EVALUATING THE RELATIONSHIPS BETWEEN GENETIC ANCESTRY AND THE CLINICAL PHENOME

Jacqueline A. Piekos[1-3^] and Jeewoo Kim,[1-3†] Jacob M. Keaton,[4] Jacklyn N. Hellwege,[1,5‡] Todd L. Edwards[1,5] and Digna R. Velez Edwards[1-3,5]

1. Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee 37203; 2. Department of Obstetrics and Gynecology, Vanderbilt University Medical Center Nashville, Tennessee 37232; 3. Department of Biomedical Informatics, Vanderbilt University Medical Center Nashville, Tennessee 37232; 4. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; 5. Department of Medicine, Vanderbilt University Medical Center Nashville, Tennessee 37203

Corresponding authors: todd.l.edwards@vumc.org, digna.r.velez.edwards@vumc.org

Abstract

There is a desire in research to move away from the concept of race as a clinical factor because it is a societal construct used as an imprecise proxy for geographic ancestry. In this study, we leverage the biobank from Vanderbilt University Medical Center, BioVU, to investigate relationships between genetic ancestry proportion and the clinical phenome. For all samples in BioVU, we calculated six ancestry proportions based on 1000 Genomes references: eastern African (EAFR), western African (WAFR), northern European (NEUR), southern European (SEUR), eastern Asian (EAS), and southern Asian (SAS). From PheWAS, we found phecode categories significantly enriched neoplasms for EAFR, WAFR, and SEUR, and pregnancy complication in SEUR, NEUR, SAS, and EAS ($p < 0.003$). We then selected phenotypes hypertension (HTN) and atrial fibrillation (AFib) to further investigate the relationships between these phenotypes and EAFR, WAFR, SEUR, and NEUR using logistic regression modeling and non-linear restricted cubic spline modeling (RCS). For EAS and SAS, we chose renal failure (RF) for further modeling. The relationships between HTN and AFib and the ancestries EAFR, WAFR, and SEUR were best fit by the linear model (beta $p < 1x10^{-4}$ for all) while the relationships with NEUR were best fit with RCS (HTN ANOVA $p = 0.001$, AFib ANOVA $p < 1x10^{-4}$). For RF, the relationship with SAS was best fit with a linear model (beta $p < 1x10^{-4}$) while RCS model was a better fit for EAS (ANOVA $p < 1x10^{-4}$). In this study, we identify relationships between genetic ancestry and phenotypes that are best fit with non-linear modeling techniques. The assumption of linearity for regression modeling is integral for proper fitting of a model and there is no knowing a priori to modeling if the relationship is truly linear.

Keywords: genetic ancestry, health disparities, PheWAS, linear modeling

*Vanderbilt University Medical Center's BioVU is supporting by institutional funding, 1S10RR025141-01 and by the CTSA grant UL1TR000445 from NCATS/NIH.
^Work partially supported by T32GM080178
†Work supported by T32GM007347 and TL1TR002244
‡Work Partially supported by K12 HD043483

1. Introduction

Race is a social construct that is an imprecise way to classify groups prevalence of heritable risk factors, therefore there is a growing consensus in clinical and population research to move away from the use of race in the context of disease risk. Some racial disparities in health condition risks documented in the epidemiological literature may be due to non-biological differences between racial groups.[1] Geographic or genetic ancestry has been proposed as a more precise approach to capture differences in disease etiology that may be due to acquired biological differences in human populations. We hypothesize that when populations have evolutionarily adapted to a specific environment encounter different circumstances, disease risks can be influenced, and disparities can arise when compared to a population that is in evolutionary equilibrium with that environment. If this hypothesis is true, then this relationship would be detectable as an association between genetically inferred proportions of ancestry and disease risk. Improved understanding of how different geographic ancestries are responding to modern environments, nutrition, and behavioral lifestyles could help us understand genetic causes of diseases and improve healthcare.

Current approaches to precision medicine focus on a patient's clinical history and are often combined with known genetic risk factors, such as causal monogenic variants and more recently polygenic risk scores. Over the last several decades, race has been incorporated into clinical risk prediction models for several conditions when racial differences have been observed in disease prevalence, particularly for estimating drug responses. Race has also been used for medical tools such as calibrating eGFR measures for assessment of kidney disease risk. However, multiple studies have shown that administratively determined race or self-reported race are imprecise estimates of an individual's genetic ancestry, and thus use of race in modeling is a flawed approach.[2,3] Imprecise racial/ancestral identification may lead to lack of response to a personalized treatment plan that depends on a strong assumption of race capturing biological differences. Furthermore, recent work by several groups have shown that for some diseases genetic ancestry (global ancestry)[4] may directly interact with a patient's clinical characteristics to modify risk for disease and that this interaction varies at specific points in their genome (local ancestry).[5-7]

Within this study we leverage the rich phenotypic information available from Vanderbilt University Medical Center's (VUMC) biobank, BioVU, to evaluate the relationship between global geographic ancestry and the clinical phenome using phenome wide association study (PheWAS). From PheWAS results, we sought to identify enriched phenotype categories for ancestry groups and selected phenotypes within them for additional modeling. Selected phenotypes were then modeled using logistic regression and restricted cubic splines (RCS) to further investigate the relationship between phenotype and ancestry group. Studies usually make the strong assumption that the relationship between genetic ancestry and disease risk is linear. We chose to explore if fitting a non-linear model better described the relationship.

2. Methods

2.1. *Study Population*

The BioVU DNA Repository is a de-identified database of electronic health records (EHR) that are linked to patient DNA samples at VUMC. A detailed description of the database and how it is maintained has been published elsewhere.[8] BioVU participant DNA samples were genotyped on a custom Illumina Multi-Ethnic Genotyping Array (MEGA-ex; Illumina Inc., San Diego, CA, USA).

Quality control included excluding samples or variants with missingness rates above 2%, excluded if consent had been revoked, sample was duplicated, or failed sex concordance checks. Imputation was performed on the Michigan Imputation Server v1.2.410 using Minimac4[9] and the Haplotype Reference Consortium (HRC) panel v1.1.[10]

2.2. *Ancestry Estimations of BioVU Participants*

Estimation of ancestry proportion for BioVU participants based upon 1000 Genomes reference data has been described elsewhere.[11] In brief, the 1000 Genome populations were grouped into six super-population by geographic ancestry of east African (EAFR), west African (WAFR), southern European (SEUR), northern European (NEUR), east Asian (EAS), and south Asian (SAS) as described in Keaton, et. al 2021[12] using ADMIXTURE.[13] The six ancestry groups were projected onto BioVU to determine proportion of the six ancestries for all samples. Ancestry proportion of samples in the cohort was visualized by plotting subjects along the x-axis and their corresponding stacked ancestry proportions on the y-axis. Subjects were sorted by increasing SEUR ancestry.

2.3. *Ancestry Phenome Wide Association Study*

We conducted hypothesis-free PheWAS analyses of evaluating phecodes in the phenome with each of the six ancestries. Each ancestry was used as the main predictor in separate analysis, adjusted for age, sex, and body mass index (BMI). PheWAS was performed with the R package 'PheWAS' version 2.[14] 1,875 clinical disease phenotypes called phecodes from Phecode Map 1.2 were evaluated.[15] A p-value of 2.7×10^{-5} was the threshold for significance to correct for multiple testing (Bonferroni correction of 0.05/1,875 phecodes tested).

2.3.1. *Hypergeometric Testing of Enrichment*

Post PheWAS, phecodes were mapped to phenotypes and the phenotypes were grouped into sixteen categories from the phecodes map. We then conducted hypergeometric testing for enrichment for each phecode category within each ancestry PheWAS result. The hypergeometric distribution function HYPGEOM.DIST from excel was used to calculate fold change and significance level for each category. Threshold for significance was 0.003 to correct for multiple testing (Bonferroni correction of 0.05/16 phecode groups tested). Hypergeometric testing results were visualized by plotting the -log(p-value) of enrichment for each category as a function of fold change. Phecode categories pregnancy complication and neoplasms were visualized by graphing each phecode in the categories by -log(p-value) as a function of effect size. Plots were made with R 4.2.2.[16]

2.3.2. *Selection of Phecodes for Modeling*

In PheWAS results, we looked for phecodes that differed in relationship between EAS and SAS, and between EAFR, WAFR and NEUR, SEUR. Renal failure (RF) was selected for further modeling in EAS and SAS. The pre-made phecode categories do not always capture all relevant codes to a certain system. To focus more on the cardiac system, we extracted phenotypes using the key terms "hypertens", "heart", "card", "valv", "fibril", "coronary", and "angina." After manual review, we excluded codes pertaining to "poisoning by agents primarily affecting the cardiovascular system" and "heartburn". Selected cardiac phecodes were visualized by plotting the -log(p-value) of the

phecodes as a function of effect size using R 4.2.2. From this cardiac systems plot, we selected phenotypes hypertension (HTN) and atrial fibrillation (AFib) for further modeling with EAFR, WAFR, NEUR, and SEUR.

2.4. *Logistic Regression Modeling of Select Phecodes*

Selected phenotypes were modeled as logistic regression and RCS using the R package "rms" version 6.2-0.[17] Each ancestry was used as the main predictor in separate models. Phenotypes were modeled as a function of ancestry proportion (ANC) using (Eq. 1) for logistic regression.

$$P\{Y = 1|X\} = \beta_0 + \beta_{ANC}X_{ANC} + \beta_{age}X_{age} + \beta_{sex}\beta_{sex} + \beta_{BMI}X_{BMI} \quad (1)$$

Odds ratios (OR) and confidence intervals (CI) calculated for each ancestry from logistic regression are given for a 10% increase in ancestry proportion. Phenotypes were modeled as a function of ancestry proportion using (Eq. 2) for RCS with three knots (a,b,c).

$$P\{Y = 1|X\} = \beta_0 + \beta_{ANC}X_{ANC} + \beta_{age}X_{age} + \beta_{sex}\beta_{sex} + \beta_{BMI}X_{BMI} + \beta_a(X_{ANC} + a)^3 + \beta_b(X_{ANC} + b)^3 + \beta_c(X_{ANC} + c)^3 \quad (2)$$

Knot positions were determined by default "rms" placement. Odds ratios for RCS were calculated using integrated "rms" functions for a quartile increase in ancestry from the 25th to 50th percentile and for the 50th to 75th percentile. Significance threshold for ANOVA tests of significant model improvement with RCS over linear was 0.004 (Bonferroni correction of 0.05/12 [six ancestries * two models]).

3. Results

3.1. *Genetic Ancestry of BioVU Participants*

There were 71,140 participants from BioVU, 59.06% of which were female, the average age was 54.09 (SD = 18.15), and the average BMI was 29.03 (SD = 7.27). (Table 1) Ancestry proportions for all individuals in BioVU are visualized in Figure 1. From the six ancestry proportions calculated, the ancestry group SEUR represented the largest proportion of genetic ancestry with a population average of 60.9%, followed by NEUR with 22.4%, WAFR with 6.41%, EAFR with 7.07%, SAS with 1.40%, and EAS with 1.76%. (Table 1)

3.2. *PheWAS Summarized with Hypergeometric Testing*

There were 404 phecodes significantly associated with EAFR, 396 with WAFR, 414 with SEUR, 150 with NEUR, 68 with SAS, and 74 with EAS. (Table 2) Hypergeometric testing of phecode categories identified enriched and de-enriched categories of phecodes. (Figure 2A) EAFR, WAFR and SEUR were de-enriched for 'injuries and poisonings' and 'musculoskeletal' and enriched for 'neoplasms.'

Table 1. Population characteristics of the BioVU cohort.

	Mean (SD) or N (%)
Age (years)	54.08 (18.15)
BMI (kg/m^2)	29.03 (7.27)
Sex (Females)	42016 (59%)
NEUR	22.43 (9.72)
SEUR	60.93 (23.29)
EAFR	7.07 (15.4)
WAFR	6.41 (14.09)
EAS	1.76 (9.6)
SAS	1.4 (6.05)

Kg: kilogram; m: meters

Phecodes significant within neoplasms showed opposite directions of effect for NEUR and SEUR groups versus WAFR and EAFR groups. (Figure 2B) Codes representing skin cancer and other skin neoplasms increased in odds with increasing NEUR and SEUR ancestry proportion but decreased in odds with increasing WAFR and EAFR ancestry proportion. Conversely uterine leiomyoma had increased odds with increased EAFR and WAFR ancestry proportion and decreased odds with increased SEUR and NEUR ancestry proportion. (Figure 2B) EAFR was additionally enriched for 'genitourinary.' 'Pregnancy complications' was enriched in NEUR, SEUR, EAS, and SAS. When investigated further, it was revealed the significant phecodes in the category were almost all in the decreased direction for NEUR and SEUR and increased direction for EAFR, WAFR, EAS, and SAS. (Figure 2C)

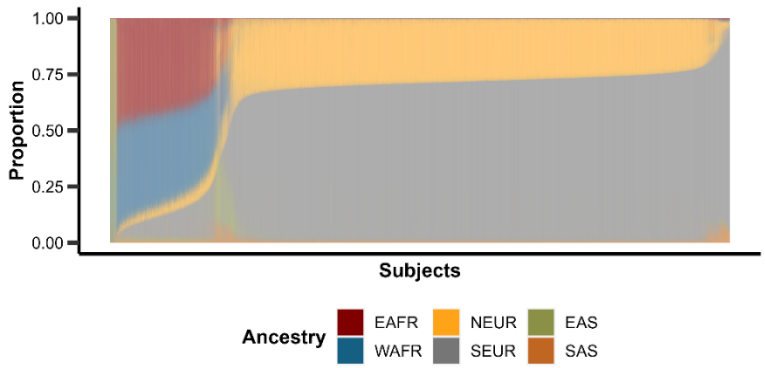

Figure 1. Structure plot of the genetic ancestry make up of BioVU Participants. Subjects are aligned on the x-axis by proportion of SEUR. NEUR: northern European; SEUR: southern European; EAFR: eastern African; WAFR: western African; EAS: eastern Asian; SAS: southern Asian ancestry.

3.3. *Modeling Ancestry Proportion*

We identified 103 phecodes that included cardiac keyword/phrases. The most significant phecodes were phecodes representing hypertension and its consequences. Increasing EAFR and WAFR ancestry proportion increases odds for the phecodes and increasing SEUR and NEUR ancestry proportion decreases odds for the conditions. Phecodes involving atrial fibrillation and related codes

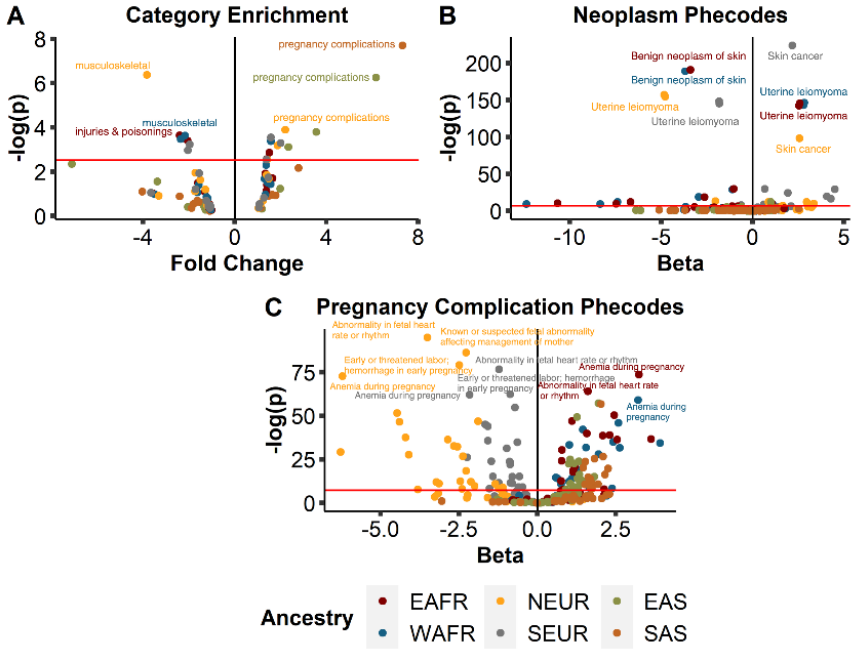

Figure 2. Volcano plots of fold change from hypergeometric testing or ancestry coefficient from PheWAS plotted against the negative log transformed p-value for A) Phecode categories B) neoplasm and C) pregnancy complications. Created with BioRender.com

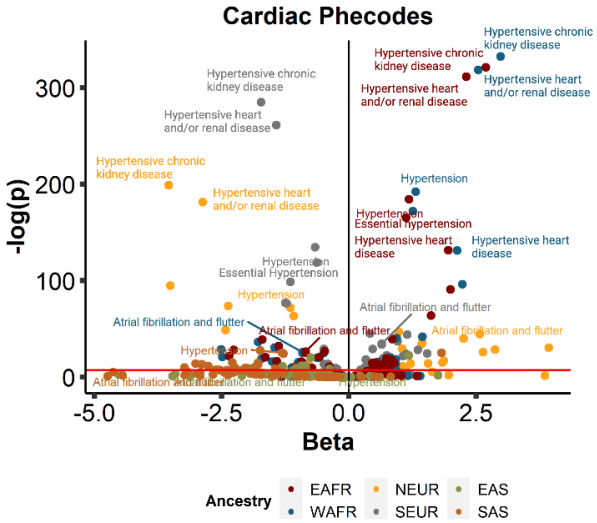

Figure 3. Volcano plot of selected phecodes related to the cardiac system. Coefficient of phecode from PheWAS is on the x-axis and the y-axis is negative log transformation of p-value. Created with BioRender.com

were significantly associated with the same ancestries, but in opposite directions: increasing NEUR and SEUR increase odds while EAFR and WAFR decrease odds. (Figure 3)

Table 2. Significant results from hypergeometric testing of phecode categories for each ancestry. Positive values indicated an enrichment of significant phecodes within that category while negative values indicate de-enrichment. Significance level is 0.003.

Ancestry ~ N Significant Codes	Fold Change	P-value
EAFR ~ 404		
genitourinary	1.39	0.003
injuries & poisonings	-2.4	2.24×10^{-4}
musculoskeletal	-2.05	3.89×10^{-4}
neoplasms	1.51	0.001
WAFR ~ 396		
injuries & poisonings	2.35	3.35×10^{-4}
musculoskeletal	-2.17	2.31×10^{-4}
neoplasms	1.57	4.13×10^{-4}
SEUR ~ 414		
injuries & poisonings	-2.05	0.001
musculoskeletal	-1.96	5.77×10^{-4}
neoplasms	1.58	2.83×10^{-4}
pregnancy complications	2	5.06×10^{-4}
NEUR ~ 150		
infectious diseases	1.87	6.38×10^{-4}
musculoskeletal	-3.83	4.17×10^{-4}
pregnancy complications	2.2	1.27×10^{-4}
SAS ~ 68		
pregnancy complications	7.32	1.97×10^{-8}
EAS ~ 74		
digestive	2.34	7.63×10^{-4}
mental disorders	3.56	1.57×10^{-4}
pregnancy complications	6.16	5.57×10^{-7}

We then investigated phecodes 401 ‘hypertension’ (HTN) and 427.2 ‘atrial fibrillation’ (AFib) with modeling in EAFR, WAFR, NEUR, and SEUR. (Figure 4A) When modeled linearly, each ancestry was associated with HTN and AFib ($p < 0.003$). (Table 3) When HTN and AFib were modeled using RCS, the ANOVA test revealed adding the complexity of non-linearity did significantly improve the model for NEUR ($p = 0.001$, $p < 1 \times 10^{-4}$ respectively) but not for EAFR, WAFR, and NEUR ($p > 0.003$). (Figure 4) Increasing ancestry proportion by 10% in the linear model gave an OR for HTN of 2.29 (95% CI: 2.11 - 2.48) for EAFR, 2.73 (95% CI: 2.48 - 3.01) for WAFR, 0.27 (95% CI: 0.22 - 0.33) for NEUR, and 0.73 (95% CI: 0.70 - 0.75) for SEUR, visualized in the top row panels of Figure 4A. For AFib, a 10% increase in ancestry proportion yields ORs of 0.58 (95% CI: 0.49 - 0.68) for EAFR, 0.53 (95% CI: 0.44 - 0.63) for WAFR, 4.39 (95% CI: 3.07 -

6.27) for NEUR, and 1.31 (95% CI: 1.22 - 1.40) for SEUR when modeled linearly and is visualized in the third row of panels in Figure 4A. Only NEUR had significant ANOVA p-values for the RCS models in both HTN and AFib. Increasing NEUR ancestry in RCS modeling of HTN from 25th to 50th percentile in NEUR ancestry proportion gave an OR of 0.96 (95% CI: 0.94 - 0.98) and the 50th to 75th percentile increase gave an OR of 0.99 (95% CI: 0.98 - 1.01). In RCS modeling of AFib, increase from 25th to 50th percentile in NEUR ancestry yielded an OR of 1.02 (95% CI: 0.99 - 1.06) and the 50th to 75th percentile increase yielded an OR of 0.98 (95% CI: 0.96 - 1.01). (Table 3) RCS models for HTN and AFib are visualized in the second and fourth row of panels in Figure 4A, respectively.

Table 3. Results of logistic regression and restricted cubic spline modeling for hypertension and atrial fibrillation in northern European, southern European, west African, and east African ancestry; and renal failure in eastern Asian and southern Asian ancestry.

	Logistic Regression		**Restricted Cubic Spline**		
	OR* (95% CI)	**P-value**	**OR± (95% CI)**	**OR ‡ (95% CI)**	**ANOVA P-value**
Atrial Fibrillation					
SEUR	1.31 (1.22-1.40)	$<1x10^{-4}$	1.00 (0.98-1.03)	1.00 (0.98-1.02)	0.06
NEUR	4.39 (3.07-6.27)	$<1x10^{-4}$	1.02 (0.99-1.06)	0.98 (0.96-1.01)	$<1x10^{-4}$
EAFR	0.58 (0.49-0.68)	$<1x10^{-4}$	0.99 (0.99-1.00)	0.97 (0.96-0.99)	0.01
WAFR	0.53 (0.44-0.63)	$<1x10^{-4}$	0.99 (0.99-1.00)	0.98 (0.96-0.99)	0.02
Hypertension					
SEUR	0.72 (0.70-0.75)	$<1x10^{-4}$	0.98 (0.96-0.99)	0.99 (0.98-1.00)	0.25
NEUR	0.27 (0.22-0.33)	$<1x10^{-4}$	0.96 (0.94-0.98)	0.99 (0.98-1.01)	0.001
EAFR	2.29 (2.11-2.48)	$<1x10^{-4}$	1.00 (0.999-1.003)	1.003 (.996-1.01)	0.30
WAFR	2.73 (2.48-3.01)	$<1x10^{-4}$	1.00 (0.998-1.002)	1.00 (0.99-1.01)	0.11
Renal Failure					
EAS	0.96 (0.73-1.26)	0.78	1.09 (1.08-1.11)	1.18 (1.15-1.21)	$<1x10^{-4}$
SAS	0.15 (0.06-0.37)	$<1x10^{-4}$	1.00 (0.98-1.03)	1.00 (0.98-1.03)	0.41

*Odds ratio given for 10% increase of ancestry proportion
±Odds ratio given for 25th to 50th percentile of ancestry proportion
‡ Odds ratio given for 50th to 75th percentile of ancestry proportion

In PheWAS results, phecode 585 ‘renal failure’ showed different relationships with EAS and SAS ancestry proportion; RF was significantly associated with SAS, but not for EAS. (Table 3) When modeled linearly, SAS ancestry proportion was significantly associated with RF ($p < 1x10^{-4}$)

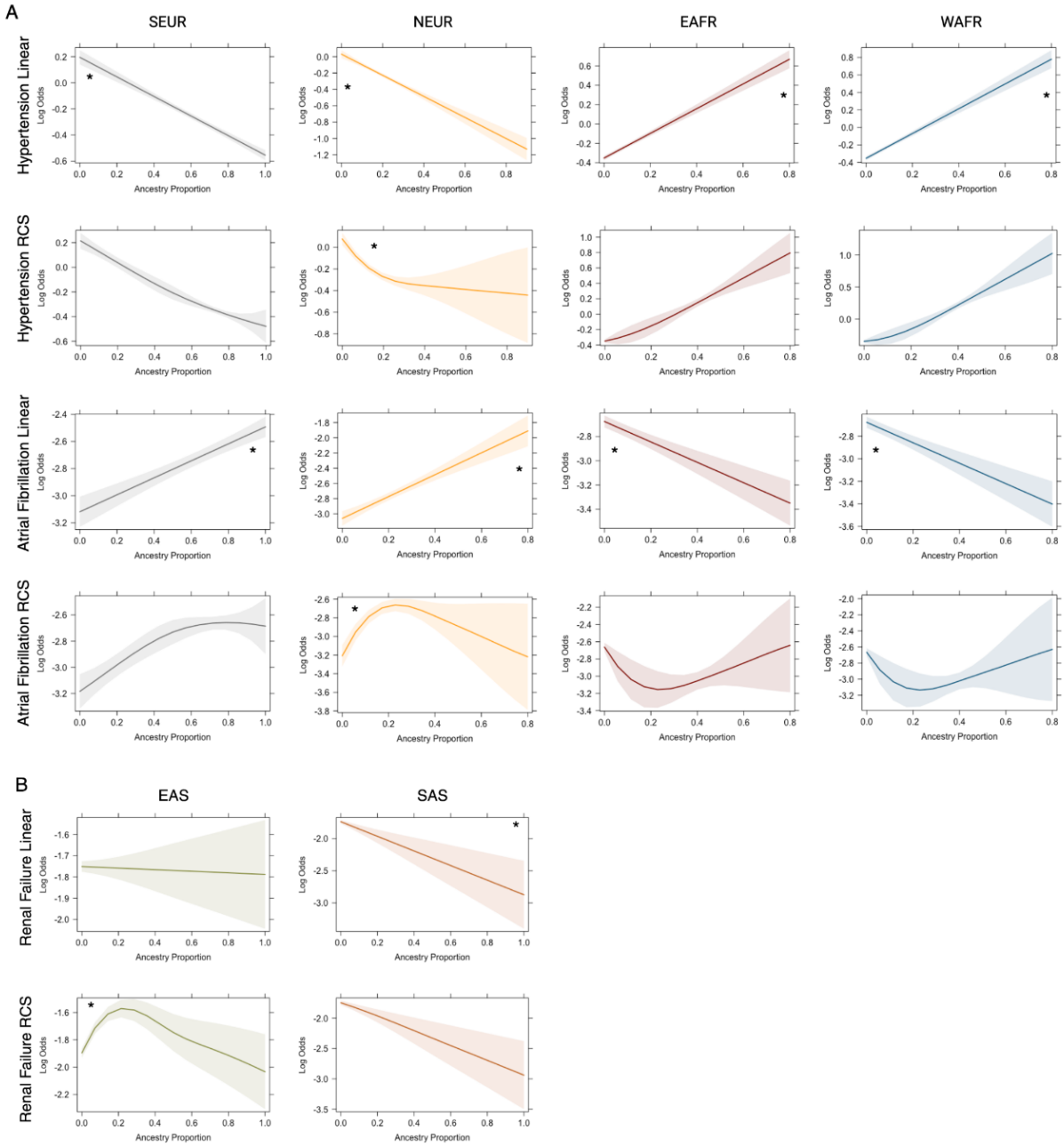

Figure 4. Linear modeling using logistic regression and restricted cubic spline (RCS) modeling of select phenotypes. A) Hypertension and atrial fibrillation risk models for SEUR, NEUR, EAFR, WAFR. B) Renal failure models for EAS and SAS. Log odds of outcome was graphed as a function of ancestry proportion adjusted for age, sex, and BMI. * = significant model. Created with BioRender.com.

but adding non-linear complexity did not significantly improve the model ($p = 0.41$). (Figure 4B) EAS was not significantly associated with RF when modeled linearly ($p = 0.78$). Modeling with the non-linear RCS revealed a significant relationship between RF and EAS ancestry proportion ($p < 1x10^{-4}$). (Figure 4B) For SAS, a 10 % increase in ancestry proportion had an OR of 0.15 (95% CI:

0.06 - 0.37) when modeled linearly. In RCS modeling, increasing from the 25th to 50th percentile of EAS ancestry proportion increases odds for RF by 1.09 (95% CI: 1.08 – 1.11) and increasing from the 50h to 75th percentile increases odds by 1.18 (95% CI: 1.15 – 1.21). (Table 3)

4. Discussion

We present an evaluation of the relationships between genetic ancestry proportions and the clinical phenome of the BioVU cohort. Our analyses revealed significantly enriched and de-enriched phecode categories for each ancestry group studied. We further evaluated the relationship between genetic ancestry and risk for HTN, AFib, and RF using linear and non-linear modeling methods.

4.1. *Relationships Between Ancestry and the Clinical Phenome*

Phecode categories that were de-enriched for PheWAS associations were 'injuries and poisonings' and 'musculoskeletal' for EAFR, WAFR, SEUR and EAFR, WAFR, SEUR, NEUR respectively. Both categories represent codes that are not conditions typically considered heritable. 'Injuries and poisonings' category comprises codes related to non-pathologic fractures, trauma injuries, and poisonings, all events caused by environment. Phecodes in musculoskeletal involve injuries or deformities of joints, bones, and muscles acquired from usage of the body. One specific phenotype to mention in this category is osteoporosis, where increasing NEUR and SEUR ancestry increased risk for codes relating to osteoporosis (phecodes 743, 743.1, 743.11) and spine curvature (737, 737.3), while the same codes have a protective effect with increasing EAFR and WAFR ancestry. Studies have shown increased bone mineral density and lower rates of osteoporosis associated in Black women compared to non-Hispanic White women.[18] Our genetic ancestry study findings support this previously observed epidemiological relationship.

In the 'neoplasm' category, many of the phecodes were in the risk direction for SEUR and NEUR ancestries and in the protective direction for WAFR and EAFR. The top significant neoplasm codes refer to skin cancer and other neoplasms of skin. The biological relationship between geographic ancestry and skin cancer has been well documented; populations in equatorial regions produce more melanin to protect against DNA damage from UV radiation while populations out towards the poles have evolved to produce less melanin due to less UV exposure.[19,20] It is possible that individuals of European genetic ancestry migrated away from the environments where they adapted to be at equilibrium and are now in new environments they are at disequilibrium with.[21]

One of the few exceptions to the pattern seen in 'neoplasms' were the phecodes 218 and 218.1, representing 'uterine leiomyoma' (or fibroids). Increasing EAFR and WAFR ancestries increases odds for fibroids while increasing SEUR and NEUR ancestries was protective against fibroids. This relationship pattern is consistent with previous epidemiology literature. Black women have been found to develop fibroids at younger ages, were more likely to have a clinical diagnosis, and to have had a hysterectomy from fibroids.[22] The overall odds of developing fibroids by age 50 were 2.9 times higher among Black women compared to White women.[22] Due to the significant racial disparities that exist for fibroids, it has been hypothesized that there is a genetic component to the condition, with a heritability estimate of ~30%.[23] Previous genetic studies have found African genetic ancestry proportion to be associated with fibroids diagnosis[12] and multiple fibroids.[24] Our study further supports the theory that African genetic ancestry may explain a portion of the risk for fibroids.

The pregnancy complication category was significantly enriched in NEUR, SEUR, EAS, and SAS. Within the category, significantly associated phecodes were all in the protective direction for NEUR and SEUR and in the risk direction for EAS and SAS. Racial disparities in maternal health outcomes have been well documented for White and Black women, with Black women having significantly higher adverse maternal outcomes compared to White women.[25] There have been many external factors posited for why Black women in US experience pregnancy complications and maternal mortality at much higher rates.[26] Trends in pregnancy complications for Asian women are less well-documented. A study of fertility treatment outcomes in Asian American women found decreased success of treatment in the forms of lower pregnancy rates and live births.[27] Using genetic ancestry proportions as a study variable may help to fill in some of the missing epidemiological gaps that still are pervasive in historically under-represented racial groups.

4.2. *Modeling Ancestry Proportion Linearly and non-Linearly*

From the phecodes grouped into the cardiac category, we saw a striking pattern. Several phecodes representing HTN and hypertensive disorders and consequences were found to be at increased risk in EAFR and WAFR and decreased risk in NEUR and SEUR. An opposite trend was seen for phecodes representing AFib and related codes; SEUR and NEUR were at increased risk while EAFR and WAFR were at decreased risk. This pattern follows what has been reported in literature.[28-30] Our study shows the trends we see for HTN and AFib are due in part to genetic ancestry.

While plenty of studies have focused on external causes and contributions to the higher prevalence of HTN in Black individuals,[31] it is known to be heritable.[32] A small (N = 998), previous study evaluated the relationship between African genetic ancestry proportion in self-identified Black individuals and hypertension and found the highest quartile of African genetic ancestry proportion had 8% higher prevalence than the lowest quartile.[33] Marden et al. used African genetic ancestry proportion to tease apart the contributions of genetics and socioeconomic status to HTN prevalence and found that their accounted socioeconomic factors only explained one-third of the difference in prevalence measured.[33] We as well sought to use genetic ancestry to determine its contribution to HTN disease risk as it helps to avoid confounders. The previous study and ours have both found African genetic ancestry to be associated with HTN risk and prevalence.

Within our evaluation of RF, we found linear modeling to be sufficient to model the relationship with SAS ancestry. EAS ancestry was not significantly associated with RF in PheWAS or when modeled individually linearly. Allowing for flexibility with non-linear RCS modeling revealed a relationship between EAS ancestry and RF. Only with the RCS model were we able to detect an OR of 1.18 (95% CI: 1.15-1.21) with an increase from 50th to 75th percentile of EAS ancestry. EAS was the ancestry group with the most skewed data density of the six groups, with the 3rd quartile ancestry proportion value being just 0.45% and one of our smallest sub-sample sizes with 760 self-identified individuals. The RCS model may have performed better due to being able to compensate for the skewness of data. Many wide-scale analyses perform only linear modeling which may not detect relationships, as seen for RF in EAS ancestry PheWAS. The risk trends for EAS and RF from RCS modeling have been reported previously in literature. Higher rates of end stage renal disease and increased risk of projected kidney failure have been observed in Far East, Southeast Asia, and Indian populations as compared to White populations.[34,35] The linear model for SAS and RCS model for EAS recapitulate these findings. Assuming linear relationships between genetics and disease may

cause associations to be missed, highlighting the need to consider non-linear modeling methods such as RCS.

4.3. *Considerations and Strengths*

While our study found some phenotype relationships that were consistent with epidemiology studies based on self-identified race, we did not evaluate the potential contribution of proportions of admixture on disease risk. On average, people had an admixture proportion of 0.33 (+/- 0.12) amongst the 6 super populations we determined with the more granular division of Southern and Northern European, Eastern and Western African, and East and South Asian. Within our cohort, those who self-identified as non-Hispanic Black had on average 78.6% African ancestry (EAFR + WAFR), 19.4% European ancestry (SEUR + NEUR), and 1.99% Asian ancestry (EAS + SAS). Those who self-identified as non-Hispanic White had on average 6.85% African ancestry, 98.0% European ancestry, and 1.26% Asian ancestry. Our study was limited in its ability to test more admixed populations where these methods may be more useful in identifying phenotypes associated with genetic ancestry.

We only used one ancestry as a predictor variable per model. Different geographic ancestries may interact differently, and this study does not account for various combinations of genetic ancestry proportions. Further investigation is needed to understand how the different genetic ancestries interact with each other and modify risk. A potential limitation of our study is the way in which some phenotypes may be diagnosed. Some phenotypes such as chronic kidney disease rely on algorithms that use self-reported race as a criterion to determine diagnosis, for example estimated glomerular filtration rate (eGFR) algorithms have historically used race as a coefficient in the equation for measuring eGFR levels which may bias diagnoses across racial and ethnic groups.[36]

In this study, we identified hundreds of traits in the clinical phenome that are associated with ancestry proportion. From our selected studies of enriched phecode categories and modeling of HTN and AFib, we observed many relationships between ancestry and phecodes that matched the epidemiology literature between self-identified race and traits. We used RCS to model a significant relationship between RF and EAS ancestry, one that was not originally identified from linear modeling. We highlighted a few phenotypes in this paper as an exploratory investigation into the potential of RCS modeling for ancestry proportion and disease risk.

Most traditional epidemiology literature notes the shortcomings of their studies revolve around using the societal construct of race, a lack of healthcare access for underrepresented groups and low-income individuals, and external environmental factors. Adjusting for race to better account for these factors like socioeconomic status or systemic discrimination in addition to using genetic ancestry proportion, which capture heritable contributions, may provide more comprehensive models. Future work controlling for genetic ancestry that demonstrates significant associations with race would highlight systemic factors affecting outcomes that are not captured by ancestry alone. In addition to utilizing genetic ancestry, we show how alternative modeling methods can be useful especially in a case of an underrepresented ancestry group where linear models may not be as successful to describe more complicated associations. Our study displays how genetic ancestry can be leveraged in furtherance of studying disease risk where traditional epidemiological studies have fallen short.

Acknowledgements

We would like to thank the participants of BioVU for their enrollments.

References

1. Yudell, M., Roberts, D., DeSalle, R., and Tishkoff, S. (2016). Taking race out of human genetics. Science (American Association for the Advancement of Science) *351*, 564-565. 10.1126/science.aac4951.
2. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine *383*, 874-882. 10.1056/NEJMms2004740.
3. Paulus, J.K., Wessler, B.S., Lundquist, C.M., and Kent, D.M. (2018). Effects of Race Are Rarely Included in Clinical Prediction Models for Cardiovascular Disease. J Gen Intern Med *33*, 1429-1430. 10.1007/s11606-018-4475-x.
4. Borrell, L.N., Elhawary, J.R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., Wu, A.H.B., Bibbins-Domingo, K., Rodríguez-Santana, J.R., Lenoir, M.A., Gavin, J.R., et al. (2021). Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. The New England journal of medicine *384*, 474-480. 10.1056/NEJMms2029562.
5. Kumar, R., Nguyen Ea Fau - Roth, L.A., Roth La Fau - Oh, S.S., Oh Ss Fau - Gignoux, C.R., Gignoux Cr Fau - Huntsman, S., Huntsman S Fau - Eng, C., Eng C Fau - Moreno-Estrada, A., Moreno-Estrada A Fau - Sandoval, K., Sandoval K Fau - Peñaloza-Espinosa, R.I., Peñaloza-Espinosa Ri Fau - López-López, M., et al. (2013). Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics (GALA II) study. J Allergy Clin Immunol *132*, 896-905. 10.1016/j.jaci.2013.02.046.
6. Neophytou, A.M., White, M.J., Oh, S.S., Thakur, N., Galanter, J.M., Nishimura, K.K., Pino-Yanes, M., Torgerson, D.G., Gignoux, C.R., Eng, C., et al. (2016). Air Pollution and Lung Function in Minority Youth with Asthma in the GALA II (Genes-Environments and Admixture in Latino Americans) and SAGE II (Study of African Americans, Asthma, Genes, and Environments) Studies. Am J Respir Crit Care Med *193*, 1271-1280. 10.1164/rccm.201508-1706OC.
7. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., and Fakim, Y.J. (2020). High-depth African genomes inform human migration and health. Nature *586*, 741-748.
8. Roden, D.M., Pulley Jm Fau - Basford, M.A., Basford Ma Fau - Bernard, G.R., Bernard Gr Fau - Clayton, E.W., Clayton Ew Fau - Balser, J.R., Balser Jr Fau - Masys, D.R., and Masys, D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther *84*, 362-369. 10.1038/clpt.2008.89.
9. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Neat Genet *48*, 1284-1287. 10.1038/ng.3656.
10. McCarthy, S.A.-O., Das, S.A.-O., Kretzschmar, W.A.-O., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Neat Genet *48*, 1279-1283. 10.1038/ng.3643.
11. Edwards, T.L., Giri, A., Hellwege, J.N., Hartmann, K.E., Stewart, E.A., Jeff, J.M., Bray, M.J., Pendergrass, S.A., Torstenson, E.S., Keaton, J.M., et al. (2019). A Trans-Ethnic Genome-Wide Association Study of Uterine Fibroids. Front Genet *10*, 511. 10.3389/fgene.2019.00511.
12. Keaton, J.M., Jasper, E.A., Hellwege, J.N., Jones, S.H., Torstenson, E.S., Edwards, T.L., and Velez Edwards, D.R. (2021). Evidence that geographic variation in genetic ancestry associates with uterine fibroids. Human genetics *140*, 1433-1440.
13. Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics *12*, 1-6.
14. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205-1210. 10.1093/bioinformatics/btq126.

15. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., and Denny, J.C. (2018). Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. BioRxiv, 462077.
16. Team, R.C. (2013). R: A language and environment for statistical computing.
17. Harrell, F.E. (2023). rms: Regression Modeling Strategies (CRAN).
18. Barrett-Connor, E., Siris, E.S., Wehren, L.E., Miller, P.D., Abbott, T.A., Berger, M.L., Santora, A.C., and Sherwood, L.M. (2005). Osteoporosis and fracture risk in women of different ethnic groups. Journal of bone and mineral research *20*, 185-194.
19. Agbai, O.N., Buster, K., Sanchez, M., Hernandez, C., Kundu, R.V., Chiu, M., Roberts, W.E., Draelos, Z.D., Bhushan, R., and Taylor, S.C. (2014). Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public. Journal of the American Academy of Dermatology *70*, 748-762.
20. Avise, J.C., and Ayala, F.J. (2010). Human skin pigmentation as an adaptation to UV radiation. In In the Light of Evolution: Volume IV: The Human Condition, (National Academies Press (US)).
21. Asadi, L.K., Khalili, A., and Wang, S.Q. (2023). The sociological basis of the skin cancer epidemic. International Journal of Dermatology *62*, 169-176.
22. Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D., and Schectman, J.M. (2003). High cumulative incidence of uterine leiomyoma in black and white women: Ultrasound evidence. American Journal of Obstetrics and Gynecology. 10.1067/mob.2003.99.
23. Bray, M.J., Davis, L.K., Torstenson, E.S., Jones, S.H., Edwards, T.L., and Velez Edwards, D.R. (2019). Estimating uterine fibroid SNP-based heritability in European American women with imaging-confirmed fibroids. Human heredity *84*, 73-81.
24. Bray, M.J., Edwards, T.L., Wellons, M.F., Jones, S.H., Hartmann, K.E., and Velez Edwards, D.R. (2017). Admixture mapping of uterine fibroid size and number in African American women. Fertility and Sterility *108*, 1034-1042.e1026. https://doi.org/10.1016/j.fertnstert.2017.09.018.
25. MacDorman, M.F., Thoma, M., Declcerq, E., and Howell, E.A. (2021). Racial and ethnic disparities in maternal mortality in the United States using enhanced vital records, 2016–2017. American journal of public health *111*, 1673-1681.
26. Saluja, B., and Bryant, Z. (2021). How implicit bias contributes to racial disparities in maternal morbidity and mortality in the United States. Journal of women's health *30*, 270-273.
27. Vu, M.H., Nguyen, A.-T.A., and Alur-Gupta, S. (2022). Asian Americans and infertility: genetic susceptibilities, sociocultural stigma, and access to care. F&S Reports *3*, 40-45.
28. Zilbermint, M., Hannah-Shmouni, F., and Stratakis, C.A. (2019). Genetics of hypertension in African Americans and others of African descent. International journal of molecular sciences *20*, 1081.
29. Keaton, J.M., Hellwege, J.N., Giri, A., Torstenson, E.S., Kovesdy, C.P., Sun, Y.V., Wilson, P.W., O'Donnell, C.J., Edwards, T.L., and Hung, A.M. (2021). Associations of biogeographic ancestry with hypertension traits. Journal of hypertension *39*, 633.
30. Marcus, G.M., Alonso, A., Peralta, C.A., Lettre, G., Vittinghoff, E., Lubitz, S.A., Fox, E.R., Levitzky, Y.S., Mehra, R., and Kerr, K.F. (2010). European ancestry as a risk factor for atrial fibrillation in African Americans. Circulation *122*, 2009-2015.
31. Usher, T., Gaskin, D.J., Bower, K., Rohde, C., and Thorpe Jr, R.J. (2018). Residential segregation and hypertension prevalence in black and white older adults. Journal of Applied Gerontology *37*, 177-202.
32. Kolifarhood, G., Daneshpour, M., Hadaegh, F., Sabour, S., Mozafar Saadati, H., Akbar Haghdoust, A., Akbarzadeh, M., Sedaghati-Khayat, B., and Khosravi, N. (2019). Heritability of blood pressure traits in diverse populations: a systematic review and meta-analysis. Journal of human hypertension *33*, 775-785.
33. Marden, J.R., Walter, S., Kaufman, J.S., and Glymour, M.M. (2016). African ancestry, social factors, and hypertension among non-Hispanic Blacks in the Health and Retirement Study. Biodemography and social biology *62*, 19-35.
34. Derose, S.F., Rutkowski, M.P., Crooks, P.W., Shi, J.M., Wang, J.Q., Kalantar-Zadeh, K., Kovesdy, C.P., Levin, N.W., and Jacobsen, S.J. (2013). Racial differences in estimated GFR decline, ESRD, and mortality in an integrated health system. American journal of kidney diseases *62*, 236-244.
35. Kataoka-Yahiro, M., Davis, J., Gandhi, K., Rhee, C.M., and Page, V. (2019). Asian Americans & chronic kidney disease in a nationally representative cohort. BMC nephrology *20*, 1-10.

36. Uppal, P., Golden, B.L., Panicker, A., Khan, O.A., and Burday, M.J. (2022). The Case Against Race-Based GFR. Delaware Journal of Public Health *8*, 86.