|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT409 | |
| **Submission Date** | 10/23/2020 | |
| **Project Title** | Development, validation, and testing polygenic risk score(s) for kidney disease | |
| **Tentative Lead Investigator** | Atlas Khan | |
| **Tentative Lead Email Address** | ak4046@cumc.columbia.edu | |
| **Tentative Senior Author** | Krzysztof Kiryluk (kk473@columbia.edu) | |
| **All Other Authors** | Columbia: Ning Shang, Cong Liu, Wendy Chung, Chunhua Weng, George Hripcsak  UAB: Nita Limdi, Hemant Tiwari, Ryan Irvin  Mt. Sinai: Eimear Kenny, Girish Nadkarni  + any other interested eMERGE investigators | |
| **Sites Participating** | All other eMERGE sites are invited to participate | |
| **Background / Significance** | Chronic kidney disease (CKD) affects 10% of the population and represents a significant expense to healthcare systems in the US and globally. Polygenic risk score (PRS) is an emerging tool that has shown a promising predictive power for risk assessment of common complex disease. PRS has been studied for a variety of different complex traits such as coronary artery disease, type 2 diabetes, body mass index, etc. We propose to develop and test a PRS for kidney disease using existing GWAS summary statistics and diverse transethnic populations enrolled in eMERGE-III and IV and external datasets. Independent testing stage will enable unbiased estimation of effect estimates for return of results to diverse participants of eMERGE-IV prospective study. | |
| **Outline of Project** | In the proposed study, we will derive kidney PRSes using different statistical methods and summary statistics from GWAS of several kidney-related traits, including the latest/largest GWASes for eGFR, eGFR slope, CKD stage 3, albuminuria, and primary kidney diseases. We will then validate and test the performance of these PRSes in the eMERGE-III/IV genotyped cohorts and additional external datasets, including the UKBB dataset. CKD cases and controls will be defined using our previously validated eMERGE phenotyping algorithms based on renal function lab tests and ICD/CPT codes. We will compare the performance of various CKD risk scores between major ancestral populations, including African, Latinx, East Asian, and European. We will perform sensitivity analyses and test the performance of the final risk score within the following subgroups: 1) age-defined (pediatric, adult) and 2) CKD subtype-defined (diabetic, hypertensive, glomerular, cystic, congenital, or unknown). We will also model the incorporation of *APOL1* risk genotype into the risk score model for patients of African ancestry. We will specifically explore the effects of our best-performing PRS on the risk of kidney disease stratified by *APOL1* risk genotype in individuals of African American and Afro-Caribbean Hispanic/Latinx ancestry. We will also test for PRS\**APOL1* genetic interactions, and if significant, we will incorporate interaction terms in the genetic risk prediction model. | |
| **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 and Phase II 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): | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list: CKD Phenotype (we plan a minor modification to include CKD subtype information)  No | |
| **Planned Statistical Analyses** | Genome-wide PRS will be derived using the Pruning & Thresholding method and LDPred algorithm by using ancestry-specific LD reference panel based on 1000 Genomes. We will also explore the impact of including admixture information on the risk score performance. Each tested PRS score will be standardized to a mean=0, standard deviation=1 prior to analysis in order to facilitate the interpretation and to standardize the comparison of PRS effects. We will optimize and test each PRS in the eMERGE-III/IV, UK Biobank, and several other available datasets. The association of each PRS and the CKD status will be tested using a multivariable logistic regression adjusting for age, sex, diabetes, eMERGE site (batch for UKBB), and principal components of ancestry. In the testing dataset, we will assess performance by comparing risk scores distributions between cases and controls, ROC AUC, precision-recall AUC, and R2 (variance explained). We will examine these metrics in predicting 3 severity-based categories of CKD: CKD of any stage, CKD stage 3 or greater, and ESRD defined by estimated GFR<15 or chronic need for renal replacement therapy (dialysis or transplant). In the analyses and subsequent publication, we will follow the general PRS reporting guidelines as proposed by the ClinGen PRS workgroup. The validated scores and their standardized performance metrics will be deposited in the PGS Catalogue. | |
| **Ethical Considerations** | N/A | |
| **Target Journal** | Depending on the results: JAMA, Ann Int Med, Nat Comm, or JASN. | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Implementation of phenotyping algorithms and data analysis by March 2021  Draft and circulate the manuscript among co-authors by April 2021  Initial journal submission (May 2021) | |

**\*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