

1 **Frequency and phenotype associations of rare variants in five monogenic cerebral small**
2 **vessel disease genes in 200,000 UK Biobank participants with whole exome sequencing**
3 **data**

4 Amy C. Ferguson¹, Sophie Thrippleton², David E. Henshall¹, Ed Whittaker², Bryan Conway³,
5 Malcolm MacLeod⁴, Rainer Malik⁵, Konrad Rawlik⁶, Albert Tenesa^{1,6,7}, Cathie Sudlow⁸,
6 Kristiina Rannikmäe¹

7 ¹ Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK

8 ² Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

9 ³ Centre for Cardiovascular Science, The Queen's Medical Research Institute, University of Edinburgh,
10 Edinburgh, UK

11 ⁴ Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

12 ⁵ Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany

13 ⁶ The Roslin Institute, University of Edinburgh, Edinburgh, UK

14 ⁷ MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General
15 Hospital, Edinburgh, UK

16 ⁸ BHF Data Science Centre, Health Death Research UK, London, UK

17 **Corresponding authors:** Amy C. Ferguson, afergus8@ed.ac.uk; Kristiina Rannikmäe,

18 kristiina.rannikmae@ed.ac.uk

19

20 **Abstract**

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

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

33 Among UKB participants, 0.5% had ≥ 1 variant(s) in studied genes. Using hospital admission and
34 death records, 4-20% of variant carriers per gene had an associated phenotype. This increased to 7-
35 55% when including primary care records. Only *COL4A1* variant carrier-status was significantly
36 associated with having ≥ 1 phenotype-of-interest and a higher phenotype score (OR=1.29, $p=0.006$).

37 While putative pathogenic rare variants in monogenic cSVD genes occur in 1:200 people in the UKB
38 population, only around half of variant carriers have a relevant disease phenotype recorded in their
39 linked health data. We could not replicate most previously reported gene-phenotype associations,
40 suggesting lower penetrance rates, overestimated pathogenicity and/or limited statistical power.

41

42 **Introduction**

43 Cerebral small vessel disease (cSVD) refers to a variety of pathological processes impacting the
44 brain's small arteries, arterioles, venules, and capillaries¹. It is estimated that 20% of ischaemic
45 strokes, and most haemorrhagic strokes, are caused by cSVD. cSVD is also the most frequent
46 pathology underlying vascular dementia and vascular cognitive impairment²⁻⁴.

47 An unknown minority of cSVD cases are considered monogenic, i.e., caused by a pathogenic
48 variant(s) in one of several genes. While *NOTCH3* (implicated in cerebral autosomal dominant

49 arteriopathy with subcortical infarcts and leukoencephalopathy, CADASIL) is the best known of
50 these, since its first description in 1996⁵, several additional cSVD genes have been identified.
51 Examples include *CTSA*, *TREX1*, *HTRA1*, *COL4A1* and *COL4A2*, but there are additional genes where
52 cSVD is either not the primary associated phenotype (e.g., *ADA2* and *GLA*) or to date there is weaker
53 causal evidence (e.g., *FOXC1*, *PITX2*, *COLGALT1*)². Many monogenic cSVD cases also show
54 overlapping systemic and neurological features^{2,6,7}. Our recent systematic literature review of 800
55 individual cases carrying putative pathogenic variants in six cSVD genes found that several
56 monogenic cSVD genes are associated with a range of extra-cerebral phenotypes affecting ocular,
57 renal, hepatic, muscle, and haematological systems⁶. These extra-cerebral phenotypes can be
58 diagnostically important through contributing to clinically recognisable syndromes informing genetic
59 testing, as well as provide mechanistic insights into the pathophysiology of cSVD^{6,8,9}.

60 To date, our knowledge of monogenic cSVD variants' frequency, penetrance and phenotype
61 associations comes primarily from case reports, small case series and family pedigree studies⁶. The
62 resulting data is therefore affected by various biases, including investigation bias (patients with
63 clinically severe and previously described manifestations are more likely to have genetic testing and
64 undergo investigations for known expected associated pathologies if a pathogenic variant in a
65 relevant gene is found), publication bias (clinicians/researchers are more likely to publish a case
66 report/series about clinically severely and/or unusually affected patients), and reporting bias
67 (published case reports/series tend to discuss previously reported or particularly unusual clinical
68 signs and symptoms rather than describe case's health in an unbiased and systematic way)^{6,10,11}.

69 There have also been few disease-based studies exploring rare variation in cSVD genes in apparently
70 sporadic cases of cSVD¹²⁻¹⁵, but the population frequency and clinical consequences of these variants
71 remains unknown.

72 Data from large-scale population-based studies collecting health outcomes in a systematic, unbiased
73 way would provide additional information on the frequency and clinical impact of monogenic cSVD

74 rare variants in a different setting. Investigating routinely collected, linked, healthcare records would
75 overcome some of the limitations of previous studies.

76 The UK Biobank (UKB) cohort is a prospective study of approximately 500,000 UK residents recruited
77 from 2006-2010, aged 40-69 at the time of recruitment. It has extensive phenotypic information
78 derived from linked healthcare and death record data, with whole exome sequencing (WES) data
79 available for around 200,000 participants as of October 2020. Full details of UKB have been
80 previously described¹⁶⁻¹⁸.

81 We aimed to use UKB to assess the population frequency of putative pathogenic rare genetic
82 variation in five known monogenic cSVD genes - *CTSA*, *TREX1*, *HTRA1*, *COL4A1* and *COL4A2*
83 (excluding *NOTCH3* since it has already been investigated in UKB¹⁹). We studied their apparent
84 penetrance (i.e., the proportion of participants with a variant manifesting a relevant cerebral and/or
85 extra-cerebral clinical phenotype) and gene-phenotype associations in the general population, not
86 selected on the basis of disease or disease risk (Figure 1).

87 **Methods**

88 ***Study population***

89 Our population of interest comprised the 200,603 UKB participants with WES data available from
90 October 2020. We excluded: (i) related individuals based on genetic relatedness pairings as provided
91 by UKB Field 22011, excluding one individual from each related pair but preferentially retaining
92 participants carrying a variant-of-interest; and (ii) participants with sex mismatch (reported sex did
93 not match genetic recorded sex). After these quality control exclusions, our study population
94 comprised 199,313 UKB participants. The following data was available for the whole sample: (i) WES;
95 (ii) coded hospital inpatient admissions and death record data [UKB Fields 41270, 40001, 40002]; (iii)
96 baseline characteristics (age at last follow up on March 2020 [derived from UKB Fields 34 and 52],
97 sex [UKB Fields 31 and 22001], self-reported ethnicity [UKB Field 21000], Townsend deprivation

98 index – a marker of socio-economic deprivation [UKB Field 189]). Coded primary care data [UKB Field
99 42040] was also available for a 48% subset (95,459 participants).

100 ***Variant selection***

101 We identified putative pathogenic rare variants in *CTSA*, *TREX1*, *HTRA1*, *COL4A1* and *COL4A2*, which
102 we refer to as ‘variants-of-interest’. We defined putative pathogenic rare variants as variants that
103 are: (i) reported as causing disease in the published literature based on our systematic review
104 (SysRev variants)⁶; and/or (ii) reported as “pathogenic” or “likely pathogenic” in the ClinVar database
105 (ClinVar variants)²⁰; and (iii) have a minor allele frequency (MAF) <1% in the UK Biobank. An
106 exception to this were *TREX1* and *CTSA* genes, which in addition to monogenic cSVD are also
107 associated with other specific monogenic disorders (Aicardi-Goutieres syndrome and
108 Galactosialidosis, respectively)^{21–23}. Hence for these two genes, variants reported to be specific for
109 conditions other than cSVD were excluded. We used the Ensembl variant effect predictor (VEP) to
110 estimate the pathogenicity and protein impact of variants-of-interest, based on SIFT, PolyPhen and
111 SnpEff²⁴.

112 ***Phenotype selection and mapping***

113 We first compiled a list of cerebral and extra-cerebral phenotypes previously attributed to *CTSA*,
114 *TREX1*, *HTRA1*, *COL4A1* and *COL4A2* variants-of-interest. We included phenotypes reported as being
115 associated with these genes in Online Mendelian Inheritance in Man (OMIM) and/or in our recent
116 systematic review^{6,23}, from here-on referred to as phenotypes-of-interest. We mapped these
117 phenotypes to the disease-coding systems used for recording hospital inpatient admissions, death
118 record and primary care data in the UKB, i.e., International Classifications of Diseases – 10 (ICD-10),
119 Read V2 and Read V3 disease-coding systems. Further detail regarding the mapping process is
120 provided in the Supplemental Methods.

121 ***Data analyses***

122 Assessing variant carrier frequencies in the UKB and their demographic characteristics

123 Using Functional Equivalence (FE)-derived PLINK files^{17,18}, we calculated the total number and
124 proportion of UKB participants with ≥ 1 variant-of-interest (from here-on referred to as “variant
125 carriers”). We used Chi-squared and two sample t-tests to assess differences between variant
126 carriers and non-carriers in age, sex, Townsend deprivation index, ethnicity, and the presence of
127 vascular risk factors (0-1 versus ≥ 2 risk factors described in Supplemental Table 6 legend).

128 Assessing the proportion of variant carriers with phenotypes-of-interest

129 As a measure of genetic variant penetrance, we calculated the proportion of variant carriers with ≥ 1
130 phenotype-of-interest in the hospital inpatient admissions and/or death record data for the whole
131 study population, and in hospital admissions, death record and/or primary care data for the 48%
132 subset of the study population with linked primary care data. We further explored the proportion of
133 variant carriers with a stroke diagnosis, one of the main clinical manifestations of cSVD²⁻⁴.

134 Assessing whether variant carrier status is associated with phenotypes-of-interest

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

138 *Association with phenotype-of-interest status*

139 For each of the five genes, we checked for differences in the proportion of participants with any
140 phenotype-of-interest, and with stroke specifically, among variant carriers compared to non-carriers.
141 We used a Chi-squared test and set a Bonferroni-corrected significance p-value of < 0.01 (corrected
142 for the 5 gene-level tests).

143 *Association with phenotype burden (phenotype score)*

144 We assessed for difference in overall phenotype burden between variant carriers and non-carriers,
145 creating for each gene and participant: binary variant scores using an unweighted gene-based
146 collapsing approach²⁵ (with carriers given a score of 1 and non-carriers a score of 0); and unweighted

147 phenotype scores (based on the number of the phenotypes-of-interest associated with the gene).
148 For example, *HTRA1* is associated with 6 phenotypes-of-interest and a participant could therefore
149 have a minimum *HTRA1* phenotype score of 0 (if they did not manifest any phenotypes-of-interest)
150 and a maximum score of 6 (if they manifested all 6 phenotypes-of-interest) (Supplemental Table 1).
151 We used Poisson regression (due to the phenotype scores being based on exact counts) to
152 investigate, for each gene, the association between carrying a rare variant and phenotype score
153 including age at last follow-up, sex, Townsend index, and 20 genetic principal components (PCs) as
154 covariates. We used a Bonferroni-corrected p-value threshold of <0.01 to determine significance
155 (corrected for 5 gene-level tests).
156 For primary analyses we additionally: (i) adjusted for vascular risk factors^{26,27} (Supplemental Table 6
157 legend); (ii) checked for interactions between variant carrier status with sex and ethnicity; and (iii)
158 ran leave-one-out sensitivity analyses, where each phenotype was removed from the phenotype
159 score one-by-one (using logistic instead of Poisson regression in case of *COL4A2* which only had two
160 phenotypes-of-interest to start with).

161 **Results**

162 We identified a total 253 variants-of-interest across the five genes to investigate in UKB: 145
163 exclusively from SysRev, 48 exclusively from ClinVar, and 60 from both sources. We found a total of
164 37 variants present in ≥ 1 of the 199,313 included UKB participants: 24 exclusively from SysRev, 2
165 exclusively from ClinVar, and 11 from both sources, but these did not include any *TREX1* or *CTSA*
166 variants-of-interest. The number of variants represented in UKB varied for each gene, ranging from 5
167 (*COL4A2*) to 17 (*COL4A1*). Across these variants, MAF in UKB ranged from 0.0005-0.14%.
168 (Supplemental tables 2 and 3).
169 VEP predicted 92% (22/24) SysRev, 100% (2/2) ClinVar and 100% (11/11) variants from both sources
170 to be missense or nonsense mutations. Of the remaining two SysRev variants, one was in the 5'
171 untranslated region and the other an intronic splice donor variant. SnpEff predicted the nonsense

172 mutations and the splice donor variant to have high impact, while all but one of the remaining
173 variants were of moderate impact. Overall, 21/37 variants (50% SysRev; 50% ClinVar, 73% both
174 sources) were predicted to have probably damaging and deleterious impacts on protein structure
175 and function (Supplemental Table 4).

176 We identified 2 to 12 phenotypes-of-interest per gene (Supplemental table 1). When mapping these
177 to disease coding systems, the number of codes per phenotype varied widely, ranging from 6 codes
178 for dry mouth to 173 codes for degenerative spine disease (Supplemental code list).

179 ***Assessing variant carrier frequencies in the UKB and their demographic characteristics***

180 Among 199,313 UKB participants, 1,050 participants (0.5%) had ≥ 1 variant-of-interest, resulting in
181 234 *HTRA1*, 481 *COL4A1* and 336 *COL4A2* variant carriers, with one participant carrying a variant in
182 both *HTRA1* and *COL4A1*. When variant carriers were not preferentially selected from related pairs,
183 their overall frequency remained the same. Most variant carriers (96%; 1,003/1,050) possessed a
184 SysRev variant, 0.3% (3/1,050) a ClinVar variant, and 4% (44/1,050) a variant represented in both
185 SysRev and ClinVar (Supplemental tables 2 and 3).

186 The mean age (at last follow-up) of all included UKB participants was 68.2 years. The mean
187 Townsend deprivation index was -1.34 and 55% of participants were female. There was no
188 significant difference in age at last follow-up, sex, or presence of vascular risk factors between
189 variant carriers and non-carriers. Carriers had significantly higher levels of deprivation than non-
190 carriers (mean Townsend index = -0.44 vs -1.34, $p < 2.2 \times 10^{-16}$), and were more likely to be of non-
191 white ethnicity ($p < 2.2 \times 10^{-16}$) (Table 1).

192 Based on these findings, we performed post-hoc analyses comparing variant-carriers and non-
193 carriers in terms of: (i) mean Townsend index stratified by manifestation of a phenotype-of-interest;
194 and, (ii) Townsend index breakdown by quintiles derived from the QResearch database^{28,29},
195 comparing the frequency of phenotypes-of-interest. Variant-carriers had significantly higher levels of
196 deprivation compared to non-carriers regardless of whether or not they manifested a phenotype-of-

197 interest (-0.31 vs -1.07, $p=0.0004$ and -0.44 vs -1.4, $p<2.2\times 10^{-16}$, respectively). Comparing the
198 frequency of participants manifesting a phenotype, among variant-carriers it was similar in the least
199 and most deprived quintiles (23% vs 24%), whereas among non-carriers, the least deprived had a
200 lower frequency compared to the most deprived (17% vs 23%) (Supplemental Table 5).

201 Exploring the ethnicity distribution further, majority of *HTRA1* and *COL4A1* variant carriers ($\geq 97\%$)
202 were of self-reported White ethnicity. For *COL4A2*, however, 57% (190/336) of variant-carriers were
203 of self-reported Black ethnicity, driven by 2 variants c.3448C>A and c.5068G>A present in 1.2% and
204 4.7% of Black participants, respectively. A further 14% of *COL4A2* variant-carriers were of mixed and
205 other ethnic groups (Table 2).

206 ***Assessing the proportion of variant carriers with phenotypes-of-interest***

207 The proportion of variant carriers (N=1,050) with ≥ 1 phenotype-of-interest in the hospital inpatient
208 admissions and/or death record data was: *HTRA1* 9% (21/234); *COL4A1* 20% (95/481); and *COL4A2*
209 4% (15/336) (Figure 2a). This proportion increased when we explored the smaller subset of the
210 variant carriers for whom primary care data was also available (N = 484): *HTRA1* 55% (64/117);
211 *COL4A1* 40% (93/236); and *COL4A2* 7% (9/132) (Figure 2b). Among variant carriers manifesting a
212 phenotype-of-interest, stroke was not always the most common phenotype: *HTRA1* 13% to 52%,
213 *COL4A1* 15% to 19% and *COL4A2* 93-100% (Figure 2a, Figure 2b).

214 ***Assessing whether variant carrier status is associated with phenotypes-of-interest***

215 Association with phenotype-of-interest status

216 For phenotypes-of-interest in the hospital inpatient admissions and/or death record data, a higher
217 proportion of *COL4A1* variant carriers compared to non-carriers had a *COL4A1*-related phenotype
218 (20% in carriers vs 15% in non-carriers, $p=0.01$). We found no significant associations for *HTRA1* and
219 *COL4A2*, and no significant associations for any gene in the secondary analyses also including
220 primary care data. There was also no significant difference in the proportion of stroke cases seen
221 between carriers and non-carriers for any of the genes.

222 Association with phenotype burden (phenotype score)

223 *COL4A1* carriers also had a greater phenotype score compared to non-carriers (OR=1.29, p=0.006).

224 We found no significant associations for *HTRA1* and *COL4A2*, nor for any gene in the secondary

225 analyses including primary care data. (Figure 3).

226 The associations remained similar after adjusting the primary analyses for the presence of vascular

227 risk factors (Supplemental Table 6). We found no significant interactions with sex (*HTRA1* p=0.41;

228 *COL4A1* p=0.14; *COL4A2* p=0.49) and ethnicity (*HTRA1* p=0.60; *COL4A1* p=0.88; *COL4A2* p=0.19) for

229 any gene.

230 Leave-one-out sensitivity analyses did not change the results significantly for *HTRA1* or *COL4A2*

231 (Supplemental Figure 1). For *COL4A1*, removing cataract, migraine or stroke from the phenotype

232 score rendered the association no longer significant, suggesting these phenotypes are important in

233 driving the association seen in the primary analyses (Figure 4).

234 Discussion

235 We found that while 1:200 UKB participants carry a previously reported putative pathogenic rare

236 variant in one of the five cSVD genes included in our study, only 4% to 20% of variant-carriers per

237 gene had an associated phenotype recorded in their hospital admission/death records, and this rose

238 moderately to 7% to 55% when also including primary care records. *COL4A1* variant carrier status

239 was associated with having phenotypes-of-interest compared to non-carriers, but we did not see

240 significant associations with expected phenotypes for other genes.

241 We are not aware of previous studies investigating these five genes in a population-based setting.

242 There has however been a study of another monogenic cSVD gene, which demonstrated that ~1:450

243 UKB participants carry a putative pathogenic (i.e., cysteine-altering) variant *NOTCH3*¹⁹. Among the

244 few disease-based studies exploring rare variation in cSVD genes in apparently sporadic cases, one

245 large study found that ~1:70 lacunar stroke patients had a monogenic cause¹³. However, this study

246 included only patients with stroke (one of several manifestations of cSVD), excluded already
247 diagnosed monogenic cSVD cases and involved an overlapping but not identical set of genes with
248 different definition of putative pathogenicity compared to our study, limiting direct comparisons.

249 We did not find any carriers of *CTSA* or *TREX1* putative pathogenic rare variants in UKB. Potential
250 reasons include: (i) variants in these genes are extremely rare in general and/or rare in a population-
251 based setting owing to their severe phenotype manifestations, this is particularly relevant for *TREX1*
252 where the majority of variants were frameshifts which were not as highly represented in UKB WES
253 data compared to single nucleotide variants¹⁸; (ii) the overall number of variants-of-interest in these
254 genes was smaller and they are not present in UKB by chance.

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

263 The majority of *COL4A2* carriers in UKB were of non-White self-reported ethnic group. The two
264 genetic variants driving the frequency among non-White participants had previously been reported
265 in the literature as causing intracerebral haemorrhage in persons of White Hispanic and African
266 American background³⁰. Earlier studies have focused mainly on investigating European populations,
267 while case reports and series are often missing ethnicity information^{6,13,31,32}, leaving monogenic cSVD
268 and relevant genetic variation prevalence estimates among non-White ethnicities largely
269 unknown^{13,32}. One study which did investigate this in the Genome Aggregation Database among
270 seven ethnic groups did not find a similar enrichment of *COL4A2* carriers among participants of Black

271 and other ethnicities, although interestingly *COL4A1* pathogenic variants were most prevalent
272 among Africans/African-Americans³¹. Furthermore, epidemiological studies have demonstrated
273 racial and ethnic differences in cSVD manifestations and burden, which may have a genetic
274 component³³. These findings underline the importance of extending future genetic studies to include
275 a broader range of ethnic groups.

276 Evidence from the existing literature summarised in a systematic review demonstrated a higher
277 proportion of variant carriers manifesting a relevant phenotype compared to our study, with
278 estimates of 59% for *COL4A2*, 75% for *COL4A1*, and 77% for *HTRA1*⁶. Considering the mean age of
279 UKB participants is older than the mean age of individuals included in the systematic review⁶, it is
280 unlikely this difference is explained by limited duration of follow-up in UKB and our analysis
281 capturing participants who will go on to develop the phenotypes in the future. Our results
282 suggesting lower frequency of phenotypic manifestations in carriers of variants associated with cSVD
283 in a population-based setting are in keeping with findings from other monogenic disease
284 investigations. Similar results have been shown for monogenic diabetes and CADASIL^{19,34-36},
285 demonstrating ascertainment context is crucial when interpreting the consequences of monogenic
286 variants. This has important implications for genetic screening and counselling in the clinical setting.

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

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

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

309 There are also limitations which need to be considered. Firstly, we used routinely collected coded
310 administrative health data to identify disease-phenotypes. While we made significant efforts to map
311 the phenotypes-of-interest systematically and transparently to relevant disease codes, driven by
312 clinically informed selection, these coded data are likely to identify some false-positive and miss
313 true-positive cases. Also, it was more challenging to map some phenotypes than others, and some
314 health conditions (e.g., alopecia, migraine or muscle cramps) are less likely to lead the person to
315 seek medical help and hence be captured by the coded data. Secondly, the UKB population is highly
316 likely affected by the 'healthy volunteer bias'¹⁶, with clinically severely affected variant carriers less
317 likely to enrol in the first place. Hence the variants identified among the UKB population may be
318 those with lower overall penetrance, variable expressivity, and weaker evidence of pathogenicity.
319 Thirdly, even when collapsing the variants across each gene, statistical power for detecting small and

320 moderate variant-level genotype-phenotype association effects remained low. Fourth, routinely
321 collected administrative disease codes did not provide data on cerebral radiological features of UKB
322 participants, which is an important (and sometimes the only) manifestation of cSVD⁴.

323 Future studies will be required to extend these findings to other populations, aiming to better
324 understand and minimise healthy volunteer biases and assessing a broader range of ages and socio-
325 economic- and ethnic groups with even larger sample sizes. As methods of disease identification
326 from routinely collected health data develop further, this will also allow more comprehensive and
327 reliable capture of phenotypes-of-interest. Future work could also explore all rare variants in these
328 genes, stratifying them by level of evidence for pathogenicity and using the richness of the routinely
329 collected health data to undertake phenome-wide association studies.

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

336 **Supplemental data**

337 Supplemental data includes six tables and one figure displaying additional variant information and
338 results.

339 Supplemental methods: details the methodology used to map the phenotypes-of-interest to ICD-10
340 and Read v2/v3 codes.

341 Supplemental code list: contains the ICD-10 and Read v2/v3 code lists used to derive the
342 phenotypes-of-interest for this study.

343 **Declaration of interests**

344 The authors declare no competing interests.

345 **Acknowledgements**

346 AF is funded by BHF award RE/18/5/34216. KR and AF are funded by Health Data Research UK
347 Rutherford fellowship (MR/S004130/1). AT is funded by HDR-UK award HDR-9004 and HDR-9003.
348 We would like to acknowledge Professor Martin Dichgans from the Institute for Stroke and
349 Dementia Research in Munich Germany for his input at the earlier stages of conception of this work.

350 **Data and code availability**

351 Original phenotype and genotype data is available from UK Biobank¹⁶. ICD-10 and Read v2/3 code
352 lists used for analysis are available in the Appendix (Supplemental code list).

353

354 **References**

- 355 1. Pantoni, L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to
356 therapeutic challenges. *The Lancet Neurology* vol. 9 689–701 (2010).
- 357 2. Marini, S., Anderson, C. D. & Rosand, J. Genetics of Cerebral Small Vessel Disease. *Stroke* **51**,
358 12–20 (2020).
- 359 3. Markus, H. S. & Schmidt, R. Genetics of Vascular Cognitive Impairment. *Stroke* **50**, 765–772
360 (2019).
- 361 4. Wardlaw, J. M., Smith, C. & Dichgans, M. Small vessel disease: mechanisms and clinical
362 implications. *Lancet Neurol.* **18**, 684–696 (2019).
- 363 5. Joutel, A. *et al.* *Notch3 mutations in CADASIL, a hereditary adult-onset condition causing*
364 *stroke and dementia.* *Nature* vol. 383 <http://www.nature>. (1996).
- 365 6. Rannikmäe, K. *et al.* Beyond the Brain: Systematic Review of Extracerebral Phenotypes
366 Associated with Monogenic Cerebral Small Vessel Disease. *Stroke* **51**, 3007–3017 (2020).
- 367 7. Yamamoto, Y., Craggs, L., Baumann, M., Kalimo, H. & Kalaria, R. N. Review: Molecular

- 368 genetics and pathology of hereditary small vessel diseases of the brain. *Neuropathol. Appl.*
369 *Neurobiol.* **37**, 94–113 (2011).
- 370 8. Sargurupremraj, M. *et al.* Cerebral small vessel disease genomics and its implications across
371 the lifespan. *Nat. Commun.* **11**, (2020).
- 372 9. Søndergaard, C. B., Nielsen, J. E., Hansen, C. K. & Christensen, H. Hereditary cerebral small
373 vessel disease and stroke. *Clin. Neurol. Neurosurg.* **155**, 45–57 (2017).
- 374 10. Murad, M. H., Sultan, S., Haffar, S. & Bazerbachi, F. Methodological quality and synthesis of
375 case series and case reports. *BMJ Evidence-Based Med.* **23**, 60–63 (2018).
- 376 11. Riley, D. S. *et al.* CARE guidelines for case reports: explanation and elaboration document. *J.*
377 *Clin. Epidemiol.* **89**, 218–235 (2017).
- 378 12. Tan, R., Traylor, M., Rutten-Jacobs, L. & Markus, H. New insights into mechanisms of small
379 vessel disease stroke from genetics. *Clin. Sci.* **131**, 515–531 (2017).
- 380 13. Tan, R. Y. Y. *et al.* How common are single gene mutations as a cause for lacunar stroke?
381 *Neurology* **93**, e2007–e2020 (2019).
- 382 14. Chong, M. *et al.* Mendelian Genes and Risk of Intracerebral Hemorrhage and Small-Vessel
383 Ischemic Stroke in Sporadic Cases. *Stroke* **48**, 2263–2265 (2017).
- 384 15. Weng, Y.-C. *et al.* COL4A1 mutations in patients with sporadic late-onset intracerebral
385 hemorrhage. *Ann. Neurol.* **71**, 470 (2012).
- 386 16. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*
387 **562**, 203–209 (2018).
- 388 17. Jia, T., Munson, B., Lango Allen, H., Ideker, T. & Majithia, A. R. Thousands of missing variants
389 in the UK Biobank are recoverable by genome realignment. *Ann. Hum. Genet.* **84**, 214–220
390 (2020).

- 391 18. Van Hout, C. V *et al.* Exome sequencing and characterization of 49,960 individuals in the UK
392 Biobank. *Nature* **586**, 749 (2020).
- 393 19. Cho, B. P. H. *et al.* NOTCH3 variants are more common than expected in the general
394 population and associated with stroke and vascular dementia: an analysis of 200 000
395 participants. *J. Neurol. Neurosurg. Psychiatry* **92**, 694–701 (2021).
- 396 20. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting
397 evidence. *Nucleic Acids Res.* **46**, D1062 (2018).
- 398 21. Yang, Y. G., Lindahl, T. & Barnes, D. E. Trex1 Exonuclease Degrades ssDNA to Prevent Chronic
399 Checkpoint Activation and Autoimmune Disease. *Cell* **131**, 873–886 (2007).
- 400 22. Caciotti, A. *et al.* Galactosialidosis: review and analysis of CTSA gene mutations. *Orphanet J.*
401 *Rare Dis.* **8**, 114 (2013).
- 402 23. OMIM - Online Mendelian Inheritance in Man. <https://omim.org/>.
- 403 24. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **2016 171** **17**, 1–14
404 (2016).
- 405 25. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and
406 applications. *Nat. Rev. Genet.* doi:10.1038/s41576-019-0177-4.
- 407 26. Boehme, A. K., Esenwa, C. & Elkind, M. S. V. Stroke Risk Factors, Genetics, and Prevention.
408 *Circulation Research* vol. 120 472–495 (2017).
- 409 27. Hauer, A. J. *et al.* Age-specific vascular risk factor profiles according to stroke subtype. *J. Am.*
410 *Heart Assoc.* **6**, (2017).
- 411 28. Home - QResearch. <https://www.qresearch.org/>.
- 412 29. Julia Hippisley-Cox. Cut offs for Townsend quintiles in England and Wales and QResearch.
413 <https://www.qresearch.org/media/1056/cut-offs-for-townsend-quintiles-in-england-and->

- 414 wales-and-qresearch.pdf (2011).
- 415 30. Jeanne, M. *et al.* COL4A2 Mutations Impair COL4A1 and COL4A2 Secretion and Cause
416 Hemorrhagic Stroke. *Am. J. Hum. Genet.* **90**, 91 (2012).
- 417 31. Grami, N. *et al.* Global Assessment of Mendelian Stroke Genetic Prevalence in 101 635
418 Individuals from 7 Ethnic Groups. *Stroke* 1290–1293 (2020)
419 doi:10.1161/STROKEAHA.119.028840.
- 420 32. Kilarski, L. L. *et al.* Prevalence of CADASIL and fabry disease in a cohort of MRI defined
421 younger onset Lacunar Stroke. *PLoS One* **10**, e0136352 (2015).
- 422 33. Castello, J. P. *et al.* Contribution of Racial and Ethnic Differences in Cerebral Small Vessel
423 Disease Subtype and Burden to Risk of Cerebral Hemorrhage Recurrence. *Neurology* **96**,
424 e2469–e2480 (2021).
- 425 34. Mirshahi, U. L. *et al.* The penetrance of age-related monogenic disease depends on
426 ascertainment context. *medRxiv* 2021.06.28.21259641 (2021)
427 doi:10.1101/2021.06.28.21259641.
- 428 35. Rutten, J. W. *et al.* Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank:
429 CADASIL to nonpenetrance. *Neurology* **95**, e1835–e1843 (2020).
- 430 36. Hack, R. J. *et al.* Cysteine-Altering NOTCH3 Variants Are a Risk Factor for Stroke in the Elderly
431 Population. *Stroke* 3562–3569 (2020) doi:10.1161/STROKEAHA.120.030343.

432

433 **Figure titles and legends**

434 **Figure 1.**

435 Title: Summary of study methods and outcomes.

436 Legend: OMIM=Online Mendelian Inheritance in Man; cSVD=cerebral small vessel disease.

437 Created with BioRender.com

438 **Figure 2.**

439 Title: Proportion of variant carriers with a phenotype-of-interest.

440 Legend: N=total number of variant-carriers; n=number of variant carriers with any phenotype-of-

441 interest or stroke.

442 **Figure 3.**

443 Title: Association of variant carrier status with phenotype burden.

444 Legend: OR=odds ratio; CI=confidence interval; *The association between *COL4A1* variant carrier

445 status and phenotype score was significant.

446 **Figure 4.**

447 Title: *COL4A1* leave-one-out analyses.

448 Legend: OR=odds ratio; CI=confidence interval; **Association no longer significant.

449

450 **Table titles and legends**

451 **Table 1.**

452 Title: Demographic characteristics of UK Biobank participants with WES data.

	All UKB participants with WES (N=199,313)	Variant carriers (N=1,050)	Non-carriers (N=198,263)	p-value
Age at last follow up ¹				
Mean (SD)	68.2 (0.018)	67.7 (0.26)	68.2 (0.018)	p=0.43
Sex				
Female	109,739 (55%)	589 (56%)	109,150 (55%)	p=0.52
Male	89,574 (45%)	461 (44%)	89,113 (45%)	
Townsend Deprivation Score				
Mean (SD)	-1.34 (0.01)	-0.44 (0.11)	-1.34 (0.01)	p<2.2x10 ⁻¹⁶
Self-reported ethnic group ²				
White	187,035 (94%)	783 (75%)	186,252 (94%)	p<2.2x10 ⁻¹⁶
Non-White	11,319 (6%)	256 (25%)	11,063 (6%)	

453

454

455 Legend: N=number of participants; UKB=UK Biobank; WES=whole exome sequencing; SD=standard

456 deviation; ¹Age at last follow up is the participant's age in March 2020, when the linked health

457 record data is complete from all sources. ²11 variant carriers and 948 non-carriers had missing

458 ethnicity information.

459 **Table 2.**

460 Title: Variant-carriers in UK Biobank by gene and ethnic group.

461

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

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

462

463 Legend: 'Other' ethnicity includes Chinese. There was one participant with a variant in both *HTRA1*

464 and *COL4A1*. There were no participants carrying variants in *TREX1* or *CTSA*.

Figure 1

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

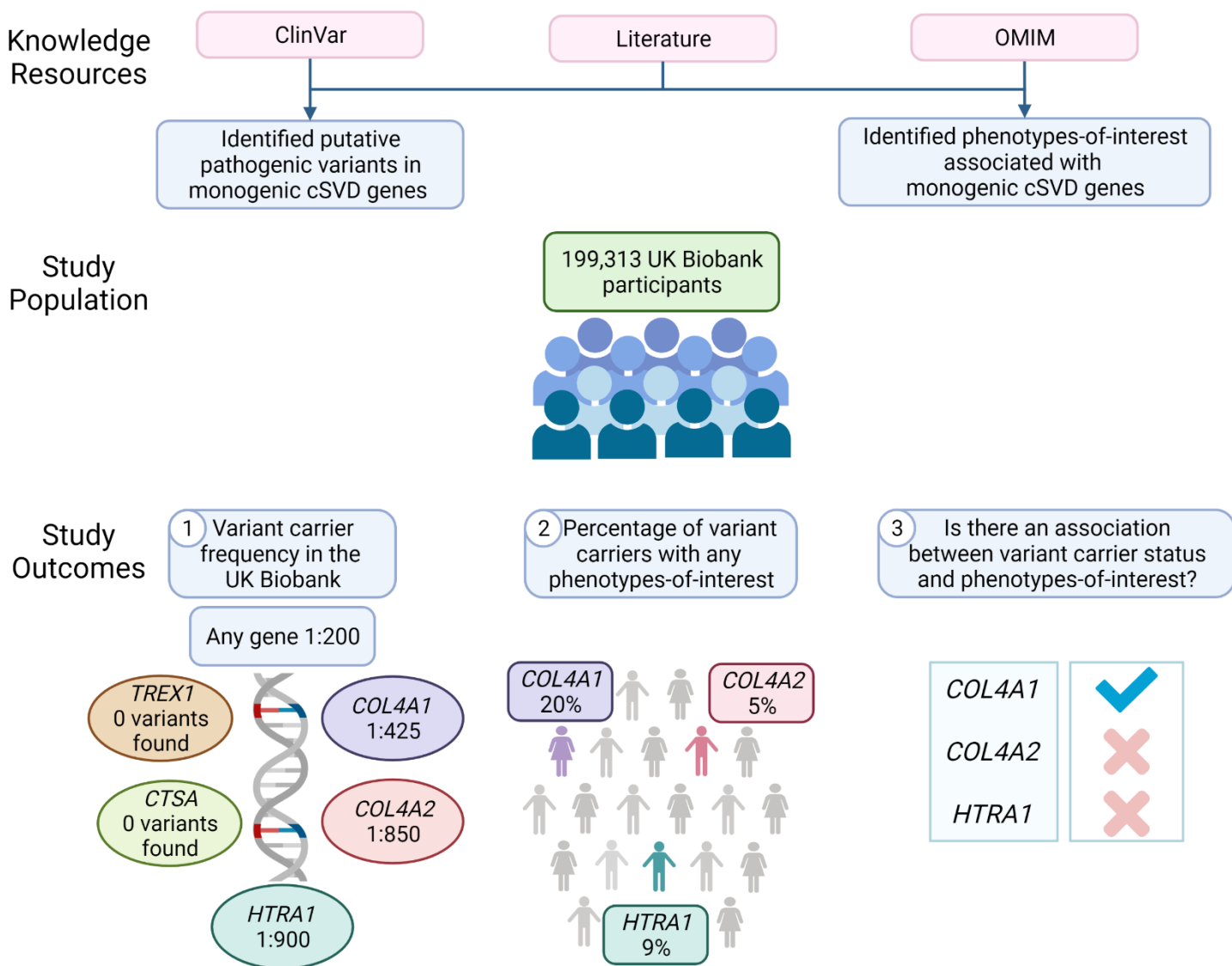


Figure 2

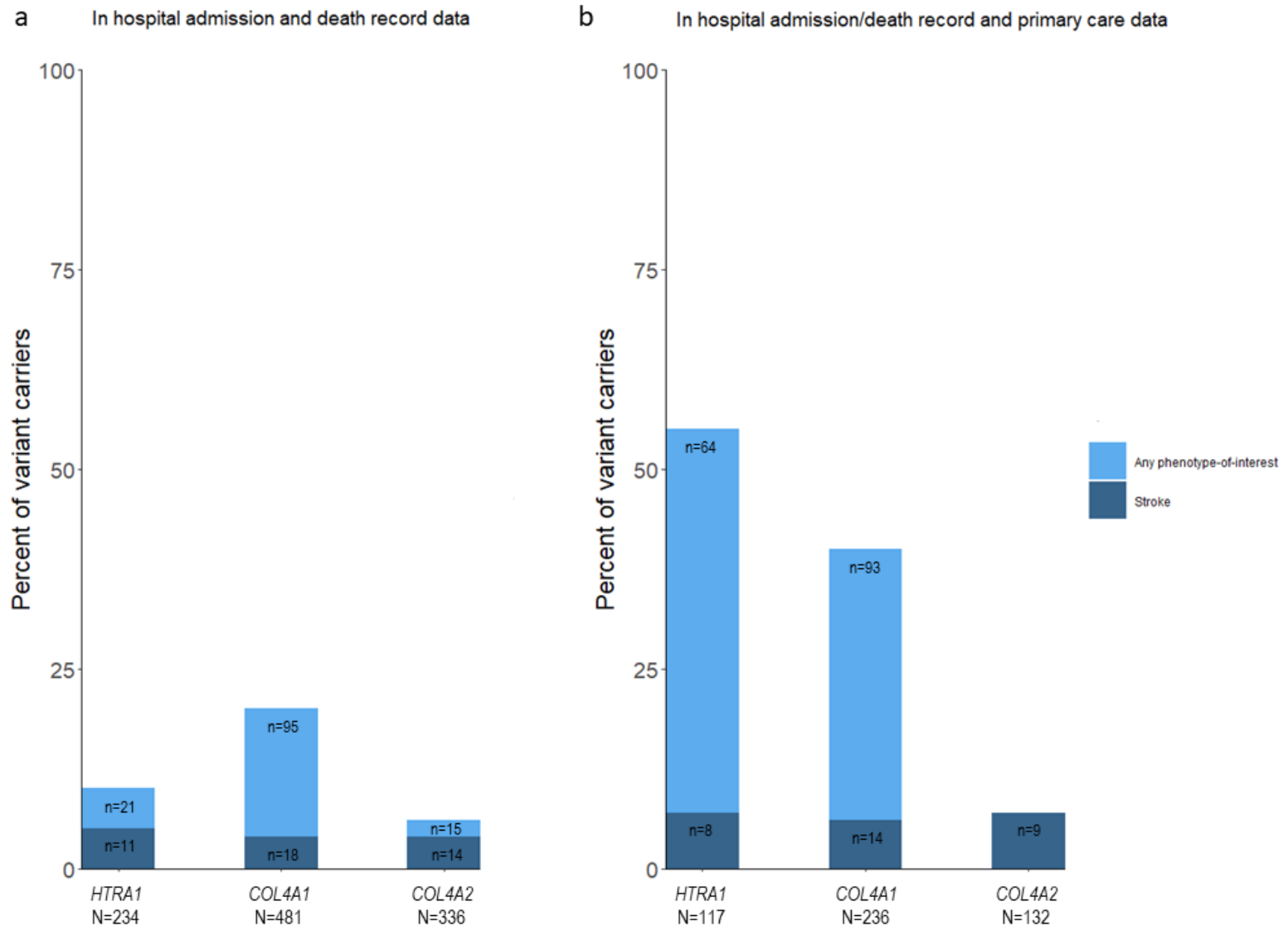


Figure 3

Association between phenotype score and variant carrier status

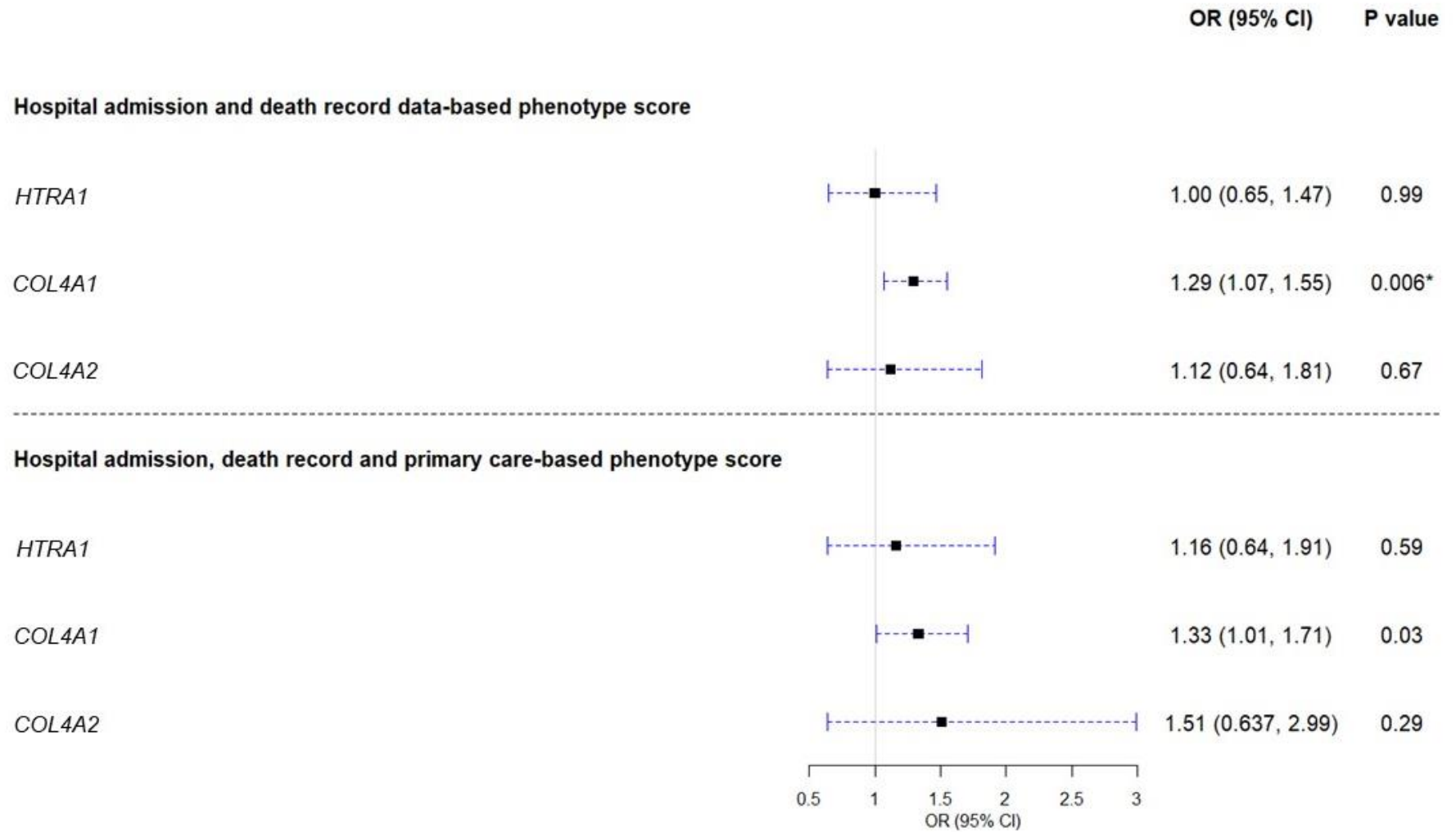


Figure 4

Leave-one-out analysis of *COL4A1* phenotype score

