**Slide 1: Introduction**

**Good morning, everyone. Thank you for coming.**

Today, I am humbled to present my thesis, **"Evaluating Disparities in End-Stage Renal Disease Risk Prediction Using the All of Us Cohort.”**

**Slide 2: Public Health Crisis**

This study addresses inequities in kidney disease prediction, an issue that resonates deeply with me as someone from The Gambia, where kidney disease is a growing public health crisis.

Chronic kidney disease affects 19.8% of the population in West Africa, with an estimated prevalence of 13% in The Gambia and a staggering 49.4% mortality rate among hemodialysis patients.

CKD affects over 850 million people globally, more than 10% of the population, surpassing diabetes and cardiovascular disease in prevalence. Over 90% of cases go undiagnosed until advanced stages like End-Stage Renal Disease (ESRD), where irreversible kidney failure occurs, requiring dialysis or a transplant to survive. The image on the bottom right, illustrates the natural progression of CKD through its five stages, beginning with normal kidney function (Stage 1) and culminating in complete kidney failure (Stage 5). This progression visually highlights the critical opportunity for early detection and treatment to halt disease progression and improve outcomes.

This immense gap in disease outcome inspired my research: how can we ensure that risk prediction models—tools meant to save lives—work equitably for vulnerable populations? Today, I'll discuss the disparities we identified, our methods, and how we can refine these tools to better support vulnerable communities.

**Slide 3: Prevalence**

Between 2001 and 2021 in the United States, ESRD cases have grown. Annual incidents now exceed 130,000 cases per year, while prevalence surpassed 800,000 total cases in 2021, reflecting rising demands for dialysis and Kidney transplants.

Racial disparities are stark. Black individuals face an incidence rate at about 1,000 cases per million persons per year—four times higher than White individuals. Prevalence among Black patients consistently exceeds 6,000 cases per million persons, more than three times that of White individuals. Native American and Hispanic patients also face elevated prevalence at about 4,000 cases per million persons.

These statistics underscore persistent racial disparities in incidence, prevalence, and outcomes.

**Slide 4: eFGR Equations and Case Study**

Moreso, One of the key factors contributing to disparities in kidney disease prediction is the role of race in calculating kidney function.

For decades, equations like MDRD and CKD-EPI 2009 ,factored race explicitly into their formulas. As shown in number 1, the MDRD equation weighs in a multiplier 1.21 for Black individuals and 1 for non-black individuals in kidney function assessment. Similarly, the CKD-EPI equation from 2009 used a race weight of 1.159 for Black and 1 for non-black.

At the bottom of the slide, we have a **case study** comparing two patients: Patient A, White, and Patient B, Black. Both are 55-year-old females with identical serum creatinine levels—1.5 Milligrams per Deciliter. Based on the 2009 CKD-EPI equation, the calculated eGFR for Patient A is 42.3, while for Patient B, it is 61.2. The clinical implications are significant. Patient A, with a lower eGFR, is flagged for referral to a nephrologist, while Patient B is classified as having normal kidney function and does not receive intervention. The only difference? Race.

In 2021, a new version of the CKD-EPI equation was introduced, eliminating race as a variable. This marked a pivotal step forward, but it raises important questions: How do we ensure that predictive models are equitable? How do we ensure the accuracy and fairness of predictions across all groups?

**Slide 5: Model Misclassification**

On the left, we have a density plot from a study by Tsai et al. (2021). This study demonstrated that including the race multiplier shifts estimated Glomerular Filtration Rate (eGFR) values higher for Black individuals, insinuating "better" kidney function. However, this might not accurately reflect actual kidney health, raising critical concerns about the fairness and accuracy of this adjustment.

The bar chart on the right from our study demonstrates how different eGFR models classify CKD stages across racial groups using the same dataset. The Race-neutral model (CKD-EPI-2021) identifies more Black individuals as having CKD or being in early stages compared to race-inclusive models. Conversely, for Asians and Whites, the race-neutral model classifies more individuals as "Normal," reducing CKD detection rates for this subpopulation.

This comparisons underscores how race adjustments in eGFR calculations affect classification outcomes, with profound implications for diagnosis, care access, and health equity.

**Slide 6: Overview**

In this slide I will go through our study objectives. In Chapter 1 of our study, we aim to assess the predictive performance of both race-inclusive and race-neutral eGFR equations for estimating ESRD risk across population groups.

In Chapter 2, our goal is to enhance ESRD risk prediction by developing models that integrate comorbidities and other predictors. Given that eGFR emerged as a particularly strong predictor of ESRD, we prioritized its inclusion in our feature set to enhance accuracy.

Finally, in Chapter 3, we work towards refining risk stratification. This includes integrating genetic markers, like APOL1 variants, and polygenic risk scores, alongside socioeconomic factors, to enhance prediction accuracy.

For this work, we’re leveraging data from the All of Us Research Program, which is a national, longitudinal dataset that aims to represent the diversity of the U.S. population. This program is one of the most comprehensive efforts to capture data from people of all racial and ethnic backgrounds, as well as varying socioeconomic status, geographic regions, and health conditions. The dataset includes biospecimens, genetic information, and survey data, making it highly suitable for studying health disparities and the factors contributing to outcomes like ESRD. Its representative nature allows us to ensure that our findings are broadly applicable across the US population, which is critical when evaluating disparities.

**Slide 7: Study Cohort**

We designed a cohort study.

On the left, we have the **inclusion and exclusion criteria** that helped define our study cohort. The process began with over 413,000 participants aged 18 and older. From there, we applied several filters. First, we excluded about 36,000 participants with congenital abnormalities—conditions that could skew kidney function measurements. Next, we removed an additional 84,000 participants of unspecified or other races, narrowing our focus to Black, White, and Asian participants to ensure we had sufficient data for meaningful subgroup analyses.

Further exclusions included participants missing creatinine lab values, and those with diseases diagnosed prior to the index date. These exclusions were vital to ensure the integrity of our analysis.

Displayed on the right is the final study cohort, comprising 127,783 participants aged 18 years and older. Sex at birth was recorded as either male or female for all included individuals.

This comprised of, over 93,000 White participants, including 794 cases and more than 92,000 controls. Around 30,000 Black participants, with 803 cases and about 29,000 controls.

Just over 4,000 Asian participants, with 58 cases and over 4,000 controls.

In the second layer of the population breakdown, we focused on participants with available albumin lab values, a key variable for assessing kidney damage. Here, the sample sizes decrease further, but this sub-cohort provides critical insights into kidney function disparities.

In the next section, I’ll outline the statistical methods and models we used, beginning with chapter 1 .

**Slide 8: Chapter 1: Methodology**

The core model used in this study is the 4-variable Kidney Failure Risk Equation (KFRE), which is widely utilized in clinical practice to predict ESRD risk. The KFRE incorporates eGFR, albumin-creatinine ratio, age, and sex to estimate the likelihood of ESRD progression. Moving to the second box, eGFR values were calculated using three different equations: MDRD, CKD-EPI 2009, and CKD-EPI 2021. Next, we assessed the assumptions underlying the 4-variable KFRE to ensure the model's validity. By addressing these assumptions, we ensured that the KFRE model was appropriately specified for accurate risk prediction.

From there, we applied the KFRE to calculate risk scores. In the risk prediction phase, we used the KFRE beta coefficient sums to estimate both 2-year and 5-year risks for each participant. These scores were further categorized into risk levels based on KDIGO guidelines, which stratify patients by eGFR and Albumin to Creatnine Ratio thresholds.

On the bottom row, the model evaluation phase assessed performance within defined risk groups. This included analyzing predictive performance for CKD medium to very high risk groups, diabetic and hypertensive patients, subpopulations at heightened risk for ESRD. To quantify the accuracy of the models, we evaluated several performance metrics including C-statistics to assess overall predictive discrimination. Finally, we employed nonparametric bootstrapping with replacement, to calculate confidence intervals for the c statistics, ensuring reliable error estimation.

Next, I will present the results.

**Slide 9: Results**

**On the left we have the KFRE model equations for 2 year risk and 5 year risk, and moving to the right,** let’s examine the predictive accuracy of the three eGFR equations when used in the KFRE in combined populations.

Looking at the **C statistic** for the 2-year prediction, the MDRD model demonstrated the highest predictive accuracy in hypertensive individuals, with a C statistic of 0.94 (95% CI: 0.89–0.98). In contrast, its accuracy was lower in medium-risk CKD populations, populations defined by KDIGO guidelines as individuals with an estimated glomerular filtration rate between 30 and 59 milliliters per minute per 1.73 square meters or mild albuminuria (A2)—where the C statistic was 0.71 (95% CI: 0.64–0.79). Similar patterns were observed with the CKD-EPI-2009 and CKD-EPI-2021 models.

The 5-year prediction, also showed comparable C statistics and overlapping confidence intervals within risk groups among models.

The overlapping confidence intervals for both 2 and 5 year predictions, indicate there are no significant differences in their predictive accuracy, demonstrating that all three models are comparable in performance when used in the combined population.

Next, I will present the subgroup analysis.

**Slide 10: Comparison of C Statistics**

Here, we present the results of the Comparison of C-Statistics Between Race-Free and Race-Based Models. The analysis is divided into two sections: 2-year predictions on the left and 5-year predictions on the right. For both time frames, we assessed predictive performance differences across key risk groups—diabetic, hypertensive, and CKD risk levels from medium to very high. The objective was to determine whether the race-free model showed significant improvement over the race-inclusive models by evaluating if the confidence intervals for the differences in C-statistics excluded zero.

In the All Populations Group, for both 2-year and 5-year predictions, confidence intervals for the difference in C-statistics generally **included** zero across all risk groups. This means that the race free model showed no superiority over the race inclusive models when looking at the population as a whole.

Shifting focus to the Black population, a different story emerges. For both 2-year and 5-year predictions, the race free model demonstrated significant superiority over the race models, in the CKD High Risk and Very High Risk groups. In these categories, the entire confidence intervals for the difference in C-statistics were above zero, indicating improved predictive accuracy.

For the Asian population, confidence intervals included zero across all risk groups for both 2-year and 5-year predictions.

Finally, In the White population, no significant differences were observed across risk groups for 2-year predictions. However, for 5-year predictions, the CKD-EPI 2009 model outperformed in the CKD very high-risk group (**point left**), with the confidence interval for the difference in C-statistics **excluding zero**. Conversely, in the medium-risk group (**point right**), the CKD-EPI 2021 model demonstrated better predictive accuracy compared to the MDRD race-based model.

These findings underscore the variability in model performance across different racial and risk subgroups. While CKD-EPI 2021 marks progress by removing race as a variable, its predictive superiority isn’t universal. In some cases, particularly for the Black population’s CKD high risk groups, it performs better, but in others, such as the White population’s CKD Very High Risk group, it may even underperform. These disparities highlight the critical need for tailored approaches in model evaluation and clinical decision-making, motivating chapter 2 of our analysis.

**Slide 12: Chapter 2 Methodology**

In Chapter 2, we retrieved 254 predictors from the ONCE framework and converted them from PheCodes into standardized SNOMED codes. These codes were then used to extract EHR data from the All of Us research benchmark, ensuring consistency in the dataset.The predictors represented chronic and acute risk factors collected over a 1-year period before the index date. The selection process, spanning four months, involved meticulous manual validation to ensure high-quality dataset tailored for the analysis.

Next, we split the data into training and testing sets using a 70-30 split, motivated from the literature. Seventy percent of the data was used to train the model, and the remaining 30 percent was allocated for independent testing to evaluate model predictive performance.

For handling missing data we applied a complete case analysis.

To address potential multicollinearity, we applied a correlation threshold of 0.7. For highly correlated variables, only the variable with the higher case count was retained, while the other was excluded. This reduced redundancy and improved model stability.

We also examined the data for perfect separation—ensuring predictors didn’t perfectly divide cases and controls, as this could lead to overfitting. This step was particularly important for achieving generalizability across populations.

Data was right censored at 2 years for 2 year risk prediction and 5 years for 5 year risk prediction. This allowed the model to focus on specific time horizons for ESRD progression, providing insights into short- and long-term risk.

In our analysis, we tested key assumptions for the Cox Proportional Hazards (Cox PH) model, including the proportional hazards assumption. Assumptions for the Boruta algorithm, were also assessed. These steps ensured the validity and reliability of our results.

**Slide 11: Feature Selection**

As mentioned in the methodology, the 254 potential predictors used in chapter 2 are derived from the ONCE framework. ONCE is a powerful tool for developing clinical prediction models.

On the right, you can see how ONCE allows you to specify your target condition, in this case, ESRD. It leverages insights from published scientific literature to propose a broad set of features, such as comorbidities, laboratory parameters, and therapeutic interventions, which are then ranked by their relevance and importance scores.

To refine these features, we carefully defined each one to ensure clinical relevance. We conducted rigorous screening to remove redundancy, prioritize statistical performance, and address missing data. This process ensures that only the most relevant, evidence-backed, and reliable features are included in our models.

Next we look at variable selection.

**Slide 13: Variable Selection**In this analysis, we utilized the Boruta Algorithm, an advanced machine learning method that uses a random forest classifier to systematically identify and prioritize relevant predictors. By refining the model-building process, this approach ensures the creation of highly accurate and effective models, ultimately driving improvements in health outcomes.

Let’s look at the Variable Importance plots:

In the Race-Specified Model, which includes racial identity as one of the variables, the most significant predictors are eGFR, age, and systolic blood pressure. Both racial identity and ethnicity were selected as key features in the model.

In the Race-Free Model, where race was excluded as a variable, the feature set remained largely consistent.

For both models, the variables identified by the Boruta algorithm were incorporated into the Cox proportional hazards model.

Next lets take a look at the results of the Cox models.

**Slide 14: Cox Model**

Below we have the model equations for both models.

Starting with the development set metrics, we see that both the race-free model and the race-specific model achieve identical predictive performance, with a C-statistic of 0.89, confidence interval 0.87 to 0.91. This indicates that both approaches have comparable and strong discrimination capabilities.

However, there are slight differences in the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, which measure model fit. The race-specific model has lower values, suggesting it captures nuances that improve model fit, due to the inclusion of race and ethnicity as variables.

The models were validated with 30% of the data across population groups, looking at the C-index on the validation set:

Across all populations, the variations in performance metrics are minimal, indicating a relatively consistent predictive capability, which is promising for the race-free approach, as it suggests it can achieve comparable performance without directly incorporating race.

Calibration was assessed using the Nam and D’Agostino chi-squared statistic, with both models achieving values around 8,222, indicating comparable overall calibration.

The line graphs provide a visual comparison of predicted versus observed risk quintiles. Observed is in grey and predicted is in blue, the race-specific model shows slightly better calibration in the highest-risk quintile, where accurate prediction is most critical for clinical intervention.

These findings highlight that the race-free model offers comparable accuracy in our data, while avoiding the ethical concerns of using race as a predictor.

Next, In Chapter 3, I’ll discuss the broader implications of these findings and propose strategies for moving towards race free models in ESRD prediction.

**Slide 15: *Integrating Genetic and Socioeconomic Determinants for Enhanced Prediction of Kidney Damage***

In **Chapter 3**, we aimed to improve risk prediction for ESRD by integrating both genetic and socioeconomic factors. The objective was to refine our risk stratification models and address health disparities in ESRD outcomes. To ensure consistency, we utilized the same methodology as outlined in Chapter 2.

Let me walk you through the key components.

**Genetic Factors:**

First, we included **APOL1 gene carrier status**. APOL1 is a gene that produces a protein involved in lipid transport and immunity. Two specific variants, G1 and G2, are strongly associated with a higher risk of CKD, especially in individuals of African ancestry. These variants originally evolved as protection against African sleeping sickness but unfortunately increase the likelihood of kidney disease.

Next, we incorporated **Polygenic Risk Scores**, or PGS, which are composite genetic measures derived from diverse ancestry groups, including African, American, European, and East Asian populations. We used 15 different PGS measures to capture broader genetic contributions to ESRD.

**Socioeconomic Factors:**

On the socioeconomic side, we used **health insurance coverage** as a key variable. This was derived from survey responses where participants were asked: *“Are you covered by health insurance or any other healthcare plan?”*

We also assessed **healthcare access and utilization**. Participants were asked if they had a usual place to seek care when they were sick. Responses were categorized into three groups: "Urgent care," "Emergency room," and "Doctor’s office."

By combining these genetic and socioeconomic factors, we were able to build a more comprehensive model. This approach not only improves prediction accuracy but also provides insights into how social and biological factors interact to influence ESRD outcomes. Ultimately, it helps address health disparities in kidney disease progression.

**Slide 16: *Survival analysis***

Starting with the survival curves based on race. The results show that Black participants experience significant lower survival probabilities compared to White and Asian participants. While this highlights the persistent disparities in kidney disease outcomes, it also underscores the limitations of using race as a variable—it reflects systemic inequities rather than intrinsic biological differences, as race is a social construct, not a biological factor.

Let’s address the role of genetics. Survival analysis was conducted to evaluate the effect of APOL1 carrier status on outcomes, using a Kaplan-Meier estimator to compare survival probabilities over time between carriers and non-carriers. The resulting survival curves demonstrate a significant difference in survival probabilities. Specifically, APOL1 gene carriers consistently exhibit lower survival probabilities compared to non-carriers, with the difference becoming more pronounced over time, with a p-value of less than 0.0001.

We conducted a comparative analysis of four models: the race-based model, the genetic model, the socioeconomic model, and the integrated model, which combines genetic and socioeconomic factors, to evaluate predictive performance.

The following section will present and analyze the results of this comparative assessment.

**Slide 17: 2 Year Prediction**

This slide presents the results of the 2-Year Prediction, comparing the performance of the four different models.

The Race model served as the baseline, achieving a C-statistic of 0.88 Confidence Interval: 0.86–0.90. The Socioeconomic model showed marginally higher discriminatory performance, with a C-statistic 0.89 Confidence Interval: 0.87–0.91. Both the Genetic and Integrated models matched the Race model’s performance.

For model fit, the Race model demonstrated the best fit among the four models, having the lowest AIC at 3,229 and BIC at 3,266.5.

In the stratified analysis, , in the validation set with 30% of the data, the models performed comparable, across population subgroups with overlapping confidence interval.

However, the Genetic and Integrated models demonstrated the best calibration with chi square statistics 6,604.5.

The calibration plots revealed that the Genetic and Integrated models showed strong alignment with observed rates in the lower to moderate quintiles but misestimated the risk in Quintiles 4 and 5.

We now evaluate the outcomes of the 5-year prediction analysis.

**Slide 18: 5 year Prediction**

Here, the Socioeconomic model showed a marginally higher but comparable discriminatory performance, with a C-statistic of 0.86 Confidence Interval: 0.84–0.88.

By Model fit, the Integrated model demonstrated the best fit among the four models, achieving the lowest AIC at 5,025.3 and BIC at 5,074.71.

In the stratified analysis, in the validation set, the models performed comparable across population subgroups.

In calibration assessment, the Genetic and Integrated models exhibited superior calibration with chi square statistics 6,504.7. Calibration plots indicate these models aligned closely with observed outcomes, particularly in the highest risk qualtiles 4 and 5, surpassing the Race model in capturing high-risk groups accurately.

Based on the chi square statistics, the 5-year prediction results suggest that the genetic and integrated model which combines socioeconomic and genetic factors can serve as a viable alternative to using race in predictive models for ESRD.

To conclude, I will discuss the study's limitations and propose directions for future research.

**Slide 19: Conclusion and Limitation:**

Beginning with the conclusions:

* First, this study shows that in eGFR calculations, the race free model, outperform traditional models that rely solely on race, in high risk black populations.
* Genetic and integrated models outperform race-based models for long term prediction of ESRD.
* These models demonstrate superior calibration and discrimination, especially in high-risk groups long term. This conclusion is supported by the Nam and D’Agostino χ² statistics and calibration plots, which indicate that these models align more closely with observed outcomes and are better at identifying individuals at higher risk.
* Finally, the results emphasize the importance of social determinants of health. Incorporating socioeconomic factors enhances the ability to address renal health disparities, especially for vulnerable populations.

Limitations:

* There are, however, a few limitations to note. First, the sample size for Asian participants in the dataset was limited, which may have affected the model's performance for this subgroup.
* Additionally, our study included only two socioeconomic determinants of health, which may not fully capture the broader impact of socioeconomic factors on health outcomes.
* Finally, while the findings are promising, broader validation in diverse clinical settings is needed to ensure the generalizability of these models.

**Slide 20: Future Work**

As I wrap up this discussion, I want to highlight how the prediction models built for this thesis can be extended in future work.

**First, Advanced Fairness-Aware Machine Learning Methods:**  
Going forward, we aim to implement fairness-aware methods to address biases. These methods are designed to identify and mitigate disparities in model predictions, ensuring equitable outcomes.

**Next, Advanced Multi-Task Learning Methods:**  
The models developed focus on specific outcomes, but healthcare challenges are rarely that isolated. Future work will involve incorporating multi-task learning approaches, allowing the models to predict multiple, related outcomes simultaneously. For example, instead of just predicting the risk of ESRD, the model could simultaneously evaluate other chronic conditions that are often comorbid, making it more comprehensive and actionable.

**Generalizing Results to Low-Income Populations:**  
Currently, the models have been trained on data from the All of Us cohort, which, while diverse, may not fully reflect the realities in low-resource settings like The Gambia. The next step is to adapt and validate these models for use in such settings. By tailoring prediction thresholds and performance metrics to the unique challenges of low-resource healthcare systems, these models can become tools for real impact in these communities.

**Slide 21: Acknowledgements**

Finally, I would like to take a moment to express my deepest gratitude to leaders in public health, biostatistics and epidemiology who have supported and guided me throughout this journey.

First and foremost, to my advisor, Dr. Duan, thank you for your mentorship, your unwavering support, and for challenging me to grow as a researcher and thinker. Your guidance has been invaluable, and this work would not have been possible without you.

To my thesis committee members, **Dr. Lake**, **Dr. Haneuse and Dr. Miller**, I am deeply grateful for the invaluable expertise, thoughtful feedback, and the generous time you dedicated to reviewing this work. Your insights have been instrumental in shaping the direction and depth of this research, elevating it to a standard I am truly proud to present. Thank you for your guidance and unwavering support throughout this journey.

It is truly an honor to stand here in the Harvard T.H. Chan School of Public Health, a global leader in building a better world through public health, as its logo so boldly declares.

This institution represents the very essence of what drives me: the pursuit of innovative solutions to address health inequities and improve lives worldwide.

Biostatistics holds immense power to transform health outcomes. It is through data, research, and rigorous analysis that we uncover disparities, inform policy, and design interventions that save lives. This journey has been as much about the knowledge I’ve gained as it has been about the incredible people I’ve had the privilege to learn from along the way.

Thank you, and I’m happy to take any questions that you may have.