Frequency and Phenotype Associations of Rare Variants in 5 Monogenic Cerebral Small Vessel Disease Genes in 200,000 UK Biobank Participants

Amy Christina Ferguson, PhD, Sophie Thrippleton, MBChB, David Henshall, MBChB, Ed Whittaker, MBChB, Bryan Conway, MBChB, Malcolm MacLeod, MBChB, PhD, Rainer Malik, PhD, Konrad Rawlik, PhD, Albert Tenesa, PhD, Cathie Sudlow, PhD, and Kristiina Rannikmae, MD, PhD

Neurol Genet 2022;8:e200015. doi:10.1212/NXG.0000000000200015

Abstract

Background and Objectives
Based on previous case reports and disease-based cohorts, a minority of patients with cerebral small vessel disease (cSVD) have a monogenic cause, with many also manifesting extracerebral phenotypes. We investigated the frequency, penetrance, and phenotype associations of putative pathogenic variants in cSVD genes in the UK Biobank (UKB), a large population-based study.

Methods
We used a systematic review of previous literature and ClinVar to identify putative pathogenic rare variants in CTSA, TREX1, HTRA1, and COL4A1/2. We mapped phenotypes previously attributed to these variants (phenotypes-of-interest) to disease coding systems used in the UKB’s linked health data from UK hospital admissions, death records, and primary care. Among 199,313 exome-sequenced UKB participants, we assessed the following: the proportion of participants carrying ≥1 variant(s); phenotype-of-interest penetrance; and the association between variant carrier status and phenotypes-of-interest using a binary (any phenotype present/absent) and phenotype burden (linear score of the number of phenotypes a participant possessed) approach.

Results
Among UKB participants, 0.5% had ≥1 variant(s) in studied genes. Using hospital admission and death records, 4%–20% of variant carriers per gene had an associated phenotype. This increased to 7%–55% when including primary care records. Only COL4A1 variant carrier status was significantly associated with having ≥1 phenotype-of-interest and a higher phenotype score (OR = 1.29, p = 0.006).

Discussion
While putative pathogenic rare variants in monogenic cSVD genes occur in 1:200 people in the UKB population, only approximately half of variant carriers have a relevant disease phenotype recorded in their linked health data. We could not replicate most previously reported gene-phenotype associations, suggesting lower penetrance rates, overestimated pathogenicity, and/or limited statistical power.
Cerebral small vessel disease (cSVD) refers to a variety of pathologic processes affecting the brain’s small arteries, arterioles, venules, and capillaries. It is estimated that 20% of ischemic strokes, and most hemorrhagic strokes, are caused by cSVD. cSVD is also the most frequent pathology underlying vascular dementia and vascular cognitive impairment.

An unknown minority of cSVD cases are considered monogenic, i.e., caused by a pathogenic variant(s) in one of several genes. While NOTCH3 (implicated in CADASIL [cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy]) is the best known of these, since its first description in 1996, several additional cSVD genes have been identified. Examples include CTSA, TREX1, HTRA1, COL4A1, and COL4A2, but there are additional genes where cSVD is either not the primary associated phenotype (e.g., ADA2 and GLA) or, to date, there is weaker causal evidence (e.g., FOXCI, PITX2, and COLGALT1). Many monogenic cSVD cases also show overlapping systemic and neurologic features.

To date, our knowledge of monogenic cSVD variants’ frequency, penetrance, and phenotype associations comes primarily from case reports, small case series, and family pedigree studies. The resulting data are therefore affected by various biases, including investigation bias (patients with clinically severe and previously described manifestations are more likely to have genetic testing and undergo investigations for known expected associated pathologies if a pathogenic variant in a relevant gene is found), publication bias (clinicians/researchers are more likely to publish a case report/series about clinically severely and/or unusually affected patients), and reporting bias (published case reports/series tend to discuss previously reported or particularly unusual clinical signs and symptoms rather than describe case’s health in an unbiased and systematic way). There have also been few disease-based studies exploring rare variation in cSVD genes in apparently sporadic cases of cSVD, but the population frequency and clinical consequences of these variants remain unknown.

Data from large-scale population-based studies collecting health outcomes in a systematic, unbiased way would provide additional information on the frequency and clinical effect of monogenic cSVD rare variants in a different setting. Investigating routinely collected, linked, health care records would overcome some of the limitations of previous studies.

We aimed to use the UK Biobank (UKB) to assess the population frequency of putative pathogenic rare genetic variation in 5 known monogenic cSVD genes—CTSA, TREX1, HTRA1, COL4A1, and COL4A2 (excluding NOTCH3 because it has already been investigated in the UKB). These genes are included in the recommended clinical framework for diagnosing monogenic cSVD by the European Academy of Neurology. We studied their apparent penetrance (i.e., the proportion of participants with a variant manifesting a relevant cerebral and/or extracerebral clinical phenotype) and gene-phenotype associations in the general population, not selected on the basis of disease or disease risk (Figure 1). Understanding the effect of these variants in the general population will aid the interpretation of variant consequences and inform the clinical response.

### Methods

#### Study Population

The UKB cohort is a prospective study of approximately 500,000 UK residents recruited from 2006 to 2010, aged 40–69 years during recruitment. It has extensive phenotypic information derived from linked health care and death record data, with whole-exome sequencing (WES) data available for approximately 200,000 participants as of October 2020. Full details of the UKB have been previously described.

Our population of interest comprised the 200,603 UKB participants with WES data available from October 2020. We excluded the following individuals: (1) related individuals based on genetic relatedness pairings as provided by the UKB Field 22011, excluding 1 individual from each related pair but preferentially retaining participants carrying a variant-of-interest and (2) participants with sex mismatch (reported sex did not match genetic sex). After these quality control exclusions, our study population comprised 199,313 UKB participants. The following data were available for the whole sample: (1) WES; (2) coded hospital inpatient admissions and death record data (UKB Fields 41270, 40001, and 40002); (3) baseline characteristics (age at the last follow-up on March 2020 [derived from UKB Fields 34 and 52], sex [UKB Fields 31 and 22001], self-reported ethnicity [UKB Field 21000], and Townsend deprivation index—a marker of socioeconomic deprivation [UKB Field 189]). Coded primary care data (UKB Field 42040) were also available for a 48% subset (95,459 participants).

#### Variant Selection

We identified putative pathogenic rare variants in CTSA, TREX1, HTRA1, COL4A1, and COL4A2, which we refer to as variants-of-
interest (Supplemental Methods Putative pathogenic variants file). We defined putative pathogenic rare variants as follows: (1) variants that are reported as causing disease in the published literature based on our systematic review (SysRev variants);6 and/or (2) variants that are reported as pathogenic or likely pathogenic in the ClinVar database (ClinVar variants);19 and (3) those that have a minor allele frequency (MAF) <1% in the UK Biobank. An exception to this was the TREX1 and CTSA genes that in addition to monogenic cSVD are also associated with other specific monogenic disorders (Aicardi-Goutieres syndrome and Galactosialidosis, respectively).20-22 Hence, for these 2 genes, variants reported to be specific for conditions other than cSVD were excluded. We used the Ensembl variant effect predictor (VEP) to describe the variants-of-interest and estimate their pathogenicity and protein effect based on SIFT (Sorting Intolerant From Tolerant), PolyPhen, and SnpEff.23

Phenotype Selection and Mapping
We first compiled a list of cerebral and extracerebral phenotypes previously attributed to CTSA, TREX1, HTRA1, COL4A1, and COL4A2 variants-of-interest (eTable 1, links.lww.com/NXG/A541). We included phenotypes reported as being associated with these genes in online mendelian inheritance in man (OMIM) and/or in our recent systematic review,6,22 from hereon referred to as phenotypes-of-interest. We mapped these phenotype descriptions to the disease coding systems used for recording hospital inpatient admissions, death record, and primary care data in the UKB, i.e., International Classifications of Diseases—10 (ICD-10) and Read V2 and Read V3 disease coding systems, using clinical expertise where needed to ensure phenotype descriptions were mapped to the most appropriate codes. Further detail regarding the mapping process is provided in the Supplemental Methods, and the code lists used for this study can be found in the eMethods (links.lww.com/NXG/A541) Mapped Phenotype codes file.

Data Analyses
Assessing Variant Carrier Frequencies in the UKB and Their Demographic Characteristics
Using Functional Equivalence–derived PLINK files,17,18 we calculated the total number and proportion of UKB participants with ≥1 variant-of-interest (from hereon referred to as variant carriers). We used the $\chi^2$ and 2 sample t tests to assess differences between variant carriers and noncarriers in age, sex, Townsend deprivation index, ethnicity, and the presence of vascular risk factors (0–1 vs ≥2 risk factors described below).

Assessing the Proportion of Variant Carriers With Phenotypes-of-Interest
As a measure of genetic variant penetrance, we calculated the proportion of variant carriers with ≥1 phenotype-of-interest in the hospital inpatient admissions and/or death record data for the
whole study population and in hospital admissions, death record, and/or primary care data for the 48% subset of the study population with linked primary care data. We further explored the proportion of variant carriers with a stroke diagnosis, one of the main clinical manifestations of cSVD.2–4

Assessing Whether Variant Carrier Status Is Associated With Phenotypes-of-Interest

We tested for statistically significant associations between variant carrier status and phenotypes-of-interest by gene. We undertook the primary analyses in the whole cohort and repeated these as secondary analyses in the subgroup with primary care data.

Association With Phenotype-of-Interest Status

For each of the 5 genes, we checked for differences in the proportion of participants with any phenotype-of-interest, and with stroke specifically, among variant carriers compared with noncarriers. We used a χ² test and set a Bonferroni-corrected p value of <0.01 (corrected for the 5 gene-level tests).

Association With Phenotype Burden (Phenotype Score)

We assessed for difference in overall phenotype burden between variant carriers and noncarriers, creating for each gene and participant: binary variant scores using an unweighted gene-based collapsing approach24 (with carriers given a score of 1 and noncarriers a score of 0) and unweighted phenotype scores (based on the number of the phenotypes-of-interest associated with the gene). For example, HTRAI is associated with 6 phenotypes-of-interest, and a participant could therefore have a minimum HTRAI phenotype score of 0 (if they did not manifest any phenotypes-of-interest) and a maximum score of 6 (if they manifested all 6 phenotypes-of-interest) (eTable 1, links.lww.com/NXG/A541). We used Poisson regression (due to the phenotype scores being based on exact counts) to investigate, for each gene, the association between carrying a rare variant and phenotype score including age at last follow-up, sex, Townsend index, and 20 genetic principal components (PCs) as covariates. We used a Bonferroni-corrected p value threshold of <0.01 to determine significance (corrected for 5 gene-level tests).

For primary analyses, we additionally performed the following: (1) adjusted for vascular risk factors, including body mass index (<30 kg/m² vs ≥30 kg/m²), smoking status (nonsmoker vs current/previous smoker), alcohol consumption (≤twice a week vs > twice a week), and hypertension and diabetes (present vs absent)25,26 (2) checked for interactions between variant carrier status with sex and ethnicity; and (3) ran leave-one-out sensitivity analyses, where each phenotype was removed from the phenotype score one by one (using logistic instead of Poisson regression in the case of COL4A2, which only had 2 phenotypes-of-interest to start with).

Standard Protocol Approvals, Registrations, and Patient Consents

All UKB participants provided informed consent as part of the UKB recruitment process, in accordance with the Declaration of Helsinki 1964 and its later amendments or comparable ethical standards.

Data Availability

Anonymized phenotype and genotype data are available from the UK Biobank;16 the UK Biobank resource can be accessed by approved researchers (ukbiobank.ac.uk/). ICD-10 and Read v2/3 code lists used for analysis are available in the eAppendix (Supplemental code list, links.lww.com/NXG/AS41).

Results

We identified a total of 260 variants-of-interest across the 5 genes to investigate in the UKB: 152 exclusively from SysRev,
Table 2 Variant Carriers in the UK Biobank by Gene and Ethnic Group

<table>
<thead>
<tr>
<th>Gene</th>
<th>Self-reported ethnic group</th>
<th>All</th>
<th>White (%)</th>
<th>Black (%)</th>
<th>Asian (%)</th>
<th>Mixed (%)</th>
<th>Other (%)</th>
<th>Unknown (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTRA1</td>
<td></td>
<td>234</td>
<td>229 (98%)</td>
<td>1 (0.4%)</td>
<td>—</td>
<td>—</td>
<td>1 (0.4%)</td>
<td>3 (1.3%)</td>
</tr>
<tr>
<td>COL4A1</td>
<td></td>
<td>481</td>
<td>465 (97%)</td>
<td>4 (0.8%)</td>
<td>5 (1%)</td>
<td>2 (0.4%)</td>
<td>5 (1%)</td>
<td>—</td>
</tr>
<tr>
<td>COL4A2</td>
<td></td>
<td>336</td>
<td>90 (27%)</td>
<td>190 (57%)</td>
<td>3 (0.9%)</td>
<td>14 (4%)</td>
<td>31 (10%)</td>
<td>8 (2%)</td>
</tr>
<tr>
<td>Any gene-of-interest</td>
<td></td>
<td>1,050</td>
<td>783 (75%)</td>
<td>195 (19%)</td>
<td>3 (0.8%)</td>
<td>16 (2%)</td>
<td>37 (4%)</td>
<td>11 (1%)</td>
</tr>
</tbody>
</table>

*Other* ethnicity includes Chinese. There was 1 participant with a variant in both HTRA1 and COL4A1. There were no participants carrying variants in TREX1 or CTSA.

52 exclusively from ClinVar, and 54 from both sources. We found a total of 37 variants present in ≥1 of the 199,313 included UKB participants: 24 exclusively from SysRev, 2 exclusively from ClinVar, and 11 from both sources, but these did not include any TREX1 or CTSA variants-of-interest. The number of variants represented in the UKB varied for each gene: 5 COL4A2 variants (3 SysRev, 1 ClinVar, and 1 from both sources), 15 HTRA1 variants (8 Sys Rev and 7 from both sources), and 17 COL4A1 variants (13 SysRev, 1 ClinVar, and 3 from both sources). Across these variants, MAF in the UKB ranged from 0.0005% to 0.14%.

VEP predicted 92% (22/24) SysRev, 100% (2/2) ClinVar, and 100% (11/11) variants from both sources to be missense or nonsense variants. Of the remaining 2 SysRev variants, one was in the 5’ untranslated region and the other an intronic splice donor variant. SnpEff predicted the nonsense variants and the splice donor variant to have a high effect, while all but one of the remaining variants were of moderate effect. Overall, 21/37 variants (50% SysRev; 50% ClinVar, and 73% both sources) were predicted to have probably damaging and deleterious effects on protein structure and function (eTable 4, links.lww.com/NXG/AS41). It is of note that VEP cannot provide potential protein structure and function effects of variation in untranslated regions, intronic regions, or splice acceptor/donor sites. We identified 2 to 12 phenotypes-of-interest per gene (eTable 1, links.lww.com/NXG/AS41). When mapping these to disease coding systems, the number of codes per phenotype varied widely, ranging from 6 codes for dry mouth to 173 codes for degenerative spine disease (Supplemental code list, links.lww.com/NXG/AS41).

Assessing Variant Carrier Frequencies in the UKB and Their Demographic Characteristics

Among 199,313 UKB participants, 1,050 participants (0.5%) had ≥1 variant-of-interest, resulting in 234 HTRA1, 481 COL4A1, and 336 COL4A2 variant carriers, with 1 participant carrying a variant in both HTRA1 and COL4A1. When variant carriers were not preferentially selected from related pairs, their overall frequency remained the same. Most variant carriers (96%; 1,003/1,050) possessed a SysRev variant, 0.3% (3/1,050) a ClinVar variant, and 4% (44/1,050) a variant represented in both SysRev and ClinVar (eTable 2, links.lww.com/NXG/AS41).

The mean age (at the last follow-up) of all included UKB participants was 68.2 years. The mean Townsend deprivation index was −1.34, and 55% of participants were female. There was no significant difference in age at the last follow-up, sex, or the presence of vascular risk factors between variant carriers and noncarriers. Carriers had significantly higher levels of deprivation than noncarriers (the mean Townsend index = −0.44 vs −1.34, p < 2.2 × 10−16) and were more likely to be of non-White ethnicity (p < 2.2 × 10−16) (Table 1).

Based on these findings, we performed post hoc analyses comparing variant carriers and noncarriers in terms of: (1) the mean Townsend index stratified by manifestation of a phenotype-of-interest and (2) the Townsend index breakdown by quintiles derived from the QResearch database,27,28 comparing the frequency of phenotypes-of-interest. Variant carriers had significantly higher levels of deprivation compared with noncarriers regardless of whether or not they manifested a phenotype-of-interest (−0.31 vs −1.07, p = 0.0004 and −0.44 vs −1.4, p < 2.2 × 10−16, respectively). Comparing the frequency of participants manifesting a phenotype, among variant carriers, it was similar in the least and most deprived quintiles (23% vs 24%), whereas among noncarriers, the least deprived had a lower frequency compared with the most deprived (17% vs 23%) (eTable 5, links.lww.com/NXG/AS41).

Exploring the ethnicity distribution further, most HTRA1 and COL4A1 variant carriers (≥97%) were of self-reported White ethnicity. For COL4A2, however, 57% (190/336) of variant carriers were of self-reported Black ethnicity, driven by 2 variants c.3448C>A and c.5068G>A present in 1.2% and 4.7% of Black participants, respectively. A further 14% of COL4A2 variant carriers were of mixed and other ethnic groups (Table 2).

Assessing the Proportion of Variant Carriers With Phenotypes-of-Interest

The proportion of variant carriers (N = 1,050) with ≥1 phenotype-of-interest in the hospital inpatient admissions and/or death record data was as follows: HTRA1 9% (21/234); COL4A1
Figure 2 Proportion of Variant Carriers With a Phenotype-of-Interest

A. In-hospital admission and death record data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype: Any phenotype of interest</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTRA1</td>
<td>21/234 (91%)</td>
<td>11/234 (47%)</td>
</tr>
<tr>
<td>COL4A1</td>
<td>18/481 (3.7%)</td>
<td>15/481 (3.1%)</td>
</tr>
<tr>
<td>COL4A2</td>
<td>10/336 (3.0%)</td>
<td>15/336 (4.5%)</td>
</tr>
</tbody>
</table>

B. In-hospital admission/death record and primary care data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype: Any phenotype of interest</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTRA1</td>
<td>64/117 (55%)</td>
<td>8/117 (7%)</td>
</tr>
<tr>
<td>COL4A1</td>
<td>93/236 (40%)</td>
<td>14/236 (6%)</td>
</tr>
<tr>
<td>COL4A2</td>
<td>9/132 (7%)</td>
<td>9/132 (7%)</td>
</tr>
</tbody>
</table>

N = total number of variant carriers; n = number of variant carriers with any phenotype-of-interest or stroke.

20% (95/481); and COL4A2 4% (15/336) (Figure 2A). This proportion increased when we explored the smaller subset of the variant carriers for whom primary care data was also available (N = 484): HTRA1 55% (64/117); COL4A1 40% (93/236); and COL4A2 7% (9/132) (Figure 2B). Among variant carriers manifesting a phenotype-of-interest, stroke was not always the most common phenotype: HTRA1 13%–52%, COL4A1 15%–19%, and COL4A2 93–100% (Figure 2, A and B).

Assessing Whether Variant Carrier Status Is Associated With Phenotypes-of-Interest

Association With Phenotype-of-Interest Status

For phenotypes-of-interest in the hospital inpatient admissions and/or death record data, a higher proportion of COL4A1 variant carriers compared with noncarriers had a COL4A1-related phenotype (20% in carriers vs 15% in noncarriers, p = 0.01). We found no significant associations for HTRA1 and COL4A2 and no significant associations for any gene in the secondary analyses, also including primary care data. There was also no significant difference in the proportion of stroke cases seen between carriers and noncarriers for any of the genes.

Association With Phenotype Burden (Phenotype Score)

COL4A1 carriers also had a greater phenotype score compared with noncarriers (OR = 1.29, p = 0.006). We found no significant associations for HTRA1 and COL4A2, or for any gene in the secondary analyses including primary care data (Figure 3). The associations remained similar after adjusting the primary analyses for the presence of vascular risk factors (eTable 6, links.lww.com/NXG/A541). We found no significant interactions with sex (HTRA1 p = 0.41; COL4A1 p = 0.14; and COL4A2 p = 0.49) and ethnicity (HTRA1 p = 0.60; COL4A1 p = 0.88; and COL4A2 p = 0.19) for any gene.

Leave-one-out sensitivity analyses did not change the results significantly for HTRA1 or COL4A2 (eFigure 1, links.lww.com/NXG/A541). For COL4A1, removing cataract, migraine, or stroke from the phenotype score rendered the association no longer significant, suggesting these phenotypes are important in driving the association seen in the primary analyses (Figure 4).

Discussion

We found that while 1:200 UKB participants carry a previously reported putative pathogenic rare variant in one of the 5 cSVD genes included in our study, only 4%–20% of variant carriers per gene had an associated phenotype recorded in their hospital admission/death records, and this rose moderately to 7%–55% when also including primary care records.

We are not aware of previous studies investigating these 5 genes in a population-based setting. There has, however, been a study of another monogenic cSVD gene, which demonstrated that 1:450 UKB participants carry a putative pathogenic (i.e., cysteine-altering) variant NOTCH3.14 Among the few disease-based studies exploring rare variation in cSVD genes in apparently sporadic cases, one large study found that ~1.70 patients with lacunar stroke had a monogenic cause.11 However, this study included only patients with stroke (one of several manifestations of cSVD), excluded already diagnosed monogenic cSVD cases, and involved an overlapping but not identical set of genes with a different definition of putative pathogenicity compared with our study, limiting direct comparisons.

We did not find any carriers of CTSA or TREX1 putative pathogenic rare variants in the UKB. Potential reasons include the following: (1) variants in these genes are extremely rare in general and/or rare in a population-based setting owing to their severe phenotype manifestations; this is particularly relevant for TREX1, where most of the variants were frameshifts that were
not as highly represented in the UKB WES data compared with single-nucleotide variants; (2) the overall number of variants-of-interest in these genes was smaller, and they are not present in the UKB by chance.

We found a significantly higher level of deprivation among variant carriers compared with that among noncarriers. Further post hoc analyses suggested that this difference is not explained by variant carriers manifesting a disease phenotype. Exploring the Townsend deprivation index distribution by quintiles suggested that carrying a putative pathogenic rare variant increases the chances of having a phenotype among the least deprived but not among the most deprived participants. One possible explanation might be that participants in the most deprived group have an already increased risk due to environmental factors, whereas in the least deprived group, genetic variation plays a more important role. However, this was not the primary aim of our study and requires further research.

Most of the COL4A2 carriers in the UKB were of non-White self-reported ethnic group. The 2 genetic variants driving the frequency among non-White participants had previously been reported in the literature as causing intracerebral hemorrhage in persons of White Hispanic and African American background. Earlier studies have focused mainly on investigating European populations, while case reports and series are often missing ethnicity information, leaving monogenic cSVD and relevant genetic variation prevalence estimates among non-White ethnicities largely unknown. One study that did investigate this in the Genome Aggregation Database among 7 ethnic groups did not find a similar enrichment of COL4A2 carriers among participants of Black and other ethnicities, although interestingly, COL4A1 pathogenic variants were most prevalent among Africans/African Americans. Furthermore, epidemiologic studies have demonstrated racial and ethnic differences in cSVD manifestations and burden, which may have a genetic component. These findings underline the importance of extending future genetic studies to include a broader range of ethnic groups.

Evidence from the existing literature summarized in a systematic review demonstrated a higher proportion of variant carriers manifesting a relevant phenotype compared with our study, with estimates of 59% for COL4A2, 75% for COL4A1, and 77% for HTRA1. Considering the mean age of UKB participants is older than the mean age of individuals included in the systematic review, it is unlikely this difference is explained by a limited duration of follow-up in the UKB and our analysis capturing participants who will go on to develop the phenotypes in the future. Our results suggesting lower frequency of phenotypic manifestations in carriers of variants associated with cSVD in a population-based setting are in keeping with findings from other monogenic disease investigations. Similar results have been shown for monogenic diabetes and CADASIL, demonstrating ascertainment context is crucial when interpreting the consequences of monogenic variants. This has important implications for genetic screening and counseling in the clinical setting, as well as potentially demonstrating the need for further investigations of some of these putative pathogenic variants to confirm their causal role in monogenic cSVD.

We identified a significant association for COL4A1 between variant carrier status and phenotype score. Leave-one-out subgroup analyses indicated that migraine, cataract, and stroke contributed most to the association seen.
We did not find significant associations between HTRA1/COL4A2 and phenotypes previously attributed to putative pathogenic variants in those genes. This may suggest that these variants: (1) have reduced penetrance; (2) have variable expressivity; (3) are not all pathogenic despite previous reports; (4) have previously been incorrectly associated with certain disease phenotypes in the literature and OMIM; and/or (5) were not present in large enough numbers limiting statistical power.

Our variants-of-interest had evidence from the literature and ClinVar to support that they are pathogenic. Using Ensemble VEP to further investigate these variants generally corroborated this assumption—all but one had a moderate or high effect according to SnpEff, and approximately two-thirds were bioinformatically predicted to be deleterious and probably damaging to protein structure and function.

Our study has several strengths. By using all available sources of routinely collected health data, we were able to systematically capture the full range of phenotypes previously reported to be associated with monogenic cSVDs, both cerebral and extracerebral. This approach limits several of the biases that affect case reports and series and is feasible among large numbers of participants. Our study is the first to explore rare variation in these 5 genes in a population-based setting and as such provides valuable information on their frequency and spectrum of clinical consequences, supplementing existing knowledge derived mainly from case reports and series. This in turn can inform clinicians of various specialties and clinical geneticist in selecting patients to test and when counseling variant carriers in the clinical setting.

There are also limitations that need to be considered. First, our efforts to ensure we have identified variants with good evidence of pathogenicity by relying on published literature and the ClinVar database may have caused some other pathogenic variants to be excluded from this study. Second, we used routinely collected coded administrative health data to identify disease phenotypes. While we made significant efforts to map the phenotypes-of-interest systematically and transparently to relevant disease codes, driven by clinically informed selection, these coded data are likely to identify some false-positive and miss true-positive cases. In addition, it was more challenging to map some phenotypes than others, and some health conditions (e.g., alopecia, migraine, or muscle cramps) are less likely to lead the person to seek medical help and hence be captured by the coded data. It is important that for the stroke phenotype, our code lists captured UKB participants who had experienced any stroke subtype, rather than cSVD-type of stroke specifically. Third, the number of phenotypes-of-interest varied between genes, introducing a potential bias when investigating the proportion of carriers with phenotypes-of-interest. Despite finding associations between variant carrier status and the presence of phenotypes, it is possible that the phenotype is the result of other causes, especially if the phenotype is prevalent in the population. Fourth, the UKB population is highly likely affected by the healthy volunteer bias, with clinically severely affected variant carriers less likely to enroll in the first place. Hence, the variants identified among the UKB population may be those with lower overall penetrance, variable expressivity, and weaker evidence of pathogenicity. Fifth, even when collapsing the variants across each gene, statistical power for detecting small and moderate variant-level genotype-phenotype

---

**Figure 4 COL4A1 Leave-One-Out Analyses**

<table>
<thead>
<tr>
<th>Hospital admission and death record data-based COL4A1 phenotype score</th>
<th>OR (95% CI)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No cataract</td>
<td>1.35 (1.06, 1.68)</td>
<td>0.011**</td>
</tr>
<tr>
<td>No kidney cyst</td>
<td>1.29 (1.07, 1.55)</td>
<td>0.007</td>
</tr>
<tr>
<td>No hematuria</td>
<td>1.33 (1.07, 1.62)</td>
<td>0.007</td>
</tr>
<tr>
<td>No muscle cramp</td>
<td>1.28 (1.05, 1.53)</td>
<td>0.010</td>
</tr>
<tr>
<td>No myalgia</td>
<td>1.28 (1.06, 1.54)</td>
<td>0.008</td>
</tr>
<tr>
<td>No arrhythmia</td>
<td>1.31 (1.07, 1.59)</td>
<td>0.006</td>
</tr>
<tr>
<td>No Raynaud phenomenon</td>
<td>1.31 (1.08, 1.56)</td>
<td>0.005</td>
</tr>
<tr>
<td>No hemolytic anemia</td>
<td>1.28 (1.06, 1.54)</td>
<td>0.009</td>
</tr>
<tr>
<td>No migraine</td>
<td>1.24 (1.02, 1.50)</td>
<td>0.030**</td>
</tr>
<tr>
<td>No stroke</td>
<td>1.29 (1.05, 1.57)</td>
<td>0.013**</td>
</tr>
<tr>
<td>No vascular dementia</td>
<td>1.28 (1.06, 1.54)</td>
<td>0.009</td>
</tr>
</tbody>
</table>

**Association no longer significant. CI = confidence interval; OR = odds ratio.**
association effects remained low. Furthermore, routinely collected administrative disease codes did not provide data on preclinical cognitive decline and cerebral radiologic features of UKB participants, important (and sometimes the only) manifestations of cSVD.4

Future studies will be required to extend these findings to other populations, aiming to better understand and minimize healthy volunteer biases and assessing a broader range of ages and socioeconomic and ethnic groups with even larger sample sizes. As methods of disease identification from routinely collected health data develop further, this will also allow more comprehensive and reliable capture of phenotypes-of-interest. Investigations including preclinical phenotypes, such as cognitive impairment, could also provide greater understanding of the overall manifestations of putative pathogenic variants. Future work could also explore all rare variants in these genes, stratifying them by level of evidence for pathogenicity and using the richness of the routinely collected health data to undertake phenotype-wide association studies. Finally, expanding this work to include a broader list of genes of importance in monogenic cSVD, as more evidence becomes available, may allow us to gain a better understanding of the phenotypic manifestations and mechanisms of cSVD.

In conclusion, putative pathogenic rare variants in 5 monogenic cSVD genes occur in the population at a frequency of 1:200, but only up to half of variant carriers have a relevant disease phenotype recorded in their linked health data. We could not replicate most previously reported gene-phenotype associations, suggesting lower penetrance rates, overestimated pathogenicity, and/or limited statistical power. We also highlight the importance of considering the wider spectrum of phenotypic manifestations in cSVD.

Acknowledgment
The authors acknowledge Professor Martin Dichgans from the Institute for Stroke and Dementia Research in Munich Germany for his input at the earlier stages of conception of this work.

Study Funding
A. Ferguson is funded by BHF award RE/18/5/34216. K. Rannikmae and A. Ferguson are funded by Health Data Research UK Rutherford fellowship (MR/S004130/1). A. Tenesa is funded by HDR-UK award HDR-9004 and HDR-9003.

Disclosure
The authors declare no competing interests. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

Publication History
Received by Neurology: Genetics February 24, 2022. Accepted in final form June 17, 2022. This manuscript was prepublished in MedRxiv doi: doi.org/10.1101/2021.11.17.21266447. Submitted and externally peer reviewed. The handling editor was Stefan M. Pulst, MD, Dr med, FAAN.

Appendix Authors

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy Christina Ferguson, PhD</td>
<td>Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data</td>
</tr>
<tr>
<td>Sophie Thripplenton, MBChB</td>
<td>Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>David Henshall, MBChB</td>
<td>Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>Ed Whittaker, MBChB</td>
<td>Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content</td>
</tr>
<tr>
<td>Bryan Conway, MBChB</td>
<td>Centre for Cardiovascular Science, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; additional contributions: clinical expertise provided</td>
</tr>
<tr>
<td>Malcolm MacLeod, MBChB, PhD</td>
<td>Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content</td>
</tr>
<tr>
<td>Rainer Malik, PhD</td>
<td>Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany</td>
<td>Drafting/revision of the article for content, including medical writing for content; study concept or design</td>
</tr>
<tr>
<td>Konrad Rawlik, PhD</td>
<td>The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data</td>
</tr>
<tr>
<td>Albert Tenesa, PhD</td>
<td>Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom; The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom; MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; study concept or design</td>
</tr>
<tr>
<td>Cathie Sudlow, PhD</td>
<td>BHF Data Science Centre, Health Death Research United Kingdom, London, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; study concept or design</td>
</tr>
<tr>
<td>Kristiina Rannikmae, MD, PhD</td>
<td>Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom</td>
<td>Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data; and additional contributions: also a corresponding author for this article.</td>
</tr>
</tbody>
</table>


Frequency and Phenotype Associations of Rare Variants in 5 Monogenic Cerebral Small Vessel Disease Genes in 200,000 UK Biobank Participants
Amy Christina Ferguson, Sophie Thrippleton, David Henshall, et al.

Neurol Genet 2022;8:
DOI 10.1212/NXG.0000000000200015

This information is current as of August 24, 2022

Updated Information & Services
including high resolution figures, can be found at:
http://ng.neurology.org/content/8/5/e200015.full.html

References
This article cites 31 articles, 5 of which you can access for free at:
http://ng.neurology.org/content/8/5/e200015.full.html##ref-list-1

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
http://ng.neurology.org/misc/about.xhtml#permissions

Reprints
Information about ordering reprints can be found online:
http://ng.neurology.org/misc/addir.xhtml#reprintsus

Neurol Genet is an official journal of the American Academy of Neurology. Published since April 2015, it is an open-access, online-only, continuous publication journal. Copyright Copyright © 2022 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology. All rights reserved. Online ISSN: 2376-7839.