|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT468 | |
| **Submission Date** | 1/6/2023 | |
| **Project Title** | **Third Generation Polygenic Scores for Kidney Disease** | |
| **Tentative Lead Investigator** | Atlas Khan | |
| **Tentative Lead Email Address** | ak4046@cumc.columbia.edu | |
| **Tentative Senior Author** | Krzysztof Kiryluk ([kk473@columbia.edu](mailto:kk473@columbia.edu)) | |
| **All Other Authors** | eMERGE CKD workgroup members  + any other interested eMERGE investigators | |
| **Sites Participating** | All other eMERGE sites/investigators are invited to participate | |
| **Background / Significance** | As part of our eMERGE-IV consortium efforts, we have recently developed, optimized, and validated a genome-wide polygenic risk score for chronic kidney disease that performs reasonably well across diverse ancestral groups (Khan et al. Nat Med 2022). We refer to this score as a “second generation” score, since it provides significant improvements over the “first generation” scores developed based on less powerful GWAS studies and involving smaller number of variants with weights not optimized for cross-ancestry performance. However, the methods for modeling polygenic risk continue to improve rapidly. Moreover, much larger GWAS studies for kidney function and for primary kidney disorders are now becoming available. Therefore, there is a need to further improve the existing polygenic risk score model for CKD by utilizing new methods and incorporating new GWAS datasets. | |
| **Outline of Project** | We propose to develop an improved “third generation” polygenic score using 1) new “discovery GWAS” meta-analyses for renal function in 2.5 million individuals (tripling the prior discovery sample size), 2) new methods for modeling polygenic risk across ancestries, 3) additional modifications to improve score performance in individuals of African ancestry, such as joint modeling of APOL1 and Sickle Cell Trait genotypes that are known to have a large effect on the risk of CKD in African Americans, 4) combining scores for renal function with scores for other related traits, such as albuminuria, and/or specific kidney disease subtypes (e.g. diabetic nephropathy, IgA nephropathy, membranous nephropathy, etc.), 5) combining polygenic predictors with monogenic risk, family history, and clinical risk factors for kidney disease to further improve risk prediction. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Geocoding 2015 ACS variables’  Other: Case/Control status on Phase I-III phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)* | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): future eMERGE-IV GWAS dataset | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list: CKD Phenotype (including CKD subtypes)  No | |
| **Planned Statistical Analyses** | 1. We will modify our CKD e-phenotype to include the latest eGFR equation (i.e., CKD EPI equation without race). We will then re-phenotype all eMERGE-III and IV individuals using our updated CKD e-phenotype. 2. We will use the new discovery GWAS for eGFR (unpublished) that involves 2.5 million individuals to derive weights for the new risk score equation. 3. We will optimize the score using multiethnic cohorts from the All-of-Us project (release 1). We have already performed genome-wide imputations in All-of-Us for this purpose. We will explore a number of newer methods including fine-mapping, machine learning, functional prioritization, and multi-phenotype analysis of kidney related traits (e.g., albuminuria, primary kidney diseases). We will select the best performing score using the metrics developed by the E-IV PRS WG. 4. We will test the performance of the best score using future releases of the All-of-Us data, and future eMERGE-IV datasets, as well as additional African American cohorts from UAB and BioMe that are not included in the discovery GWAS (similar to the second generation score). | |
| **Ethical Considerations** | N/A | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Risk score development and optimization: completed by December 2023  Risk score testing in eMERGE and External Cohorts: completed by December 2024  Study completion and journal submission” End of 2024. | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants