UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

UNIVERSITY OF THE WITWATERSRAND

FACULTY OF HEALTH SCIENCES

SCHOOL OF PUBLIC HEALTH

Title: Evaluating the accuracy of the CKD-EPI equations in estimating the glomerular filtration rate among adult Africans in Malawi, Uganda, and South Africa

Name: Esnath Tatenda Mudiwa (2392604)

Research submitted for the Degree: MSc in Epidemiology (Biostatistics)

Supervisor 1: Professor Eustasius Musenge[1]

Supervisor 2: Dr June Fabian[2]

[1] University of the Witwatersrand, School of Public Health,

[2] University of the Witwatersrand, Internal Medicine Department

October 2024

Declaration

I, Esnath Tatenda Mudiwa, declare that this Research Report is my own, unaided work. It is being submitted for the Degree of Master of Science degree in Epidemiology (Biostatistics) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Esnath Tatenda Mudiwa

2 October 2024

Dedication

This research report is dedicated to my family. Special mention goes to my mother, Ms. Rejoice Mlambo, and my brother, Dr. Benson-Munyaradzi Mudiwa, for their unwavering emotional and financial support throughout this journey and to my niece, Motsi Mudiwa. And to God who makes everything possible.

Abstract

Background

Chronic kidney disease (CKD) is a significant global health concern with a growing burden, affecting millions of individuals worldwide. CKD was ranked 18th amongst the highest non-communicable diseases cause of death worldwide. CKD is a global health issue, which reduces kidney function and increases cardiovascular risk. Accurate glomerular filtration rate (GFR) estimation is crucial for CKD management. The commonly used CKD-EPI (Chronic Kidney Disease-Epidemiology Collaboration) equations were primarily developed for Caucasian and African American populations. This study aimed to evaluate existing equations and develop new ones specific to the African population using data from Malawi, Uganda, and South Africa.

Methods

This study was a secondary analysis of data collected in three African countries collectively referred to as the ARK (African Research on Kidney Disease) Consortium. The study used data from 2433 participants, with plasma iohexol clearance as an indication of GFR (mGFR). Estimating equations were developed using serum cystatin C and or in combination with serum creatinine, mirroring the CKD-EPI equations by adopting a non-linear modelling approach. The study developed a predictive model using supervised machine learning techniques. Bland-Altman plots were used to assess linearity and agreement of the eGFR methods. Accuracy within 10% and 30% of mGFR, bias, and precision were assessed overall and by CKD stage.

Results

Analysis of 2433 participants from the three African countries revealed significant differences in mean measured glomerular filtration rate (mGFR) by country and sex. New serum cystatin C and creatinine-cystatin C-based equations for estimating GFR were developed, showing high accuracy ranging between (94-95%) and (93-95%), respectively for GFR ≥90 ml/min/1.73m^2. The equations however had lower precision (2.07 – 2.12) compared to existing ARKM (African Research on Kidney Disease Model) equations (2.36 – 2.37). Six machine learning (ML) classification models were evaluated, with Random Forest emerging as the top performer, followed by Logistic Regression. ML approaches demonstrated higher F1 score measures (89%-100%) than eGFR equations, accuracies ranging between 75% and 95% and less bias. Overall, it was concluded for this study that ML techniques provide better performing models in comparison to the existing and developed eGFR models. This was evident as AUC measures for all ML models were higher (93% - 100%) than the accuracy measures of the eGFR equations (75% - 95%).

Conclusion

The results highlight the value of cystatin C as a biomarker for improving GFR estimation and underscore the importance of population-specific GFR estimation tools for African populations. While no single method is perfect across all GFR levels, the findings demonstrate the potential of both refined eGFR equations and machine learning models in enhancing GFR estimation accuracy. However, confirmation in broader populations is needed, and regular monitoring and adaptation of ML models will be required to maintain predictive performance over time. These findings can inform efforts to improve GFR estimation and CKD evaluation in Africa.

Table of Contents

Acknowledgements

I extend my sincere gratitude to Professor Eustasius Musenge and Dr June Fabian whose supervision, guidance, patience, and support contributed immensely to the success of this research report.

I am also grateful to Dr Zvifadzo Matsena-Zingoni and Dr Chipo Mudzingwa for their guidance, thank you.

Thank you to the 2021 Biostatistics cohort for their support.

List of figures

List of tables

List of Equations

Definitions

ARK	African Research on Kidney Disease
ARKM 1	African Research on Kidney Disease Model 1 from primary study
ARKM 2	African Research on Kidney Disease Model 2 from primary study
ARKM 3	African Research on Kidney Disease Model 3 from primary study
AUC	Area under the Curve
CKD	Chronic Kidney Disease
CKD-EPI	Chronic Kidney Disease-Epidemiology Collaboration
CKD-EPI 2009	Existing CKD-EPI Creatinine based equation
CKD-EPI 2012	Existing CKD-EPI Cystatin C based equation
CKD-EPI 2021	Existing CKD-EPI (creatinine-Cystatin C) based equation
eGFR	estimated Glomerular Filtration Rate
GA	Genomic Ancestry
HIV	Human Immunodeficiency Virus
IDMS	Isotope Dilution Mass Spectrometry
KDIGO	Kidney Disease: Improving Global Outcome
LOOCV	Leave One Out Cross-Validation
mGFR	measured Glomerular Filtration Rate
MNHSRC	Malawi National Health Sciences Research Committee
MSE	Mean Square Error

PLHIV	People Living with HIV
ROC	Receiver-Operating Characteristic
uACR	urine albumin: creatinine ratio
UNCST	Uganda National Council for Science and Technology
UVRI-REC	Uganda Virus Research Institute, Research Ethics Committee
VSA	Validation Set Approach
Wits-HREC	University of Witwatersrand – Human Research Ethics Committee

1. Introduction

1.1 Background of study

Chronic kidney disease (CKD) is a condition that occurs when the kidneys are damaged and aren't working properly over time. It can lead to kidney failure, also known as end-stage kidney disease. CKD was ranked 18th amongst the highest non-communicable diseases cause of death worldwide in 2010, doubling as a cause of death between 1990 and 2017 (3).

Globally, estimates of CKD prevalence range between 8% and 16%, with obesity-induced diabetes as the most common cause in high-income settings. In 2017, there were approximately 697.5 million reported cases of CKD, marking a significant 29.3% rise in prevalence since 1990. In the same year, 1.2 million people died from CKD, indicating an increase in the all-age mortality rate of 41.5% from 1990 (2). CKD prevalence in Sub-Saharan Africa (SSA) is estimated at 13.9%, with hypertension, diabetes, and HIV as the most commonly identified causes (3). In sub-Saharan Africa, where over 20 million people live with HIV, CKD is a highly anticipated burden. (5).

1.1.1 Stages of Chronic Kidney Disease

According to the Kidney Disease: Improving Global Outcome (KDIGO) 2012 Clinical Practice Guideline for evaluating and managing CKD, it occurs when the estimated glomerular filtration rate (eGFR) is 60 ml/min per 1.73m^2; or (ii) a urinary albumin/creatinine ratio (UACR) of 3.0mg/mmol has been determined by subsequent testing after a minimum of three months. (1). As shown in figure 1, KDIGO further classifies kidney disease into stages of severity for eGFR and albuminuria, the so-called “all-stage” classification. For eGFR

(measured in ml/min per 1.73m^2) stage G1: ≥90; stage G2: 60-89; stage G3a: 45-59; stage G3b: 30-44; stage G4: 15-29; and stage 5: <15; where stage G1 is least severe and stage G5 is most severe requiring kidney replacement therapy to sustain life. For albuminuria, using uACR (mg/mmol) stage A1: <3.0; stage A2: 3.0-30; stage A3: > 30 (2).

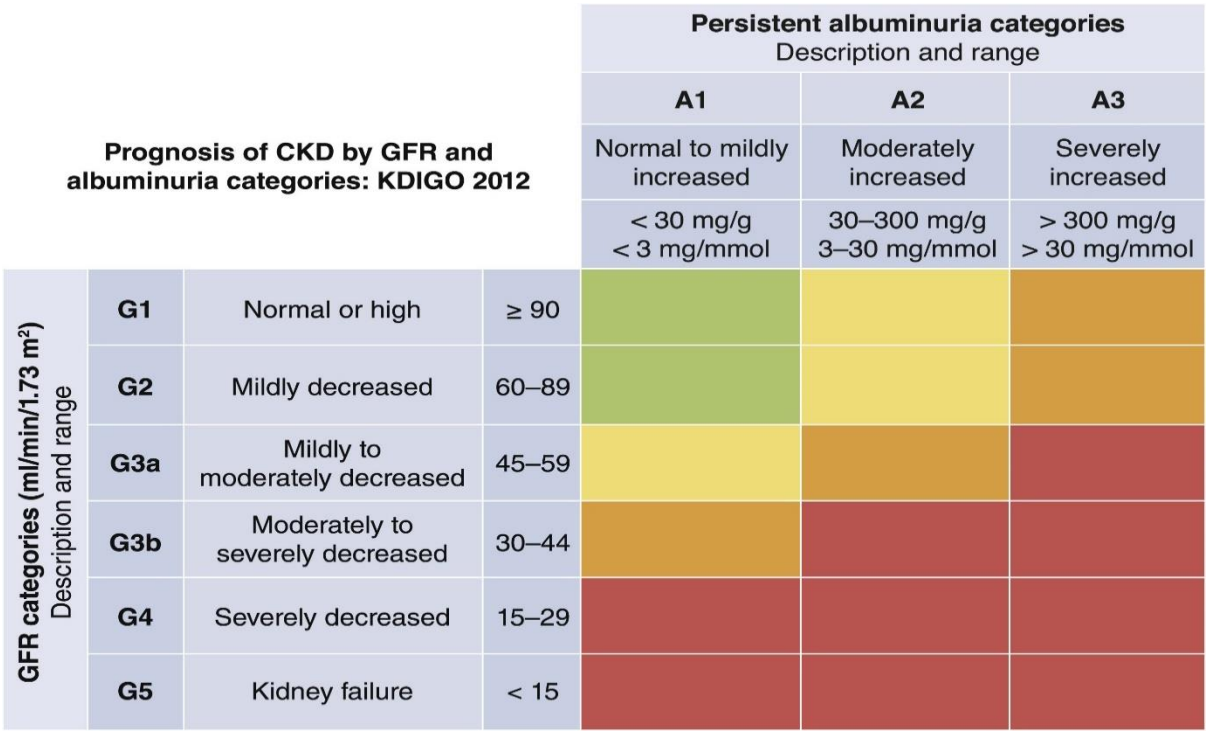


Figure 1: *Adapted from the 'KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease'. Risk categories for CKD according to the 2012 KDIGO. Green indicates low risk; Yellow indicates moderate risk; Orange indicates high risk; Red indicates very high risk (1)*

1.1.2 Measuring kidney function

Accurate estimates of the kidney function are essential in diagnosis and treatment of CKD. The glomerular filtration rate is a specific measure of how much blood is filtered through the glomeruli of the kidneys per minute, defined as the volume of plasma cleared of an ideal substance per unit of time. Therefore, measuring glomerular filtration rate (mGFR) is considered the best way to assess kidney function (6). Among the markers used to measure renal function are inulin, 51chromium ethylene diamine tetra acetic acid (51Cr-EDTA), and 99mTc-DTPA (diethylene triamine pentaacetate) marked with technetium. It is, however,

impractical to use these exogenous markers in routine clinical practice due to their limited accessibility and high costs. As a pragmatic but less accurate alternative, GFR is estimated (eGFR) using endogenous biomarkers such as serum creatinine (SCr) and serum Cystatin C (SCysC). SCr is the most commonly used marker in the clinical laboratory to estimate GFR (7).

As described, mGFR can be a clinically complicated and lengthy procedure; hence various formulae have been developed over the years to estimate GFR. Each eGFR equation is developed and validated in a discrete study sample using mGFR as the reference. An endogenous biomarker for kidney function such as serum creatinine or, serum Cystatin C concentration, and clinical variables that influence the concentration of the biomarker, for example, age and sex are used to model the estimating equation. Ideally, before using an eGFR equation, its performance should be validated in different settings and different populations. Unfortunately, this doesn't always occur, particularly in low and middle-income countries where costs and resources for validation are limited.

In the case of serum creatinine, a by-product of dietary protein digestion and the normal breakdown of muscle tissue, its serum concentration is affected by muscle mass, hence the adjustment for sex, as men have higher muscle mass than women (on average) (8). Because muscle mass varies between age, sex and level of fitness, it's concentration can differ between individuals with the same mGFR (9). In the development of the Modification of Diet in Renal Disease (MDRD) eGFR equation in the USA, it was observed that those of African American ethnicity had higher mGFR for a given creatinine, and an adjustment factor was modeled based on "race." The scientific rationale for adjusting eGFR based on self-identified ethnicity (often referred to as "race" in the medical literature) has been questioned (8). A recent study has modelled a new eGFR equation that incorporates Cystatin C and

creatinine but omits adjustments for African American ethnicity (10). For simplicity, hereafter, the adjustment for African American ethnicity in eGFR equations will be referred to as the "ethnicity factor".

Serum Cystatin C was first proposed as a GFR filtration biomarker in 1985. Unlike serum creatinine, a steady rate of serum cystatin is produced by nucleated cells and filtered by the kidneys. These characteristics make it a more accurate biomarker for kidney function as it is less affected by sex, age, body size, and muscle mass. Serum Cystatin C production varies less than creatinine production between individuals, and blood concentrations of Cystatin C are fairly similar between individuals who have the same mGFR (9). As such, the KDIGO 2012 CKD guidelines recommend use of cystatin-C based equations as a confirmatory test for adults having a creatinine based eGFR between 45 and 60mL/min/1.73m^2 (1). Cystatin C based equations were found to have better accuracy among those with better preserved kidney function, namely mGFR > 60mL/min/1.73m^2. This means that Cystatin C could be better for early detection of kidney impairment (11). However, the assay is relatively more expensive than creatinine and has limited availability in resource-limited settings.

The most commonly used eGFR equations were developed from predominantly European populations diagnosed with CKD in Europe and the USA (12). In Canada in 1976, Cockcroft and Gault developed the Cockcroft-Gault equation to estimate creatinine clearance based on plasma creatinine, age, height, weight, and sex (13). Later, in the USA, the MDRD study developed the 6-variable MDRD (6-v MDRD) equation for estimating GFR, followed by a 4 variable MDRD equation (4-v MDRD) based on age, sex, plasma creatinine, taking ethnicity into account (14). Following the MDRD equation, a representative US population database was used to develop the Chronic Kidney Disease – Epidemiology (CKD-EPI) equation in 2009. The CKD-EPI Cystatin C equation was developed in 2012.

KDIGO currently recommends the CKD-EPI equation as the reference equation for assessing CKD worldwide because it is more accurate and could replace the MDRD equation (8). However, these recommendations were not informed by validation studies in low and middle-income countries, and in continental African populations. Studies in SSA revealed that when the African American ethnicity adjusted coefficient is used for MDRD and CKD-EPI equations, eGFR tends to be overestimated, thus underdiagnosing those with CKD (15).

1.2 Literature Review

Since the KDIGO guidelines recommend use of the CKD-EPI (creatinine) equation for CKD screening (and it is the most widely used eGFR equation globally), the development and validation of this equation, with particular focus on its performance in African countries, will be the focus of this literature review. The study explored how machine learning techniques have been used in the development of eGFR equations.

1.2.1 Factors associated with Chronic Kidney Disease

CKD causes substantial global morbidity and increases cardiovascular and all-cause mortality (16). It is associated with diabetes, high blood pressure, heart disease, smoking, obesity, African American ethnicity, family history of kidney disease, atypical kidney structure, taking medications that can harm the kidneys, and older age. Irrespective of the cause, CKD has been linked to increased risk for morbidity and premature all-cause and cardiovascular mortality (17,18). Type 2 diabetes is regarded as the leading cause of CKD in the most developed countries. Subsequently, in low-income countries it is replacing communicable diseases as the leading cause of CKD (19).

1.2.2 Development of CKD-EPI equations

Efforts have been made globally to develop GFR estimating equations that can be used to clinically assess kidney functions. Existing equations had limited precision and underestimated mGFR at higher levels. The Chronic Kidney Disease Epidemiology Collaboration then developed a new estimating equation for GFR, (CKD-EPI). The research aimed to develop an equation that estimates GFR as the MDRD estimating equation at lower GFR's (≤ 60ml/min per 1.73 m^2;) and at the same time yielding better accuracy than the

MDRD estimating equation at higher GFR values (> 60ml/min per 1.73 m^2;). The CKD-EPI equations adjust for body surface area (BSA), sex, ethnicity, and age (8).

1.2.2.1 CKD-EPI (creatinine) equation

The CKD-EPI (creatinine) equations include an additional adjustment for the ethnicity factor. Using pooled samples from various studies in the USA and Europe, (N=8 254) the CKD-EPI (creatinine) equation was developed in a subset (N= 5 504) and validated internally in a separate subset (N=2 750) of adult participants. The equation was also then externally validated using samples from 16 other studies, (N= 3 896). GFR in the development and internal validation of the equation was measured using the method of iothalamate clearance (mGFR), with a mean of 68 mL/min/1.73m^2 and standard deviation of 40 mL/min/1.73m^2. Participants used for external validation of the equation included iothalamate and other filtration markers. The equation developed from this study was as follows:

$$\text{GFR} = 141\times\min\left(\frac{\text{Scr}}{\kappa},1\right)^{\alpha}\times\max\left(\frac{\text{Scr}}{\kappa},1\right)^{-1.209}\times 0.993^{Age}\ [\times 1.018 \text{ if female}]\ \_\ [\times 1.159 \text{ if black}] \quad (1)$$

Equation 1.1: *CKD-EPI (creatinine) equation*

where SCr is serum creatinine, κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males, min indicates the minimum of SCr/κ or 1, and max indicates the maximum of SCr/κ or 1.

In equation 1.1, the ethnicity adjustment for the ethnicity factor was 1.159. Implying that, for individuals of the same sex, age, and serum creatinine value, the estimated GFR will be

expected to be 15.9% higher for African Americans compared to those of other ethnicities. The race adjustment was validated for the US population.

1.2.2.2.CKD-EPI (Cystatin C) equation

The CKD-EPI (Cystatin C) estimating equation was developed using cross-sectional analyses from diverse populations totaling 5 352 participants from 13 studies. The participants were randomly divided into data sets for development (N=3 522) and internal validation (N= 1 830). External validation was done using participants (N= 1 119) from 5 other studies (20). The CKD-EPI (Cystatin C) equation was developed in 2012 as in Equation 1.2:

$$\mathrm{GFR} = 133 \times \min(\mathrm{SCys}/0.8, 1)^{-0.499} \times \max(\mathrm{SCys}/0.8, 1)^{-1.328} \times 0.996^{\mathrm{Age}} [\times 0.932 \text{ if female}] \quad (2)$$

***Equation 1.2**: CKD-EPI (Cystatin C) equation*

where SCys is serum Cystatin C (in mg/l), min indicates the minimum of SCys/0.8 or 1, and max indicates the maximum of SCysC/0.8 or 1.

1.2.2.3 CKD-EPI (creatinine-Cystatin C) equation

An eGFR equation based on the combination of creatinine and Cystatin C was developed within the same settings by CKD-EPI as in equation 1.3 (20):

$$\text{GFR} = 135 \times \min(\frac{\text{SCr}}{\text{k}}, 1)^{\alpha} \times \max(\frac{\text{SCr}}{\text{k}}, 1)^{-0.601} \times \min(\frac{\text{SCys}}{0.8}, 1)^{-0.375} \times \max(\frac{\text{SCys}}{0.8}, 1)^{-0.711}, \times 0.995^{\text{Age}} [\times 0.969 \text{ if female}] [\times 1.08 \text{ if black}] \quad (3)$$

Equation 1.3: *CKD-EPI (creatinine-Cystatin C) equation*

where SCr is serum creatinine (in mg/dl), SCys is serum Cystatin C (in mg/l), k is 0.7 for females and 0.9 for males, α is -0.248 for females and -0.207 for males, min(SCr/k, 1) indicates the minimum of SCr/k or 1, and max(SCr/k, 1) indicates the maximum of SCr/k or 1; min(SCysC/0.8, 1) indicates the minimum of SCysC/0.8 or 1 and max(SCysC/ 0.8, 1) indicates the maximum of SCysC/0.8 or 1.

The CKD-EPI equation (creatinine and Cystatin C) performed better than those that contained either creatinine or Cystatin C alone. Although the bias was relatively the same for all three equations.

1.2.3 Validation of the CKD-EPI equations evaluations

Small studies have been performed to validate the CKD-EPI equations in continental African populations. A measure of how many eGFR fall within $\pm$ 30% of mGFRs, termed P30, known as accuracy, was used in these studies to determine accuracy of the CKD-EPI equations without the African American ethnicity adjustment. A P30 $\geq$ 70% is considered accurate enough for clinical decision making.

In 2010, the accuracy of the CKD-EPI eGFR equations were evaluated accounting for the ethnicity factor in black South Africans. For this study, (n=100) participants were enrolled.

They were all black South Africans older than the age of 18. The patients in this study were either residents or outpatients at Chris Hani Baragwanath Hospital or being followed up at the renal unit of the hospital. All participants were either diagnosed with established CKD or identified as being at a higher risk of developing CKD by the assessing physician. The study developed SCr and SCysC based equations to estimate GFR among their study population. The SCysC based equation had a P30 of 84%, while the SCr based eGFR equation had a P30 of 70%. The CKD-EPI equation had a P30 of 72% (21).

The performance of CKD-EPI equations (with and without the ethnicity factor) in comparison with the MDRD equation (with and without the ethnicity factor) was examined in a cross-sectional study in Kinshasa, DRC on a cohort of (n=93) subjects, for both subjects with an mGFR ≥ 60 mL/min/1.73m^2 and mGFR < 60 mL/min/1.73m^2. Overall, the study found that accuracy (P30) improved when ethnicity coefficients were eliminated for all the equations. For example, for the CKD-EPI (creatinine) equation, accuracy (P30) with ethnicity factor (64.6%) was lower than without (77.7%); and for the combined CKD-EPI (creatinine and Cystatin C) equation, accuracy with the ethnicity factor (76.7%) was lower than without (79.2%). The accuracy of the CKD-EPI (Cystatin C) equation was 78.7% (22) (Appendix 1).

An analysis of cross-sectional data collected from Kinshasa's adults (n=210) and Ivory Coast (n=284) evaluated performance of eGFR equations using plasma clearance of iohexol as the reference standard for mGFR. It concluded that the ethnicity factor adjustment was not an improvement of the equations. Comparison of performance of CKD-EPI equations was done for both with and without the ethnicity factor for the SCr, SCysC and combined based equations. The CKD-EPI combined estimating equation without the adjustment of the ethnicity factor outshined all the other equations with a P30 of 79.2%. The CKD-EPI SCr

without the adjustment of the ethnicity factor had a P30 of 77.7 whilst the adjusted one had a P30 of 64.6%. The SCysC estimating equation had a P30 of 78.7%. (Appendix 2) (15).

It is estimated that more than two thirds of the world's population of people living with HIV (PLHIV) are in Africa. HIV-positive participants (n=100) were enrolled for a study in Kenya to determine the function of their kidney using the GFR estimating equations based on serum creatinine. Their accuracy was measured within 30% of the mGFR, measured by iohexol clearance. Similarly, the CKD-EPI estimating equation, unadjusted for the ethnicity factor for mGFR ≥ 90 mL/min/1.73/m2 had an accuracy (P30) of 92% as compared to 91% for the equation adjusted for the ethnicity factor. For mGFR <90 mL/min/1.73/m2, the CKD-EPI equation adjusted for the ethnicity factor performed had a P30 of 50% compared to the P30 of 64% for the equation without the adjustment. Findings of this study indicate that the CKD-EPI estimating equation is more accurate for higher ranges of mGFR as compared to the lower ranges as well that accuracy is better when the equation is unadjusted for the ethnicity factor (23).

An assessment of the CKD-EPI equation, with and without the ethnicity factor adjustment in adults in Kwa Zulu Natal, South Africa was done. The study had (n=287) participants for the analysis consisting of (n=107) Black African females, (n=81) Black African males, (n=56) Indian females, and (n=43) Indian male patients. None of the estimating equations showed an accuracy within 30% of the mGFR for at least 90% of the population. In the Black African participants, the CKD-EPI equation without the ethnicity factor correction performed best. The study results did not attain the 2002 KDOQI benchmark of 30% accuracy as only 11.1% – 17% of the individuals were correctly classified (24).

A comprehensive systematic review on how to estimate GFR in Africa concluded that with regards to estimating equations to mGFR, six studies demonstrated that African American coefficients in MDRD4-v and CKD-EPI equations were overestimating GFR (17).

This is a growing concern within public health as it hinders accurate reporting of CKD prevalence in African populations, hence poor preparedness, and management of CKD.

1.2.4 Use of machine learning techniques in detecting Chronic Kidney Disease

Chronic kidney disease (CKD) poses as a major emerging public health threat, contributing largely to morbidity and mortality rate increase for non-communicable diseases over the years (25). Globally, machine learning techniques have been used widely to improve detection and management of chronic kidney disease. This has been done with the aim of improving early detection of CKD which will improve and inform better management of the disease.

The paper by Debal and Sitote presents a machine learning approach to predict chronic kidney disease (CKD) stages using a dataset from Ethiopia (25). The authors compared several classifiers - Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) - with and without feature selection techniques like ANOVA and recursive feature elimination with cross-validation (RFECV). The dataset contained 1718 instances with 19 features like age, blood pressure, serum creatinine etc. Both binary classification (CKD or not) and multi-class classification (5 disease stages) were evaluated.

For binary classification, RF with RFECV gave the highest accuracy of 99.8% using 8 selected features. SVM with hyperparameter tuning also reached 99.83% accuracy. For multi-class classification, XGBoost gave the best accuracy of 82.56% with all features. RF with RFECV achieved 79% accuracy with 9 features. Important predictive features identified were serum creatinine, blood urea nitrogen, hemoglobin, and specific gravity. The authors conclude that the RF, SVM and XGBoost models show promising performance for assisting diagnosis of CKD stages. They recommend future work using unsupervised learning and deep learning methods.

A study was done where machine learning models were developed to predict CKD using a dataset of 400 patients. Their goal was to identify the optimal model and feature set for early CKD diagnosis (26). They pre-processed the dataset, applied feature selection techniques like recursive feature elimination and chi-square test, and built models using logistic regression, SVM, KNN, naïve Bayes and neural networks. The logistic regression model with chi-square feature selection achieved the highest accuracy of 98.75%, using 12 key features like blood, urea, creatinine, hemoglobin, age, hypertension status, etc. This study significantly advanced CKD prediction research by thoroughly evaluating machine learning algorithms and feature selection methods. The high accuracy achieved highlights the potential of data-driven approaches for improving early CKD diagnosis and preventing adverse outcomes.

1.3 Problem Statement

The CKD-EPI (creatinine) equation is the globally accepted eGFR equation for initial CKD screening. Despite being widely accepted, this equation is inaccurate in African populations. The inaccuracies are, in part, related to adjustment using the ethnicity factor, and even if

omitted, this equation still overestimates GFR in African populations. The choice of biomarker also influences GFR estimation. Cystatin C performs better than creatinine. For screening for CKD in patients with abnormal creatinine excretion, low muscle mass, and cardiac failure and liver failure, and those undergoing chemotherapy, the CKD-EPI (Cystatin C) equation has been a more accurate biomarker (27).

This study estimated GFR among African study participants using the CKD-EPI equations with and without adjusting for the ethnicity factor. We compared the performance of serum creatinine and serum Cystatin C in these equations. Conclusions from this research will be beneficial as they will improve CKD classification among African populations, which may assist with managing CKD.

1.4 Justification

We evaluated the performance of CKD-EPI equations using serum Cystatin C and serum creatinine in a sample of African participants compared to mGFR (using plasma excretion of iohexol) and model an equation that better estimates GFR in the sample of African participants. Improved accuracy of eGFR equations in African populations will aid in the diagnosis of those with CKD. Improved accuracy in the diagnosis of CKD aids the public health fraternity in policy development. It is of utmost importance that CKD management policies be informed by accurate information from African populations.

1.5 Research Question

The study sought to ascertain:

(i) Whether the ethnicity adjustment in the CKD-EPI equations estimates GFR accurately among adults from the three African countries, South Africa, Malawi, and Uganda; and

(ii) Whether the performance of the CKD-EPI (Cystatin C) equation is better than the CKD-EPI (creatinine) equation, irrespective of ethnicity correction among black adults from the three countries, South Africa, Malawi, and Uganda.

(iii) The diagnostic accuracy of serum Cystatin C alone compared to the combined use of serum Cystatin C and serum creatinine in classifying chronic kidney disease (CKD) status among study participants

1.6 Aim

To evaluate the accuracy of CKD-EPI eGFR equations based on serum creatinine and Cystatin C in estimating GFR among black adults from South Africa, Malawi, and Uganda.

1.7 Objectives

1. To describe the characteristics of those classified as having CKD among the ARK study participants.
2. To develop and validate a model that classifies study participants as either having CKD (CKD present) or not (CKD absent), using serum Cystatin C and a combination of both serum Cystatin C and serum creatinine.
3. To evaluate the performance of the glomerular filtration rate estimating equations against the reference mGFR using three metrics: accuracy, bias, precision.

2. Methods

2.1 Introduction

In this chapter, the study design, study site, study population, data collection and sampling methods are described for both the primary study and the current study. Study variables and data management for the study are also described in this chapter.

Descriptive statistics were used to describe the characteristics presented by patients in both the training data set and validation data sets. Logistics regression methods were used to develop the Cystatin-C and the Cystatin-C_Creatinine estimating equations. Resampling methods of the cross-validation approach were used to validate the performance equations.

Lastly, the CKD-EPI equations and the developed equation's performances were evaluated against the reference mGFR using three metrics: accuracy, bias, and precision.

2.2 Study design

This study was based on a secondary data analysis, derived from a cross-sectional study embedded from ongoing cohorts in South Africa, Uganda, and Malawi, collectively referred to as the African Research on Kidney Disease (ARK) Consortium.

The primary aim of the ARK Consortium was to establish the optimal approach to estimate GFR in Malawi, South Africa, and Uganda by way of evaluating the performance of existing CKD-EPI equations to mGFR (using iohexol), and if necessary, model a new equation with improved performance. The focus of the ARK Consortium was for creatinine-based equations, given limited availability of cystatin C testing.

In this study, the focus was on the utility of cystatin C and creatinine, or cystatin C alone and aimed to develop a new cystatin C based equation and a creatinine-cystatin C based equation accommodating between country differences. Furthermore, the study sought to

make use of machine learning techniques to model and validate the equations using the data, comparing its performance against existing estimating equations for this study population.

2.3 Study setting

The primary study sites were: The Kalunga district in rural south-western Uganda; Lilongwe (urban) and Karonga (rural) districts in Malawi; and the rural Bushbuckridge in Mpumalanga Province, South Africa.

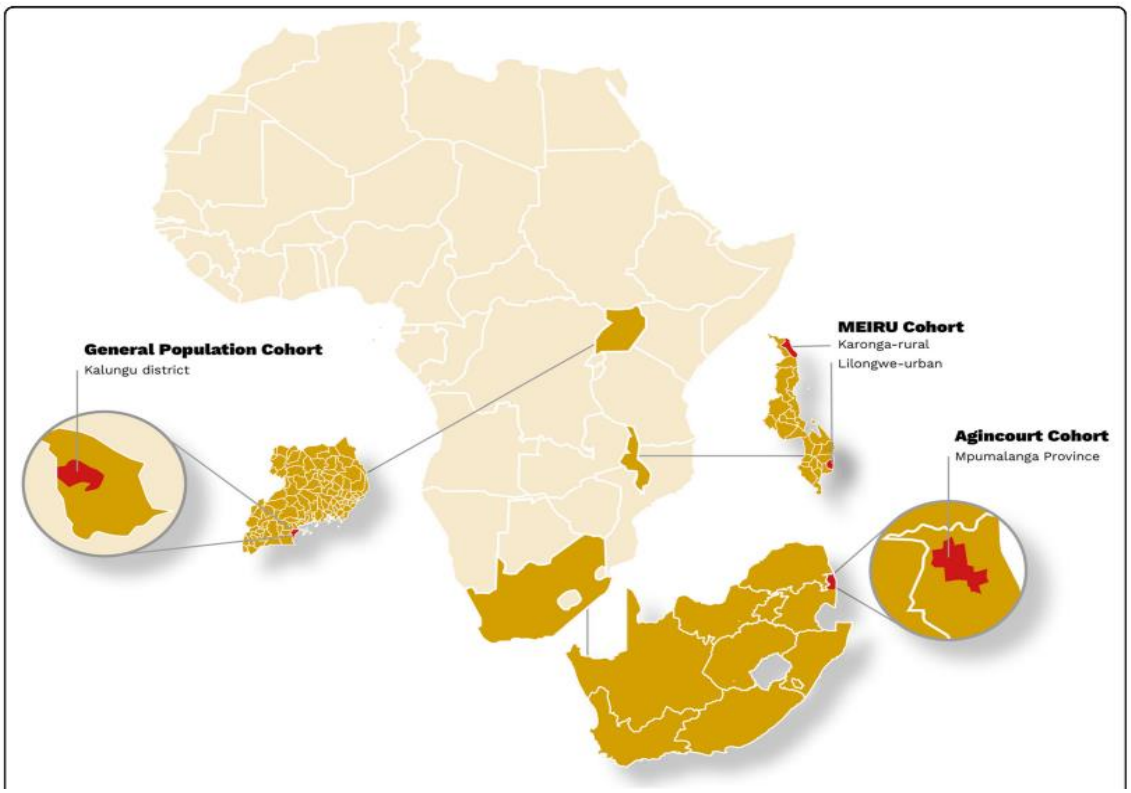


***Figure 2**: sourced from Open source, Map of Africa showing the ARK centres from Malawi, South Africa, and Uganda(28)*

2.3.1 Uganda

The study in Uganda was based on the 1986 General Population Cohort. The study area is situated in a rural sub-county of Kalunga district in rural south-western Uganda, approximately 120 km from Entebbe. Residents of 25 adjacent villages comprising of approximately 22 000 men and women made up the cohort. An established general clinic within the area provided medical care.

2.3.2 Malawi

Participants recruited for the study from Malawi were drawn from ongoing epidemiological studies conducted by The Malawi Epidemiology and Intervention Research Unit (MEIRU) which were based at two sites, Karonga district and Area 25 of Lilongwe. Karonga district is in rural northern Malawi and its population mainly comprises of informal traders, farmers, and fishermen. There are a variety of socioeconomic groups in Lilongwe's Area 25 including government employees, traders, and casual workers. A stratified sample from over 5000 individuals whose serum creatinine was measured in a more comprehensive chronic condition survey (n = 33,000) was selected for use in the study. An established chronic care Non-Communicable Disease (NCD) clinic provided medical care.

2.3.4 South Africa

In South Africa, the study was conducted in Agincourt, Bushbuckridge sub-district of the Mpumalanga province as part of the Medical Research Council/University of the Witwatersrand Rural Public Health and Health Transitions Research Unit. Based on an annual update of resident status and vital events, the health and socio-demographic surveillance system (HDSS) covers approximately 115,000 people in the area. Three district

hospitals, two health centers, and six clinics were part of the primary health care system that managed patients.

2.4 Study population

The ARK study defined the inclusion criteria as adults of age greater than 18 from the three countries, that were able to give written informed consent.

The exclusion criteria were defined as having a systolic blood pressure measurement greater than 180 mmHg and a diastolic blood pressure measurement greater than 110 mmHg, pregnant, breastfeeding, known allergy to iodine−containing substances, acute febrile illness and or having uncontrolled seizures (within the last 12 months).

2.5 Sampling methods

Participants recruitment for the ARK study took place in all the three participating countries. For Uganda, the first recruitment phase consisted of serum creatinine measurements from stored sera. This phase included the collection of data from community surveys and house-to-house mobilizations conducted by village-based hubs to select 5,979 participants. The study team explained to the selected participants the study details, and individual written consent was sought from the participants. Completion of a questionnaire was done by participants who consented to participate in the study. The second phase comprised selecting a sub-sample of participants stratified by eGFR and sex for participation in the iohexol measured GFR study.

In Malawi, the first recruitment phase used creatinine assays obtained from a prior survey. As is in the Uganda recruitment plan, the study team explained the details of the study to the participants. Informed consent was sought, and then the participants were subjected to a questionnaire that collected their demographic information and known history of risk factors that could cause CKD. The second phase comprised selecting a sub-sample of participants stratified by eGFR and sex for participation in the iohexol measured GFR study (17).

For South Africa, 2 021 participants from a stratified random sample were used for baseline creatinine measures. As in the other countries, the second phase comprised selecting a sub-sample of participants stratified by eGFR and sex for participation in the iohexol measured GFR study (17).

For the prospective iohexol mGFR study, each partner study site aimed to recruit 1000 participants so that the pooled sample would contain 3000 participants for the final analysis.

2.6 Data collection for the iohexol mGFR study

Primary data was collected from consenting adults of age greater than 18 years from the three countries. Measurement of GFR was obtained using iohexol plasma excretion. Methods and design of the ARK study have been published for ease of reference (17).

Data for the study were collected from the clinic visit: age, sex, height, weight, temperature, blood pressure, baseline serum creatinine and serum Cystatin C. Iohexol plasma excretion was measured with after initial intravenous bolus administration of a prefilled syringe of iohexol (dose iohexol calculated based on weight) and serial measures of plasma iohexol

sampling at 120-, 180-, and 240-min intervals. Figure 3 illustrates the study procedures and data sets used for analyses for this study.

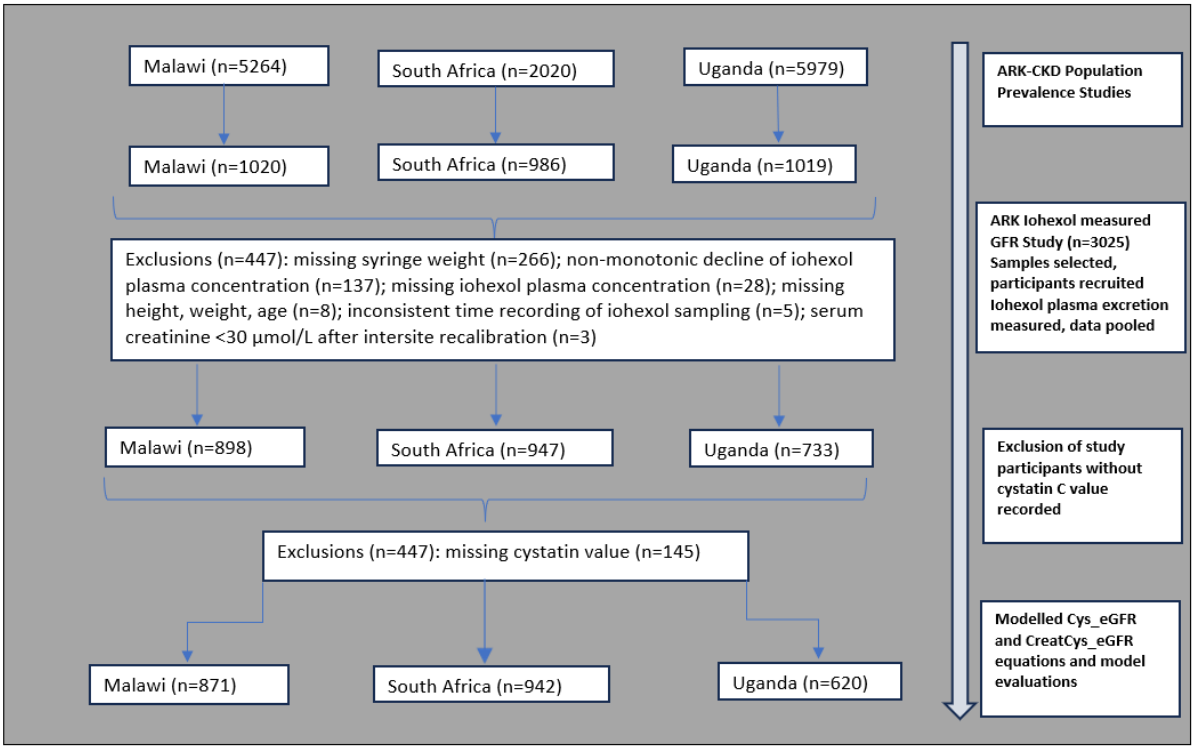


Figure 3: *sourced from CC commons open access license : Study procedures and datasets used for the Iohexol-Measured GFR Study (28)*

2.7 Data management

Data cleaning procedures were performed on the final pooled ARK iohexol mGFR data set.

This included extraction of data on variables that we will use for our study. A new variable, mGFR, was developed, from finding intercept of the iohexol plasma concentration at the three intervals, 120, 180, and 240 minutes. This new variable was then used for assessment of accuracy of the eGFR models.

The study intended to develop an estimating equation using the variable Cystatin C and 6%, (n=145), of our dataset was missing the Cystatin C value. Assumption to draw meaningful conclusions from 94%, (n= 2433) of the dataset which was complete was made, hence the

decision to drop the missing values. Therefore, of the 2 548 participants, Analysis was done on data for 2 433 participants who had complete data.

2.8 Statistical Analysis Plan

Objective 1: *To describe the characteristics of those classified as having CKD from those without CKD among the ARK study participants*

Descriptive statistics were used to describe the characteristics presented by the study participants in both the training set and the validation set. Normally distributed continuous variables were described using their means mean and standard deviations whilst categorical variables were described using their frequencies and proportions.

Objective 2: *To develop and validate a model that classifies study participants as either having CKD (CKD present) or not (CKD absent), using serum Cystatin C, using machine learning techniques of resampling methods*.

The hypothesis for this study was that the eGFR model developed will better estimate GFR among African populations as compared to previously developed eGFR equations and that the equation with both creatinine and Cystatin C will better estimate eGFR as compared to that with either of the two biomarkers.

2.8.1 Calculation of measured GFR using the Slope Intercept Method

In order to perform analysis, I calculated the measured GFR (mGFR) using the slope intercept method for iohexol plasma clearance based on three blood samples drawn at 120mins 180mins and 240mins time intervals, using the exact times at which the blood sample was drawn (29). Area under the plasma-concentration curve is needed for this calculation.

The decrease in plasma concentration evidenced by periodic measurement from the blood samples drawn at time intervals, (t), follows an exponential decay. This decrease in plasma concentration is denoted as, $P(t)$. Assuming that the initial plasma concentration after injection is denote by (P_0), and is decreasing exponentially at time, (t), at a constant rate, (L):

$$P(t) = P_0 \cdot e^{-L.t} \quad (4)$$

Equation 2.1: *Decrease in Plasma concentration*

The area under the curve will therefore be represented as:

$$A = \frac{P_0}{L} \quad (5)$$

Equation 2.2: *Area under the Curve*

Use of the last three blood samples was utilized to achieve the best quality and reliability of results. AUC correction and BSA correction (normalized to the “standard man”) were effected to the final mGFR calculations obtained (30). The mGFR will be used as the reference standard for assessing the accuracy of the eGFR equations.

2.8.1.1 Calculation of volume injected dose of iohexol

Of essence in the calculation of the mGFR is the applied dose of the iohexol plasma concentration. I calculated the volume of iohexol injected by converting the weight of the syringe from grams to mL by dividing it by the gravity for iohexol, 1.406, as shown in equation 2.3:

$$mL\ injected = \frac{iohexol\ weight}{1.406} \quad (6)$$

Equation 2.3: *Volume of iohexol injected*

The iohexol solution injected contained 755mg of iohexol per mL, therefore calculation of the total dose administered was calculated as follows:

$$Dose = mL\ injected * 755 \quad (7)$$

Equation 2.4: *Total iohexol dose administered*

Calculation of the volume of distribution (Vd) which is defined as proportionality constant that relates the total amount of drug in the body to the plasma concentration of the drug at a given time,(31) was done as follows:

$$Vd = \frac{Dose}{expconstant} \quad (8)$$

Equation 2.5: *Volume of distribution*

2.8.1.2 Calculation of the measured GFR

Having gathered all necessary variables, mGFR was calculated based on the Slope Intercept method, ($mGFR_{SI}$), as follows:

$$mGFR_{SI} = -(slope) * Vd * 1000 \quad (9)$$

Equation 2.6: *mGFR calculation using the Slope-Intercept Method*

2.8.1.3 Adjusting for Body Surface Area (BSA)

After having obtained the measured GFR, based on the slope-intercept method, the equation was adjusted for Body Surface Area (BSA). This correction is done because GFR varies according to BSA. mGFR will therefore be presented as a pure value, corrected for BSA, and normalized to the "standard man" (body surface $1.73m^2$), adjusted for (BSA) (mL/min/1.73m^2) using the Dubois Dubois formula as in equation 2.7:

$$mGFR_{BSA} = 0.00718 * Weight^{0.425} * Height^{0.725} \quad (10)$$

***Equation 2.7**: mGFR calculation equation adjusting for Body Surface Area*

2.8.2 Model development

The focus of this study was to develop a glomerular filtration rate estimating equation based on Cystatin C which will be used to compare performance against the existing CKD_EPI equations and the ARK models. Model development was performed using the log scale of $mGFR_{BSA}$ as evidence shows that derivation of eGFR equations has multiplicative error (28), using the $CKD - EPI_{Cys}(2012)$ in development of this estimating equation.

For model developmensimple random sampling was performed on the data set in Stata17, to randomly divide the sample into 2 data sets using the splitsample command, using the ratio (0.8: 0.2) for sampling the data. The first dataset with (N = 1946) was treated as the training data set, whilst the remaining observations, (N = 487) made up the testing data set.

Calculation of eGFR for use in this development was done using non-linear least squares estimation which was used to estimate glomerular filtration rate which was in turn used to categorize study participants into either having CKD (CKD present) or not having CKD (CKD absent), where having CKD was described as the probability of having a GFR reading of >60mL/min/1.73.

2.8.3 Using machine learning techniques to detect CKD

This study focused on CKD prediction using unsupervised machine learning models based on the ARK study data set. With binary outcome classes, 0 and 1, with 1 representing CKD present and 0 representing CKD absent. The data was pre-processed, missing data was handled, followed by feature selection using ranking method and correlation analysis to come up with the features to predict mGFR and came up with the final data set for use in our models.

Based on literature, Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Neural Networks, Naive Bayes Classifiers, and K-Nearest Neighbors were used to predict CKD. LR was selected because it is a simple but effective classification algorithm that predicts a binary target variable, (CKD present or not). It also works well with a mix of numerical and categorical features. RF is an ensemble method that combines multiple decision trees. Handles nonlinear relationships well. Also provides feature importance metrics. SVM is a powerful algorithm that handles complex datasets with many features, and is effective at classification tasks like detecting CKD. Naive Bayes Classifiers are a probabilistic model making strong independence assumptions. Very fast to train and useful as a baseline for comparison.

Objective 3: *To evaluate the performance of the new models and the existing CKD-EPI equations against the reference mGFR using three metrics: accuracy, bias, and precision.*

The Bland-Altman (B&A) plots, a graphical technique was used to evaluate agreement between two quantitative methods of measurement, in our case eGFR equations and mGFR. In Bland-Altman plot analysis, a bias between the mean differences is evaluated, and an agreement interval is estimated, within which 95% of the differences between the second and first methods fall. (32). For performance evaluation of the estimating equations, bias, precision, and accuracy metrics were assessed on the internal validation dataset (N=1217). These metrics were assessed on five eGFR equations, namely the CKD-EPI_Creat (2009), CKD-EPI_Cys (2012), CKD-EPI_CreatCys (2012), ARK models equations, and the newly developed cystatin C based equation and the Creatinine-Cystatin C based equation.

Bias

In all estimating equations, bias was calculated as the median difference between measured and estimated GFR. If θˆ = T(X) is an estimator of θ, then the bias of θˆ is the difference between its expectation and the 'true' value: i.e.

$$Bias(\hat{\theta}) = E_{\theta}(\hat{\theta}) - \theta \quad (11)$$

Equation 2.8: *Bias calculation*

Where $\hat{\theta}$ is eGFR and θ is mGFR.

Precision

Precision was measured by the interquartile range of the difference between mGFR and eGFR. True values and their estimates are measured by precision, which is a measure of how close they are. It is therefore calculated as follows:

$$Precision = \frac{True\ positives}{True\ positives + False\ positives} \quad (12)$$

Equation 2.9: *Precision calculation*

Accuracy

As a measure of accuracy, the root-mean-square error (RMSE) was calculated as well as the percentage of estimates that differed from the mGFR by more than 30% (1-P30) or more than 20% (1-P20). Accuracy is described as the measure of how close a given set of measurements is to their true values. To calculate the accuracy, the error rate was calculated as follows:

$$Error\ rate = \frac{|eGFR - mGFR|}{mGFR} * 100 \quad (13)$$

Equation 2.10: *Error calculation*

The accuracy was then calculated as follows:

$$Accuracy = 100\% - Error\ rate \quad (14)$$

Equation 2.11: *Accuracy calculation*

2.9 Ethical considerations

Ethical approval for this study was granted by the University of Witwatersrand Human Research Ethics Committee (#M220506).

Ethics approval for the primary study was sought from each relevant institution for each partner country in the ARK Consortium, namely: Uganda Virus Research Institute, Research

Ethics Committee (UVRI-REC-#HS 1978) the Uganda National Council for Science and Technology (UNCST-#SS 4283), the Malawi National Health Sciences Research Committee (#1072) and the University of Witwatersrand Human Research Ethics Committee (#M160938)

Written informed consent was sought from each individual before they participated in the study at each study site.

Participants diagnosed with advanced CKD were referred to get appropriate medical attention under the direction of the study physicians in each center.

3. Results

3.1 Descriptive and Exploratory Statistics

Table 1 shows the characteristics of the participants recruited into the ARK study from Malawi, South Africa, and Uganda. Malawi participants both males and females were older when compared to ages of participants in other two countries. The middle age group, comprising individuals between 40 and 60 years old, represented 46% of the participants. Female South African participants had the highest BMI readings across the participants as well as body surface area. However, eGFR measurements using the CKD-EPI (cystatin) equation, South African participants, both males and females had the highest mean eGFR.

Table 1: *Socio demographic and health variables for enrolled study participants by sex and country*

Characteristics of Study Participants by sex and by country (N=2433)							
	Malawi		South Africa		Uganda		Overall
Variable	Females (N = 457)	Males (N = 414)	Females (N = 631)	Males (N = 311)	Females (N = 347)	Males (N = 273)	Overall (N = 2433)
Age (years), Mean (SD)	53.09 (13.61)	52.75 (16.22)	45.57 (14.40)	44.50 (16.2)	50.76 (14.72)	51.63 (15.13)	49.49 (15.34)
Height (cms), Mean (SD)	155.72 (6.00)	165.33 (6.85)	161.54 (5.89)	172.60 (7.144)	154.94 (6.60)	164.26 (6.45)	161.86 (8.51)
Weight (kgs), Mean (SD)	64.96 (15.48)	63.14 (11.83)	78. 72 (17.38)	74.64 (16.71)	57.42 (12.33)	56.76 (9.80)	67.46 (16.91)
Serum Creatinine, Mean (SD)	74.33 (17.86)	91.17 (26.11)	63.37 (13.91)	82.73 (20.16)	65.87 (18.50)	78.83 (29.07)	74.73 (22.83)
Serum Cystatin C, Mean (SD)	1.07 (0.31)	1.11 (0.28)	0.99 (0.27)	1.01 (0.28)	0.99 (0.21)	1.02 (0.28)	1.03 (0.28)
Body Mass Index, Mean (SD)	26.72 (5.83)	23.06 (3.85)	30.15 (6.41)	25.01 (5.08)	23.85 (4.54)	21.00 (3.15)	25.72 (6.04)
*Body Surface Area, Mean (SD)	1.63 (0.18)	1.69 (0.16)	1.82 (0.19)	1.87 (0.20)	1.54 (0.16)	1.61 (0.14)	1.71 (0.21)
mGFR (mL/min/1.73m^2), Mean (SD)	86.74 (26.69)	90.19 (26.42)	93.02 (34.11)	95.30 (35.83)	99.78 (34.25)	114.13 (44.17)	94.98 (34.13)
**eGFR (mL/min/1.73m^2), Mean (SD)	73.29 (22.04)	74.75 (21.90)	83.24 (23.69)	88.04 (26.42)	80.24 (19.79)	84.09 (21.44)	80.21 (23.24)
*Body surface area calculated using the Dubois Dubois formula. **Estimated GFR calculated using the CKD-EPI (cystatin C) equation 2019							

Country Effect

***Table 2**: Post-hoc analyses for comparing means of mGFR between countries*

	SOUTH AFRICA		UGANDA	
BONFERRONNI	**P-value**	**DIFF**	**P-value**	**DIFF**
MALAWI	0.002	5.39031	<0.001	17.7178
SOUTH AFRICA			<0.001	12.3275

SCHEFFE	**P-value**	**DIFF**	**P-value**	**DIFF**
MALAWI	0.003	5.39031	<0.001	17.7178
SOUTH AFRICA			<0.001	12.3275

SIDAK	**P-value**	**DIFF**	**P-value**	**DIFF**
MALAWI	0.002	5.39031	<0.001	17.7178
SOUTH AFRICA			<0.001	12.3275

Post-hoc analyses were performed to compare the means of mGFR between the three countries. The analysis of variance indicated differences existed between the three countries ($F=51.87$, $p<0.001$). Figure 4 shows the same difference. The post-hoc tests (Bonferroni, Scheffe, Sidak) as illustrated in Table 2; all show that the mean mGFR for Malawi is significantly lower than South Africa ($p=0.002$). The mean mGFR for Malawi is significantly lower than Uganda ($p<0.001$) and the mean mGFR for South Africa is significantly lower than Uganda ($p<0.001$). All differences between pairs of means are statistically significant at $p<0.05$ after adjusting for multiple comparisons. This aligns with the marginal means from the ANOVA model and provides confirmation that mGFR differs significantly between each pair of countries.

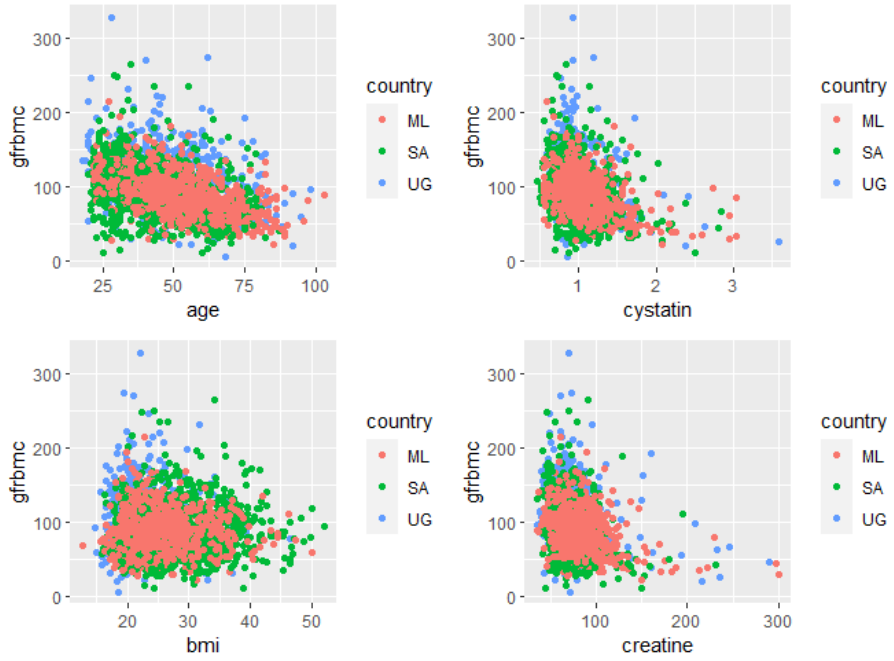


Figure 4: *Country Differences in mGFR Relationships with Cystatin, Creatinine, Age, and BMI*

Sex Effect

Table 3: *Post-hoc analyses for comparing means of mGFR between sex*

	FEMALES	
BONFERRONNI	**P-value**	**DIFF**
MALES	<0.001	5.67512

SCHEFFE	**P-value**	**DIFF**
MALES	<0.001	5.67512

SIDAK	**P-value**	**DIFF**
MALES	<0.001	5.67512

Table 3 shows the post-hoc tests (Bonferroni, Scheffe, Sidak) results, illustrating the mean difference in mGFR between sex. The ANOVA comparing mGFR between the two sex groups is statistically significant (F=16.38, p=0.0001). All the post-hoc tests (Bonferroni, Scheffe, Sidak) indicate that the mean mGFR for males is significantly higher than that of females, with a mean difference of 5.68 (p<0.001). This confirms that males have

significantly higher mGFR compared to females. As such, figure 5 illustrates the effect one's sex has on distribution of Cystatin C, serum creatinine, age and BMI. The distribution between males and females is somewhat even across the increasing value of Cystatin C. There is however a notable concentration between 0.5mg/dl and 1.5mg/dl of both males and females. This means that sex has no impact on the amount of Cystatin C concentration. Serum creatine, however, displays a different picture, with an increase in the number of males as serum creatinine increases whilst most females are seen to be concentrating between 50$\mu mol/L$ and 100$\mu mol/L$.

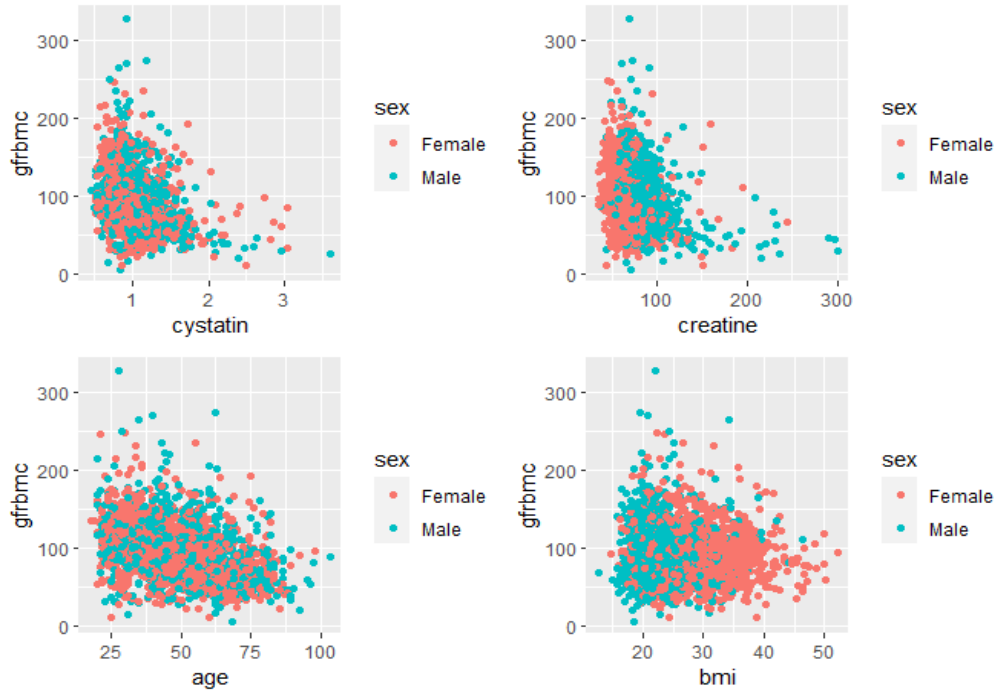


Figure 5: *Sex Differences in mGFR Relationships with Cystatin, Creatinine, Age, and BMI*

3.2 Model development

The study developed two sets of glomerular filtration rate estimating equations using our ARK Study data set. Development of the estimating equations was based on mirroring the CKD-EPI equations. Two sets of equations were developed with one set being of equations

based on serum Cystatin C only and the other set based on both markers, serum Cystatin C and serum creatinine.

In each set, two separate equations were developed, one for females and another set for males. The study did another non-linear least square estimation set of equations adjusting for the age parameter to see if it improves the models or not.

3.2.1 Cystatin C based estimating equations

As discussed earlier, the study mirrored the CKD-EPIcys (2012). It used the non-linear least squares estimation method in the development of these equations.

For the females' equations an eGFR non-linear estimating function was fit as shown in equation 3.1:

$$Y_i = 133 \times \min(\frac{\rho_i}{0.8}, 1)^{-\beta_0} \times \max(\frac{\rho_i}{0.8}, 1)^{-\beta_1} \times 0.996^{\sigma_i} [\times \beta_2 \text{ if female}] + E_i \quad (15)$$

Equation 3.1: *Developed non-linear eGFR estimating function for females based on Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ρ_i and σ_i are the participants serum Cystatin C level and age respectively. Because β_2 multiplying the whole equation, zero is not a feasible starting point as it would turn the equation to zero, with the starting point for β_2 as 0.5.

Another non-linear estimation function was developed adjusting for age to check the effect in improvement of the model as shown in equation 3.2:

$$Y_i = 133 \times \min(\frac{\rho_i}{0.8}, 1)^{-\beta_0} \times \max(\frac{\rho_i}{0.8}, 1)^{-\beta_1} \times (\beta_2)^{\sigma_i} [\times \beta_3 \text{ if female}] + E_i \quad (16)$$

Equation 3.2: *Developed non-linear eGFR estimating function for females adjusted for Age based on Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i and ρ_i and σ_i are the participants serum Cystatin C level and age respectively. The starting point for the age parameter, β_2, was set to 0.8. Because β_3 multiplying the whole equation, zero is not a feasible starting point as it would turn the equation to zero, hence the starting point for β_3 was set to 0.5.

Table 4 shows the list of estimating equations based on serum cystatin for females.

Table 4: *Female Cystatin C based estimating Equations*

Female Cystatin estimating equation without Age adjustment
133 * $(\min(\frac{\text{cystatin}}{0.8}, 1)^{-0.411})$ * $(\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.481})$ * (0.996^{Age}) * 0.932
Female Cystatin estimating equation with Age adjustment
133 * $(\min(\frac{\text{cystatin}}{0.8}, 1)^{-0.401})$ * $(\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.374})$ * (0.993^{Age}) * 1.061

For the males' equations based on serum Cystatin C, an eGFR non-linear estimating function was fitted, as shown in equation 3.3:

$$Y_i = 133 \times \min(\frac{\rho_i}{0.8}, 1)^{-\beta_0} \times \max(\frac{\rho_i}{0.8}, 1)^{-\beta_1} \times 0.996^{\sigma_i} + E_i \quad (17)$$

Equation 3.3: *Developed non-linear eGFR estimating function for males based on Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ρ_i and σ_i are the participants serum Cystatin C level and age respectively.

As done with the female participants, the study further developed another non-linear estimation function based on serum cystatin for male participants adjusting for age to check the effect in improvement of the model as shown in equation 3.4:

$$Y_i = 133 \times \min(\frac{\rho_i}{0.8}, 1)^{-\beta_0} \times \max(\frac{\rho_i}{0.8}, 1)^{-\beta_1} \times {\beta_2}^{\sigma_i} + E_i \quad (18)$$

Equation 3.4: *Developed non-linear eGFR estimating function for males adjusted for Age based on Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ρ_i and σ_i are the participants serum Cystatin C level and age respectively. The starting set point for $\beta_2 = 0.8$ for our non-linear estimating equation.

Table 5 shows the set of equations for males based on serum cystatin. Both the equation further adjusted for the age parameter and the equation without the age adjustment.

Table 5: *Male Cystatin C based estimating equations*

Male Cystatin estimating equation without Age adjustment
$133 * (\min(\frac{\text{cystatin}}{0.8}, 1)^{0.206}) * (\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.476}) * (0.996^{\text{Age}})$
Male Cystatin estimating equation with Age adjustment
$133 * (\min(\frac{\text{cystatin}}{0.8}, 1)^{0.344}) * (\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.577}) * (0.997^{\text{Age}})$

A comparison on whether the age adjustment had an effect of improving the model was done with the positive effect largely visible amongst females as compared to males (Appendix 3).

Overall, the age adjustment had a positive effect on the estimation equation. As shown in Figure 6, the graph to the left represents the modelled cystatin C – based equation without the age adjustment, whilst the graph to the right represents the modelled cystatin C -based equation with the age adjustment. The blue line represents mGFR and the red line represents the modelled eGFR equation based on cystatin C.

In the graph to the right, it is evident that the lines are better imposed on each other as compared to the graph on the left. However, in both equations, the estimating equations underestimate GFR at lower levels of age and overestimates GFR at higher levels of age.

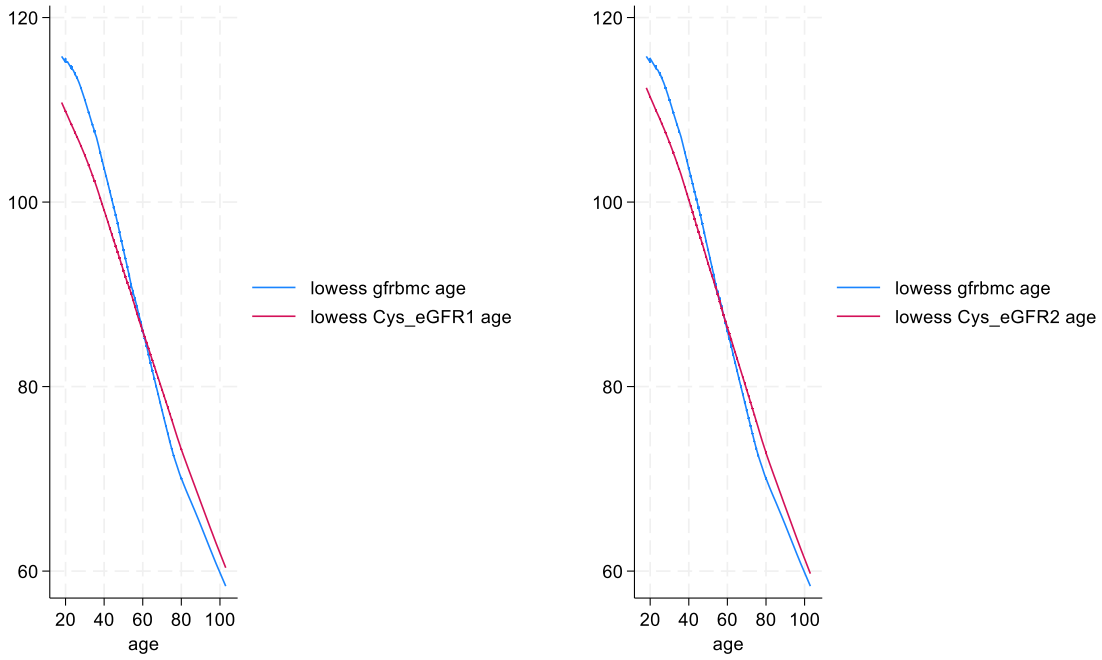


***Figure 6**: Comparison of the modelled Cystatin C based equations*

Pearson correlation coefficients between mGFR and the developed cystatin C based equations both estimating equations is 0.4476 and they are both statistically significant with p-value <0.001. Overall, this indicates that there is a moderate positive and statistically significant linear relationship between mGFR and both Cys_eGFR1 and Cys_eGFR2.

3.2.2 CystatinC-Creatinine based estimating equations

In development of the estimating equation based on a combination of serum Cystatin C and serum creatinine, non-linear estimation methods in Stata 17 were used and mirrored the CKD_EPI (creatinine-Cystatin C) equation. The study further developed the set of equations adjusting for the age parameter to check for improvement in the model.

For females the nonlinear estimation function was fit as shown in equation 3.5:

$$Y_i = 135 \times \min\left(\frac{\omega_i}{0.7}, 1\right)^{-0.248} \times \max\left(\frac{\omega_i}{0.7}, 1\right)^{-\alpha_1} \times \min\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_2} \times \max\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_3} \times 0.995^{\lambda_i} [\times \alpha_4 \text{ if female}] + E_i \quad (19)$$

***Equation 3.5**: Developed non-linear eGFR estimating function for females based on both serum Creatinine and Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ω_i, ψ_i and λ_i are the serum creatinine, serum Cystatin C level and age for participant, i, respectively. Because α_4 is multiplying the whole equation, zero is not a feasible starting point as it would turn the equation to zero, hence the starting point for α_4 was chosen at 0.5.

Another non-linear estimation equation was developed, adjusting for age using the non-linear estimation as shown in equation 3.6:

$$Y_i = 135 \times \min\left(\frac{\omega_i}{0.7}, 1\right)^{-0.248} \times \max\left(\frac{\omega_i}{0.7}, 1\right)^{-\alpha_1} \times \min\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_2} \times \max\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_3} \times \alpha_4{}^{\lambda_i} [\times \alpha_5 \text{ if female}] + E_i \quad (20)$$

Equation 3.6: *Developed non-linear eGFR estimating function for females adjusting for age, based on both serum Creatinine and Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ω_i, ψ_i and λ_i are the serum creatinine, serum Cystatin C level and age for participant, i, respectively. Because α_4 multiplying the whole equation, zero is not a feasible starting point as it would turn the equation to zero, hence the starting point for α_5 was set to 0.5, and for $\alpha_4 = 0.8$.

Table 7 shows the modelled estimating equations based on both serum creatinine and serum cystatin for females.

Table 6: *Female modelled Creatinine-Cystatin C based estimating Equations*

Female Creatinine-Cystatin estimating equation without Age adjustment
$135 * (\min(\frac{\text{creatinine}}{0.7}, 1)^{-0.248}) * (\max\left(\frac{\text{creatinine}}{0.7}, 1\right)^{-0.296}) * (\min(\frac{\text{cystatin}}{0.8}, 1)^{-0.382}) * (\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.310}) * (0.995^{\text{Age}}) * 3.541$
Female Creatinine-Cystatin estimating equation with Age adjustment
$135 * (\min(\frac{\text{creatinine}}{0.7}, 1)^{-0.248}) * (\max\left(\frac{\text{creatinine}}{0.7}, 1\right)^{-0.288}) * (\min(\frac{\text{cystatin}}{0.8}, 1)^{-0.377}) * (\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.250}) * (0.993^{\text{Age}}) * 3.702$

For males the nonlinear estimation function was fit as shown in equation 3.7:

$$Y_i = 135 \times \min(\frac{\omega_i}{0.9}, 1)^{-0.207} \times \max(\frac{\omega_i}{0.9}, 1)^{-\alpha_1} \times \min(\frac{\psi_i}{0.8}, 1)^{-\alpha_2} \times \max\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_3} 0.995^{\lambda_i} + E_i \quad (21)$$

Equation 3.7: *Developed non-linear eGFR estimating function for males based on both serum Creatinine and Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ω_i, ψ_i and λ_i are the serum creatinine, serum Cystatin C level and age for participant, i, respectively.

Adjusting for the age parameter, we fit the non-linear estimating equation for male participants based on both infiltration markers as shown in equation 3.8:

$$Y_i = 135 \times \min(\omega_i/0.9, 1)^{-0.207} \times \max(\omega_i/0.9, 1)^{-\alpha_1} \times \min(\psi_i/0.8, 1)^{-\alpha_2} \times \max\left(\frac{\psi_i}{0.8}, 1\right)^{-\alpha_3} {\alpha_4}^{\lambda_i} + E_i \quad (22)$$

Equation 3.8: *Developed non-linear eGFR estimating function for males adjusting for age, based on both serum Creatinine and Cystatin C*

Where Y_i is the measured glomerular filtration rate of the participant, i, and ω_i, ψ_i and λ_i are the serum creatinine, serum Cystatin C level and age for participant, i, respectively. I chose the starting point for $\alpha_4 = 0.8$

Table 7 shows the modelled estimating equations based on both serum creatinine and serum cystatin C for males.

Table 7 : Male modelled Creatinine-Cystatin C based estimating Equations

Male Creatinine-Cystatin estimating equation without Age adjustment
135 * $(\min(\frac{\text{creatinine}}{0.9}, 1)^{-0.207})$ * $(\max\left(\frac{\text{creatinine}}{0.9}, 1\right)^{0.017})$ * $(\min(\frac{\text{cystatin}}{0.8}, 1)^{0.411})$ * $(\max(\frac{\text{cystatin}}{0.8}, 1)^{-0.579})$ * (0.995^{Age})
Male Creatinine-Cystatin estimating equation with Age adjustment
135 * $(\min(\frac{\text{creatinine}}{0.7}, 1)^{-0.207})$ * $(\max\left(\frac{\text{creatinine}}{0.7}, 1\right)^{-0.406})$ * $(\min(\frac{\text{cystatin}}{0.8}, 1)^{0.409})$ * $(\max(\frac{\text{cystatin}}{0.7}, 1)^{-0.353})$ * (0.994^{Age}) * 6.577

Figure 7 illustrates the estimation pattern of the modelled Creatinine-Cystatin C estimating equations against mGFR. In the case of the Creatinine-Cystatin C equations, the age adjustment had an improvement on the estimation especially between 40 and 80 years of age where the estimations were almost super-imposed against those of the measured gfr. However, over estimation and underestimation of the extreme ages was more than that of the estimating equation which did not have the age adjustment factor.

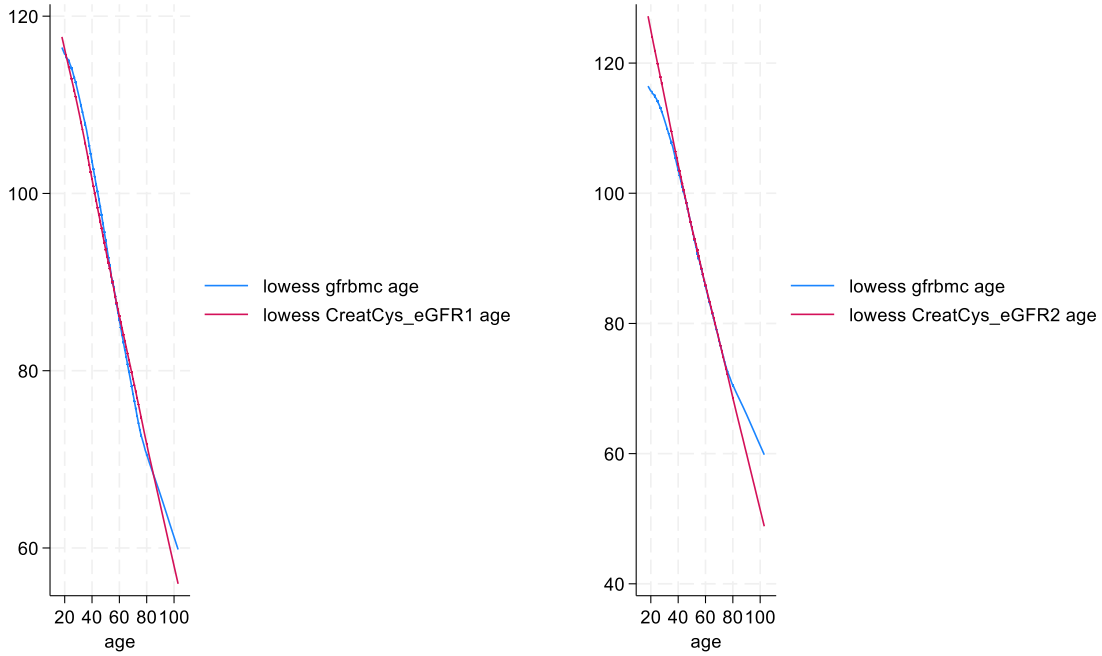


Figure 7: *Comparison of modelled Creatinine-Cystatin C based equations*

Pearson correlation showed that both CreatCys_eGFR1 and CreatCys_eGFR2 have positive and moderate correlations with mGFR, with correlation coefficients of 0.4664 and 0.4890, respectively. The correlation between mGFR and CreatCys_eGFR2 (0.4890) is slightly higher than the correlation with CreatCys_eGFR1 (0.4664), suggesting a slightly stronger linear relationship with CreatCys_eGFR2. All the correlations are statistically significant at any conventional significance level, as indicated by the $p<0.0001$ significance levels. Compared to the previous correlations with Cys_eGFR1 and Cys_eGFR2, the

correlations with CreatCys_eGFR1 and CreatCys_eGFR2 are slightly higher, suggesting that these variables may have a stronger linear relationship with mGFR.

3.3 Evaluation of performance of models

Evaluation of the performance of the modelled equations was measured against mGFR. To assess the performance and validity of the estimating equations, the correlation was assessed between all eGFR equations (both the modelled and existing equations) and mGFR via scatter plots.

Figure 8 shows these scatterplots that allow visual assessment of agreement and correlation between the different eGFR estimation methods and mGFR. The modelled Cystatin C and Creatinine-Cystatin C based methods demonstrate the strongest agreement based on the tight clustering along the identity lines. This suggests that the modelled Cystatin C and Creatinine-Cystatin C eGFR equations have the strongest correlation and agreement with mGFR. The various CKD-EPI equations perform similarly whilst ARK model methods appear to have weaker correlation against mGFR in comparison to the other equations performance.

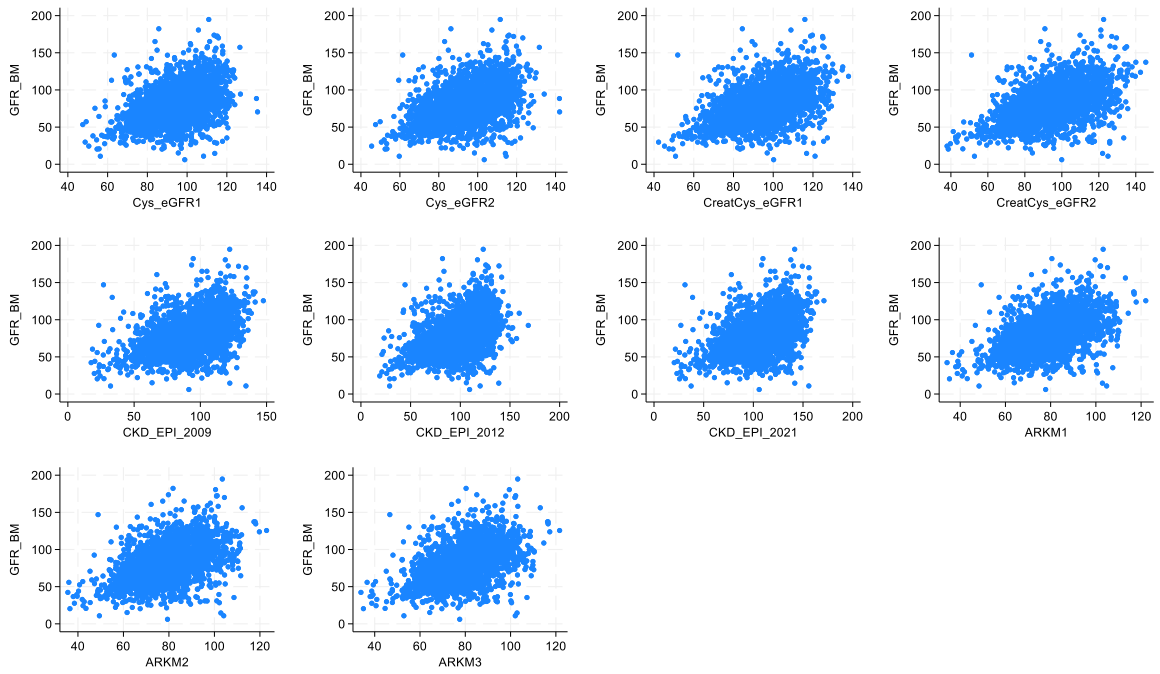


Figure 8: *Scatter plots depicting assessment of linearity on the eGFR models (both the modelled and the existing equations*

The Bland-Altman analysis provides a quantitative visualization of agreement. Figure 9 illustrates the agreement between the eGFR equations (both the modelled and the existing equations) and mGFR. The modelled cystatin C based, and creatinine-cystatin C based eGFR equations show the strongest agreement with mGFR based on their narrow levels of agreements (-100 -100), and small mean differences. The CKD-EPI equations plots show wider limits of agreement compared to the modelled equations plots. They exhibit weaker overall agreement with mGFR. Some slight proportional bias is evident with more points below 0 difference at lower averages. Lastly, the ARKM equations display the poorest agreement with mGFR.

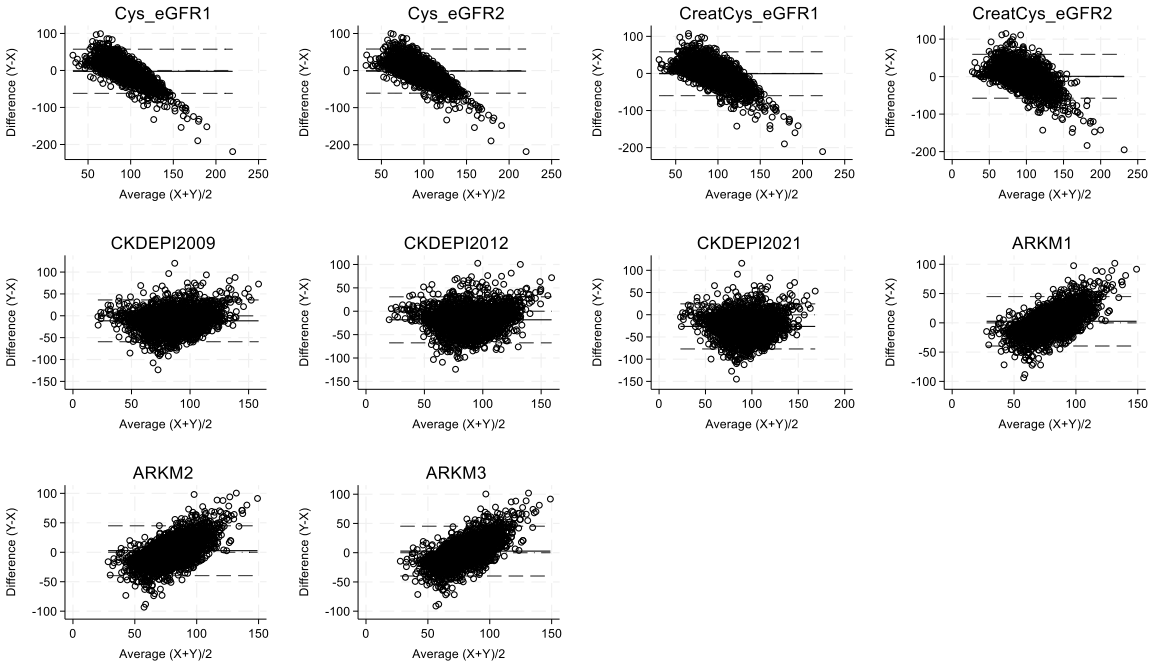


Figure 9: *Bland-Altman plots illustrating agreement between eGFR equations (both the modelled and existing equations) and mGFR*

Accuracy, bias, and precision were used to evaluate the performance of the eGFR equations. The values of the three-performance metrics for the modelled eGFR equation were compared against the same metrics calculated for existing eGFR equations. The reference standard used to compute all metrics was the mGFR. Table 8 compares the performance of the different eGFR equations against mGFR across CKD stages.

Table 8: *Comparing eGFR estimating equations to iohexol-measured GFR (mGFR) by GFR range and stage of chronic kidney disease*

	≥90 mL/min per 1.73m² of BSA, CKD stage 1	**60-89 mL/min per 1.73m² of BSA, CKD stage 2**	**45-59 mL/min per 1.73m² of BSA, CKD stage 3a**	**<45 mL/min per 1.73m² of BSA, CKD stage 3b-5**	**Overall**
Iohexol GFR, mL/min per 1·73 m2 of BSA	838(34%)	1121(46%)	320(13%)	154(6%)	n=2433
Categories by GFR stage for GFR estimating equations					

CKD-EPI 2009	1794(74%)	472(19%)	114(5%)	53(2%)	1067(44%)
CKD-EPI 2012	1675(69%)	635(26%)	79(3%)	44(2%)	1135(47%)
CKD-EPI 2021	1868(77%)	475(20%)	55(2%)	32(1%)	1554(64%)
ARKM1	471(19%)	1811(74%)	137(6%)	14(1%)	1336(55%)
ARKM2	470(19%)	1815(75%)	134(6%)	14(1%)	1336(55%)
ARKM3	457(19%)	1836(75%)	127(5%)	13(1%)	1323(54%)
Cys-based_1	1520(62%)	896(37%)	17(1%)	0%	1202(49%)
Cys-based_2	1496(61%)	917(38%)	20(1%)	0%	1226(50%)
CreatCys-based_1	1502(62%)	904(37%)	26(1%)	1(0.04%)	1221(50%)
CreatCys-based_2	1472(61%)	913(38%)	39(2%)	9(0.4%)	1262(52%)
Bias, mL/min per 1·73 m2					
CKD-EPI 2009	-4(-6 - -3)	15(14-16)	26(24-29)	39(34-43)	11(10-12)
CKD-EPI 2012	2(1-4)	22(21-23)	33(31-36)	42(38-47)	18(17-19)
CKD-EPI 2021	12(11-14)	30(28-31)	39(36-42)	51(46-56)	-3(-3.5--1.8)
ARKM1	-22(-23 - -21)	1(0.6-2)	17(16-18)	32(29-34)	-3(-3.5--1.8)
ARKM2	-22(-23 - -21)	1(0.5-2)	17(16-18)	32(29-35)	-3(-3.5--1.8)
ARKM3	-22(-23 - -21)	1(0.7-2)	17(16-18)	32(30-35)	-3(-3.5--1.8)
Cys-based_1	-7(-8 - -5)	17(16.6-18)	33(32-35)	47(44-50)	13(12-14)
Cys-based_2	-5(-6 - -4)	17(16.5-18)	33(31-34)	46(43-49)	13(12-14)
CreatCys-based_1	-4(-6 - -3)	17(16.7-18)	32(31-34)	46(42-49)	13(12-14)
CreatCys-based_2	-5(-7 - -4)	17(16.6-18)	33(31-34)	46(43-49)	13(12-14)
Relative Bias					
CKD-EPI 2009	-2%(-4 - -2)	21%(19-22)	50%(46-55)	130%(104-156)	23%(21-23)
CKD-EPI 2012	4%(3-5)	30%(29-32)	63%(58-68)	144%(114-173)	33%(30-35)
CKD-EPI 2021	13%(12-15)	40%(38-42)	74%(69-79)	167%(136-197)	43%(40-46)
ARKM1	-19%(-20 - -18)	24%(16-33)	33%(30-35)	108%(87-128)	6%(4-8)
ARKM2	-19%(-19 - -18)	24%(15-33)	33%(30-35)	109%(88-129)	6%(4-8)
ARKM3	-19%(-20 - -18)	26%(17-35)	33%(30-36)	110%(89-130)	6%(4-8)
Cys-based_1	-4%(-5 - -3)	24%(23-25)	63%(60-66)	154%(129-178)	27%(25-30)
Cys-based_2	-3%(-4 - -2)	24%(23-25)	62%(59-65)	152%(126-178)	28%(25-30)
CreatCys-based_1	-3%(-4 - -2)	24%(23-25)	61%(58-64)	151%(126-177)	28%(25-30)

CreatCys-based_2	-1%(-2 - 0)	24%(22-25)	58%(55-61)	146%(120-172)	28%(25-30)
Precision, log RMSE					
CKD-EPI 2009	2.19(2.16-2.22)	2.28(2.26-2.30)	2.42(2.37-2.47)	2.57(2.50-2.64)	2.29(2.27-2.31)
CKD-EPI 2012	2.22(2.19-2.25)	2.40(2.38-2.43)	2.55(2.50-2.59)	2.57(2.50-2.64)	2.37(2.35-2.39)
CKD-EPI 2021	2.36(2.33-2.38)	2.51(2.49-2.53)	2.61(2.57-2.64)	2.68(2.62-2.74)	2.48(2.46-2.50)
ARKM1	2.36(2.33-2.39)	1.94(1.91-1.96)	2.26(2.22-2.31)	2.54(2.49-2.59)	2.17(2.15-2.19)
ARKM2	2.37(2.34-2.39)	1.94(1.91-1.96)	2.28(2.24-2.32)	2.55(2.50-2.59)	2.17(2.15-2.19)
ARKM3	2.36(2.33-2.38)	1.93(1.90-1.95)	2.26(2.22-2.30)	2.55(2.50-2.60)	2.17(2.14-2.18)
Cys-based_1	2.07(2.04-2.11)	2.29(2.26-2.31)	2.61(2.50-2.63)	2.74(2.72-2.76)	2.29(2.27-2.31)
Cys-based_2	2.09(2.06-2.12)	2.29(2.26-2.31)	2.59(2.57-2.61)	2.72(2.70-2.75)	2.29(2.27-2.31)
CreatCys-based_1	2.10(2.06-2.13)	2.29(2.27-2.32)	2.59(2.57-2.61)	2.72(2.69-2.74)	2.29(2.27-2.31)
CreatCys-based_2	2.12(2.09-2.15)	2.28(2.25-2.290	2.55(2.53-2.58)	2.67(2.63-2.71)	2.29(2.27-2.30)
Accuracy					
CKD-EPI 2009	89%(86-91)	64%(61-66)	34%(29-39)	16%(10-22)	65%(63-67)
CKD-EPI 2012	86%(84-89)	48%(45-51)	21%(17-26)	14%(9-21)	56%(54-58)
CKD-EPI 2021	75%(72-78)	35%(33-38)	17%(13-22)	12%(7-18)	45%(43-47)
ARKM1	80%(78-83)	94%(93-96)	49%(43-55)	10%(5-15)	78%(77-80)
ARKM2	79%(77-83)	94%(93-96)	50%(44-55)	9%(5-15)	78%(76-80)
ARKM3	80%(77-83)	95%(93-96)	50%(45-56)	8%(5-14)	78%(77-80)
Cys-based_1	95%(93-96)	65%(62-68)	6%(4-9)	0%(0-2.4)	63%(61-65)
Cys-based_2	94%(93-96)	64%(61-67)	11%(8-15)	0.65%(0-4)	64%(62-65)
CreatCys-based_1	95%(93-96)	64%(61-67)	12%(8-16)	1.3%(0.16-5)	64%(62-65)
CreatCys-based_2	93%(91-95)	66%(63-68)	20%(16-25)	3.9%(1-8)	93%(91-95)

CKD-EPI 2009 and 2012 equations have negative bias, underestimating GFR, while CKD-EPI 2021 overestimates. ARKM equations have minimal overall bias. There is a positive bias in the modelled Cystatin C based equations and the Creatinine – Cystatin C based equations, resulting in an overestimation of GFR. Bias increases in magnitude with later CKD stages for all equations.

CKD-EPI and ARKM equations have higher precision errors (RMSE) than the modelled equations. Accuracy is highest for the modelled estimating equations. CKD-EPI equations and ARKM equations accuracy decreases in later CKD stages. The modelled Cystatin C and creatinine- cystatin C based equations have the best combination of minimal bias and good accuracy across all CKD stages.

Modelled estimating equations, both based on cystatin C and a combination of cystatin C and creatinine, outperform the existing CKD-EPI equations and the ARKM equations overall, with less bias and greater accuracy. However, performance declines for all equations in later CKD.

3.4 Machine Learning techniques in detecting CKD

Table 9 and Table 10 shows the ranks and correlation coefficients associated with the features in our data and how they are related to mGFR prediction. Based on the ranks and correlation statistics, the most useful features are cystatin, age, and serum creatinine. These have moderately strong correlations with mGFR and decent information gain scores.

Including these features would help predict mGFR accurately. Country, sex, and bmi have extremely low information gain scores, indicating they are not useful predictors of mGFR. Height has a small positive correlation with mGFR, so may provide signal. Weight has almost no correlation and was excluded.

The optimal model included cystatin C, age, and serum creatinine as the primary predictive features, with country and sex as secondary features.

Ranks

Table 9: Ranks of the features used for feature selection

		#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
1	N mGFR		0.325	0.162	0.111	NA	607.630	0.117	0.337
2	N cystatin C		0.040	0.020	0.014	NA	91.919	0.009	0.000
3	N age		0.033	0.017	0.012	NA	86.854	0.043	0.000
4	N creatinine		0.018	0.009	0.006	NA	46.682	0.014	0.000
5	C country	3.0	0.004	0.002	0.001	NA	1.648	0.000	0.000
6	N bmi		0.002	0.001	0.001	NA	1.965	0.026	0.000
7	N height		0.002	0.001	0.001	NA	4.569	0.023	0.000
8	N weight		0.001	0.000	0.000	NA	0.256	0.025	0.000
9	C sex	2.0	0.000	0.000	0.000	NA	0.145	0.000	0.000

Spearman correlation

Table 10 : Spearman correlation coefficients of the features of the ARK study data set

1	-0.405	age	mGFR
2	-0.369	cystatin	mGFR
3	-0.248	creatinine	mGFR
4	+0.120	gfrbmc	height
5	-0.069	bmi	gfrbmc
6	-0.020	gfrbmc	weight

Figure 10 shows the workflow chat in development of the machine learning techniques used in this study for detecting CKD with a binary outcome (CKD-present and CKD-absent). Six machine learning algorithms were developed as shown in figure 10.

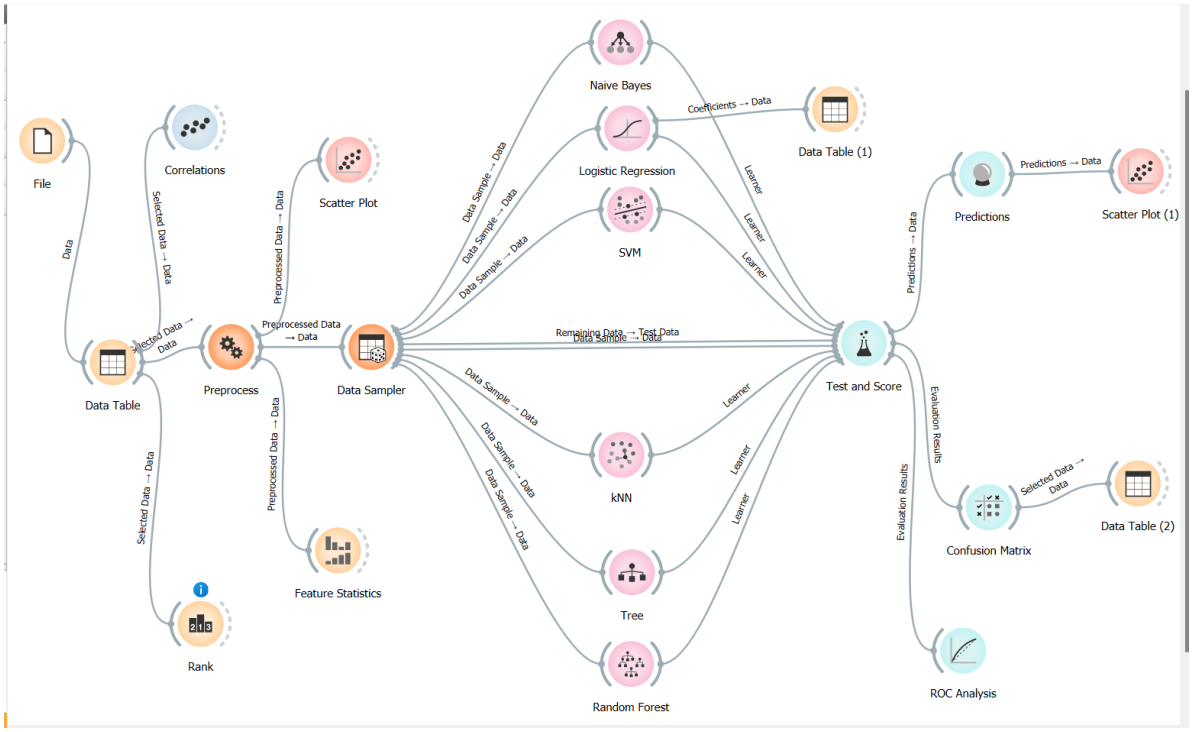


Figure 10: *Workflow chat for the machine learning techniques developed in Orange*

The data set was split using the ratio 75:25, into training data set (N=1825), and the testing data set (N=608). The study developed the models for all the machine learning techniques displayed in Figure 9 using the training data set. Evaluation of the performance based on the model predictions by computing the tests and scores of each technique using the test data set and the results of the performances are shown in Table 11.

Table 11: *Performance evaluation of machine learning models on the Training data set*

MODEL	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
Naïve Bayes	0.945	0.887	0.890	0.895	0.887	0.564	0.713	0.207
Logistic Regression	1.000	0.991	0.991	0.991	0.991	0.961	0.943	0.050
Tree	0.998	0.999	0.999	0.999	0.999	0.998	0.997	0.019
SVM	0.989	0.947	0.943	0.947	0.947	0.765	0.706	0.143
Random Forest	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.009**
Knn	0.956	0.938	0.933	0.935	0.938	0.721	0.689	0.357

Table 12: *Performance evaluation of machine learning models on the Testing data set*

MODEL	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
Naïve Bayes	0.935	0.880	0.888	0.902	0.880	0.516	0.715	0.225
Logistic Regression	1.000	0.993	0.993	0.993	0.993	0.968	0.962	0.048
Tree	0.993	0.998	0.998	0.998	0.998	0.992	0.988	0.057
SVM	0.992	0.951	0.947	0.949	0.951	0.740	0.688	0.106
Random Forest	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.007**
Knn	0.981	0.946	0.944	0.943	0.946	0.724	0.736	0.105

Random Forest has perfect scores of 1.0 across all metrics in the training data set. This indicates that it is accurately classifying the learner data with a high degree of confidence. It also has perfect scores across all metrics on the training set indicating that it has clearly fit the training data very well.

Logistic Regression also performs very well on the training set with AUC of 1.0 and other metrics above 0.99. It is a strong model but slightly behind Random Forest. It again does quite well in predicting using the test data set, with AUC of 1.0 and other metrics above 0.99, indicating slightly higher performance than the learner set.

Naive Bayes lags on the training data set with AUC of 0.945 and metrics in the 0.88-0.89 range. Its high log loss of 0.207 shows uncertain predictions. Similarly, it again lags with AUC of 0.935 and metrics in the 0.88 range in the testing data set. Log loss of 0.225 is however an improvement over the training data set

The other models (Tree, SVM, KNN) achieve reasonable but not standout performance on the training data set. Their metrics are in the 0.90-0.95 range. They, however, perform noticeably better on the training set implying some overfitting to the training data.

In summary, Random Forest is the top performer. Logistic Regression does very well on prediction of mGFR. The other models appear to have overfitted the training data more than the learner data. Comparing performance on the two datasets provides insights into generalization capabilities.

4. Conclusion and Recommendations

4.1 Discussion

This study evaluated the accuracy of CKD-EPI eGFR equations and developed new equations for estimating GFR among black adults from South Africa, Malawi, and Uganda. The analysis provided several key insights regarding the performance of existing and newly developed eGFR equations in this African population.

The descriptive statistics for this study reported that the differences in socio demographic and health variables for participants were not statistically significant. The results were similar to those reported in the primary study (33). Strong evidence of the difference in means of mGFR by country and sex were established in post-hoc analyses, highlighting the importance of considering these factors when estimating GFR in diverse African populations. This finding aligns with previous studies that have noted variations in kidney function across different African ethnic groups (34)

The newly developed serum cystatin C-based and combined creatinine-cystatin C equations showed improved performance compared to existing equations. The cystatin C-based equation performed particularly well, which is consistent with other research suggesting cystatin C may be a more reliable GFR marker than creatinine in some populations (21). However, it's important to note that our equations were developed and validated on the same dataset, which may lead to overly optimistic performance estimates. External validation in independent African cohorts is needed to confirm the generalizability of these new equations.

The development and evaluation of new GFR estimating equations in the African population provides important insights into performance across CKD stages. The analysis showed that the developed eGFR equations based solely on serum cystatin C performed better than those based on a combination of both serum creatinine and serum cystatin C. The modelled equations both had higher correlation with mGFR across CKD stages based on scatterplot analysis. Bland-Altman plots also exhibited narrower limits of agreement for the modelled equations, indicating better quantitative agreement versus mGFR. Finally, the modelled equations displayed lower bias and higher accuracy compared to existing CKD-EPI equations and ARKM equations.

These findings were consistent across CKD stages, with the modelled equations maintaining minimal bias and good accuracy even in later stage 3-5 CKD. In contrast, existing equations showed declining performance in advanced CKD. This provides evidence that the new modelled equations address limitations of current methods in the African population across disease severity.

In evaluating equation performance, the new equations had lower bias and higher accuracy across CKD stages compared to existing CKD-EPI and ARKM equations. This was especially true in later stage CKD, where existing equations tend to perform poorly. These findings support the need for population-specific eGFR equations in African settings, as noted by other researchers (35).The machine learning analysis provided an innovative approach to CKD detection, with Random Forest and Logistic Regression models showing excellent performance. However, the perfect scores achieved by these models, especially on the test set, are unusual and may indicate some data leakage or overfitting issues. Further investigation and external validation would be prudent before drawing strong conclusions about the superiority of these ML approaches.

The evaluation of 6 machine learning classification techniques on a training dataset and testing dataset provides insights into their comparative predictive capabilities on this problem, in this case, classification. Random Forest is clearly the top performing model, with perfect scores across all metrics on both datasets. This indicates a strong generalization ability. Logistic Regression is a close second, also achieving an AUC of 1.0 and scores above 0.99 on both sets.

The other models - Naive Bayes, Tree, SVM, and KNN - lag Random Forest and Logistic Regression. Their performance is noticeably better on the training set versus the learner set, implying some overfitting has occurred. Naive Bayes in particular struggles with this data, with the lowest scores and highest log loss on both sets.

4.2 Strengths and limitations of nonlinear models

The strength for this study was based on the flexibility demonstrated by non-linear models to model complex non-linear relationships between predictors and outcome. The association between GFR, cystatin C, creatinine and age is non-linear, so non-linear models are better suited. The study avoided assumptions of linearity which may not apply as non-linear models make no assumptions about linear associations. Non-linear models can fit data more precisely across full range of values, not just locally around mean, hence they became useful for modeling entire spectrum of CKD severity. They also allow inclusion of interaction terms and modifiers like age and sex in flexible non-linear ways. The modelled equations demonstrated improved performance metrics like lower bias and higher accuracy versus linear models.

In the context of existing literature, our findings generally align with other studies showing the need for improved eGFR estimation in African populations (33,36). The superior performance of cystatin C-based equations is also consistent with some previous research (37). However, the degree of improvement we observed with our new equations is larger than in some other studies, which could be due to our specific population or methodological differences.

The weaknesses of this study are that more complex specification and fitting, requiring non-linear regression methods may be associated with risk of overfitting if not careful. Non-linear models are harder to interpret parameters compared to linear coefficients. The models still showed some inaccuracies at very high and low mGFR values, suggesting room for improvement. Limited external validation so far compared to well-established linear CKD-EPI model.

Several limitations should be considered when interpreting these results. First, the study population, while diverse, may not be fully representative of all African populations. Second, the use of iohexol plasma clearance as the mGFR reference standard, while accepted, has its own potential sources of error (29). Third, the multiple analyses performed (equation development, traditional statistical evaluation, and machine learning) increase the risk of type I errors. Finally, the lack of external validation limits our ability to assess the true generalizability of the new equations and ML models

Overall, the non-linear modeling approach provided flexibility to develop equations tailored to the African population that addressed limitations of traditional linear models. The improved

performance metrics validate this approach. Further refinement and external validation will be important next steps to realize the full potential of non-linear models for GFR estimation.

4.3 Conclusion

The evaluation of the performance of traditional and newer GFR estimating equations yields important insights. The analysis showed cystatin C and creatinine-cystatin C based equations demonstrate quantitatively stronger performance compared to existing methods across CKD severity in the African population. The modelled equations displayed lower bias, higher accuracy, and better agreement versus measured GFR.

These findings provide evidence that the modelled equations address limitations of current estimating equations in the African population across CKD stages. This supports transitioning to the modelled cystatin C and creatinine-cystatin C equations as superior methods for GFR assessment in African patients clinically.

However, no equation perfectly estimated GFR across all ranges. The ARKM equations showed the lowest overall bias, while tending to underestimate higher GFR and the CKD-EPI 2021 equation overestimated function in CKD patients. These findings confirm limitations in relying solely on creatinine as a filtration marker in current clinical practice for all patients.

Overall, the results highlight the promise of novel cysteine and cysteine-creatinine estimating equations to address gaps with traditional formulas in patients with normal kidney function or hyperfiltration. Accuracy in these populations has been a previously known shortcoming of

widely used equations like CKD-EPI and MDRD. The integration of filtration markers such as Cystatin C helps improve performance particularly for those without impaired kidney function.

The analysis and evaluation of different machine learning models on the task of predicting CKD shows that advanced techniques like Random Forest and Logistic Regression can significantly outperform traditional statistical models. Traditional models like logistic regression have been widely used in healthcare to predict outcomes like CKD onset. However, machine learning models that can capture nonlinear relationships and complex interactions can improve predictive performance.

In conclusion, this study provides valuable insights into GFR estimation in African populations and offers promising new equations for improved accuracy. However, external validation and further investigation of the machine learning approaches are necessary next steps. Future research should also explore the clinical impact of using these new equations on CKD diagnosis and management in African healthcare settings.

4.4 Recommendations

Based on the current findings regarding performance of GFR estimating equations, the following recommendations can be made:

1. Consider preferential use of Cystatin-based equations for patients without known CKD, particularly in settings where true GFR measurement is not feasible. This may improve accuracy in this cohort specifically.

2. Further studies should evaluate precision and accuracy trends across full age and disease spectrums to refine optimal equation selection approaches in both kidney health and disease states.

3. GFR estimates ideally should be confirmed periodically using measured methods when clinical context warrants, especially with estimates at odds with overall patient presentation. No equation is considered sufficiently accurate to replace true GFR assessment via testing when precision is essential. Random Forest should be the primary model deployed for this classification task, given its dominant performance. It provides accurate predictions with a high degree of certainty.

4. Logistic Regression is a strong secondary model to consider, given its excellent metrics just slightly behind Random Forest. It may be a suitable fallback if Random Forest were to encounter issues.

5. References

1. Andrassy KM. Comments on 'KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease'. Kidney Int. 2013;84(3):622–3.

2. Bill F, Foundation MG. Global , regional , and national burden of chronic kidney disease , 1990 – 2017 : a systematic analysis for the Global Burden of Disease Study 2017. 2017;

3. Stanifer JW, Jing B, Tolan S, Helmke N, Mukerjee R, Naicker S, et al. The epidemiology of chronic kidney disease in sub-Saharan Africa: a systematic review and meta-analysis. Lancet Glob Health. 2014 Mar;2(3):e174-81.

4. Stanifer JW, Jing B, Tolan S, Helmke N, Mukerjee R, Naicker S, et al. The epidemiology of chronic kidney disease in sub-Saharan Africa: a systematic review and meta-analysis. Lancet Glob Health. 2014 Mar;2(3):e174-81.

5. UNAIDS. Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic.

6. Swedish Council on Health Technologies Assesment. Methods to Estimate and Measure Renal Function (Glomerular Filtration Rate). A Systematic Review.

7. Van Deventer HE, Paiker JE, Katz IJ, George JA. A comparison of cystatin C-and creatinine-based prediction equations for the estimation of glomerular filtration rate in black South Africans. Nephrol Dial Transplant. 2011;26(5):1553–8.

8. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May;150(9):604–12.

9. AACC Programs and partners. Cystatin C and Creatinine: Complementary Markers of GFR Expert John C. Lieske, MD.

10. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. N Engl J Med. 2021 Sep 23;385(19):1737–49.

11. Muiru AN, Charlebois ED, Balzer LB, Kwarisiima D, Elly A, Black D, et al. The epidemiology of chronic kidney disease (CKD) in rural East Africa: A population-based study. PloS One. 2020;15(3):e0229649.

12. Glaser N, Deckert A, Phiri S, Rothenbacher D, Neuhann F. Comparison of Various Equations for Estimating GFR in Malawi: How to Determine Renal Function in Resource Limited Settings? PloS One. 2015;10(6):e0130453.

13. Omuse G, Maina D, Mwangi J, Wambua C, Kanyua A, Kagotho E, et al. Comparison of equations for estimating glomerular filtration rate in screening for chronic kidney disease in asymptomatic black Africans: a cross sectional study. BMC Nephrol. 2017 Dec;18(1):369.

14. Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med. 2006 Aug;145(4):247–54.

15. Monnet D, Pottel H, Bukabau JB, Yayo E, Gnionsahe A, Cavalier E, et al. Performance of creatinine- or cystatin C – based equations to estimate glomerular fi ltration. :1181–9.

16. Shlipak MG, Tummalapalli SL, Boulware LE, Grams ME, Ix JH, Jha V, et al. The case for early identification and intervention of chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2021;99(1):34–47.

17. Kalyesubula R, Fabian J, Nakanga W, Newton R, Ssebunnya B, Prynn J, et al. How to estimate glomerular filtration rate in sub-Saharan Africa: design and methods of the African Research into Kidney Diseases (ARK) study. BMC Nephrol. 2020;21(1):20.

18. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, et al. Chronic kidney disease: global dimension and perspectives. The Lancet. 2013;382(9888):260–72.

19. Koye DN, Magliano DJ, Nelson RG, Pavkov ME. The Global Epidemiology of Diabetes and Kidney Disease. Adv Chronic Kidney Dis. 2018;25(2):121–32.

20. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, et al. Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C. N Engl J Med. 2012 Jul 5;367(1):20–9.

21. Van Deventer HE, Paiker JE, Katz IJ, George JA. A comparison of cystatin C-and creatinine-based prediction equations for the estimation of glomerular filtration rate in black South Africans. Nephrol Dial Transplant. 2011;26(5):1553–8.

22. Justine B., Bukabau, Ernest K., Sumaili, Etienne, Cavalier, Hans, Pottel, Bejos, Kifakiou, Aliocha, Nkodila, Jean, Robert R. Makulo, Vieux M., Mokoli, Chantal V., Zinga, Augustin L., Longo, Yannick M., Engole, Yannick M., Nlandu, François B., Lepira, Naza D. Performance of glomerular filtration rate estimation equations in Congolese healthy adults: The inopportunity of the ethnic correction. PloS One. 2018;

23. Wyatt CM, Schwartz GJ, Owino Ong'or W, Abuya J, Abraham AG, Mboku C, et al. Estimating Kidney Function in HIV-Infected Adults in Kenya: Comparison to a Direct Measure of Glomerular Filtration Rate by Iohexol Clearance. PLoS ONE. 2013;8(8).

24. Moodley N, Hariparshad S, Peer F, Gounden V. Evaluation of the CKD-EPI creatinine based glomerular filtration rate estimating equation in Black African and Indian adults in KwaZulu-Natal, South Africa. Clin Biochem. 2018;59:43–9.

25. Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. J Big Data. 2022 Nov 20;9(1):109.

26. Healthcare | Free Full-Text | Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease [Internet]. [cited 2024 Feb 18]. Available from: https://www.mdpi.com/2227-9032/10/2/371

27. Benoit SW, Ciccia EA, Devarajan P. Cystatin C as a biomarker of chronic kidney disease: latest developments. Expert Rev Mol Diagn. 2020;00(00):1–8.

28. Fabian J, Kalyesubula R, Mkandawire J, Hansen CH, Nitsch D, Musenge E, et al. Measurement of kidney function in Malawi, South Africa, and Uganda: a multicentre cohort study. Lancet Glob Health. 2022 Aug 1;10(8):e1159–69.

29. Pottel H, Schaeffner E, Ebert N, van der Giet M, Delanaye P. Iohexol plasma clearance for measuring glomerular filtration rate: effect of different ways to calculate the area under the curve. BMC Nephrol. 2021 Dec;22(1):1–8.

30. Geist B. Calculation of GFR via the Slope-Intercept Method in Nuclear Medicine. In 2019.

31. Drug distribution and binding - PubMed [Internet]. [cited 2023 Mar 20]. Available from: https://pubmed.ncbi.nlm.nih.gov/3793947/

32. Giavarina D. Understanding Bland Altman analysis. Biochem Medica. 2015 Jun 5;25(2):141–51.

33. Kalyesubula R, Fabian J, Nakanga W, Newton R, Ssebunnya B, Prynn J, et al. How to estimate glomerular filtration rate in sub-Saharan Africa: design and methods of the African Research into Kidney Diseases (ARK) study. BMC Nephrol. 2020;21(1):20.

34. Peralta CA, Katz R, DeBoer I, Ix J, Sarnak M, Kramer H, et al. Racial and Ethnic Differences in Kidney Function Decline among Persons without Chronic Kidney Disease. J Am Soc Nephrol JASN. 2011 Jul;22(7):1327–34.

35. Measuring kidney function: the voice of Africa - The Lancet Global Health [Internet]. [cited 2024 Jul 16]. Available from: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(22)00279-0/fulltext

36. Estimated glomerular filtration rate equations in people of self-reported black ethnicity in the United Kingdom: Inappropriate adjustment for ethnicity may lead to reduced access to care | PLOS ONE [Internet]. [cited 2024 Jul 16]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0255869

37. Performance of Cystatin C-Based Equations for Estimation of Glomerular Filtration Rate in Diabetes Patients: A Prisma-Compliant Systematic Review and Meta-Analysis | Scientific Reports [Internet]. [cited 2024 Jul 16]. Available from: https://www.nature.com/articles/s41598-018-38286-9

6. Appendices

Appendix 1: Plagiarism

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

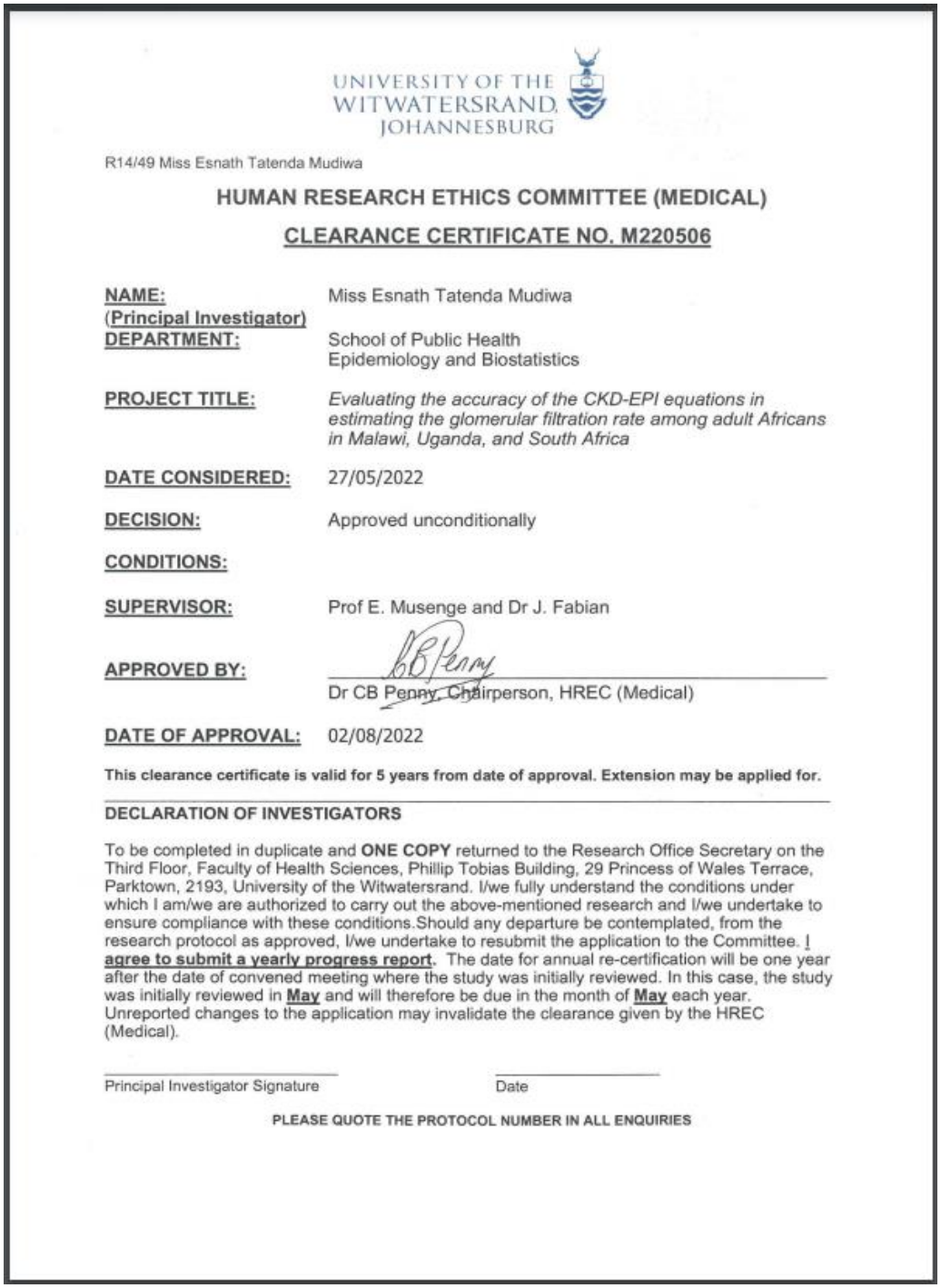
I ESNATH T MUDIWA (Student number: 2392604) am a student registered for the degree of MSc Biostatistics in the academic year 2021.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 02 April 2024

Appendix 2: Human Research Ethics Clearance certificate



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Esnath Tatenda Mudiwa

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M220506

NAME: (Principal Investigator)	Miss Esnath Tatenda Mudiwa
DEPARTMENT:	School of Public Health Epidemiology and Biostatistics
PROJECT TITLE:	*Evaluating the accuracy of the CKD-EPI equations in estimating the glomerular filtration rate among adult Africans in Malawi, Uganda, and South Africa*
DATE CONSIDERED:	27/05/2022
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof E. Musenge and Dr J. Fabian
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	02/08/2022

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. I **agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **May** and will therefore be due in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 3: Comparison of accuracy within 30% and within 10% of estimating equations by exact McNemar tests of the DRC Kinshasa study (34)

Table 4. Comparison of accuracy within 30% and within 10% of estimating equations by exact McNemar tests.

		P30%	P30%	p-value	P10%	P10%	p-value
MDRD nef		86.0			41.9		
	CKD-EPI SCr nef		81.7	0.2891		51.6	0.136
	CKD-EPI SCys		91.4	0.3018		51.6	0.211
	CKD-EPI SCrCys nef		92.5	0.1094		57.0	**0.0385**
	CKD-EPI SCr		73.1	**0.0169**		20.4	**0.0091**
	CKD-EPI SCrCys		87.1	1.0000		37.6	0.6587
	MDRD		79.6	0.2101			0.4426
MDRD		79.6			34.4		
	CKD-EPI SCys		91.4	**0.0127**		51.6	**0.0226**
	CKD-EPI SCrCys nef		92.5	**0.0005**		57.0	**0.0002**
	CKD-EPI SCr		73.1	0.0703		20.4	**0.0106**
	CKD-EPI SCrCys		87.1	**0.0156**		37.6	0.7201
CKD-EPI SCr nef		81.7			51.6		
	CKD-EPI SCys		91.4	**0.0352**		51.6	1.0000
	CKD-EPI SCrCys nef		92.5	**0.0020**		57.0	0.4421
	CKD-EPI SCr		73.1	0.0574		20.4	**<0.0001**
	CKD-EPI SCrCys		87.1	0.1250		37.6	0.0533
	MDRD		79.6	0.7266		34.4	**0.0090**
CKD-EPI SCr		73.1			20.4		
	CKD-EPI SCys		91.4	**0.0002**		51.6	**<0.0001**
	CKD-EPI SCrCys nef		92.5	**<0.0001**		57.0	**<0.0001**
	CKD-EPI SCrCys		87.1	**0.0002**		37.6	**0.0004**
CKD-EPICys		91.4			51.6		
	CKD-EPI SCrCys		87.1	0.3437		37.6	0.0596
	CKD-EPI SCrCys nef		92.5	1		57	0.4049
CKD-EPI SCrCys		87.1			37.6		
	CKD-EPI SCrCys nef		92.5	**0.0039**		57	0.0625

CKD-EPI SCr: Chronic Kidney Disease-Epidemiology Collaboration equation based on serum creatinine only, with ethnic factor; CKD-EPI SCr nef: CKD-EPI without ethnic factor; CKD-EPI SCys: CKD-EPI equation based on cystatin C only; CKD-EPI SCrCys: CKD-EPI combining creatinine and cystatin C with ethnic factor. CKD-EPI SCrCys nef: CKD-EPI combining serum creatinine and cystatin C without ethnic factorMDRD: Modification of Diet in Renal Disease study equation with ethnic factor; MDRD nef: MDRD without ethnic factor. P10: accuracy within 10%; P30: accuracy within 30%; RMSE: Root mean square error.

Appendix 4: Performance of equations in the whole cohort for the Kinshasa and Ivory Coast study(34)

Table 3 | Performance of equations in the whole cohort (N = 494)

Equation	Absolute bias (95% CI)	Absolute SD	Accuracy within 30% (95% CI)	Lin's CCC (95% CI)
MDRD	−7.8 (−9.5 to −6.1)	19.4	76.1 (72.3 to 79.9)	0.73 (0.67 to 0.78)
MDRD ef	8.2 (6.1 to 10.2)	23.3	73.3 (69.4 to 77.2)	0.70 (0.61 to 0.77)
CKD-EPI	0.0 (−1.6 to 1.6)	18.1	77.7 (74.1 to 81.4)	0.81 (0.76 to 0.84)
CKD-EPI ef	13.3 (11.4 to 15.2)	21.3	64.6 (60.3 to 68.8)	0.71 (0.66 to 0.76)
FAS SCr	−1.5 (−3.1 to 0.2)	18.7	82.4 (79.0 to 85.8)	0.78 (0.72 to 0.82)
FAS SCr af	2.5 (0.7 to 4.2)	19.6	80.6 (77.1 to 84.1)	0.77 (0.70 to 0.82)
CKD-EPI CysC	3.2 (1.5 to 4.8)	18.5	78.7 (75.1 to 82.4)	0.80 (0.75 to 0.83)
FAS CysC	3.7 (2.0 to 5.3)	18.7	84.2 (81.0 to 87.4)	0.78 (0.72 to 0.82)
CKD-EPI combined	1.1 (−0.4 to 2.5)	16.6	79.2 (75.6 to 82.7)	0.83 (0.79 to 0.86)
CKD-EPI combined ef	7.8 (6.2 to 9.4)	18.1	76.7 (73.0 to 80.5)	0.80 (0.75 to 0.83)
FAS combined	0.1 (−1.3 to 1.5)	15.8	86.4 (83.4 to 89.5)	0.84 (0.79 to 0.87)
FAS combined af	2.1 (0.7 to 3.5)	16.1	85.4 (82.3 to 88.6)	0.83 (0.78 to 0.86)

CI, confidence interval; CKD-EPI, Chronic Kidney Disease Epidemiology Collaboration; CKD-EPI combined, Chronic Kidney Disease Epidemiology Collaboration equation combining creatinine and cystatin C; CKD-EPI combined ef, Chronic Kidney Disease Epidemiology Collaboration equation with the ethnic factor; CKD-EPI CysC, Chronic Kidney Disease Epidemiology Collaboration equation based on cystatin C; FAS combined, Full Age Spectrum equation based on creatinine and cystatin C with White Q; FAS combined af, Full Age Spectrum equation combined with African Q; FAS CysC, Full Age Spectrum equation based on cystatin C; FAS SCr, Full Age Spectrum equation based on serum creatinine with White Q; FAS SCr af, Full Age Spectrum equation based on serum creatinine with African Q; Lin's CCC , Lin's concordance correlation coefficient; MDRD, Modification of Diet in Renal Disease; MDRD ef, Modification of Diet in Renal Disease with the ethnic factor.

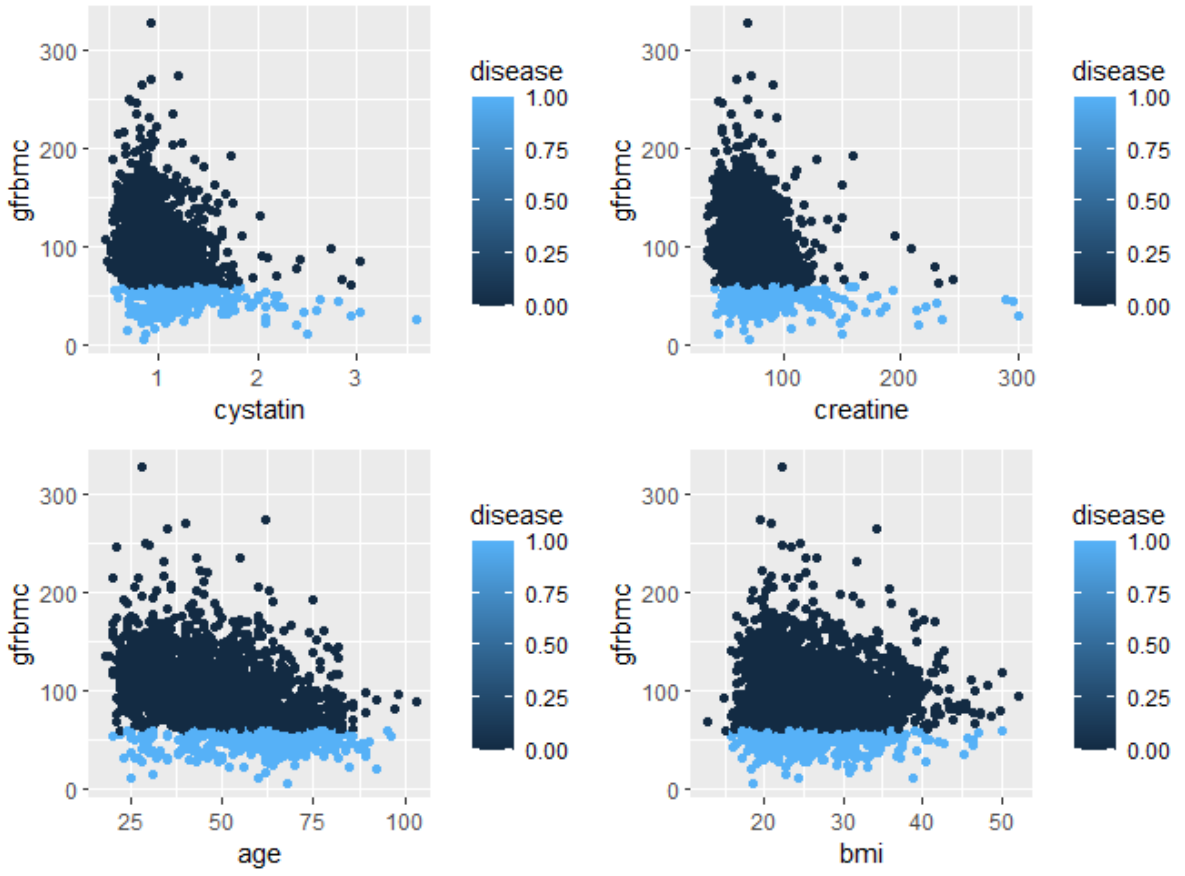
Appendix 5: Summary of the performance of the different equations in the DRC Kinshasa study(34)

Table 3. Performance of the MDRD and CKD-EPI (with and without ethnic factors).

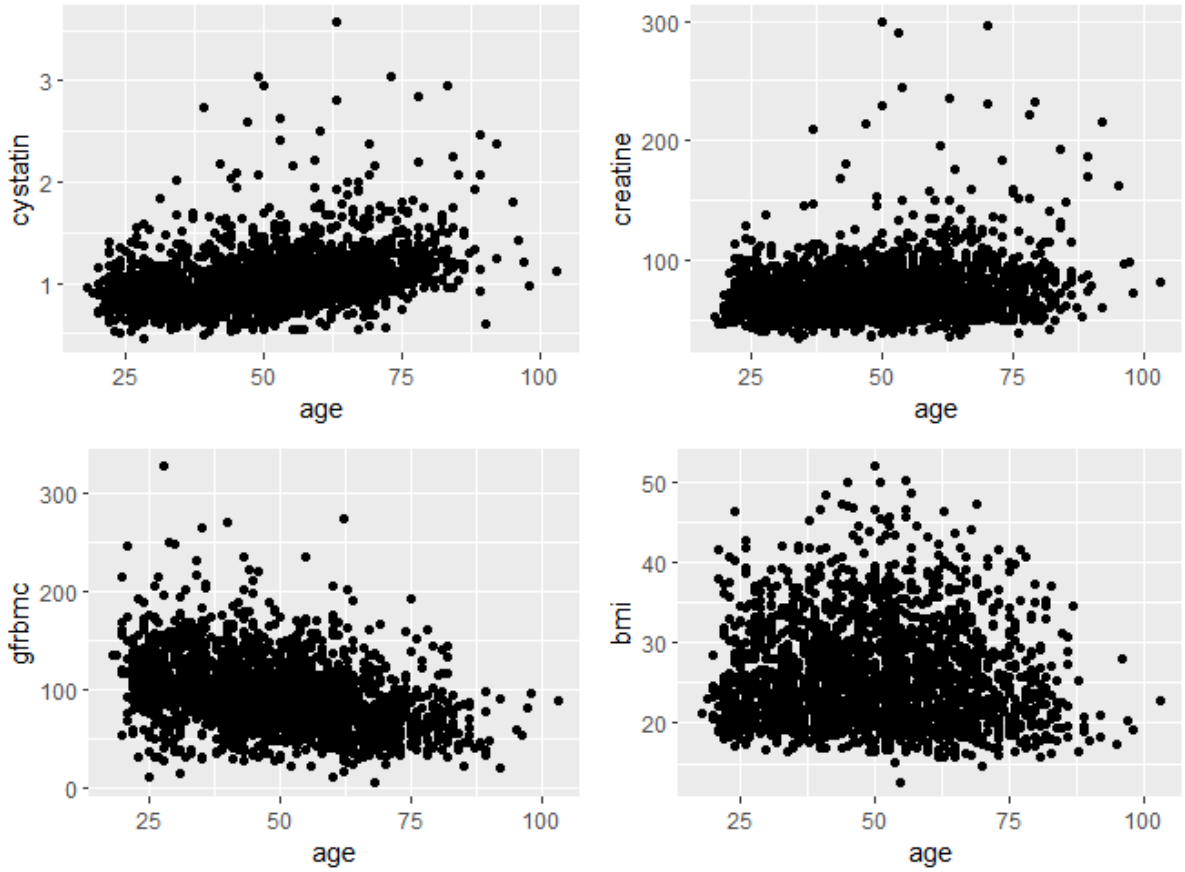
Equations	Lin's CCC [95%CI]	P10 [95%CI]	P30 [95%CI]	Bias [95%CI]	SD	Proportional bias [95%CI]	RMSE [95%CI]
MDRD	0.338 [0.198; 0.465]	34.4 [24.6–44.2]	79.6 [71.2–87.9]	13.6 [8.0; 19.2]	27.0	1.16 [1.10; 1.22]	30.1 [11.2; 41.1]
MDRD nef	0.414 [0.250; 0.555]	41.9 [31.7–52.2]	86.0 [78.8–94.0]	-4.9 [-9.6; -0.2]	22.9	0.96 [0.91; 1.01]	23.3 [11.0; 31.0]
CKD-EPI SCr	0.429 [0.302; 0.540]	20.4 [12.1–28.8]	73.1 [63.9–82.3]	17.2 [13.3; 21.2]	19.3	1.20 [1.15; 1.25]	25.8 [21.9; 29.2]
CKD-EPI SCr nef	0.594 [0.450; 0.707]	51.6 [41.3–62.0]	81.7 [73.7–89.7]	2.3 [-1.3; 5.8]	17.1	1.03 [0.99; 1.08]	17.1 [13.8; 20.0]
CKD-EPI SCys	0.605 [0.459; 0.718]	51.6 [41.3–62.0]	91.4 [85.6–97.2]	1.5 [-1.8; 4.7]	15.9	1.03 [0.99; 1.07]	15.9 [13.4; 18.0]
CKD-EPI SCrCys	0.607 [0.479; 0.710]	37.6 [27.6–47.7]	87.1 [80.2–94.0]	9.0 [5.9; 12.0]	14.7	1.11 [1.07; 1.15]	17.2 [14.5; 19.5]
CKD-EPI SCrCys nef	0.682 [0.557; 0.777]	57.0 [46.7–67.2]	92.5 [87.0–97.9]	1.5 [-1.4; 4.4]	14.0	1.03 [0.99; 1.06]	14.0 [11.4; 16.2]

CCC: Concordance Correlation Coefficient; CKD-EPI SCr: Chronic Kidney Disease-Epidemiology Collaboration equation based on serum creatinine only, with ethnic factor; CKD-EPI SCr nef: CKD-EPI without ethnic factor; CKD-EPI SCys: CKD-EPI equation based on cystatin C only; CKD-EPI SCrCys: CKD-EPI combining creatinine and cystatin C with ethnic factor; CKD-EPI SCrCys nef: CKD-EPI combining serum creatinine and cystatin C without ethnic factor; MDRD: Modification of Diet in Renal Disease study equation with ethnic factor; MDRD nef: MDRD without ethnic factor; P10: accuracy within 10%; P30: accuracy within 30%; RMSE: Root mean square error.

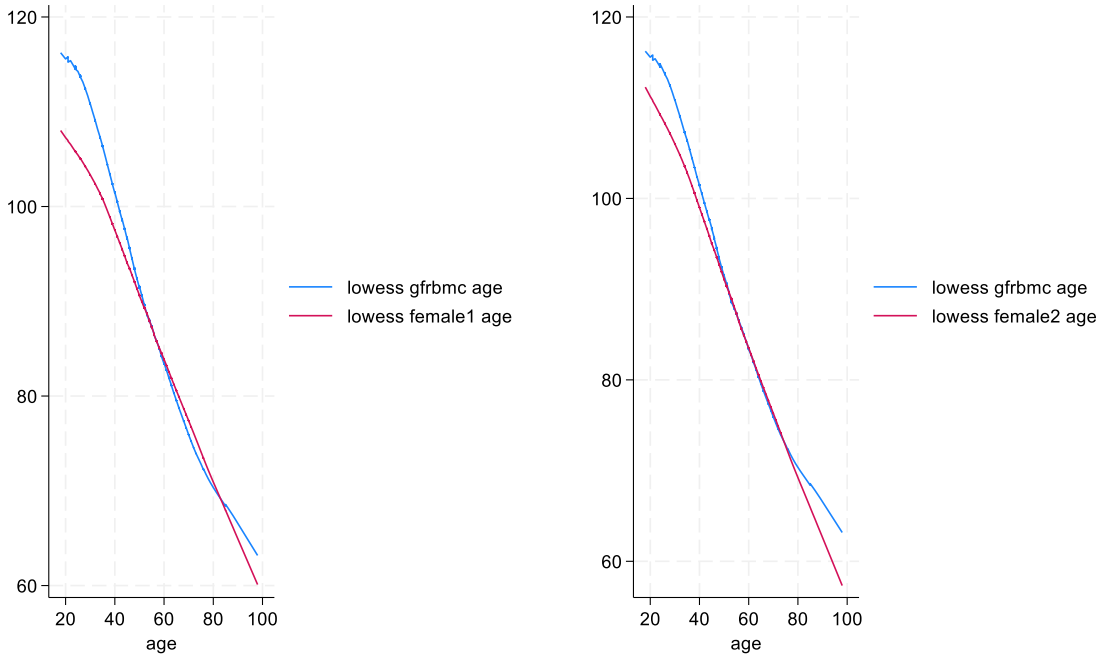
Appendix 6: Disease Effect



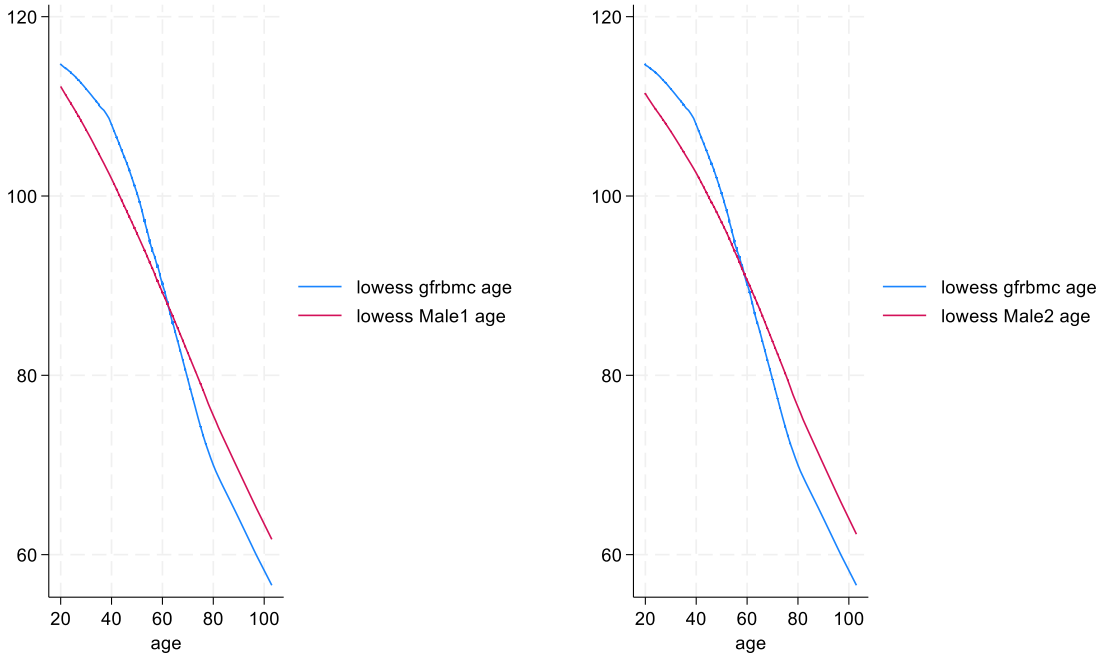
Appendix 7: Age Effect



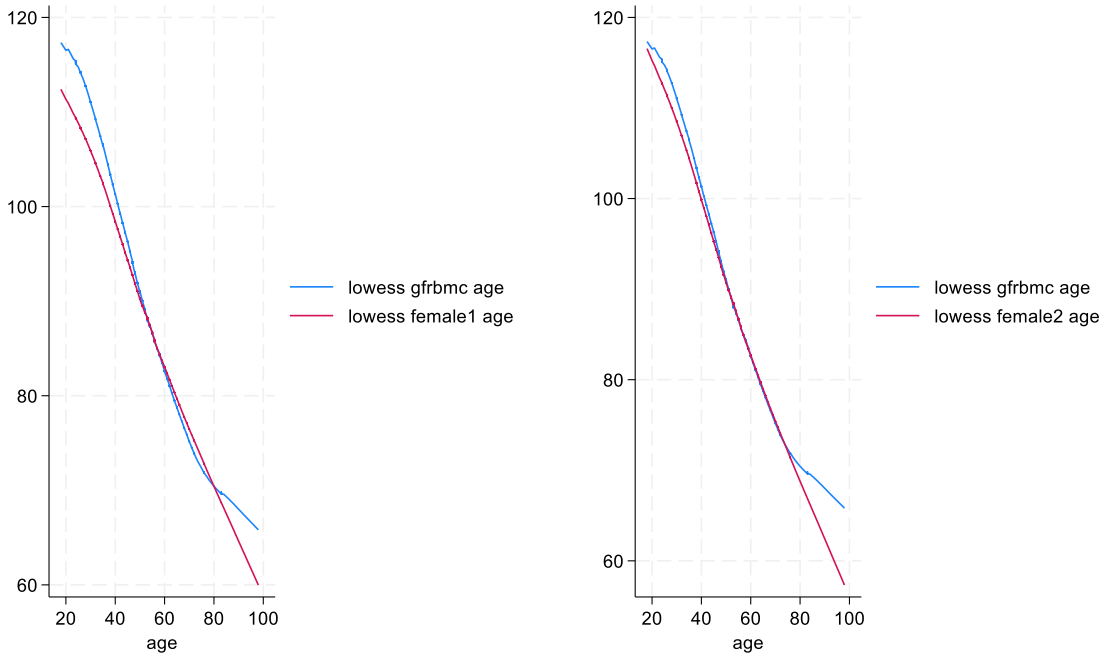
Appendix 8: Comparison of the modelled Cystatin C based equations for females. The right graph representing the modelled equation with the age adjustment whilst the left graph is representing the modelled equation without the age adjustment



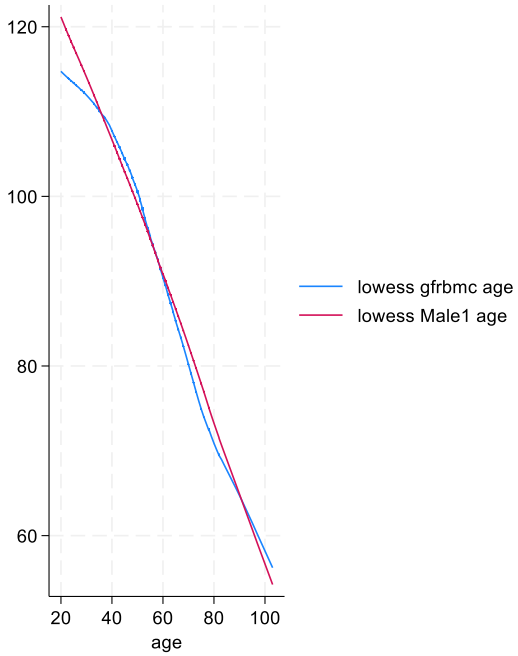
Appendix 9: Comparison of modelled Cystatin C based equations for males. The right graph representing the modelled equation with the age adjustment whilst the left graph is representing the modelled equation without the age adjustment

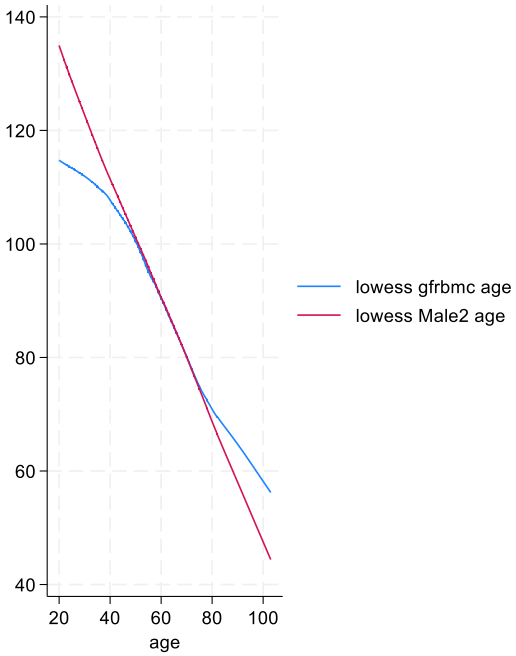


Appendix 10: Comparison of the modelled Creatinine-Cystatin C based equations for females against the mGFR. The right graph representing the modelled equation with the age adjustment whilst the left graph is representing the modelled equation without the age adjustment



Appendix 11: Comparison of the modelled Creatinine-Cystatin C based equations for males. The right graph representing the modelled equation with the age adjustment whilst the left graph is representing the modelled equation without the age adjustment





Appendix 12 Signed Turn it In Report

Turnitin Originality Report

Processed on: 02-Apr-2024 10:44 PM SAST
ID: 2338216262
Word Count: 12820
Submitted: 1

Research Report _ 2392604.pdf By Esnath Mudiwa

Similarity Index	Similarity by Source	
17%	Internet Sources:	12%
	Publications:	14%
	Student Papers:	4%

1% match ()
Kalyesubula, Robert, Fabian, June et al. "How to estimate glomerular filtration rate in sub-Saharan Africa : design and methods of the African Research into Kidney Diseases (ARK) study", 'Springer Science and Business Media LLC', 2020

1% match (Nicole Aumann, Sebastian E. Baumeister, Rainer Rettig, Wolfgang Lieb et al. "Regional Variation of Chronic Kidney Disease in Germany: Results From Two Population-Based Surveys", Kidney and Blood Pressure Research, 2015)
Nicole Aumann, Sebastian E. Baumeister, Rainer Rettig, Wolfgang Lieb et al. "Regional Variation of Chronic Kidney Disease in Germany: Results From Two Population-Based Surveys", Kidney and Blood Pressure Research, 2015

1% match (Internet from 08-Mar-2024)
https://www.researchonline.lshtm.ac.uk/id/eprint/4663553/2/2021_EPH_PhD_Kalyesubula_R-SR.pdf

1% match (June Fabian, Robert Kalyesubula, Joseph Mkandawire, Christian Holm Hansen et al. "Measurement of kidney function in Malawi, South Africa, and Uganda: a multicentre cohort study", The Lancet Global Health, 2022)
June Fabian, Robert Kalyesubula, Joseph Mkandawire, Christian Holm Hansen et al. "Measurement of kidney function in Malawi, South Africa, and Uganda: a multicentre cohort study", The Lancet Global Health, 2022

< 1% match (Internet from 12-Jun-2022)
https://www.coursehero.com/file/149775996/FINAL-PK-variation-Group-1-16Bpptx/

< 1% match (Internet from 15-Mar-2024)
https://WWW.coursehero.com/file/139805537/Chronic-Kidney-Failure-1docx/

< 1% match (Internet from 14-Mar-2022)
https://www.coursehero.com/file/129897807/The-Kidneysdocx/

< 1% match (Internet from 15-Sep-2023)
https://www.coursehero.com/file/205239435/Intradialytic-Blood-pressure-Changes-Among-patients-on-intermittent-Hemodialysisdoc/

< 1% match (Mindikoglu, Ayse L., Thomas C. Dowling, Matthew R. Weir, Stephen L. Seliger, Robert H. Christenson, and Laurence S. Magder. "Performance of chronic kidney disease epidemiology collaboration creatinine-cystatin C equation for estimating kidney function in cirrhosis", Hepatology, 2013.)
Mindikoglu, Ayse L., Thomas C. Dowling, Matthew R. Weir, Stephen L. Seliger, Robert H. Christenson, and Laurence S. Magder. "Performance of chronic kidney disease epidemiology collaboration creatinine-cystatin C equation for estimating kidney function in cirrhosis", Hepatology, 2013.

< 1% match (student papers from 11-Apr-2016)
Submitted to The University of Manchester on 2016-04-11

< 1% match (Xiao-Hua Pei, Juan He, Qiao Liu, Bei Zhu, Li-Hua Bao, Chen-Jing Yan, Jian-Qing Wu, Wei-Hong Zhao. "Evaluation of serum creatinine- and cystatin C-based equations for the estimation of glomerular filtration rate in a Chinese population", Scandinavian Journal of Urology and Nephrology, 2012)
Xiao-Hua Pei, Juan He, Qiao Liu, Bei Zhu, Li-Hua Bao, Chen-Jing Yan, Jian-Qing Wu, Wei-Hong Zhao. "Evaluation of serum creatinine- and cystatin C-based equations for the estimation of glomerular filtration rate in a Chinese population", Scandinavian Journal of Urology and Nephrology, 2012

< 1% match (Internet from 11-Oct-2021)
http://etd.astu.edu.et/bitstream/handle/123456789/1678/Dibaba%20Adeba.pdf?isAllowed=y&sequence=1

< 1% match (Lesley A. Inker, Andrew S. Levey. "Measurement and Estimation of Kidney Function", Elsevier BV, 2019)
Lesley A. Inker, Andrew S. Levey. "Measurement and Estimation of Kidney Function", Elsevier BV, 2019

< 1% match (Internet from 16-Sep-2021)
https://pafmj.org/index.php/PAFMJ/article/download/921/766/

< 1% match (Internet from 25-Jul-2020)
https://auth.aacc.org/publications/cln/articles/2016/april/cystatin-c-and-creatinine-complementary-markers-of-gfr-expert-