|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT403 | |
| **Submission Date** | 8/18/2020 | |
| **Project Title** | Targeted validation of MR instruments for varicose veins | |
| **Tentative Lead Investigator** *(first author)* | Pradeep Suri, MD, MS (external collaborator, University of Washington) | |
| **Tentative Lead Investigator Email Address** | psuri@uw.edu | |
| **Tentative Senior Author**  *(last author)* | Gail Jarvik, MD, PhD | |
| **eMERGE Site Sponsor & Contact** | Gail Jarvik, MD, PhD (Sponsor); Ian Stanaway (Contact) | |
| **All Other Authors** | Ian Stanaway, PhD; Yakov Tsepilov, PhD; Alexandra Shadrina, PhD | |
| **Sites Participating** | All | |
| **Background / Significance** | Varicose veins (VVs) is a common condition affecting nearly 30% of the population. Despite intensive research efforts, pathogenesis and risk factors for this condition are still not fully understood.  2-sample Mendelian randomization (2SMR) is a technique which allows to infer causal relationships between phenotypes without the need to conduct randomized controlled trials. 2SMR is performed using GWAS results for potential “exposure” and “outcome” phenotypes (<https://doi.org/10.7554/eLife.34408>).  In a recent study, our collaborators Drs. Tsepilov and Shadrina applied 2SMR and performed a hypothesis-free search for causal relationships between 2,221 phenotypes and VVs (<https://doi.org/10.1371/journal.pgen.1008110>). Using UK Biobank data and GWAS results for blood circulating proteins measured with the SOMAscan platform, they showed that increased plasma levels of the MICB and CD209 (DC-SIGN) proteins may be risk factors for VVs. Therefore, we conclude that these proteins may be involved in the pathogenesis of VVs.  Validation of the results of *in silico* analysis in independent cohorts is an important step towards providing robust and clinically useful evidence. In the current analysis, we will attempt to replicate the Mendelian randomization results for MICB and CD209 levels as an “exposure” and VVs as an “outcome” phenotype. We have already selected instrumental variables (SNPs) for the analysis and obtained data on their associations with MICB and CD209 plasma levels in cohorts different from those used in our previous study. The next step is obtaining data on their association with VVs to combine these data in Mendelian randomization tests. Thus, this MCS seeks to conduct genetic association analyses to examine the association of 25 SNPs with VVs. | |
| **Outline of Project** | For this work, we plan to use eMERGE3 data that is currently available from the coordinating center.   1. We will use phecodes, ICD9/10 and CPT codes to define VV phenotypes. The target phenotype of interest is “varicose veins of the lower extremities” or “varicose veins”. 2. We will perform genetic association analyses for a limited list of 25 variants as described below, using the 105,108 imputed merged data set | |
| **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)*** | None other than 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  No | |
| **Planned Statistical Analyses** | * We will use existing ancestry principal components analysis across all samples, and within the largest ancestry subsets * We will define phenotypes using phecodes. * We may alter phecode algorithms using ICD9/10 codes or CPT codes. * The association of the following SNPs (selected as instrumental variables) with VVs will be examined:   + rs519477   + rs2281438   + rs2203043   + rs11653587   + rs505922   + rs10120246   + rs8106657   + rs10405336   + rs11905584   + rs2236570   + rs4787101   + rs3094005   + rs1264347   + rs7775397   + rs1573296   + rs404240   + rs3130837   + rs9276731   + rs34218844   + rs41269265   + rs35555795   + rs6940698   + rs11966717   + rs1107675   + rs7229090 * We will conduct logistic regressions of imputed SNPs with an additive genotype model in R   + With all samples, including adults age 18 years or older   + Adjusting for sex, age, ancestry principal components   + We will analyze EAs only   + We will provide the following results to our external collaborators * effective/reference allele * effective allele frequency * beta of the logistic regression * SE (beta) * *P*-value * Number of patients in case/control group * *P* (Hardy-Weinberg equilibrium) * Call rate * Info metric of imputation quality | |
| **Ethical Considerations** | Using de-identified data only, there are no ethical concerns. | |
| **Available Funding or Resources** | None | |
| **Target Journal** | We expect 1 manuscript related to these results. We hope to submit to PLOS Genetics, but this expected destination may change. | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Dates of expected completion   1. Perform analysis –September 2020. 2. Await results of other analyses by collaborators and draft manuscript – November 2020 3. Submit manuscript for publication- January 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