Race- and sex-specific disease trajectory in individuals with rare variants in seven cerebrovascular small vessel disease genes: a genotype-first population study

Jiang Li[1], Vida Abedi[1,2], Durgesh Chaudhary[3], Regeneron Genetics Center[7], Oded Goren[4], David J. Carey[1], Karen E. Wain[5], Ramin Zand[3,6]

[1]Department of Molecular and Functional Genomics, Weis Center for Research, Geisinger Health System, Danville, Danville, PA 17822, USA

[2]Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA.

[3]Neuroscience Institute, Geisinger Health System, Danville, PA 17822, USA.

[4]Department of Neurosurgery, Geisinger Health System, Danville, PA 17822, USA

[5]Autism & Developmental Medicine Institute, Geisinger Health System, Danville, PA 17822, USA.

[7]Regeneron Genetics Center, LLC, Tarrytown, NY 10591, USA

[6]Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA

Corresponding Author:

Ramin Zand MD, MPH, FAHA.

Email: ramin.zand@gmail.com

Jiang Li Ph.D

Email: jli@geisinger.edu

Conflict of Interest Disclosures: J Li, D Chaudhary, DJ Carey, O Goren, KE Wain, R Zand, and V Abedi report no disclosures relevant to the manuscript.

Word count:

Figures: 5

Tables: 1

Supplementary Figures and Tables: 9 and 1

Supplementary Material for Method.

Keywords: monogenic disease, RV, IS, cerebral small vessel disease, age of onset, relative risk, percent relative effect, burden test, survival analysis, Cox proportional hazards regression, hazard ratio, electronic health records.

Glossary

ACMG = American College of Medical Genetics and Genomics; **AFR** = African ancestry; **CI** = confidence interval; **CoxPH** = Cox proportional-hazards model; **CSVD** = cerebral small vessel disease; **EHR** = Electronic Health Record; **eQTL** = expression quantitative trait loci; **EUR** = European ancestry**; GO** = Gene Ontology; **GWAS** = genome-wide association study; **Hazards ratio** = HR; **HGMD** = Human Gene Mutation Database; **HWE** = Hardy-Weinberg Equilibrium; **ICD** = International Classification of Disease; **IS** = IS; **LAS** = large-artery strokes; **LD** = linkage disequilibrium; **MAF** = minor allele frequency; **ML** = machine learning; **OR** = odds ratio; **PSEA** = PheCode Set Enrichment Analysis; **PTV =** protein truncation variants; **RD** = risk difference**; RR** = relative risk; **RV** = RV; **PCA** = Principal Component Analysis; **PRE** = percent relative effect; **PRS** = polygenic risk score; **SNP** = single nucleotide polymorphism; **TIA** = transient ischemic attack; **VEP** = Ensembl Variant Effect Predictor; **VUS** = a variant of uncertain significance.

Key Points:

Question Is there an increased risk for a lifetime or early onset of cardiovascular diseases and others within or beyond the circulatory system observed in rare variant carriers of seven well-characterized monogenic cerebral small vessel disease (CSVD) risk genes (*COL4A1, COL4A2, NOTCH3, HTRA1, TREX1, CTC1,* and *GLA*)?

Findings In this retrospective longitudinal study of 175K individuals, which were temporally split into discovery (2738/5490) and replication cohorts (1695/3410) for carriers/matched noncarriers, *NOTCH3* ($HR_{earlyonset}$ = *2.175* for White), *TREX1* ($HR_{earlyonset}$ = *4.006* for White), and COL4A1/2 ($HR_{earlyonset}$ = *2.163* for Black) were associated with ischemic stroke, particularly in female participants. *HTRA1* also showed a similar trend. Circulatory diseases were overrepresented in discovery and replication cohorts with an early onset of cardiovascular phenotypes irrespective of race.

Meaning Carriers for monogenic CSVD risk genes demonstrated the increased risk for the lifetime or early onset of cardiovascular disease and diseases within or beyond the circulatory system, some of which are race- and sex- dependent.

Abstract:

Background: This study is aimed to explore the risk for a lifetime or early onset of cardiovascular diseases and diseases within or beyond the circulatory system observed in rare variant(RV) carriers of seven well-characterized monogenic cerebral small vessel disease (CSVD) risk genes (*COL4A1, COL4A2, NOTCH3, HTRA1, TREX1, CTC1,* and *GLA*) using a genotype-first approach within a hospital system population-based biobank cohort.

Method: This was a retrospective longitudinal study. MyCode participants with sequenced exomes were temporally split into discovery (n=92,445) and replication (n=81,130) cohorts. A workflow was created to prioritize potentially pathogenic variants by integrating three variant annotation pipelines. After propensity score matching, the number of participants in both discovery and replication cohorts was 2738/5490 and 1695/3410 for carriers/noncarriers, respectively.

Result: Most of the RVs identified were in a heterozygous form and disproportionately present in participants with African ancestry. Carriers showed an increased risk for early signs and symptoms of cerebrovascular disease. Cox regression model showed *NOTCH3* (European), *TREX1* (European), and COL4A1/2 (African) were associated with ischemic stroke. Circulatory diseases were overrepresented in discovery and replication cohorts with an early-onset of cardiovascular phenotypes irrespective of race. Sex- and race-dependent effect for cerebrovascular disease risk was clearly detectable, particularly for *NOTCH3(*$HR_{earlyonset}$ = 2.175[1.391-3.403], p = 0.001*),* and

TREX1($HR_{earlyonset}$ = 4.006[1.797-8.931], p = 0.001). *COL4A1/2*($HR_{earlyonset}$ = 2.163[0.87-5.38], p = 0.097). *HTRA1* (European) also showed a similar trend of association.

Conclusion: Carriers for monogenic CSVD risk genes demonstrated the increased risk for the lifetime or early onset of cerebrovascular disease and diseases within or beyond the circulatory system, some of which in a race- and sex-dependent manner. Our findings support the concept of developing a gene panel of CSVD for population screening of patients with early-onset circulatory diseases.

1. Introduction

Exome sequencing has made it possible to identify rare variants(RVs) in known genes causing monogenic stroke at both familial and population levels[1-4]. Previous attempts have been made to compile a comprehensive gene panel based on earlier linkage studies of monogenic stroke with Mendelian inheritance[5,6]. Whether used for familial[5] or population[7,8] screening, this technology, which combines exome capture and targeted sequencing, has demonstrated its potential to identify risk for monogenic stroke with early[5] or late onset[7,8].

However, some obstacles to promoting exome sequencing in stroke prediction include: 1) Limited knowledge of lifetime incidence of disease among individuals with RVs identified from risk genes; 2) Subtype dependency of risk genes leading to varied effect size[9]; 3) Variable expressivity of phenotypes in patients with rare heterozygous variants for monogenic diseases[10,11]; 4) Low population-based penetrance of pathogenic variants which could be modified by age for late-onset diseases, even for autosomal dominant disorders[10,12-14]; 5) Race- and sex-dependency of cerebral small vessel disease (CSVD) risk genes in patients harboring RVs and complex disease profiles[6,15]; 6) Early signs, symptoms (e.g. transient ischemic attack (TIA)[16], hypercoagulability[17], atrial fibrillation[18]), and patterns of comorbidities that have not been explored in individuals with RVs – leading to opportunities for personalized prevention strategy; 7) Development of a variety of small vessel-related phenotypes beyond the brain[19] and circulatory system[20-22] at the early stage in carriers. Finally, stroke is an ultimate expression of a systemic failure in arteriolar function[23].

Geisinger's comprehensive Electronic Health Record (EHR) intersected with MyCode exome sequencing for genomic screening make it possible to identify potentially pathogenic variants in biobank-leveled population and associated disease profile and trajectory[1,10,24,25].

The purpose of this study is to identify rare, potentially pathogenic variants of seven well-characterized monogenic CSVD risk genes (*COL4A1, COL4A2, NOTCH3, HTRA1, TREX1, CTC1,* and *GLA*) using a genotype-first approach within a healthcare system population-based biobank cohort. The goal is to explore the clinical manifestations observed in variant carriers and assess the risk for the lifetime or early onset of CSVD and other diseases within or beyond the circulatory system.

2. Method:

2.1 Participating cohorts

This was a retrospective longitudinal study with a two-step design. The Geisinger patient population with comprehensive EHR was intersected with the MyCode initiative for exome sequencing of blood samples, which was performed through the DiscovEHR collaboration between Geisinger and the Regeneron Genetics Center (Tarrytown, New York) [1,24,26,27]. The Geisinger Institutional Review Board approved this study to meet "Non-human subject research" using de-identified information. All research was performed in accordance with relevant guidelines/regulations. Informed consent was obtained from all participants and/or their legal guardian(s). The participants and analysis pipeline are illustrated in **Figure 1**.

The MyCode participants, a predominantly adult population with European ancestry, were temporally split into discovery (n=92,445) and replication (n=81,130) based on the time of dataset availability.

2.2 Genetic data

Probes (xGen Exome Research Panel v1.0.) from Integrated DNA Technologies (IDT) were used for the exome capture. Sequencing was performed by 75bp paired-end reads on Illumina HiSeq2500. Coverage depth was sufficient to provide more than 90% coverage for 99% of samples. Samples with low-quality sequence data or showing discordance of sex from self-reported sex or genotypes from array-based data were removed.

Alignments and variant calling were based on the human GRCh38 reference genome. Variant calls were produced by the WeCall variant caller (Genomics PLC, Version 1.1.2). A project variant-call file (pVCF) was compiled using the GLnexus[28] joint genotyping tool (version 1.1.3-4). Sample level variant calls with site read depth (DP) <15, Phred-scaled genotype quality score (GQ) < 40, and alternate allele balance <15% (or fewer than 5 alternate reads) were removed. The call rate was 99.97%. Participants with the missing genotyping call (<0.001%) were considered null for the rare allele.

2.3 Prioritization of potentially damaging rare variants

RVs from seven risk genes for monogenic CSVD were identified from pVCFs by three variant annotation pipelines [29]. They were *COL4A1* (OMIM:120130), *COL4A2* (OMIM:120090), *NOTCH3* (OMIM:600276), *HTRA1* (OMIM:602194), *GLA* (OMIM: 300644), *CTC1* (OMIM:613129), and *TREX1* (OMIM:606609) as shown in **Supplementary Figure 1**. Both *COL4A1* and *COL4A2* are located on 13q34. They are structurally expressed and typically function together. Thus, we treated them as one in this study. Variants were initially filtered by minor allele frequency (MAF) because the rate of pathogenic variation was lower due to selection constraints. Only variants with MAF < 0.001 from gnomAD and MAF < 0.01 from the MyCode population were considered as rare and subjected to the following variants annotation process. The MAF cutoffs were empirically determined based on the characteristics of each study population.

VEP annotation was conducted on 12/29/2019 using Ensemble web interface (https://useast.ensembl.org/info/docs/tools/vep/index.html), and the result in Supplementary Table 1 was updated on 01/24/2022.

Pipeline 1(P1): RVs with functional impact predicted to be 'moderate' by VEP annotation were subsequently filtered by HGMD (professional 2018.2 released) annotation[30]. Only variants predicted to be 'disease modifying' (DM) entered the final variant assessment process.

Pipeline 2(P2): RVs with functional impact predicted to be 'high' by VEP annotation were retained[31]. All of them are predicted loss-of-function (LoF) variants with enriched protein truncation variants (PTVs).

Pipeline 3(P3): RVs in ClinVar annotated as 'pathogenic' (P) or 'likely pathogenic' (LP), without conflicting assertions (as of 12/29/2019 database and updated on 01/24/2022 database) were selected[32]. ACMG interpretation criteria were applied[32] and no VUS variants were included.

For variants with minor allele count (MAC) < 3 in the entire MyCode population, participant-level quality control (QC) data (GT, DP, AD, GQ, SBPV) from pVCF file or individual CRAM file for mapping/alignment were accessed by Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv/home) to determine the final genotyping call.

Based on the above RV selection criteria(P1-3), we identified a total of 188 RVs in the discovery cohort. We further filtered out some RVs (n = 121) due to 1) conflict reports of significance based on ACMG criteria[32]; 2) more stringent sequencing QC score for rare frameshift variants; 3) clear evidence of benign variants reported by ACMG but identified by HGMD as 'disease modifying'; 4) prior knowledge based on a literature search (*in vitro* functional studies as well as pedigree studies). Our preliminary study

using *NOTCH3* as an example showed that adding this filtering process would help to enrich the pathogenic variants with higher penetrance (data not shown).

The prioritized 67 RVs, including their genotype counts as well as functional consequence (HGVSp) are illustrated in **Supplementary Figure 1** using UCSC Genome Browser (GRCh38/hg38). Additional tracks from ClinVar, HGMD, OMIM, and Pfam Domains (GENCODE database) were mapped to those variants.

2.4 Phenotype data

Geisinger billing codes for diagnostics at any encounter setting were mapped to the *International Classification of Disease (ICD)*-related ICD9 and ICD10 codes. Demographic information was extracted from the structured EHR. ICD-9 and ICD-10 diagnostic codes were combined and mapped to 1,866 hierarchical PheCodes[33], each representing a specific disease phenotype. Rollup PheCodes without clinical definition of disease were removed from the subsequent analyses. All the PheCodes were grouped into 18 disease categories, such as the Circulatory system, the Genitourinary system, and more, as shown in a longitudinal layout (Figure 1). Study participants were labeled with a PheCode if they had one or more PheCode-specific ICD codes. “Case” was defined as all study participants with the PheCode of interest, and “control” was defined as all study participants without the PheCode of interest during their lifetime or within a predefined observation window for early-onset (e.g., ≤65yrs). To avoid an uneven sampling of the disease trajectory, which might lead to temporal bias in the longitudinal study, we also tested the age cutoff at 60 years or 55 years. Sex, as a

biological variable, was included during the mapping process so that sex-specific PheCodes would not be mislabeled.

2.5 Comparability of variant carriers and non-carriers

Propensity score matching (PSM) to control confounding was applied to obtain the 'counterfactual' control group. Race, sex, and index age, which were denoted as the last active (last encounter) date, were identified as potential confounding factors. Other unmeasured confounding factors were assumed to be correlated with these factors and their confounding effects would be largely reduced after matching. They were selected as covariates in a Logistic Regression model to create propensity scores (R 'MatchIt' package). We chose "nearest neighbor matching without replacement" to create a more balanced carrier:noncarrier ratio at 1:2, as shown in **Supplementary Figure 2**; this step ensured that the standardized mean differences (SMD) for the matched variables were within ± 0.1. The number of participants in both cohorts was 2738/5490 and 1695/3410, respectively. For noncardiovascular-related common diseases, Risk Differences (RD) were < 0.1 for lifetime and early onset (≤65yr) in both cohorts, suggesting no selection bias between the carrier and noncarrier groups[34].

2.6 Outcome measures

Major outcome variables in this study included age of onset, relative risk (RR) and percent relative effect (PRE).[35]

Risk Ratio (Relative Risk, RR) is defined as, $Risk\ Ratio = \frac{Cl_e}{Cl_u}$, where ($Cl_e$) = cumulative incidence among the exposed subjects, and (Cl_u) is the cumulative incidence among

unexposed subjects. The relative risk (or risk ratio, RR) is an intuitive way to compare the risks for the two groups.

$RR - 1$ is defined as the percent change in the exposed group (carriers). We consider the unexposed group (noncarriers) as having 100% of the risk, and the phenotype expression of the exposed group, expressivity, is relative to that. PRE, or $RR - 1$, is alternatively interpreted as a normalized RD with a positive or negative value. The PRE for each phenotype after exposure is similar to the percentage change in gene expression after exposure. This parameter will be used for PheCode set enrichment analysis, a counterpart of gene set enrichment analysis, to calculate the enrichment score.

$$Percent\ Relative\ Effect = \frac{Risk\ Difference}{Cl_u} = \frac{Cl_e - Cl_u}{Cl_u} = RR - 1$$

A circo plot was created based on the top 50 PheCodes ranked by PRE and clustered by the processes. This representation shows the extracerebral disease manifestations and overrepresentation while the rainbow links represent the association between PheCodes in carriers.

2.7 Statistics

Categorical data are reported as number (percentage); continuous data are presented as mean (SD), as indicated. Statistical analyses were performed using R, version 3.6.

All mapped PheCodes were divided into common (with a better mapping performance[36]) or rare diseases by the prevalence (<1%) of the disease estimated from noncarriers. Fisher's exact test was conducted to determine the frequency difference between

PheCodes in carriers and noncarriers. Due to imbalanced case-control ratios for rare diseases which may lead to inflated statistical significance, only the associations between common diseases were reported. There were 756 common PheCodes mapped for 4433 participants, in which 695 or 580 phenotypes had case-control ratios greater than 1:100 (regarded as common diseases) for both cohorts, respectively.

Survival analyses and Kaplan-Meier estimator for cumulative incidence probability were conducted (R ‘survival’ and ‘survminer’ packages). Participants were considered at risk of diseases at lifetime (for up to 85 years of follow-up) or early onset (age of onset > 65 years was considered as null and were censored at the last active date or the date of data extraction (06/12/2021). Multivariate Cox proportional hazards model, adjusted for confounding variables (sex, index age, and 5 main principal components from PCA) was also performed and the hazard ratio for phenotypes of interest at lifetime as well as early-onset (≤65 years) was calculated. The proportional hazards assumption was verified using Schoenfield residuals. For all analyses, $p<0.05$ was considered statistically significant. For all post-hoc pairwise tests, p-values were present as raw.

2.8 Subgroup analyses

Subgroup analyses were conducted by stratifying the participants based on race (EUR or AFR), sex, and gene. The early onset was defined as patients with the first diagnosis before 65 years old. We also tested 60 and 55 years as cutoffs to determine if the trend of association or expressivity (PRE) remained valid. We evaluated the effect size (Hazard Ratio, HR) of RVs associated with PheCodes or PheCode sets in a multivariate CoxPH model in all samples as well as subgroups stratified by sex, race, and gene. All subgroup analyses were considered exploratory and p-values were present as raw.

2.9 PheCode Set Enrichment Analysis (PSEA)

PheCode Set Enrichment Analysis was conducted to determine which PheCode set from a disease category (e.g., circulatory system) was significantly enriched against a randomly selected PheCode set (same number) from that category. In this study, we only focused on common diseases, the prevalence of which in the MyCode population ≥ 1%. Using pre-ranked differential gene expression data to conduct gene-set enrichment analysis (GSEA) inspired us to develop the PSEA. The general idea of the PSEA method is to test whether the distribution of PheCodes in the PheCode-set differed from a uniform distribution. To achieve this goal, we used the established ranking metric strategy used in GSEA – namely, signed statistic (e.g., disease-level PRE) was selected as a substitute to the statistics of differential gene expression used in GSEA.

fgsea[37] is an R-package for fast pre-ranked gene set enrichment analysis. (https://github.com/ctlab/fgsea). Here we adopt the idea and procedure of GSEA. The preranked PheCode-level statistics S (PRE) for the PheCode U = {1, 2, …, N} and a list of query PheCode sets, P, resulting in Max(ES) per disease category. This analysis aims to determine which of the PheCode sets from P has a non-random behavior. Fgsea used a gene set enrichment score function $s_r(p)$ originally defined by Subramanian et al[38]. The more positive is the value of $s_r(p)$ the more enriched the gene set is in the positively-regulated genes (with $S_i > 0$). Accordingly, negative $s_r(p)$ corresponds to enrichment in the negatively regulated genes. Value of $s_r(p)$ can be calculated as follows. Let $k = |p|, NS = \sum_{i \in p} |S_i|$. Let also ES be an array specified by the following formula[37]:

$$ES_i = \begin{cases} 0 & if\ i = 0 \\ ES_{i-1} + \frac{1}{NS}|S_i| & if\ 1 \le i \le N\ and\ i \in p, \\ ES_{i-1} - \frac{1}{N-k} & if\ 1 \le i \le N\ and\ i \notin p. \end{cases}$$

The value of $s_r(p)$ correspond to the largest by the absolute value entry of ES[37]:

$$s_r(p) = ES_{i^*}, where\ i^* = argmax|ES_i|$$

Following Subramanian et al. [38] for a pathway, p, Korotkevich et al. define GSEA p-value as[37]:

$$P_{value}(p) = \begin{cases} \frac{P_k(s_r(q) \ge s_r(p))}{P_k(s_r(q) \ge 0)} & if\ s_r(p) > 0 \\ \frac{P_k(s_r(q) \le s_r(p))}{P_k(s_r(q) \le 0)} & if\ s_r(p) < 0 \end{cases}$$

Where q is a random gene set of size k.

NES = enrichment score normalized to mean enrichment of random samples of the same size of tested PheCode-set**.** P-value estimation is based on an adaptive multi-level split Monte-Carlo scheme, which allows for estimating a lower p-value (< 10^{-5}). To speed up the process, fgsea divides the entire ranked PheCodes into n blocks and tries all the thresholds to get the minimum p-value. P-value adjustment for multiple testing of a number of PheCode sets was conducted using the Benjamin-Hochberg procedure.

Result:

Participants Demographics

The entire dataset was temporally split based on the availability of the sequencing data. Demographic and clinical features for the discovery and replication cohorts are listed in

Table 1a. We identified a total of 4,446 RV carriers of at least one of the seven candidate genes out of 173,575 MyCode participants. Initial EHR data extraction was completed on 12/29/2019 for a pilot study of the discovery cohort; the second EHR data extraction was completed on 12/06/2021 for both discovery and replication cohorts. After removing carriers without longitudinal EHR data (at least two encounters separated by at least one month), we end up with 4433 (2738/1695) carriers. Five-fold noncarriers, 22,250 (13,725/8525) were randomly sampled from noncarriers in each cohort. The prevalence of any diseases in the MyCode population was estimated by these randomly selected noncarriers. After PSM at a 1:2 ratio, there was no significant difference in sex, race (EUR or AFR), or the index age between carriers and noncarriers. The replication cohort was, on average, five years younger than the discovery cohort. This difference may lead to some findings from the discovery cohort not being replicated in the younger cohort, particularly for late-onset diseases.

Distribution of RVs identified by three pipelines

A total of 24 participants had RVs from any two genes. Around 20.18% (569/2738 and 329/1695) or 40.61% (1244/2738 and 563/1695) of participants had at least second/third-degree relatedness based on PiHAT score of 0.20 or 0.10, respectively. Most participants (94.45%/86.96%) are self-claimed White and confirmed with EUR ancestry by Principal Component Analysis (PCA). A smaller portion of participants (4.77%/11.79%) are self-claimed Black and confirmed with AFR ancestry by PCA. Participants from other ethnical groups were < 2% and were not included for further analysis.

Missense (38 distinct variants), stop gain (12), NMD (nonsense-mediated decay), transcript variant (8), and splice donor (5)/acceptor (7) variants were the top categories of VEP annotated consequence (**Supplementary Table 1**), suggesting these potential pathogenic variants may have an impact on protein function at the various levels.

In **Table 1a**, most participants with RVs were female in both cohorts. RVs were mainly identified in *COL4A1/2* and *NOTCH3* genes (**Table 1b**).

According to PSM plots (**Supplementary Figure 2**), RVs were observed disproportionately more in participants with AFR ancestry (for the unadjusted sample, SMD deviated to the right side for AFR, whereas SMD deviated to the left side for EUR) in both cohorts. The majority of identified RVs were in heterozygous form. Compared to the replication cohort, the discovery cohort had more participants (EUR: 60.70% vs. 38.54%; AFR: 95.42% vs. 89.05%) with *COL4A1/2* variants but fewer participants (EUR:22.37% vs. 34.57%; AFR: 0% vs. 2.49%) with *NOTCH3* variants. The distributions of RVs from the other four genes across the two cohorts were comparable ($p > 0.05$ by Fisher's exact test) (**Table 1b**).

Three variant annotation pipelines were applied to capture potentially damaging RVs. RVs derived from P2 were enriched with putative loss of function variants affecting RNA splicing (splice donor and acceptor variants) or translation (start lost, stop gained, and stop lost). However, the replication cohort had fewer participants identified from pipeline 2 with these consequences (**Supplementary Figure 3**). It was noted that many of the RVs prioritized by P2 were singleton or doubleton with acceptable QC data. Using PheCodes related to the circulatory system as examples, we explored the performance of the three pipelines. These PheCodes showed a stronger percent relative effect (PRE)

in either discovery or replication cohorts (**Supplementary Figure 3**). P1 showed the largest PRE (green bar) in most of the PheCodes than other pipelines or the intersect pipeline (red bar), suggesting that P1 augmented the performance of the P3 (purple bar), the ClinVar annotation. A similar observation was reported by others[39]. In some PheCodes, P2 (blue bar) showed the largest $PREs$, particularly in the discovery cohort but less obvious in the replication cohort, perhaps due to fewer carriers (44% versus 10%) exclusively identified by P2. Since three pipelines show their capability of identifying positive cases ($PRE \geq 0$) for some but not all tested PheCodes at varying degrees, carriers identified from all pipelines were combined for the following study.

Top PheCodes with the largest PRE were overrepresented in early-onset patients

The circo plot (**Supplementary Figure 4**) demonstrated the top 50 PheCodes ranked by PRE, grouped by the process. Unlike AFR participants with overrepresented phenotypes in the circulatory system, EUR participants did not show any disease categories represented excessively in the top 50 PheCodes with increased lifetime risk. Only in early-onset patients, diseases from the circulatory system were exclusively overrepresented in the top 50 PheCodes in both EUR and AFR participants (**Figure 2**). Primary cardiomyopathy was the only shared PheCode in the circulatory system in early-onset EUR between discovery and replication cohorts, whereas several PheCodes related to cerebral vascular diseases were top-ranked in the lifetime and early-onset AFR cohorts.

It was noted that all PheCode links between and within categories were specific to carriers in both AFR cohorts since no links were identified in noncarriers. However, some links identified in both EUR cohorts were shared by carriers and noncarriers. For

example, the links between cardiac and circulatory congenital anomalies and heart valve disease were identified in both carriers (Figure 3), particularly for *COL4A1/2* **(Supplementary Figure 5)** and noncarriers.

For individual genes (e.g., *COL4A1/2* and *NOTCH3*) in the early-onset study, PheCodes from the circulatory system were overrepresented in the top 50 common diseases in both discovery and replication cohorts (**Supplementary Figure 5**). However, this observation was valid only in *TREX1 and CTC1* in the replication cohort. The increased risk for cardiac and circulatory congenital anomalies[20,21 40] was identified in the EUR discovery cohort (**Figure 2**), particularly for *COL4A1/2* **(Supplementary Figure 5)**. Some connections between CSVD genes with phenotypes beyond cardiovascular systems were captured. These secondary findings included: 1) *GLA* and thyroid diseases; It was previously suggested that a physiological concentration of thyroid hormone directly facilitates matrix Gla protein gene expression in smooth muscle cells via thyroid hormone nuclear receptors, leading to prevention of vascular calcification [22]; 2) *COL4A1/2 and HTRA1* with renal findings (hematuria)[21], and *COL4A1/2* with abnormal function of bladder[41]; 3) *HTRA1* and musculoskeletal disease[42].

Notably, some common cerebrovascular disease-related phenotypes (such as TIA, primary cardiomyopathy, atrial fibrillation, and hypercoagulability) were top ranked with an increased relative risk in at least one cohort. The links between TIA and cardiomyopathy and cerebral ischemia and carditis were captured in the discovery cohort of EUR carriers but not in noncarriers. The link between occlusion of cerebral arteries and chronic venous insufficiency was identified only in AFR carriers but not in noncarriers in the replication cohort.

Race-, and Sex- dependent association

The entire gene-specific survival analysis is illustrated in **Figure 3**. In the discovery cohort, *COL4A1/2* or *NOTCH3* variants showed the (trend of) association with IS in AFR or EUR participants, respectively. In the replication cohort, *COL4A1/2 or TREX1* variants showed a similar (trend of) association with IS in AFR or EUR participants, respectively. When combined both AFR cohorts (**Figure 4C**), *COL4A1/2* was significantly associated with IS ($HR_{lifetime}$=2.032(1.065-3.875), p=0.031). It was noted that *HTRA1* carriers had an increased cumulated incidence of IS in the late-onset participants from both discovery and replication cohorts (the green curve spikes after 65yr in **Figure 3A-3B**).

Sex-dependent association was demonstrated in **Figure 4**. In the discovery cohort, the effect sizes were larger in female participants ($HR_{earlyonset}$=1.499[1.046-2.148], p=0.027 for *COL4A1/2* and 2.175[1.39-3.401], p < 0.001 for *NOTCH3* in EUR; 2.343[0.713-7.701], p=0.161 for *COL4A1* in AFR). In the replication cohort, female patients with early-onset IS showed larger effect sizes ($HR_{earlyonset}$=2.445[0.584-10.232], p=0.221 for *COL4A1/2* in AFR; 4.006[1.797-8.931], p<0.001 for *TREX1* in EUR). However, these findings for *NOTCH3* or *TREX1* were not observed in the replication or discovery cohort, respectively. AFR participants had no RVs identified for *NOTCH3*. This sex-dependent effect also appeared in *HTRA1* carriers after combining samples from both discovery and replication cohorts with a trend of increased effect size in female carriers ($HR_{lifetime}$=1.758[0.654-4.722], p=0.263) but not in male carriers $HR_{lifetime}$=0.98[0.377-2.72], p=0.98).

COL4A1/2 variants were associated with hypercoagulability in EUR participants in the discovery cohort ($HR_{lifetime}$=1.49[1.065-2.084], p=0.020). The trend of association was observed in the replication cohort ($HR_{lifetime}$=1.626[0.856-3.085], p=0.137) (**Supplementary Figure 7A**). Effect sizes were larger for *COL4A1/2* ($HR_{earlyonset}$=1.963[1.285-2.998], p = 0.002) in EUR female participants of hypercoagulability, but not for *NOTCH3* ($HR_{earlyonset}$=1.194[0.57-2.501], p = 0.638) in the discovery cohort. However, this significant association was only observed in the replication male cohort ($HR_{earlyonset}$= 5.113[1.649-15.854], p = 0.005). The association between *COL4A1/2* variants and hypercoagulability was also identified in male AFR participants ($HR_{earlyonset}$=5.113[1.649-15.854], p = 0.005). No connection between hypercoagulability and cerebrovascular disease was observed by Fisher's exact test in all genes (**Figure 2**) or *COL4A1/2*-only (**Supplementary Figure 5**) in the discovery cohort. The connection between hypercoagulability and heart valve disorders was only observed in carriers (**Figure 2 and Supplementary Figure 5**) but not in noncarriers (data not shown) in the EUR discovery cohort. The association between HTRA1 and hypercoagulability was identified in the discovery EUR cohort and validated in the replication EUR cohort (**Supplementary Figure 7**).

CTC1 had a significant association with primary cardiomyopathies[425.1] with a stronger signal in the replication EUR cohort ($HR_{earlyonset}$=8.463[2.608-27.466], p=0.0006; $HR_{lifetime}$=4.601[1.433-14.674], p=0.010) than in the discovery EUR cohort ($HR_{earlyonset}$=4.568[1.118-18.663], p=0.034; $HR_{lifetime}$=2.779[0.883-8.752], p=0.081) (**Supplementary Figure 7B)**. *TREX1* only demonstrated a significant association with

primary cardiomyopathies in the replication EUR cohort ($HR_{earlyonset}$=3.146[1.403-7.054], p = 0.005; $HR_{lifetime}$=2.386[1.184-4.810], p=0.015).

PheCode-set Enrichment Analysis in early-onset patients

The top enriched disease category was circulatory system in both discovery (N = 18, NES = 2.95, $p_{adj} = 2.8x10^{-10}$ for EUR; N = 20, NES = 2.43, $p_{adj} = 2.7x10^{-9}$ for AFR) and replication cohorts (N = 20, NES = 2.22, $p_{adj} = 2.8x10^{-7}$ for EUR; N = 20, NES=2.36, $p_{adj} = 8.5x10^{-8}$ for AFR (**Figure 5 and Supplementary Figure 8**). Validation across cohorts or across racial groups was also performed using the PheCode set derived from circulatory system as an example (**Figure 5D**). Neither the replication EUR cohort (NES = 1.10, p_{adj} = 0.97) nor the discovery AFR cohort (NES = 1.61, p_{adj} = 0.12) validated the enriched PheCode set from the discovery EUR cohort. The enriched PheCode set from the discovery AFR cohort was marginally replicated in the replication AFR cohort (NES = 1.72, p_{adj} = 0.083), but was not replicated in the discovery EUR cohort (NES = 0.77, p_{adj} = 0.96). The enriched PheCode set from the replication EUR or AFR cohort was mutually replicated each other (NES = 2.02, $p_{adj} = 7.1x10^{-4}$ or NES = 1.89, $p_{adj} = 1.6x10^{-3}$ in the replication AFR or EUR cohort, respectively).

Discussion:

There is a lack of evidence to strongly support population-based screening for monogenic stroke from an expert panel consensus. There is an unmet need to evaluate the lifetime risk and the variability in age of onset for cerebrovascular disease by leveraging large-scale biobank datasets linked to comprehensive EHRs, and to identify potentially actionable genomic variants that may contribute to monogenic IS. Although some efforts have been made[25,39] in this field, there are ample opportunities to improve the analytical procedures and reduce systemic biases.

In this study, using a workflow to prioritize potentially pathogenic variants by an integration of three variant annotation pipelines, we showed that 1) Most of the RVs identified were in a heterozygous form and disproportionately present in AFR participants; 2) Inclusion of participants with ultra-RVs, which predicted to have a 'high' functional impact and have not been observed in ClinVar and/or HGMD, may augment the disease profiling; 3) *NOTCH3* , *TREX1, and COL4A1/2* were associated with IS in a race- and sex-dependent manner (more frequent among women); 4) Circulatory diseases were overrepresented in discovery and replication cohorts with early-onset of cardiovascular phenotypes irrespective of race; 5) Carriers of the seven genes showed increased risk for early signs and symptoms of cardiovascular disease such as, TIA, cardiomyopathy, atrial fibrillation, and hypercoagulability; and 6) Major findings from the PheCode enrichment analysis were replicated among AFR participants or the replication cohort($p_{adj} < 0.1$ for PheCode sets derived from circulatory system from both cohorts).

Our phenome-wide analyses on RVs carriers with mostly heterozygous genotypes confirmed the pleiotropic and non-additive effects which were reported on homozygous carriers of PTVs, so-called “human knockouts” in a UK Biobank study[43]. We also confirmed that the *HTRA1* carriers had an increased risk for IS, particularly in the late-onset EUR participants. There is a conflict report between *in vitro* functional study[2,44] and *in silicon*-based prediction[45] about whether or not the *HTRA1*-p.Gln151Lys variant is deleterious. Most *HTRA1* carriers with EUR ancestry (103/120) were heterozygous for p.Gln151Lys, located at N-terminal domain with IGFBP and Kazal homology[44]. According to the Kaplan-Meier analyses (**Supplementary Figure 9),** p.Gln151Lys carries alone also had an increased risk for late-onset IS, suggesting its damaging effect. We argued that it is still valid to develop a knockout model to determine the functional effect of *HTRA1* IGFBP/Kazal domain within and beyond the cardiovascular system despite its *in vitro* null effect on protease activity[44].

Our finding indicated that female carriers had a significantly increased risk for IS during menopause and the elderly stage than male carriers(Figure 4), suggesting it is important to recognize the sex differences in precision medicine for IS[46].

Some of the findings identified in EUR participants were not replicated, likely due to their small effect sizes in the EUR, overrepresentation of participants of younger age, and some RVs being over- or under-represented in the replication cohort[9]. Given the extensive connection among PheCodes observed in the discovery EUR but not in the replication EUR (**Figure 2**), the disease trajectories for carriers in the discovery EUR could be more complicated than those observed in replication cohorts. This complexity

may determine if the PheCode sets derived from one cohort show significant enrichment in the other cohort.

Various factors may account for the possibility of replication in an independent cohort. The validity of the results is determined by empirical variant curation (ACMG Clinvar annotation[47]) or machine learning algorithms (VEP annotation[48] + additional approach[30]), which are used to prioritize functional variants. Using a comparative approach, we identified some discrepancies due to the variant annotation pipelines (**Supplementary Figure 3**). This discrepancy could result from the varied functional impact of each variant, their allele frequency and conservation, and their interaction with environmental factors (lifestyle, social determinants, and more) at the patient level.

Observations and conclusions made from monogenic diseases:

1. RVs with large effects:

The main driver for some variants with extremely low frequency is due to natural selection; however, the 'reversal rare' theory for protective alleles also exists[43]. Thus, the success of this study depends on the number of genuinely causal variants, the number of neutral variants or protective variants diluting signals, and whether the direction of effect is consistent within the gene. As we mentioned earlier, the additional filtering process for RVs would help enrich true deleterious RVs with higher penetrance. Type II error may occur due to including fewer RVs for some candidate genes.

2. The omnigenic model – Core and peripheral genes:

Most genes affect disease risk through highly connected gene regulatory/expression networks[49]. While RVs from the core genes demonstrate a larger effect size, peripheral

genes exert their effects through interaction with core genes. For an omnigenic disease model, adding more ‘peripheral’ genes into the ‘core’ genes could lead to the diluted impacts with an increase in the false-positive results[9,49]. “The omnigenic model predicts that virtually any variant with regulatory effects in a given tissue is likely to have (weak) effects on all diseases that are modulated through that tissue”[49]. Instead of testing RVs individually, grouping variants with a similar function or belonging to genes within a similar functional network could boost the power of the association study. Pathway enrichment analysis of *COL4A1, NOTCH3,* and *HTRA1* coexpression networks in approximately 20 transcriptomic datasets from cardiovascular diseases showed ‘extracellular matrix organization’ as the top GO Biological Process terms (https://seek.princeton.edu/seek/). *COL4A1, NOTCH3,* and *HTRA1* are also highly correlated at the transcriptional level. In contrast, *CTC1* (T cell receptor signaling pathway), *GLA* (cellular response to lipopolysaccharide), and *TREX1*(cellular response to interferon) were mostly related to inflammation and immune response. Therefore, the overall effect from the seven candidate genes mainly reflected the disease profile due to the damage of the extracellular matrix or matrisome disfunction[50].

3. Pleiotropy and disease trajectory:

Horizontal and vertical pleiotropy of monogenic variants results in diverged disease profile and trajectory. Extracerebral phenotypes are common in monogenic CSVD risk genes. The large-scale population-based phenome-wide longitudinal study of health outcomes could promote a better understanding of under-investigated disease trajectory beyond the brain[19]. Deep Phenotyping and Phenotypes conversion (stacking, orthogonal, deconvolution) enhance computing trajectories from wellness to disease for

genetically predisposed individuals[51]. We must be cautious when the PTVs are enriched by the pipeline 2 (P2) selection; the latter may lead to patients with different disease trajectories or more patients with certain early-onset diseases compared to damaging missense variants enriched by P1 and P3 selections (**Supplementary Figure 3**). A similar observation of the rare or ultra-rare PTVs versus damaging missense variants across the phenotypic spectrum was also reported[52]. Genetic screening of patients with an early-onset of circulatory diseases for RVs of CSVD risk genes would prioritize individuals for continuous granular monitoring and promote personalized management.

Study strength and limitation:

This study had some strengths, such as leveraging rich longitudinal EHR, comparing three variants-annotation pipelines, and conducting PheCodes enrichment and sensitivity analyses to determine the sex- and age-dependent association. However, the study also has some limitations. First, the small sample size for participants reporting AFR ancestry and other ethnicities may prevent us from adequately estimating the effect size from those populations. Second, a potential underdiagnosis for carriers with mild phenotype is an inherent challenge of EHR-based observational study[53]. This would lead to an underestimation of the prevalence of genetically predisposed traits as the early sign or symptoms of advanced phenotypes (e.g., stroke or heart failure). Pathogenic variants resulting in mild phenotypes may survive natural selection during reproductive age but increase the risk for mid- and late-onset diseases. Third, RVs identified by the three pipelines were limited to missense, splicing, and stop gain/lost sequence variants. Finally, the pleiotropic effect and varied effect sizes for each RV complicated the overall disease profiles and trajectories assessment. The varied

expressivity of carriers could be due to an interplay between complex (polygenic) and Mendelian (monogenic) genetics[13] for IS; the common disease is more complex than the implied "core genes" model[54]. Polygenic Risk Scores (PRS) derived from common variants could modify the outcome of patients with a monogenic disease[13,55,56]. How PRS modify the outcome of patients with RVs of monogenic risk genes will be the future focus.

Conclusion:

- Some common cardiovascular disease-related phenotypes (TIA, primary cardiomyopathy, atrial fibrillation, and hypercoagulability) were identified at an increased relative risk in RV carriers of seven CSVD risk genes.
- RVs from *COL4A1/2* or *NOTCH3* and *TREX1* were associated with IS in participants with AFR or EUR ancestry, respectively. Effect sizes are significantly larger in female participants. *HTRA1* carriers with EUR ancestry also showed a similar trend of association.
- Diseases from the circulatory system were significantly enriched in carriers with both EUR and AFR ancestry.
- Our findings support the concept of developing a gene panel of CSVD for population screening of patients with early-onset circulatory diseases. Whether the inclusion of CSVD genes on population-based genomic screening/ACMG secondary findings lists (which would be phenotype-agnostic) remains debatable.

Reference:

1 Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, doi:10.1126/science.aaf6814 (2016).

2 Malik, R. *et al.* Whole-exome sequencing reveals a role of HTRA1 and EGFL8 in brain white matter hyperintensities. *Brain* **144**, 2670-2682, doi:10.1093/brain/awab253 (2021).

3 Mishra, A. *et al.* Association of variants in HTRA1 and NOTCH3 with MRI-defined extremes of cerebral small vessel disease in older subjects. *Brain* **142**, 1009-1023, doi:10.1093/brain/awz024 (2019).

4 Rutten, J. W. *et al.* The effect of NOTCH3 pathogenic variant position on CADASIL disease severity: NOTCH3 EGFr 1-6 pathogenic variant are associated with a more severe phenotype and lower survival compared with EGFr 7-34 pathogenic variant. *Genet Med* **21**, 676-682, doi:10.1038/s41436-018-0088-3 (2019).

5 Ilinca, A. *et al.* A stroke gene panel for whole-exome sequencing. *Eur J Hum Genet* **27**, 317-324, doi:10.1038/s41431-018-0274-4 (2019).

6 Grami, N. *et al.* Global Assessment of Mendelian Stroke Genetic Prevalence in 101 635 Individuals From 7 Ethnic Groups. *Stroke* **51**, 1290-1293, doi:10.1161/STROKEAHA.119.028840 (2020).

7 Hack, R. J. *et al.* Cysteine-Altering NOTCH3 Variants Are a Risk Factor for Stroke in the Elderly Population. *Stroke* **51**, 3562-3569, doi:10.1161/STROKEAHA.120.030343 (2020).

8 Rutten, J. W. *et al.* Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank: CADASIL to nonpenetrance. *Neurology* **95**, e1835-e1843, doi:10.1212/WNL.0000000000010525 (2020).

9 Nicolae, D. L. Association Tests for Rare Variants. *Annu Rev Genomics Hum Genet* **17**, 117-130, doi:10.1146/annurev-genom-083115-022609 (2016).

10 Oetjens, M. T., Kelly, M. A., Sturm, A. C., Martin, C. L. & Ledbetter, D. H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat Commun* **10**, 4897, doi:10.1038/s41467-019-12869-0 (2019).

11 Goodrich, J. K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat Commun* **12**, 3505, doi:10.1038/s41467-021-23556-4 (2021).

12 Forrest, I. S. *et al.* Population-Based Penetrance of Deleterious Clinical Variants. *JAMA* **327**, 350-359, doi:10.1001/jama.2021.23686 (2022).

13 Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231 e1211, doi:10.1016/j.cell.2020.08.008 (2020).

14 Choi, S. H. *et al.* Association Between Titin Loss-of-Function Variants and Early-Onset Atrial Fibrillation. *JAMA* **320**, 2354-2364, doi:10.1001/jama.2018.18179 (2018).

15 Hall, M. A. *et al.* Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS Genet* **10**, e1004678, doi:10.1371/journal.pgen.1004678 (2014).

16 Shahjouei, S. *et al.* Risk of Subsequent Stroke Among Patients Receiving Outpatient vs Inpatient Care for Transient Ischemic Attack: A Systematic Review and Meta-analysis. *JAMA Netw Open* **5**, e2136644, doi:10.1001/jamanetworkopen.2021.36644 (2022).

17 Siegerink, B., Maino, A., Algra, A. & Rosendaal, F. R. Hypercoagulability and the risk of myocardial infarction and ischemic stroke in young women. *J Thromb Haemost* **13**, 1568-1575, doi:10.1111/jth.13045 (2015).

18 Diaz, J. *et al.* Adherence to anticoagulant guideline for atrial fibrillation: A large care gap among stroke patients in a rural population. *J Neurol Sci* **424**, 117410, doi:10.1016/j.jns.2021.117410 (2021).

19 Rannikmae, K. *et al.* Beyond the Brain: Systematic Review of Extracerebral Phenotypes Associated With Monogenic Cerebral Small Vessel Disease. *Stroke* **51**, 3007-3017, doi:10.1161/STROKEAHA.120.029517 (2020).

20 Kuo, D. S., Labelle-Dumais, C. & Gould, D. B. COL4A1 and COL4A2 mutations and disease: insights into pathogenic mechanisms and potential therapeutic targets. *Hum Mol Genet* **21**, R97-110, doi:10.1093/hmg/dds346 (2012).

21 Meuwissen, M. E. *et al.* The expanding phenotype of COL4A1 and COL4A2 mutations: clinical data on 13 newly identified families and a review of the literature. *Genet Med* **17**, 843-853, doi:10.1038/gim.2014.210 (2015).

22 Sato, Y. *et al.* Thyroid hormone targets matrix Gla protein gene associated with vascular smooth muscle calcification. *Circ Res* **97**, 550-557, doi:10.1161/01.RES.0000181431.04290.bd (2005).

23 Thompson, C. S. & Hakim, A. M. Living beyond our physiological means: small vessel disease of the brain is an expression of a systemic failure in arteriolar function: a unifying hypothesis. *Stroke* **40**, e322-330, doi:10.1161/STROKEAHA.108.542266 (2009).

24 Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* **18**, 906-913, doi:10.1038/gim.2015.187 (2016).

25 Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat Genet*, doi:10.1038/s41588-021-01011-w (2022).

26 Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, doi:10.1126/science.aaf7000 (2016).

27 Li, J. *et al.* Polygenic Risk Scores Augment Stroke Subtyping. *Neurol Genet* **7**, e560, doi:10.1212/NXG.0000000000000560 (2021).

28 Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, doi:10.1093/bioinformatics/btaa1081 (2021).

29 Kelly, M. A. *et al.* Leveraging population-based exome screening to impact clinical care: The evolution of variant assessment in the Geisinger MyCode research project. *Am J Med Genet C Semin Med Genet* **187**, 83-94, doi:10.1002/ajmg.c.31887 (2021).

30 Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677, doi:10.1007/s00439-017-1779-6 (2017).

31 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).

32 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424, doi:10.1038/gim.2015.30 (2015).

33 Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325, doi:10.2196/14325 (2019).

34 Schuemie, M. J. *et al.* A plea to stop using the case-control design in retrospective database studies. *Stat Med* **38**, 4199-4208, doi:10.1002/sim.8215 (2019).

35 Monaghan, T. F. *et al.* Foundational Statistical Principles in Medical Research: A Tutorial on Odds Ratios, Relative Risk, Absolute Risk, and Number Needed to Treat. *Int J Environ Res Public Health* **18**, doi:10.3390/ijerph18115669 (2021).

36 Wei, W. Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508, doi:10.1371/journal.pone.0175508 (2017).
37 Korotkevich, G. *et al.*, doi:10.1101/060012 (2021).
38 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
39 Park, J. *et al.* Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. *Nat Med* **27**, 66-72, doi:10.1038/s41591-020-1133-8 (2021).
40 Steffensen, L. B. & Rasmussen, L. M. A role for collagen type IV in cardiovascular disease? *Am J Physiol Heart Circ Physiol* **315**, H610-H625, doi:10.1152/ajpheart.00070.2018 (2018).
41 Rouaud, T. *et al.* Acute urinary retention due to a novel collagen COL4A1 mutation. *Neurology* **75**, 747-749, doi:10.1212/WNL.0b013e3181eee440 (2010).
42 Tiaden, A. N. & Richards, P. J. The emerging roles of HTRA1 in musculoskeletal disease. *Am J Pathol* **182**, 1482-1488, doi:10.1016/j.ajpath.2013.02.003 (2013).
43 DeBoever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat Commun* **9**, 1612, doi:10.1038/s41467-018-03910-9 (2018).
44 Eigenbrot, C. *et al.* Structural and functional analysis of HtrA1 and its subdomains. *Structure* **20**, 1040-1050, doi:10.1016/j.str.2012.03.021 (2012).
45 Di Donato, I. *et al.* Heterozygous mutations of HTRA1 gene in patients with familial cerebral small vessel disease. *CNS Neurosci Ther* **23**, 759-765, doi:10.1111/cns.12722 (2017).
46 Bushnell, C. *et al.* Sex differences in the evaluation and treatment of acute ischaemic stroke. *Lancet Neurol* **17**, 641-650, doi:10.1016/S1474-4422(18)30201-1 (2018).
47 Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet* **98**, 1067-1076, doi:10.1016/j.ajhg.2016.03.024 (2016).
48 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
49 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186, doi:10.1016/j.cell.2017.05.038 (2017).
50 Debette, S. & Markus, H. S. Stroke Genetics: Discovery, Insight Into Mechanisms, and Clinical Perspectives. *Circ Res* **130**, 1095-1111, doi:10.1161/CIRCRESAHA.122.319950 (2022).
51 MacRae, C. A. & Califf, R. M. Reimagining What We Measure in Atherosclerosis-a "Phenotype Stack". *Circ Res* **126**, 1146-1158, doi:10.1161/CIRCRESAHA.120.315941 (2020).
52 Ganna, A. *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am J Hum Genet* **102**, 1204-1211, doi:10.1016/j.ajhg.2018.05.002 (2018).
53 Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* **37**, 61-81, doi:10.1146/annurev-publhealth-032315-021353 (2016).
54 Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573-1580, doi:10.1016/j.cell.2018.05.051 (2018).
55 Giudicessi, J. R. & Ackerman, M. J. Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. *Transl Res* **161**, 1-14, doi:10.1016/j.trsl.2012.08.005 (2013).
56 Li, J. *et al.* Predicting mortality among ischemic stroke patients using pathways-derived polygenic risk scores. *Sci Rep* **12**, 12358, doi:10.1038/s41598-022-16510-x (2022).

Table 1. Patient characteristics of RVs carriers or noncarriers in both Discovery and Replication cohorts.

The study cohort was split into Discovery and Replication datasets and further split according to having RVs or not. All continuous variables were reported as mean (standard deviation), and dichotomous variables were reported as absolute count (frequency). For disease phenotype, we reported the absolute count of patients having lifetime and early-onset of the corresponding PheCode from Geisinger EHR. For the lifetime observation, patients were followed up to 85 years; Patients with age of onset after 65yrs considered null for the early-onset cases. ANOVA or Fisher's exact test was selected to determine the significant difference between groups for quantitative or qualitative measures, respectively.

Figure legends:

Figure 1. A flowchart of study design and analytic pipeline for the genotype-first phenome-wide longitudinal study.

A diagnostic panel of RVs (n=67) prioritized by three pipelines (P1-3) was created for the purpose of interpreting exome sequencing results for monogenic stroke candidate genes (*COL4A1, COL4A2, NOTCH3, HTRA1, TREX1, CTC1,* and *GLA)*. Simulated longitudinal disease profile for a patient with RV(s) as an example to illustrate the data extraction and PheCode mapping. The colors represent different PheCodes from the same disease category. Cases for the specific PheCode were defined by 'Rule of Two' from any encounter types in the EHR.

Figure 2. Circo plots with interlinks of the top 50 common diseases of which RVs from seven CSVD risk genes having increased relative risk in early-onset patients.

Participants with the age of onset for a specific disease after 65yrs were considered null for the early-onset phenotype; A circo plot was created based on the top 50 PheCodes ranked by precent relative effect (PRE) and clustered by the processes. The links represent the significant association between two PheCodes in RV carriers calculated by Fisher's exact test after correction for multiple testing ($p < 0.001$ given 50 phecodes being tested)). Significant associations for adjacent PheCodes were labeled with spike but not linked. Y axis represents PRE; The area of the dot was proportional to the significant association between carriers and the corresponding disease according to Fisher's exact test. Diseases from the circulatory system were overrepresented in early-

onset carriers in both cohorts with EUR or AFR ancestry. Disease categories with the number of PheCodes >= 2 were plotted.

Figure 3. Kaplan-Meier estimated the cumulative probability of ischemic cerebrovascular disease for RV carriers compared to matched noncarriers, a gene- and race- dependent association.

A. EUR in discovery cohort; **B**. EUR in replication cohort; **C**. AFR in discovery/replication cohort. Carriers for genes other than COL4A1/2 were too low to show in AFR subgroups. Only patients with European (EUR) or African (AFR) ancestry were included in this study. Ischemic cerebrovascular disease was determined by six PheCodes. The follow-up was censored on 12/06/2021. A Lifelong (up to 85yrs) Log-rank test was applied to comparing carriers for each gene to the shared noncarriers for all seven genes. For post-hoc pairwise tests, p-values were present as raw. Cox proportional-hazards regression model, adjusted for sex, index age, and five principal components, was conducted to determine the association between collapsed RVs from each gene or combined and the IS. Adjusted Hazard Ratios (HR) estimates and their 95% confidence intervals (Cis) for a lifetime (up to 85yrs) or early-onset (≤65yrs) risk were present in the forest plot.

Figure 4. Kaplan-Meier estimated the cumulative probability of IS for carriers compared to matched noncarriers, a sex-dependent analysis.

For post-hoc pairwise Log-rank tests, p-values were present as raw. The effect size (HR) was observed to be larger in Female than Male carriers. A. NOTCH3 in the discovery EUR cohort; B. TREX1 in the replication EUR cohort; C. COL4A1 in both discovery and replication AFR cohort. Cox proportional-hazards regression model,

adjusted for sex, index age, and five principal components, was conducted to determine the association between collapsed RVs from each gene and the IS. Adjusted Hazard Ratios (HR) estimates and their 95% confidence intervals (Cis) for a lifetime (up to 85yrs) or early-onset (≤65yrs) risk were present in the forest plot.

Figure 5. PheCode Set Enrichment Analysis was conducted to determine which disease categories were significantly enriched.

A. A graph that corresponds to a calculation of enrichment score. Each breakpoint on a graph corresponds to a PheCode present in the disease category (e.g., Circulatory system). The curve goes rightward when phecode $\notin$ disease category and goes upward when $\in$ phecode sets**. B.** For example, in Discovery(EUR), the point corresponds to PheCode ranked the 71th from the entire PheCode of common diseases of any category, where the maximal value of the enrichment score is reached. The vertical bars along the x-axis represent 69 PheCodes belonging to the interrogated disease category(e.g., the circulatory system). The vertical dash line defined “leading edge subset” as all those PheCodes (in bracket) ranked as least as high as the enriched set, and the enriched score reached the maximum. **C.** The forest plot was created to show the effect size of significantly enriched PheCode sets from each disease category in Discovery(EUR), where the common diseases were ranked based on Percent Relative Effect (PRE). The raw and adjusted p-value for multiple testing using BH procedure was listed. **D**. Validation across cohorts or across races using the PheCode set derived from the circulatory system as examples**.** The PheCode set enrichment analysis has two input objects: 1) an array of pre-ranked PheCode-level statistics $S\ (PRE)$ for the PheCode $U\ =\ \{1,\ 2,\ .\ .\ .\ ,\ N\}$, shown as Rank List (y-axis); 2) a list of query PheCode

sets P which consist of 18 enriched PheCode sets from the corresponding 18 disease categories. The goal of the analysis is to determine which of the 18 PheCode sets enriched from one dataset also shows a non-random overrepresentation in the other dataset. Here we only reported the result of PheCode Sets derived from the Circulatory system (x-axis).

Table 1. Patient characteristics of RVs carriers or noncarriers in both Discovery and Replication cohorts.

Cohort	Discovery				Replication				Prevalence of common diseases estimated from noncarriers (%)
Characteristic	Carriers		NonCarriers		Carriers		NonCarriers		
	Number of cases	cumulative incidence (‰)	Number of cases	cumulative incidence (‰)	Number of cases	cumulative incidence (‰)	Number of cases	cumulative incidence (‰)	
sample size	2738		5490		1695		3410		
Index_age	57.18(18.34)		56.99(18.35)		52.04(18.73)		52.52(19.45)		
Male(%)	1028(37.55%)		2053(37.4%)		612(36.11%)		1185(34.75%)		
American Indian Or Alaska Native	2 (0.07%)		5 (0.09%)		5 (0.29%)		17 (0.50%)		
Asian	9 (0.33%)		14 (0.26%)		1 (0.06%)		2 (0.06%)		
Native Hawaiian Or Other Pacific Islander	2 (0.07%)		3 (0.05%)		5 (0.29%)		28 (0.82%)		
Unknown(Multi-ancestry)	8 (0.29%)		16 (0.29%)		9 (0.53%)		28 (0.82%)		
White	2586 (94.45%)		5242 (95.46%)		1474(86.96%)		3036 (89.03%)		
Black Or African American	131 (4.77%)		211 (3.84%)		201 (11.79%)		299 (8.77%)		
Other and unspecified coagulation defects(286.7)	42/34*	15.34/12.42	70/44	12.75/8.01	22*/18*	12.98/10.62	25/20	7.33/5.87	1.1
Hypercoagulable state(286.8)	63***/50**	23.01/18.26	74/63	13.48/11.48	22/17	12.98/10.03	39/30	11.44/8.8	1.33
Primary hypercoagulable state(286.81)	61***/49**	22.28/17.9	74/63	13.48/11.48	22/17	12.98/10.03	38/30	11.14/8.8	1.3
Heart valve disorders(395)	213/108	77.79/39.44	453/180	82.51/32.79	89/42	52.51/24.78	198/87	58.06/25.51	7.43
Abnormal heart sounds(396)	74/50	27.03/18.26	142/84	25.87/15.3	32/26	18.88/15.34	93/55	27.27/16.13	2.59
Hypertension(401)	1419/1084	518.26/395.91	2898/2211	527.87/402.73	720/556	424.78/328.02	1465/1105	429.62/324.05	50.06
Myocardial infarction(411.2)	133/78	48.58/28.49	336/174	61.2/31.69	47/30	27.73/17.7	88/48	25.81/14.08	4.99
Coronary atherosclerosis(411.4)	448/234	163.62/85.46	954/532	173.77/96.9	169/99**	99.71/58.41	319/149	93.55/43.7	15.06
Pulmonary heart disease(415)	153/72	55.88/26.3	323/166	58.83/30.24	54/32	31.86/18.88	101/51	29.62/14.96	4.7
Cardiomyopathy(425)	80/49	29.22/17.9	156/86	28.42/15.66	51/39	30.09/23.01	88/52	25.81/15.25	3
Primary/intrinsic cardiomyopathies(425.1)	72/43	26.3/15.7	137/76	24.95/13.84	49/37**	28.91/21.83	75/44	21.99/12.9	2.71
Cardiac conduction disorders(426)	260/141	94.96/51.5	512/282	93.26/51.37	142/88**	83.78/51.92	253/125	74.19/36.66	10.29
Atrial fibrillation and flutter(427.2)	263/92	96.06/33.6	568/191	103.46/34.79	105/55	61.95/32.45	238/84	69.79/24.63	9.23
Intracranial hemorrhage(430)	33/13	12.05/4.75	69/29	12.57/5.28	16/6	9.44/3.54	24/15	7.04/4.4	1.08
Cerebrovascular disease(433)	349/176	127.47/64.28	663/315	120.77/57.38	138/76**	81.42/44.84	243/114	71.26/33.43	10.33
Occlusion and stenosis of precerebral arteries(433.1)	209/92	76.33/33.6	409/175	74.5/31.88	69/30	40.71/17.7	142/56	41.64/16.42	6.35
Occlusion of cerebral arteries with cerebral infarction(433.11)	22/11	8.04/4.02	41/21	7.47/3.83	5/3	2.95/1.77	14/6	4.11/1.76	0.56
Cerebral atherosclerosis(433.12)	83/35	30.31/12.78	165/62	30.05/11.29	24/12	14.16/7.08	52/27	15.25/7.92	2.37
Occlusion of cerebral arteries(433.2)	93/49	33.97/17.9	161/74	29.33/13.48	37/19	21.83/11.21	65/28	19.06/8.21	2.53

Cerebral artery occlusion with cerebral infarction(433.21)	93/49	33.97/17.9	161/74	29.33/13.48	36/18	21.24/10.62	63/26	18.48/7.62	2.47
Cerebral ischemia(433.3)	188/102*	68.66/37.25	358/160	65.21/29.14	62/36	36.58/21.24	124/56	36.36/16.42	5.33
Transient cerebral ischemia(433.31)	177/88*	64.65/32.14	325/136	59.2/24.77	61/35	35.99/20.65	119/54	34.9/15.84	4.92
Cerebral aneurysm(433.5)	22/14	8.04/5.11	44/34	8.01/6.19	16/11	9.44/6.49	19/13	5.57/3.81	0.67
Late effects of cerebrovascular disease(433.8)	67/30	24.47/10.96	110/51	20.04/9.29	13/7	7.67/4.13	31/15	9.09/4.4	1.64
Peripheral vascular disease unspecified(443.9)	165/83	60.26/30.31	384/169	69.95/30.78	63/37**	37.17/21.83	105/45	30.79/13.2	5.36
Deep vein thrombosis [DVT](452.2)	91/53	33.24/19.36	151/86	27.5/15.66	25/15	14.75/8.85	76/43	22.29/12.61	2.49

25

Figure 1. A flowchart of study design and analytic pipeline for the genotype-first phenome-wide longitudinal study.

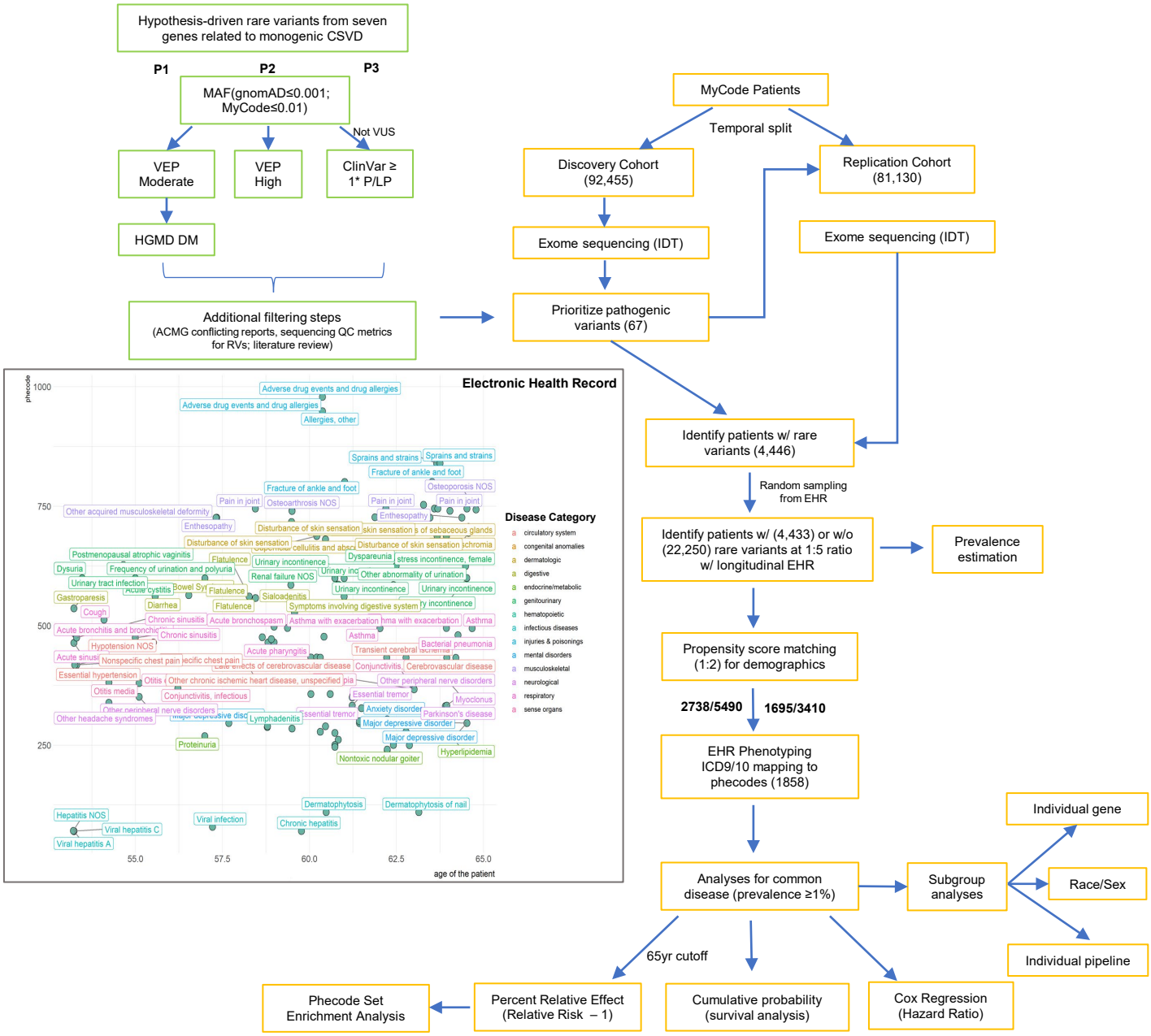


Figure 2. Circo plots with interlinks of the top 50 common diseases of which rare variants from seven CSVD risk genes having increased relative risk in early-onset patients.

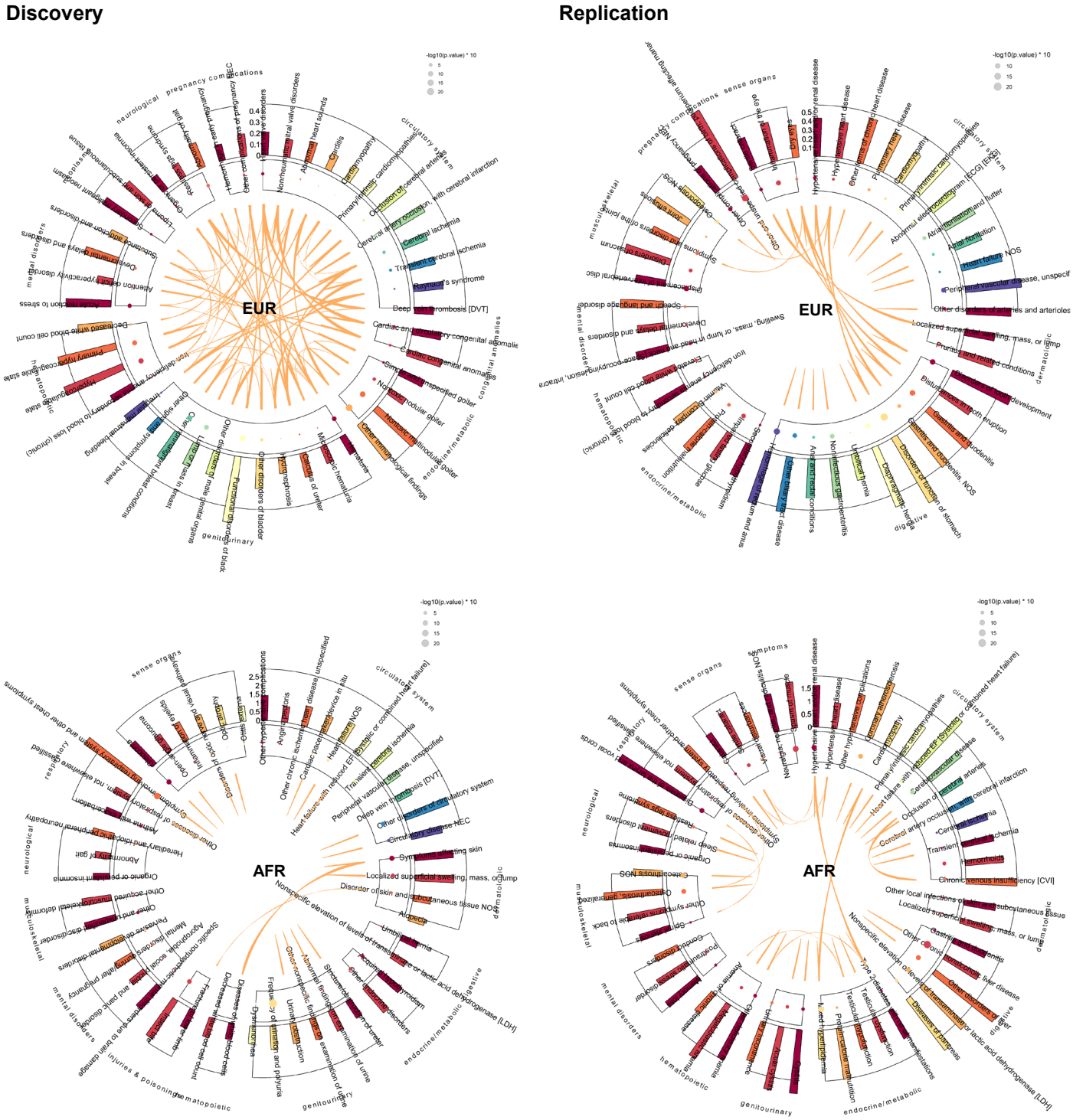
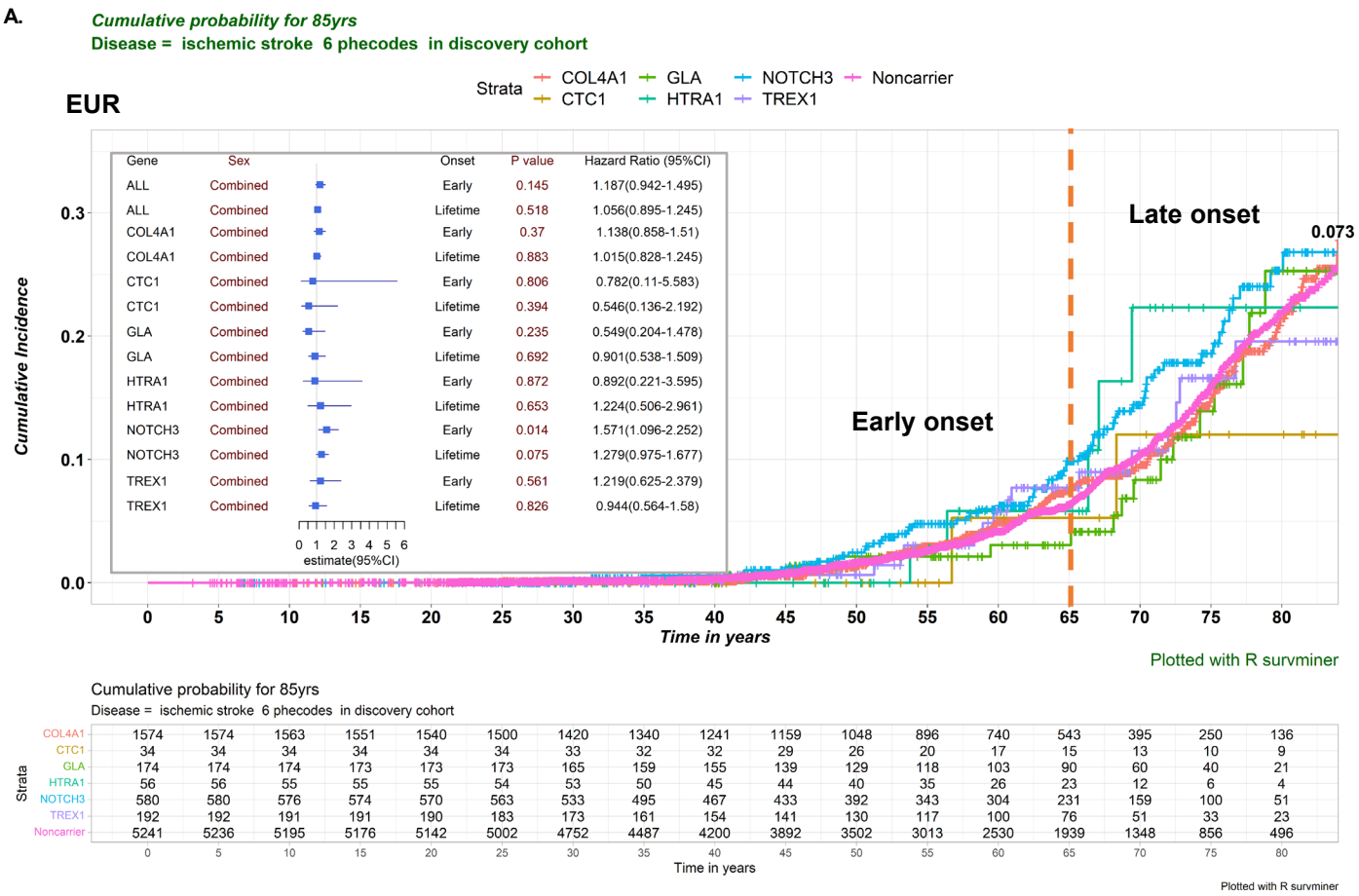


Figure 3. Kaplan-Meier estimated the cumulative probability of ischemic cerebrovascular disease for rare variant carriers comparing to matched noncarriers, a gene- and race- dependent association.

A.



Cumulative probability for 85yrs

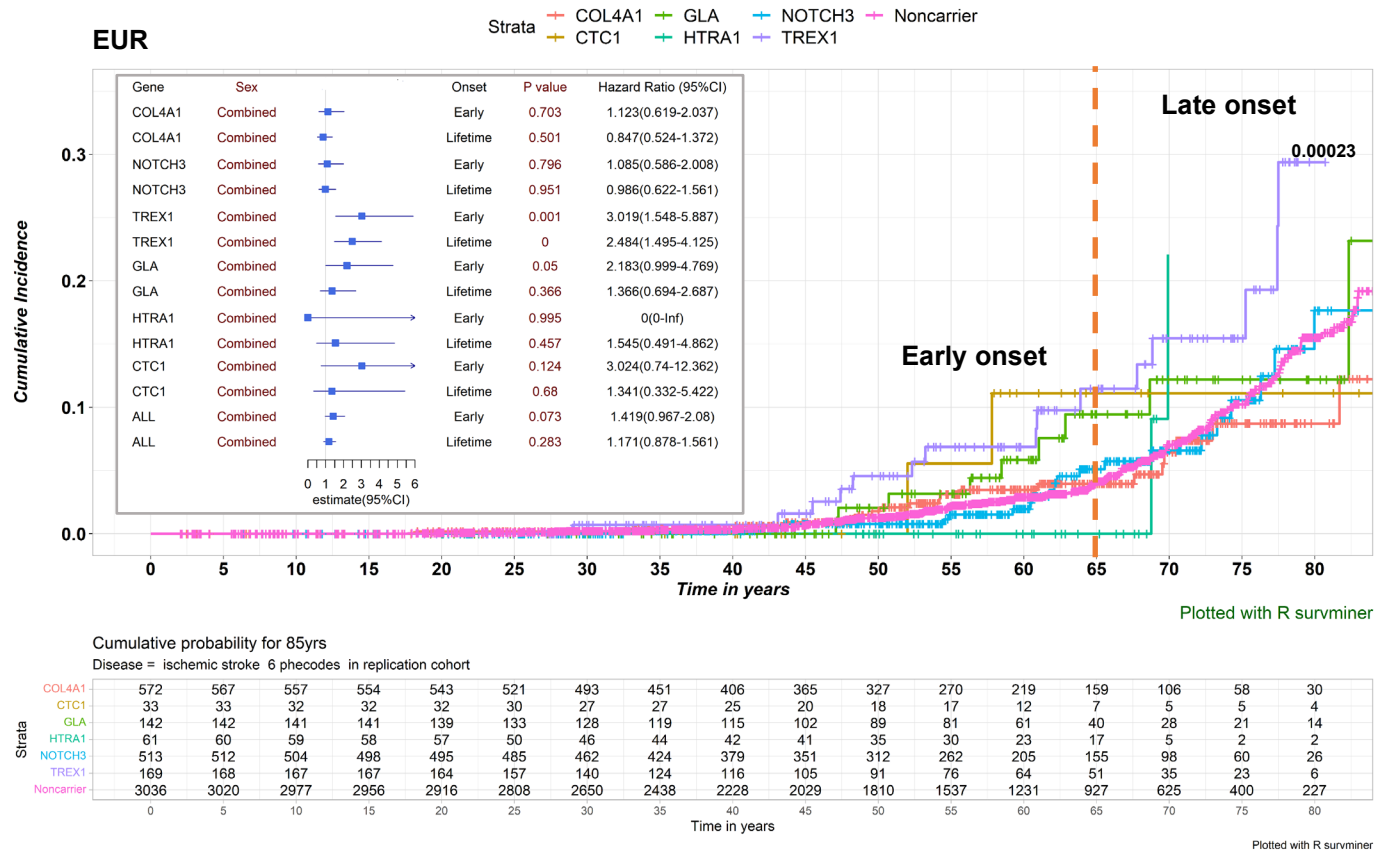
Disease = ischemic stroke 6 phecodes in discovery cohort

Strata	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
COL4A1	1574	1574	1563	1551	1540	1500	1420	1340	1241	1159	1048	896	740	543	395	250	136
CTC1	34	34	34	34	34	34	33	32	32	29	26	20	17	15	13	10	9
GLA	174	174	174	173	173	173	165	159	155	139	129	118	103	90	60	40	21
HTRA1	56	56	55	55	55	54	53	50	45	44	40	35	26	23	12	6	4
NOTCH3	580	580	576	574	570	563	533	495	467	433	392	343	304	231	159	100	51
TREX1	192	192	191	191	190	183	173	161	154	141	130	117	100	76	51	33	23
Noncarrier	5241	5236	5195	5176	5142	5002	4752	4487	4200	3892	3502	3013	2530	1939	1348	856	496

Time in years

Plotted with R survminer

B.



Cumulative probability for 85yrs

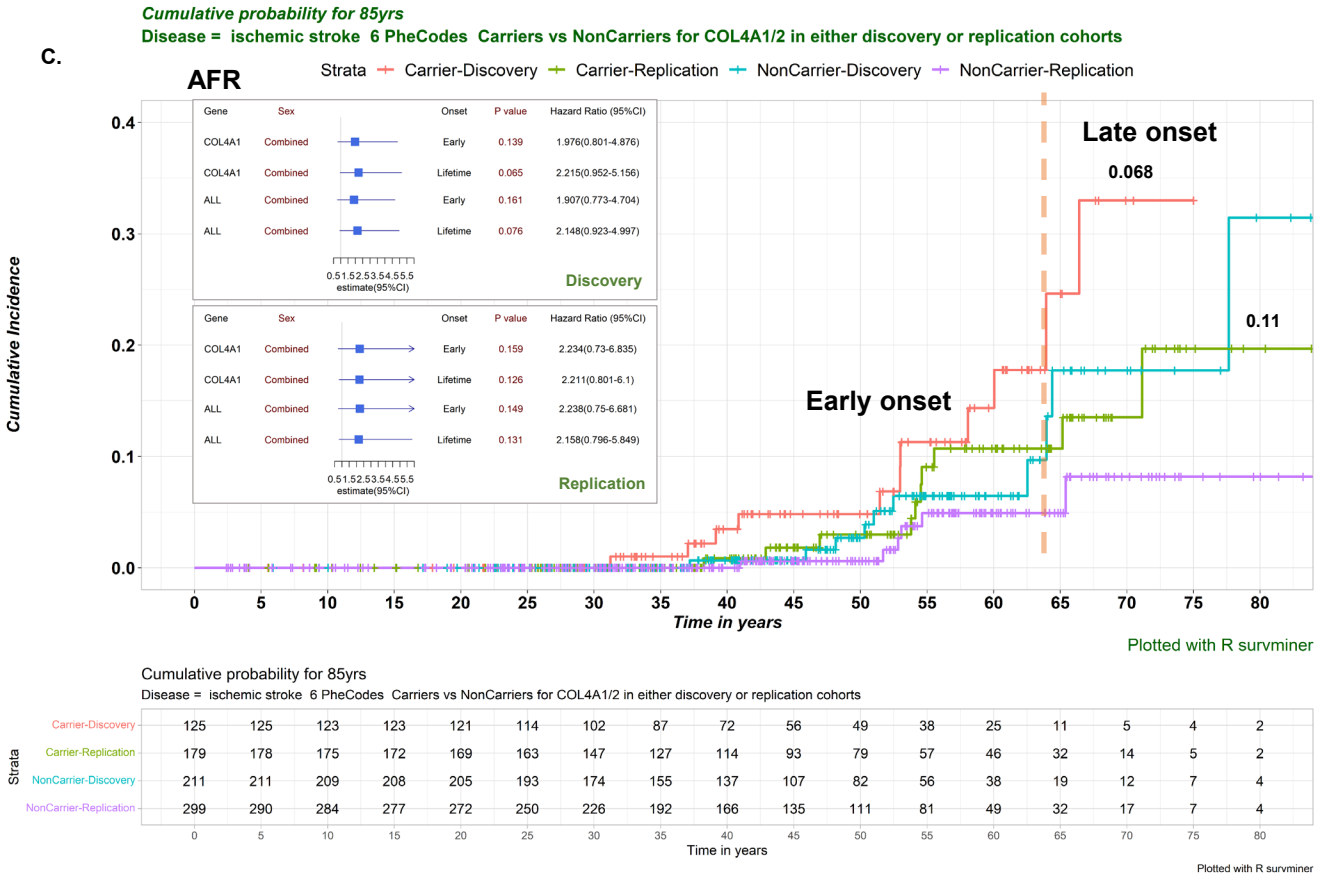
Disease = ischemic stroke 6 phecodes in replication cohort

Strata	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
COL4A1	572	567	557	554	543	521	493	451	406	365	327	270	219	159	106	58	30
CTC1	33	33	32	32	32	30	27	27	25	20	18	17	12	7	5	5	4
GLA	142	142	141	141	139	133	128	119	115	102	89	81	61	40	28	21	14
HTRA1	61	60	59	58	57	50	46	44	42	41	35	30	23	17	5	2	2
NOTCH3	513	512	504	498	495	485	462	424	379	351	312	262	205	155	98	60	26
TREX1	169	168	167	167	164	157	140	124	116	105	91	76	64	51	35	23	6
Noncarrier	3036	3020	2977	2956	2916	2808	2650	2438	2228	2029	1810	1537	1231	927	625	400	227

Time in years

Plotted with R survminer

C.



Cumulative probability for 85yrs

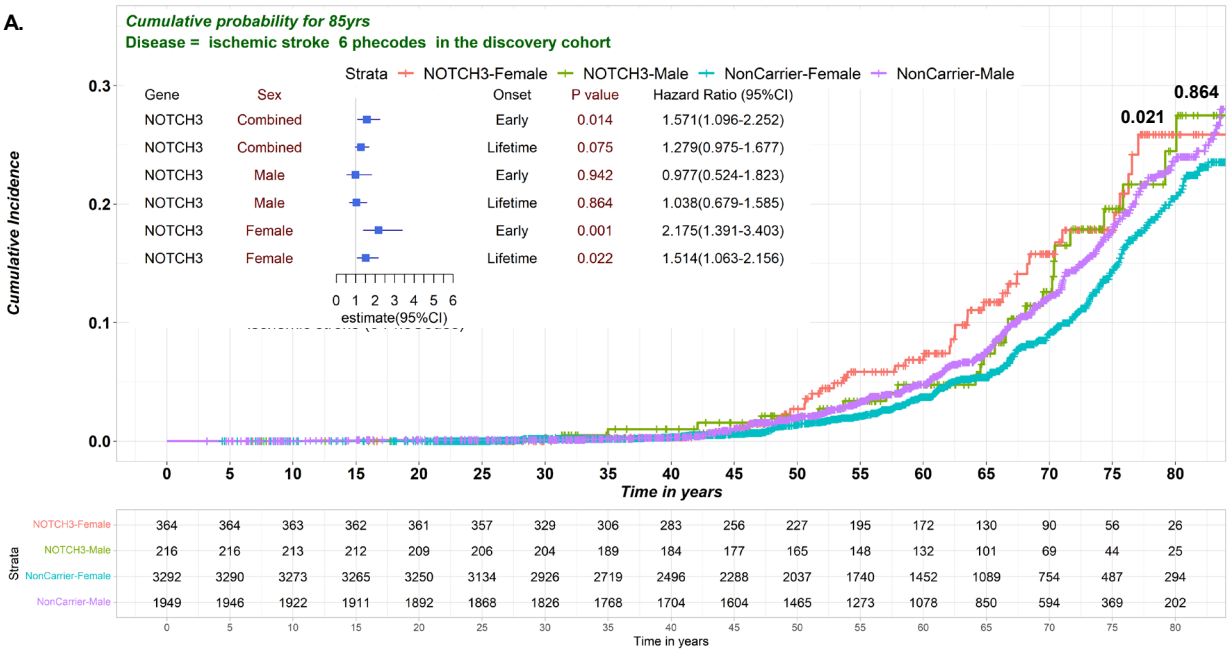
Disease = ischemic stroke 6 PheCodes Carriers vs NonCarriers for COL4A1/2 in either discovery or replication cohorts

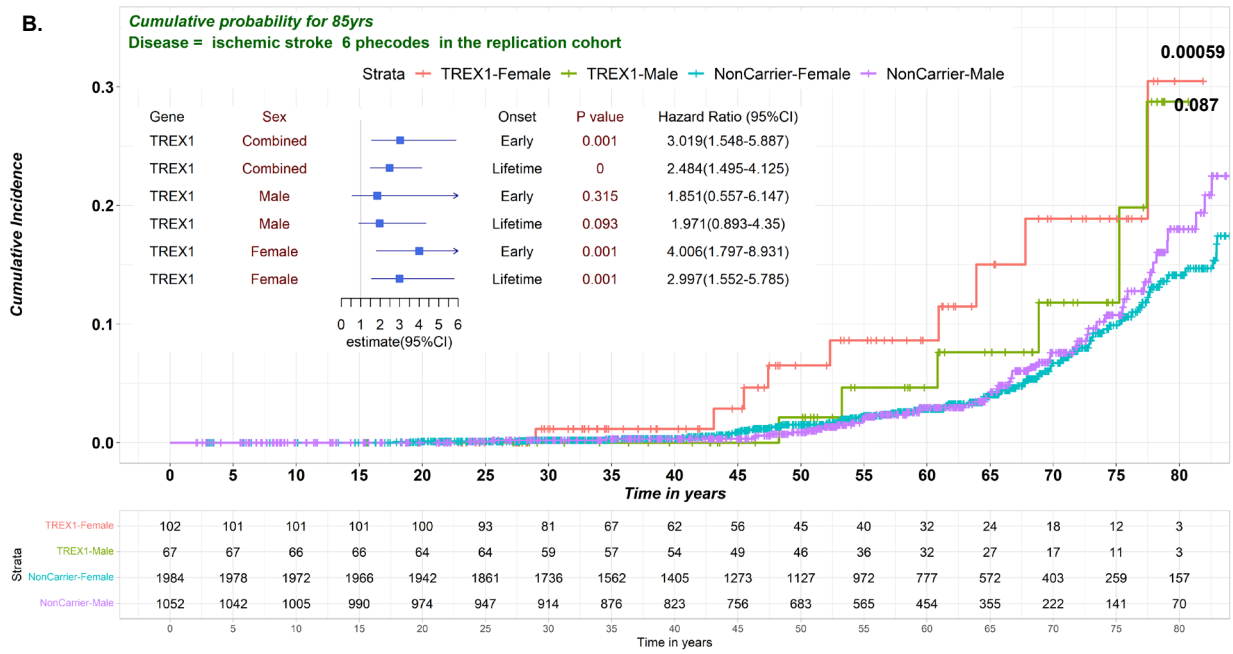
Strata	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
Carrier-Discovery	125	125	123	123	121	114	102	87	72	56	49	38	25	11	5	4	2
Carrier-Replication	179	178	175	172	169	163	147	127	114	93	79	57	46	32	14	5	2
NonCarrier-Discovery	211	211	209	208	205	193	174	155	137	107	82	56	38	19	12	7	4
NonCarrier-Replication	299	290	284	277	272	250	226	192	166	135	111	81	49	32	17	7	4

Time in years

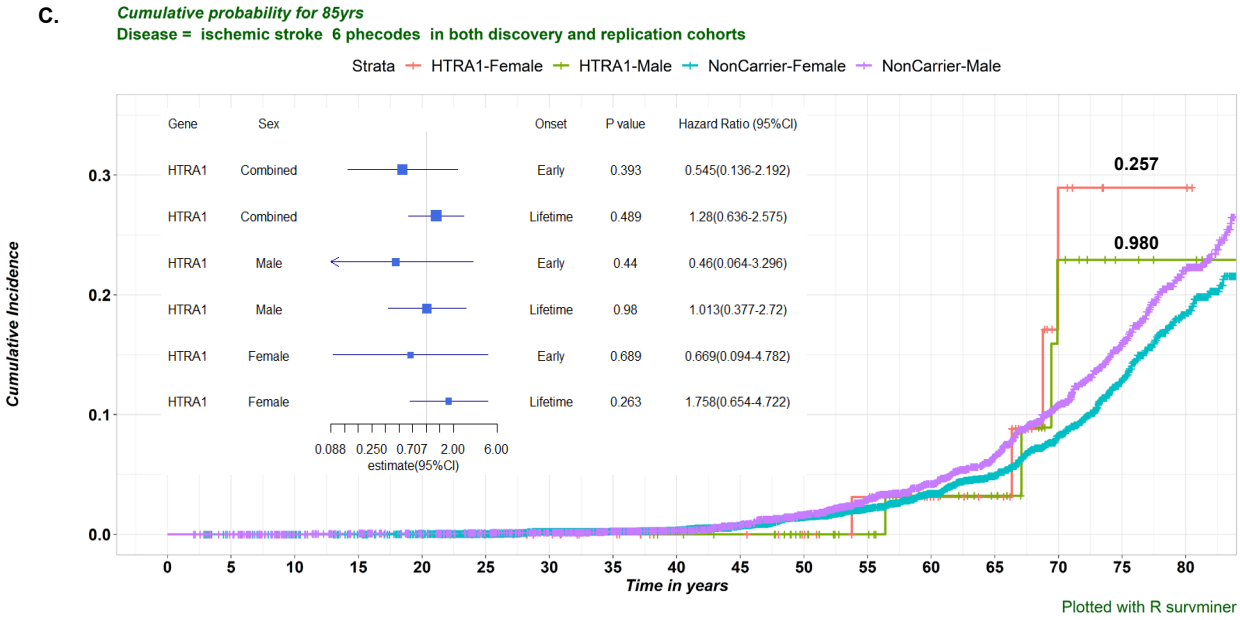
Plotted with R survminer

Figure 4. Kaplan-Meier estimated the cumulative probability of ischemic stroke for carriers compared to matched noncarriers, a sex-dependent analysis.





C.



Cumulative probability for 85yrs

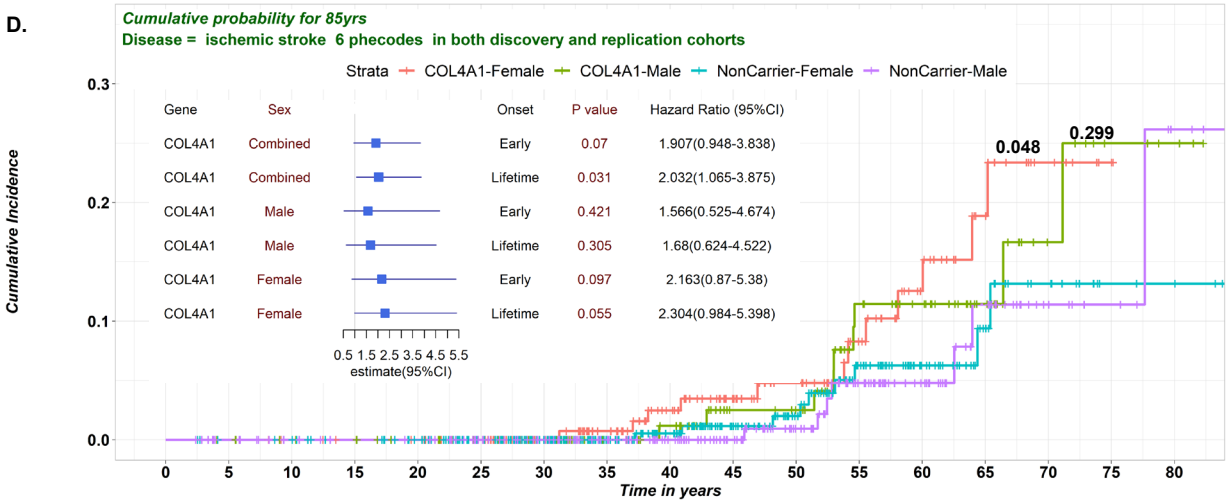
Disease = ischemic stroke 6 phecodes in both discovery and replication cohorts

Strata	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
HTRA1-Female	62	62	61	60	60	55	50	45	40	40	36	31	23	19	6	2	2
HTRA1-Male	55	54	53	53	52	49	49	49	47	45	39	34	26	21	11	6	4
NonCarrier-Female	5276	5268	5245	5231	5192	4995	4662	4281	3901	3561	3164	2712	2229	1661	1157	746	451
NonCarrier-Male	3001	2988	2927	2901	2866	2815	2740	2644	2527	2360	2148	1838	1532	1205	816	510	272

Time in years

Plotted with R survminer

D.

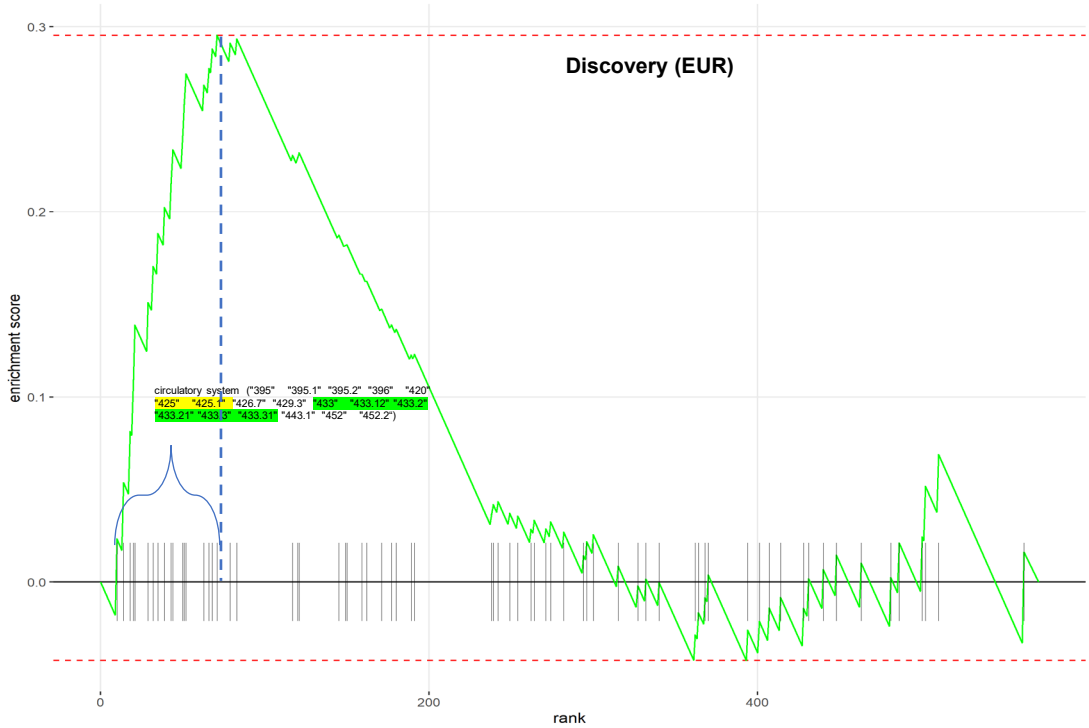


Strata	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
COL4A1-Female	179	179	177	175	174	165	144	123	103	83	65	49	33	19	9	4	1
COL4A1-Male	125	124	121	120	116	112	105	91	83	66	63	46	38	24	10	5	3
NonCarrier-Female	306	303	302	297	290	266	234	198	171	130	104	74	45	26	17	6	5
NonCarrier-Male	204	198	191	188	187	177	166	149	132	112	89	63	42	25	12	8	3

Time in years

Figure 5. PheCode Set Enrichment Analysis was conducted to determine which disease categories were significantly enriched.

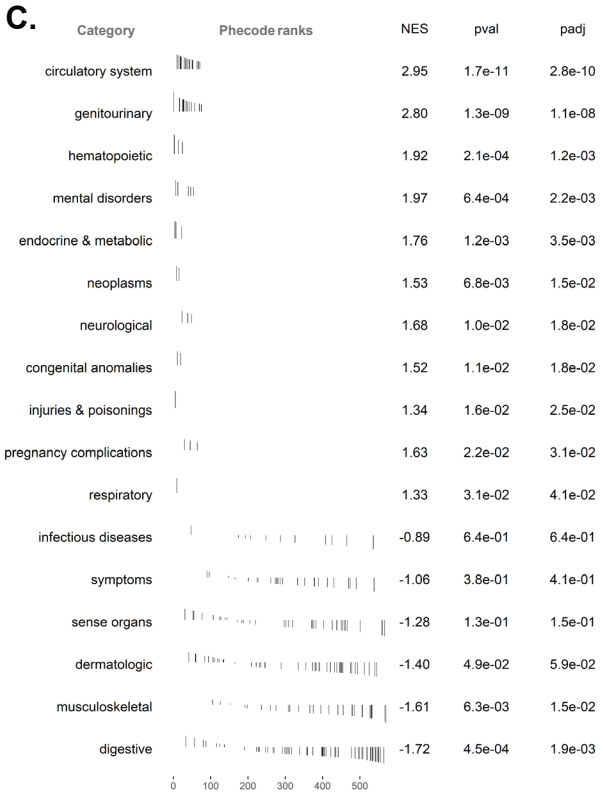
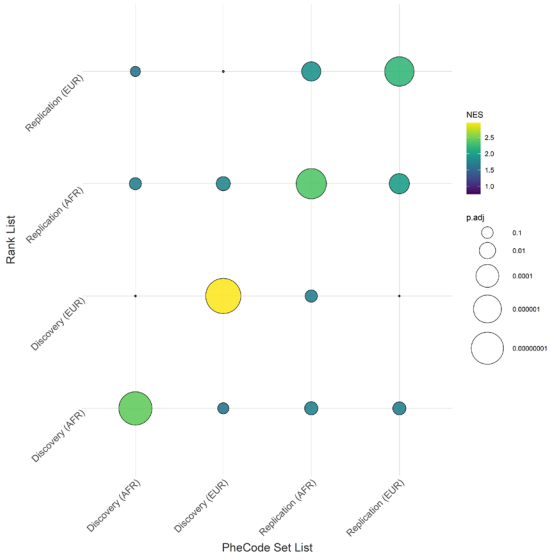
A.



B.

Cut_point	Enrichment_score_max	Category	TotalNum	Num
536	0.061	infectious diseases	11	11
15	0.361	neoplasms	10	2
22	0.096	endocrine & metabolic	64	3
24	0.364	hematopoietic	19	4
54	0.267	mental disorders	29	5
49	0.212	neurological	21	3
566	0.007	sense organs	35	35
71	0.295	circulatory system	69	18
9	0.054	respiratory	40	1
564	0.012	digestive	56	56
75	0.230	genitourinary	68	16
64	0.204	pregnancy complications	13	3
545	0.047	dermatologic	43	43
0	0.000	musculoskeletal	39	28
570	0.000	musculoskeletal	39	28
571	0.000	musculoskeletal	39	28
19	0.491	congenital anomalies	6	2
538	0.059	symptoms	26	26
5	0.213	injuries & poisonings	21	1

C.



D.