI)

All

Black

White

Asian

AIC

BIC

**5 years**

**Model= All**

**1 2 3 4**

**Metric**

C Statistics (CI)

All

Black

White

Asian

AIC

BIC

**Table 3 ( Rac Specifice vs Race Free Models):**

**Evaluation Dataset**

**Model= All**

**1 2 3 4 5**

**Metric**

C Statistics

All

Black

White

Asian

Nam and D’Agostino chi-sq statistic

**5 years**

**Model= All**

**1 2 3 4**

**Metric**

C Statistics

All

Black

White

Asian

Nam and D’Agostino chi-sq statistic

DISCUSSION

External validation, also To evaluate the effect of definition of risk categories on reclassification , also evaluate the effect of the competing risk of mortality before kidney failure on risk prediction,

Model 1 (Socioeconomic): eGFR + age\_precise + Health\_insurance\_yes + raceBlack + Essential hypertension + Systolic blood pressure + Type 2 diabetic mellitius +

Polyneuropathy due to diabetic mellitius + chronic kidney disease stage 3 + BMI + Gout + Anemia in chronic kidney disease

Model 2 (race free): eGFR + age\_precise + Essential hypertension + Systolic blood pressure + Type 2 diabetic mellitius +

Polyneuropathy due to diabetic mellitius + chronic kidney disease stage 3 + BMI + Gout + Anemia in chronic kidney disease

Model 3 (lab free): Sex\_at\_birth + age\_precise + Essential hypertension + Systolic blood pressure + Type 2 diabetic mellitius +

Polyneuropathy due to diabetic mellitius + chronic kidney disease stage 3 + BMI + Gout + Anemia in chronic kidney disease

Model 3 (Lasso): eGFR + age

**2 years**

1. Use my\_dataframe\_insured\_routine with sistolic blood pressure
2. Remove all labs, remove eGFR, remove race
3. Data preprocessing, One variable cox, take out variables not significant at 0.10
4. Test Data preprocessing, One variable cox,
5. Run lasso for variable selection
6. **Put lasso variables** into tien code, assess in All, Black, White, Asian
7. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model
8. **Put** eGFR + Lasso lab free into tien code, assess in All, Black, White, Asian
9. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model
10. **Put** raceBlack + raceWhite + eGFR + Lasso lab free into tien code, assess in All, Black, White, Asian
11. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model

**5 years**

1. Use my\_dataframe\_insured\_routine
2. Remove all labs, remove eGFR, remove race
3. Data preprocessing, One variable cox, take out variables not significant at 0.10
4. Test Data preprocessing, One variable cox,
5. Run lasso for variable selection
6. **Put lasso variables** into tien code, assess in All, Black, White, Asian
7. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model
8. **Put** eGFR + Lasso lab free into tien code, assess in All, Black, White, Asian
9. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model
10. **Put** raceBlack + raceWhite + eGFR + Lasso lab free into tien code, assess in All, Black, White, Asian
11. Calculate Nam and D’Agostino chi-sq statistic, AIC, Sigma for full model

Put results in Table, write the results section and Discussion.

# Chapter 3

**INTRODUCTION: Assessing the effect of socioeconomic variables and APOL1 genes**

APOL1 is associated with ESRD, but not all people who have APOL1 gene hhave ESRD, so we set out to investigate Among people who had the APOL1 gene, what was the difference between the cases and controls? What other comobid conditions or genetic factors did they have.

**METHODS**

**Study population and data source**

This retrospective cohort study utilized data from the "All of Us Research Program Controlled Tier Dataset v7," a comprehensive national database designed to capture health information from diverse populations across the United States

to advance precision medicine. Data extraction for this study was conducted up to July 1, 2022 (version C2022Q4R11,released on December 5, 2023). The study cohort consisted of 127,783 participants with recorded serum creatinine values. Within this cohort, 1,655 participants were identified as cases of End-Stage Renal Disease (ESRD), with 803 cases among Black participants, 794 among White participants, and 58 among Asian participants. The control group included 126,128 participants without

ESRD, comprising 29,199 Black participants, 92,844 White participants, and 4,085 Asian participants. In our analysis, we included 125,006 observations who had positive time to event.

**Cohort Description**

Table 1: With the gene vs without the gene

**Prediction model construction**

The outcome of interest was kidney failure defined by SNOMED code 46177005. The time horizons for risk prediction were 2 and 5 years. The dataset was developed using SNOMED standardized variables from the All of Us Research Program. A comprehensive manual compilation was conducted to identify candidate predictors, guided by the ONCE framework and existing literature. Predictors were selected based on individual item counts, ensuring that the analysis captured the frequency of specific events or measurements. Systolic blood pressure was recorded as the closest value with measurement date time closest to index date. The index date for each participant was defined as the date of the first recorded serum creatinine

measurement. A 1-year retrospective feature window prior to the index date was used to generate predictors, with only data available during this period included. This feature window allowed us to capture count-based data for each predictor, reflecting the occurrence rates of specific conditions or exposures. Participants with end-stage renal disease (ESRD), defined by SNOMED code 46177005, were excluded if the diagnosis occurred before or on the the index date to get only positive time to event. The disease date is defined as the date of first diagnosis of ESRD. For the controls, the censored date is defined as the latest date of condition end date, device exposure end date, drug exposure end date, lab measurement date, observation date or procedure date.

We also included Socioeconomic data from the ALL of us platform, the socioeconomic indicators used were the Basics : Are you covered by health insurance or some other kind of health care plan? - Yes or No, and Health care access and utilization: Is there a place that you USUALLY go to when you are sick or need advice about your health? – Yes, Urgent\_care; Yes, Emergency room, Yes, Doctors\_office. These were treated as binary variables.

Additionally we included a binary indicator for APOL1 gene carrier defined by the SNP/Indel Variants: Variant 22-36265860-A-G or Variant 22-36265988-T-G or Variant 22-36265995-AATAATT-A.

We also included poly genic risk scores from African, American, Euroipean, East asian and matched them to our dataset to collect polygenic risk scores of our cohort. We included 15 Polygenic Score (PGS) Variables: "PGS002757" "PGS003988" "PGS004004" "PGS004016" "PGS004030" "PGS004045" "PGS004058"

"PGS004074" "PGS004088" "PGS004101" "PGS004112" "PGS004128" "PGS004142" "PGS004158" "PGS004889".

. Each of these represents a specific polygenic score (PGS) for a particular trait, condition, or set of genetic variants. These polygenic scores are typically calculated based on the presence or absence of certain genetic markers (e.g., SNPs) and the weights assigned to these markers based on their association with specific traits or conditions. Each score represents a composite measure of risk based on multiple genetic variants. wE HAD 21824 observations with missing polygenic risk scores, 105959 had polygenic risk scores.

**Statistical Analysis**

**Model Development**

The data preprocessing involved several critical steps. The First, a 70-30 split was performed, randomly assigning 70 percentof dataset on white and black participants to training and the remaining 30 percent to testing. 100 percent of the data on Asians was assigned to testing as there was not enough data on this group. The dataset was split to maintain a balanced distribution across training and testing sets based on race and case-control status. For the training set, Black or African American individuals had 19,953 controls and 489 cases, while White individuals had 63,798 controls and 507 cases. In the testing set, Black or African American individuals included 8,538 controls and 223 cases, White individuals had 27,322 controls and 238 cases, and Asian individuals comprised 3,894 controls and 44 cases. This stratification was implemented to ensure adequate representation across racial groups in both the training and testing phases, supporting robust model evaluation across diverse demographics. To address missing data, all NA values in the count data were replaced with zero ensuring data consistency and treating missing observations as zero counts. Multicollinearity was handled by setting a correlation threshold of 0.7, where only the variable with the higher case count was retained among correlated variables, while the other was excluded. Predictors were then examined for perfect separation to ensure clear differentiation between cases and controls. Variables with NA variance were excluded to avoid computational instability, ensuring that only stable and informative features were retained to enhance model robustness. Right truncation was applied, limiting data to two years for 2-year risk prediction and five years for 5-year risk prediction. Lastly, a log(x + 1) transformation followed by standard scaling was applied to normalize numeric variables, address skewness, and enhance interpretability. The addition

of 1 in the transformation prevented zero-value issues and ensured consistency across the data set.

To focus on a high-risk population, we restricted the dataset to individuals with an estimated glomerular filtration rate (eGFR) below 60, that is Patients with CKD stages 3 to 5 (estimated GFR, 60 mL/min/1.73 m2) calculated using the eGFR CKD-EPI 2021 equation, thereby concentrating the analysis on those with reduced kidney function. Each variable was then assessed using univariable Cox proportional hazards models, with non-significant variables (p > 0.10) (tangari 2011) excluded to refine the model. We used cox proportional hazards model to develop a sequencial series of models for 2 year and for 5 year predictions, comparing simple models to more complicated models. All cox proportional hazard model assumptions were assessed and the proportional hazards assumption was tested and on variables that it was violated we stratified on them. All final models satisfied the proportional hazards assumption on each variable and on the global test. We also compared the cox models to XGBoost, Random Forest, Lasso, Ridge and Elastic net models.

**Prediction Model Assessment**

We used a series of methods to evaluate the performance of the models in the de- velopment and the validation data sets. The baseline hazard function and coefficients from the de- veloped model were fixed and applied to the validation data set (tangari 2011).

***Discrimination***: Discrimination refers to the ability of a model to correctly distinguish between 2 classes of outcomes (kidney failure vs no kidney failure). Concordance statistics (C statistics) and integrated discrimina- tion improvement were computed as measures of discrimination.

***Calibration*:** Calibration describes how closely the predicted probabilities agree numerically with the observed out- comes. We compared the observed vs predicted risk of kidney failure for each quintile of predicted risk and determined the magnitude of the deviation using the Nam and D’Agostino chi-sq statistic.

***Goodness of Fit*:** Overall model fit for sequential models was compared using the Akaike Information Criterion (AIC), which takes into account both the sta- tistical goodness of fit and the number of parameters required to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters.

“***Shaply***”: Check prof Duan notes from meeting

All statistical analyses were performed using R on the ALL of us platform. Two- sided *P*  .05 was considered statistically significant. (Note: try shaply value, scale, + impact and – impact)

**RESULTS**

Figure 1: Observed vs Predicted Probability of Kidney Failure for 2 yr( Lasso and stepwise) and 2 yr( Lasso and stepwise) on the full test data

**Table 2 ( Rac Specifice vs Race Free Models):**

**Development Dataset**

**2 years**

**Model= All**

**1 2 3 4 5**

**Metric**

C Statistics (CI)

All

Black

White

Asian

AIC

BIC

**5 years**

**Model= All**

**1 2 3 4**

**Metric**

C Statistics (CI)

All

Black

White

Asian

AIC

BIC

**Table 3 ( Rac Specifice vs Race Free Models):**

**Evaluation Dataset**

**Model= All**

**1 2 3 4 5**

**Metric**

C Statistics

All

Black

White

Asian

Nam and D’Agostino chi-sq statistic

**5 years**

**Model= All**

**1 2 3 4**

**Metric**

C Statistics

All

Black

White

Asian

Nam and D’Agostino chi-sq statistic

DISCUSSION

External validation, also To evaluate the effect of definition of risk categories on reclassification , also evaluate the effect of the competing risk of mortality before kidney failure on risk prediction,

selected\_columns <- c(

"time\_to\_event\_2", "outcome","race",

"eGFR\_ckd\_epi\_2021",

"age\_precise",

"Anemia",

"Hypothyroidism",

"Iron.deficiency.anemia",

"Gout",

"Proteinuria",

"Acidosis",

"Hyperkalemia"

)

selected\_columns <- c(

"time\_to\_event\_2", "outcome", "race",

"raceBlack",

"ethnicity",

"eGFR\_ckd\_epi\_2021",

"age\_precise",

"sex\_at\_birth",

"Gout",

"Congestive.heart.failure",

"Type.2.diabetes.mellitus",

"BMI",

"Anemia",

"Atherosclerosis.of.coronary.artery.without.angina.pectoris",

"Hypothyroidism",

"Essential.hypertension"

)

<- c(

"time\_to\_event\_2", "outcome","race",

"eGFR\_ckd\_epi\_2021",

"age\_precise",

"sex\_at\_birth",

"Type.2.diabetes.mellitus",

"Essential.hypertension",

"Proteinuria",

"Anemia",

"Hypothyroidism",

"Iron.deficiency.anemia",

"Gout",

"Acidosis"

)