

# Appendix

## 1. Overall Approach

The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) is a research group funded by the National Institute of Diabetes, Digestive and Kidney Disease (NIDDK) to develop and validate improved glomerular filtration rate (GFR) estimating equations by pooling data from research studies and clinical populations (hereafter referred to as “studies”). Studies include individuals with diverse clinical characteristics, with and without kidney disease, and across a wide range of GFR. CKD-EPI developed estimating equations to relate measured GFR to predictor variables including serum levels of endogenous filtration markers and clinical characteristics. Equations were developed and internally validated in a database of 10 studies (6 research studies and 4 clinical populations) with a total of 8,254 participants, divided randomly into separate datasets for development (n=5,504) and internal validation (n=2,750). The equations were then externally validated in a database of 16 other studies with a total of 3,896 participants. This appendix reviews the sources of data and methods for model development and validation for equations using serum creatinine as the endogenous filtration marker.

## 2. Sources of Data

We identified studies by searching Medline as well as use of investigators’ and collaborators’ contacts. Inclusion criteria comprised: measurement of GFR using exogenous filtration markers; ability to calibrate the serum creatinine assay; experience of collaborators in GFR measurement, creatinine assay, and clinical investigation; and willingness of collaborators to share individual patient data. **Appendix Figure 1** shows the search and retrieval strategies for the studies and their allocation to development, interval validation or external validation datasets. We restricted the development and internal validation datasets to ten studies using urinary clearance of iothalamate for GFR measurements (“iothalamate studies”). A random selection of 2/3 of the data in this first set were used for development and the remaining 1/3 for internal validation. Data from some additional iothalamate studies as well as studies using filtration markers other than iothalamate (“non-iothalamate studies”) to form a second dataset for external validation so as to include similar populations in both datasets. Both sets of studies included randomized studies as well as clinical populations. Clinical populations were subdivided into people with known or suspected CKD and healthy individuals being evaluated for potential kidney donation. The first GFR measurement was taken if studies included more than one measurement. **Appendix Tables 1** and **2** show the study and participant characteristics for each study.

## 3. Model Development

We sought to develop a new equation to predict measured GFR that would perform better with respect to bias, precision and accuracy compared to the MDRD Study study equation. To reflect the known multiplicative relationship between GFR and serum creatinine as well as to satisfy the stable variance assumption of linear regression, we transformed both GFR and serum creatinine to the natural logarithmic scale for all model development. All models included forms of the four predictor variables included in the MDRD Study study equation (age, sex, race and creatinine) as well as additional variables and interactions among all variables.

Initial model development focused on finding optimal forms of the continuous predictors (age and serum creatinine). Using parametric approximations to non-parametric smoothing splines in generalized additive models, we explored linear, polynomial, logarithmic and spline transformations of age and linear, polynomial and spline transformation of log serum creatinine. Parametric transformations were considered adequate representations of the non-parametric variables if the difference in model deviance was not significant by the approximate chi-square test of nonlinearity (31). We ranked different parametric models by their goodness-of-fit using  $R^2$  and likelihood ratio tests (for nested models), favoring simpler forms over more complex ones.

We determined that age expressed on the natural scale performed better than age expressed on the log scale, as had been used in the MDRD Study study equation. Therefore, we used the natural scale thereafter. We determined that the best forms for age were linear and a piecewise linear spline with one knot at 40 years. For log serum creatinine, the best forms were linear and a piecewise linear spline with a single knot that differed for men and women. For men, the knot was at a serum creatinine value 0.9 mg/dL and for women, the knot was at a value of 0.7 mg/dL. These two transformations (linear and spline) for age on the natural scale and for log serum creatinine were combined to form four families of models as shown in **Appendix Table 3**.

Within each model family, types of models were defined by the presence or absence of interactions and/or new variables (**Appendix Table 4**). The first type of model included only variables from the original MDRD study

equation: serum creatinine, age, sex, and race without interactions (main effects). Second, more complex type of models were developed adding new variables and/or interactions to the original four variable base models. The new variables included diabetes, transplant status, and weight because of their hypothesized relationship with non-GFR determinants of serum creatinine, in particular creatinine generation, and because they were uniformly available in all studies. As a result of the concern that inclusion of weight in models might hinder reporting by clinical laboratories, models were developed both with all three new variables, and also with only diabetes and transplant.

In each model family, we first tested each new variable alone and only considered it further if it met two conditions. First, its main effect or its interaction with serum creatinine had to be significant at a 0.01 level of significance; and second, it had to improve model performance compared to the model excluding it under the criteria described below in Section 4. All variables that met these criteria were included along with the terms from the original four variables in a backward stepwise selection process, with a separate model development process for each family of models. New variables significant in any one model family were retained in all families. Interaction models were developed by testing pair-wise interactions among all variables included in each model. A backward selection procedure was used to select significant individual interaction terms at a significance level of  $< 0.001$ . We also tested three-way interactions among sets of three variables which had significant two-way interactions among all three subsets of pairs of variables.

#### **4. Model Performance**

We assessed model performance in the development dataset overall and in subgroups defined by ranges of estimated GFR and clinical characteristics. Ranges of estimated GFR were  $<15$ , 15-29, 30-59, 60-89, 90-119, and  $\geq 120$  ml/min/1.73 m<sup>2</sup> (multiply by 0.0167 for GFR in mL/s/m<sup>2</sup>). Clinical characteristics included: age ( $<40$ , 40-65,  $>65$  years); sex; race (Black or White and other); diabetes (yes, no); prior organ transplant (yes, no); and body mass index (BMI,  $<20$ , 20 to 25, 25-30 and  $>30$  kg/m<sup>2</sup>).

Root mean square error (RMSE) was used as the primary measure of model fit because it measures precision (bias is expected to be zero in the development dataset), is not affected by the ranges of the predictor and outcome variables (unlike  $R^2$ ), and is computed on the same scale as the predictor and outcome variables (unlike bias, absolute bias, or P30). We selected a minimum relative change in RMSE of 2% as the criterion defining improvement because a smaller change is not likely to be associated with a clinically important change in performance, even in combination with other significant variables.

Within each model family, we compared the more complex models to simpler models that included only the original four variables. We brought models forward to internal validation if the more complex models if their performance improved in the overall dataset or in a subgroup based on estimated GFR or relevant clinical characteristics (defined by the variables involved in the main effect or interaction). In addition, we examined confounding by study using two methods: cross-validation and testing of the statistical significance of study terms and study interactions included in models. If the relative performance of the models was different with removal of multiple studies (tested one at a time) or within multiple individual studies, then such models would not have been brought forward to internal validation.

#### **5. Internal Validation**

Models developed were internally validated in the remaining 1/3 of the data. Validation consisted of assessing consistency of the significance of coefficients and of performance and between the development and internal validation data. Model performance was compared within a model family overall and within subgroups using RMSE, difference between 5<sup>th</sup> and 95<sup>th</sup> percentile of the difference between measured GFR and estimated GFR, and percent bias. Models with consistent performance in the two datasets were refit in the combined development and internal validation datasets to determine final coefficients and were brought forward into external validation.

#### **6. External Validation**

In external validation, we sought to assess generalizability of the developed models when applied in populations of similar characteristics and to select the best model for use in clinical practice. Five steps were used to select models, as described below.

1. *Validation.* Validity of models was assessed by comparing performance in the combined development/internal validation dataset to that in the external validation dataset and to the MDRD Study study equation. RMSE was the

primary metric used. If model performance remained as good as or better than the MDRD Study study equation, we then considered models as valid

*2. Pooling of iothalamate and non-iothalamate studies.* In order to determine the appropriateness of combining iothalamate and non-iothalamate studies, we compared model performance in the two sets of studies. To maximize the similarity between the source populations in the iothalamate and non-iothalamate studies, we restricted this comparison to the subset of patients with CKD excluding organ transplant recipients (n=1,876). Performance was evaluated as bias in addition to the RMSE as we anticipated that differences among markers would be more likely to affect bias than precision. As the differences in bias and RMSE were less than 3 mL/min per 1.73 m<sup>2</sup> and 3%, respectively, we considered the two sets of data to be combinable.

*3. Selection of Models with Alternative Forms of Age and Creatinine.* To determine the most appropriate form of the creatinine variable, we compared the RMSE and bias of models of the spline to linear log serum creatinine in estimated GFR subgroups 60-89 and 90-119 mL/min/1.73 m<sup>2</sup>. Similarly, because the spline transformation of age most affected the younger participants, we compared the spline to linear age in the group of patients under 40 years of age. For both variables, the most appropriate form was determined if the form improved performance in the subgroups by at least 2% for RMSE and 2 ml/min per 1.73 m<sup>2</sup> for bias. The best model family was determined by assessment of the improved performance combination of the creatinine and age forms. In addition to all models within the selected model family, we also carried forward the simplest models that used linear relationships of both log serum creatinine and age on the natural scale.

*4. Comparison of Main Effects Models to Interaction Models.* After choosing the appropriate forms of the creatinine and age variables, we investigated the need to include interactions in models that included: (1) age, sex, race and creatinine; (2) age, sex, race, creatinine, diabetes, transplant and weight; and (3) age, sex, race, creatinine, diabetes and transplant. Here, models were compared across model families. Bias was the primary metric; and RSME and other measures of performance were secondary metrics. Performance was first examined overall and then in subgroups defined by estimated GFR and clinical characteristics, with emphasis placed on the subgroups of estimated GFR 60-89 and 90-119 mL/min/1.73 m<sup>2</sup> and subgroups defined by clinical characteristics modeled by the new variables. Improved performance was evaluated using the same criteria described in step 3. The best model from each comparison was carried forward to the next phase.

*5. Ranking Models.* We next compared selected models to the MDRD Study study equation as well as the main effects models for log creatinine and age on the natural scale. Performance was assessed overall and within subgroups defined by estimated GFR and clinical characteristics using RMSE and bias. The best model was that model with the lowest bias and RMSE. Models that were “close” to the best model were defined as those within 2% of the RMSE and 0.5 mL/min/1.73 m<sup>2</sup> of the bias compared to the best model. Final ranking among these selected models were made on the basis of the combination of the ranking for the whole dataset, ranking in subgroups, and ease of application in clinical practice.

For all steps, we performed sensitivity analyses to evaluate the robustness of the results to potential confounding by study. These analyses included investigating the consistency of model form within each study and the use of cross-validation to determine if the removal of one study changed the decisions regarding model selection.

## 7. Equation Performance

**Appendix Table 5** compares measures of equation performance in the development, internal and external validation dataset, stratified by level of GFR.

## 8. Prevalence Estimates

**Appendix Table 6** compares classification of eGFR categories in NHANES using the CKD-EPI and MDRD Study study equations. **Appendix Table 7** provides prevalence estimates for GFR stages in NHANES and the US population, stratified by level of albuminuria. **Appendix Table 8** provides prevalence estimates for CKD overall and by stages, stratified by sex, race and age. Number of subjects in NHANES was 16,032 (including two with CKD-EPI estimated GFR < 15) and in subgroups was as follows: Men (8,312), women (7,720), Non-Hispanic Whites (8,199), Non-Hispanic Blacks (3,179), ages 20-39 (5,131), 40-59 (5,091), 60-69 (2,601) and >70 (3,209). US population base for prevalence estimates was 200,948,641 and in subgroups was as follows: Men (96,808,352), women (104,140,289), Whites (143,101,175), Blacks (19,968,057), ages 20-39 (81,562,389), 40-59 (73,589,052), 60-69 (20,338,992) and >70 (25,458,208).

## **Investigators of CKD-EPI Aim 1 and Aim 2:**

**Tufts Medical Center, Boston, MA:** Andrew S. Levey, MD, Lesley A. Stevens, MD MS FRCP(C), Christopher H. Schmid, PhD, Yaping (Lucy) Zhang, MS

**Cleveland Clinic Foundation, Cleveland, OH:** Frederick Van Lente, PhD, Liang Li, PhD

**University of Utah, Salt Lake City, UT:** Tom Greene, PhD

**John Hopkins University, Baltimore, MD:** Josef Coresh, MD PhD MHS, Jane Manzi, PhD, Brad Astor PhD MPH, Elizabeth Selvin, PhD MPH

**University of Pennsylvania, Philadelphia, PA:** Harold I. Feldman, MD MSCE, J. Richard Landis, PhD, Marshall Joffe, MD MPH PhD

**National Institute of Diabetes and Digestive and Kidney Diseases:** John W. Kusek, PhD, Paul W. Eggers, PhD  
Collaborators of CKD-EPI Aim 1 and Aim 2:

**African American Study of Kidney Disease and Hypertension (AASK):** Gabriel Contreras, MD MPH, Julia B. Lewis, MD

**Baylor University Medical Center:** Goran B. Klintmalm, MD PhD

**Captopril in Diabetic Nephropathy Study (CSG):** Roger A. Rodby, MD, Richard D. Rohde, BS

**Chronic Renal Insufficiency Cohort (CRIC):** Harold I. Feldman, MD MSCE, Lawrence J. Appel, MD MPH, Jing Chen, MD MS, Alan S. Go, MD, Lee Hamm, MD, Chi-yuan Hsu, MD, James P. Lash, MD, Akinlolu O. Ojo, MD, Mahboob Rahman, MD, Raymond R. Townsend, MD, Matthew R. Weir, MD, Jackson T. Wright, MD

**Cleveland Clinic Foundation (CCF):** Phillip Hall, MD, Emilio Poggio, MD

**Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP):** Vicente Torres, MD PhD

**Dallas Nephrology Associates (DNA):** Larry Melton, MD PhD FACP, Ruben Velez, MD

**Diabetes Control and Complications Trial (DCCT):** Saul Genuth, MD, Michael W. Steffes, MD PhD

**Diabetic Renal Disease Study Group (DRDS):** Robert G. Nelson, MD PhD

**Groningen Renal Hemodynamic Cohort Study Group (GRECO):** Gerjan Navis, MD PhD

**International Diabetic Nephropathy Study Group (IDNSG) and the Renin-Angiotensin System Study (RASS):** Michael Mauer, MD

**Departments of Clinical Chemistry and Nephrology, University Hospital, Lund, Sweden:** Anders Grubb, MD PhD, Omran Bakoush, MD PhD

**Mayo Clinic:** Andrew D. Rule, MD MS, Timothy Larson, MD, Fernando Cosio, MD

**Modification of Diet in Renal Disease (MDRD) Study:** Gerald Beck, PhD

**NephroTest Cohort:** Jerome Rossert, MD PhD, Marc Froissart, MD PhD

**Steno Diabetes Center:** Hans-Henrik Parving, MD PhD, Peter Rossing, MD DMSc

## **Scientific Advisory Committee for CKD-EPI:**

Allan Collins, MD FACP, University of Minnesota, Minneapolis, MN

Gary Curhan, MD ScD, Brigham and Women's Hospital, Boston, MA

Ralph D'Agostino, PhD, Boston University, Boston, MA

John Eckfeldt, MD PhD, University of Minnesota, Minneapolis, MN

Adeera Levin, MD FRCP(C), University of British Columbia, Vancouver, BC