1	Genetically predicted levels of the human plasma proteome and risk of stroke: a Mendelian
2	Randomization study
3	
4	Lingyan Chen ¹⁺ , James E. Peters ² , Bram Prins ¹ , Elodie Persyn ^{1,3} , Matthew Traylor ⁴⁺ , Praveen
5	Surendran ^{1,5} , Savita Karthikeyan ¹ , Ekaterina Yonova-Doing ¹⁺ , Emanuele Di Angelantonio ^{1,6,7,8} , David
6	J. Roberts ^{7,9,10} , Nicholas A. Watkins ¹¹ , Willem H. Ouwehand ^{6,11,12,13} , John Danesh ^{1,6,7,8,14} , Cathryn M.
7	Lewis ^{3,15} , Paola G. Bronson ¹⁶ , Hugh S. Markus ¹⁷ , Stephen Burgess ^{1,6,18} , Adam S. Butterworth ^{1,6,7,8} ,
8	Joanna M. M. Howson ¹ * ⁺ .
9	1. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and
10	Primary Care, University of Cambridge, Cambridge, UK
11	2. Department of Immunology and Inflammation, Faculty of Medicine, Imperial College London,
12	London, UK
13	3. Department of Medical & Molecular Genetics, King's College London, London, UK
14	4. Clinical Pharmacology, William Harvey Research Institute, Queen Mary University of London,
15	London, UK
16	5. Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge,
17	UK
18	6. British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK
19	7. National Institute for Health Research Blood and Transplant Research Unit in Donor Health and
20	Genomics, University of Cambridge, Cambridge, UK
21	8. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge,
22	Cambridge, UK
23	9. NHS Blood and Transplant-Oxford Centre, Level 2, John Radcliffe Hospital, Oxford, UK
24	10. Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK
25	11. NHS Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, UK

- 26 12. Department of Haematology, University of Cambridge, Cambridge, UK
- 27 13. Wellcome Sanger Institute, Hinxton, UK
- 28 14. Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK
- 29 15. Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK
- 30 16. R&D Translational Biology, Biogen, Inc., Cambridge, MA, USA
- 31 17. Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK
- 32 18. Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, University of
- 33 Cambridge, Cambridge, UK
- 34 *Correspondence to: Joanna M. M. Howson jmmh2@medschl.cam.ac.uk
- ⁺ Current address: Novo Nordisk Research Centre Oxford, Innovation Building, Old Road Campus,
- 36 Roosevelt Drive, Oxford, UK.

37

38 Abstract

39	Proteins are the effector molecules of biology and are the target of most drugs. To identify proteins
40	and related pathways that may play a causal role in stroke pathogenesis, we used Mendelian
41	randomisation (MR). We tested potential causal effects of 308 plasma proteins (measured in 4,994
42	blood donors from the INTERVAL study) on stroke outcomes (derived from the MEGASTROKE
43	GWAS) in a two-sample MR framework and assessed whether these associations could be mediated
44	by cardiovascular risk factors. We extended the analysis to identify whether pharmacological
45	targeting of these proteins might have potential adverse side-effects or beneficial effects for other
46	conditions through Phenome-wide MR (Phe-MR) in UK Biobank.
47	MR showed an association between stroke and genetically predicted plasma levels of TFPI, IL6RA,
48	MMP12, CD40, TMPRSS5 and CD6 ($P \le 1.62 \times 10^{-4}$). We identified six risk factors (atrial fibrillation,
49	body mass index, smoking, blood pressure, white matter hyperintensities and type 2 diabetes) that
50	were associated with stroke ($P \le 0.0071$) using MR. The association of TFPI, IL6RA and TMPRSS5
51	with stroke could be mediated by these risk factors, such as body mass index, white matter
52	hyperintensity and atrial fibrillation. Thirty-six additional proteins were potentially causal for one or
53	more of these risk factors. The Phe-MR suggested that targeting TFPI could have potential beneficial
54	effects on other disorders of arteries and hyperlipidaemia in addition to stroke. Our results highlight
55	novel causal pathways and potential therapeutic targets for stroke.

56

57 Key Words

- 58 Mendelian Randomization; genetics; proteomics; stroke; cardiovascular disease; drug target;
- 59 Phenome-wide MR (PheMR)

60

61

62 Introduction

63	Stroke is the second leading cause of death worldwide, estimated to cause ~5.5 million deaths
64	annually and is the leading cause of long-term disability, with a growing burden on global health ¹ .
65	Therefore, there is a need for new and improved treatments and prevention strategies for stroke. While
66	conventional risk factors, such as hypertension 2 , account for ~50% of stroke risk, there remains a
67	need to identify new risk factors, biomarkers and therapies for stroke ³ . In 2017, ~75% of FDA-
68	approved drugs were targeted at human proteins ⁴ . Plasma proteins play a central role in a range of
69	biological processes frequently dysregulated in diseases ⁵⁻⁸ , and represent a major source of
70	therapeutic targets for many indications ^{4,9,10} . In particular, plasma proteins are particularly relevant
71	for circulatory diseases such as stroke as they are in physical contact with the blood vessels
72	(compared to tissue-specific diseases, e.g. inflammatory bowel disease).
73	Genome-wide association studies (GWAS) of plasma protein levels have identified genetic variants
74	that are associated with proteins, usually referred to as 'protein quantitative trait loci (pQTLs)' 11-13,
75	offering an opportunity to test the causal effect of potential drug targets on the human disease
76	phenome using Mendelian randomization (MR) ^{14,15} . Briefly, MR can be thought of as nature's
77	randomized trial, by capitalising on the random allocation of genetic variants at conception to separate
78	individuals into subgroups (one equivalent to placebo and the other to intervention in a randomized
79	control trial, RCT) and so allows testing of the potential causal association of risk factors (e.g. plasma
80	proteins) with disease outcomes (e.g. stroke) as confounders should also be randomised.
81	Here, we perform a two-sample MR to estimate the causal effects of plasma proteins on stroke, where
82	we derived genetic instrumental variables of 308 circulating plasma proteins from 4,994 participants
83	¹⁶ and obtained genetic associations of stroke subtypes, (any stroke (AS), any ischemic stroke (IS),
84	large-artery-stroke (LAS), cardio-embolic-stroke (CES) and small-vessel-stroke (SVS)) from the
85	MEGASTROKE GWAS ¹⁷ . Then, to verify the robustness of the proteins' instrumental variables, we
86	perform colocalization analyses. We evaluate the causal relationship of plasma proteins on stroke
87	risk factors and assess potential safety effects of targeting the proteins for stroke therapy by
88	performing a phenome-wide MR in UK Biobank GWASs ¹⁸ .

89

90 Methods

91 The overall study design is illustrated in **Figure 1**. Details of the methods and study participants are

- 92 provided below.
- 93

94 **Proteomic profiling and quality control**

A subset of 4,994 blood donors at mean age of 61 years (SD 6.7 years) enrolled in the INTERVAL

96 BioResource¹⁶, were processed for proteomic profiling using the Olink Proseek® Multiplex platform

97 by 4 high-throughput, multiplex immunoassays: Inflammatory I (INF1), Cardiovascular II (CVD2),

98 Cardiovascular III (CVD3) and Neurology I (NEURO) (Olink Bioscience, Uppsala, Sweden). Each

99 panel enables the simultaneous measurement of 92 proteins through relative quantification using the

100 Proximity Extension Assay (PEA) Technology ¹⁹, in which each pair of oligonucleotide-labelled

101 antibodies ("probes") are allowed to bind to their respective target present in the sample and trigger

102 extension by DNA polymerase. DNA barcodes unique to each protein are then amplified and

103 quantified using a standard real time polymerase chain reaction (PCR). Default pre-processing of

104 the proteomic data by Olink included applying median centring normalization between plates, where

105 the median is centred to the overall median for all plates, followed by log2 transformation to provide

106 normalised protein expression (NPX) values. Further details on the Olink proteomic data processing

107 can be found at <u>http://www.olink.com</u>. Probes were labelled using Uniprot identifiers, which we re-

108 mapped to HUGO gene name nomenclature for the (cis-) gene encoding the relevant protein. All

109 protein names and descriptions are provided in **Supplementary Table 1**.

110 Samples that failed standard Olink quality control metrics were removed. 4,902, 4,947, 4,987, and

111 4,660 samples passed quality control for the 'INF1', 'CVD2', 'CVD3' and 'NEURO' panels,

112 respectively. According to the manufacturer's recommendation, we also removed four proteins

113 (HAGH, BDNF, GDNF and CSF3) in the 'NEURO' panel and one protein (GDNF) in the 'INF1'

114 panel due to high levels of missingness.

115

116 **Proteome GWAS**

- 117 The INTERVAL study¹⁶ was genotyped using the UK Biobank Affymetrix Axiom array
- 118 (http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/), and imputed to 1000 Genomes
- 119 Phase 3-UK10K combined reference panel, employing the PBWT imputation algorithm ²⁰. Genetic
- 120 data for the ~5000 participants with Olink proteomic profiling were extracted to test for association of
- 121 the genetic variants with plasma proteins. More details regarding the INTERVAL genetic data QC
- 122 can be found here 21 . Within the ~5,000 participant subset, we removed six related individuals (those
- 123 individuals with pairwise values of twice the kinship coefficient (PI_HAT) > 0.1875 (removing the
- 124 individuals with the lowest call rate from each pair). The final imputed dataset was additionally
- 125 filtered for imputation quality (only retaining variants with an info score > 0.4) and Hardy-Weinberg
- 126 equilibrium (retaining variants with $P_{\text{HWE}} > 1 \times 10^{-4}$).
- 127 354 proteins (of 363) that passed quality control were taken forward for the GWAS. Normalized
- 128 protein levels ('NPX') were regressed on sex, age, plate, time from blood draw to processing (in
- days), season (as a categorical variable: 'Spring', 'Summer', 'Autumn', 'Winter'), and batch when
- 130 appropriate. The residuals were then rank-inverse normalized. Linear regression of the rank-inversed
- 131 normalize residuals on genotype was carried out in SNPTEST v.2.5.2²², with the first three
- 132 components of multi-dimensional scaling as covariates to adjust for ancestry. Only proteins with at
- 133 least one SNP with an association *P*-value passing the genome-wide significant threshold ($P \le 5.0 \times 10^{-10}$
- ⁸) were kept, resulting in 308 proteins for MR analyses.
- 135

136 Genetic variants associated with proteins

- 137 For each plasma protein, cis- and trans- pQTLs from its corresponding GWAS were used as genetic
- 138 instruments. We followed these steps to select pQTL instruments: (i) we obtained SNPs that were
- 139 also tested in the MEGASTROKE GWAS of stroke outcomes (see below), (ii) we performed linkage
- 140 disequilibrium (LD) clumping using PLINK 1.90 (<u>www.cog-genomics.org/plink/1.9/</u>)²³ to obtain

141 approximately independent SNPs for each protein. In brief, the LD clumping algorithm groups SNPs in LD ($r^2 \ge 0.1$ in 4.994 European ancestry participants from the INTERVAL study ^{16,21}) within +/-142 143 1MB of an index SNP (SNPs with association $P \le 5 \times 10^{-8}$). Analyses assessing sensitivity to the $r^2 \ge$ 144 0.1 LD threshold are detailed below. The algorithm loops through all index SNPs, beginning with the 145 smallest *P*-value and only allowing each SNP to appear in one clump. The final output therefore 146 contains the most significant protein-associated SNPs for each LD-based clump across the genome. 147 We split pOTL variants into cis-pOTLs (+/-1MB window of the gene encoding the target protein) and 148 trans-pQTLs (outside the $\pm/-1MB$ window). We then performed MR in a two-step approach. Our 149 primary analysis was restricted to cis-pQTLs. Having performed MR restricted to cis-pQTL only as 150 IVs, we broadened the analysis to consider the effects of adding in trans-pOTLs as IVs. We estimated the variance of each protein explained by its IVs through calculating the $R^{2/24}$ and the strength of each 151 IV by the *F*-statistic²⁵. Summary association statistics of all the instrumental variables (IVs) for the 152 153 15 significant proteins are provided in **Supplementary Table 2**. 154 To assess the robustness of the $r^2 > 0.1$ threshold for IV selection, we performed two additional 155 sensitivity analyses (Supplementary Table 11) for proteins of interest to verify the robustness of 156 MR causal relationship: 1) we performed conditional analysis to derive conditionally independent variants as IVs using the FINEMAP software package ²⁶ with --cond flag; 2) we performed fine-157 mapping to obtain variants in the 95% credible set as IVs using FINEMAP software package ²⁶ with -158

160

--sss flag.

159

161 Genetic variants associated with stroke and its risk factors

162 The primary outcomes were the risk of stroke and its subtypes. Genetic association estimates for

163 stroke outcomes were obtained from the MEGASTROKE consortium, a large-scale international

- 164 collaboration launched by the International Stroke Genetics Consortium (ISGC). A detailed
- 165 description of the study design and characteristics of study participants were provided in the original
- 166 publication ¹⁷. To reduce confounding by population stratification, we extracted estimates for the
- 167 associations of the protein IVs with stroke and its subtypes restricted to individuals of European

- 7 -

168 ancestry (40,585 cases and 406,111 controls). The primary outcomes for this study were any stroke 169 (including both ischemic and haemorrhagic stroke; AS, $N_{cases} = 40,585$), any ischemic stroke (IS, 170 $N_{cases} = 34,217$), and the three etiologic ischemic stroke subtypes: large-artery stroke (LAS, $N_{cases} =$ 171 4,373), cardio-embolic stroke (CES, $N_{cases} = 7,193$) and small-vessel stroke (SVS, $N_{cases} = 5,386$). 172 Summary-level data (beta coefficients and standard errors) for the associations of the five stroke 173 outcomes were obtained from the MEGASTROKE GWAS http://www.megastroke.org/index.html. 174 The secondary outcomes we considered were stroke risk factors, including blood pressure (BP)²⁷, atrial fibrillation (AF)²⁸, type 2 diabetes (T2D)²⁹, white matter hyperintensity (WMH)³⁰, body mass 175 index (BMI)³¹, alcohol consumption and smoking behaviour³². We used the same pQTLs as IVs for 176 177 the secondary outcomes as for the primary outcomes. The SNP-outcome effects for all the above risk 178 factors were obtained from previously published GWASs when available. Table 1 provides full 179 details of the data sources and sample size for these GWASs.

180

181 Phenome-wide MR (Phe-MR) analysis of 784 phenotypes for target proteins

182 We expanded the exploration of side-effects for the six stroke-associated proteins to include non-

183 stroke phenotypes by performing Phe-MR analyses for a range of diseases. We used summary

184 statistics for SNP-outcome effects calculated using the UK Biobank cohort (N \leq 408,961) by Zhou et

185 al. ³³, who performed GWAS using the Scalable and Accurate Implementation of GEneralized mixed

186 model (SAIGE v.0.29) method ³³ to account for unbalanced case-control ratios. They defined disease

- 187 outcomes based on "PheCodes", a system developed to organize International Classification of
- 188 Diseases and Related Health Problems (ICD-9/-10) codes into phenotypic outcomes suitable for
- 189 systematic genetic analysis of numerous disease traits ^{18,33}. Outcomes with fewer than 500 cases were
- 190 excluded due to statistical power, leaving 784 diseases for Phe-MR analyses (Supplementary Table

191 **8**). SNP-outcome associations were downloaded from SAIGE GWAS ³³

192 (https://www.leelabsg.org/resources). pQTLs were derived from the same proteome GWAS as in the

193 primary analysis with stroke subtypes.

194 Phe-MR findings can be interpreted as the risk/protective effect per-SD increase in the plasma protein

195 level, same as with primary stroke outcomes. That is, if the effect direction of the additional

196 indication is consistent with the effect direction in Stroke, the identified protein that is therapeutically

197 targeted for the treatment of stroke may also be "beneficial" for the additional indication, and vice

198 versa. MR causal effects are considered statistically significant at $P \le 1.06 \times 10^{-5}$ (Bonferroni-adjusted

199 for 6 proteins and 784 phenotypes: 0.05/6/784).

200

201 Systematic MR screening for causal proteins of stroke and stroke risk factors

202 We used two-sample MR³⁴⁻³⁶ to estimate the associations between genetically-predicted protein

203 levels and target outcomes (stroke, stroke risk factors, and potential adverse effects or additional

204 indications). Two sample MR³⁷ is where the genetic associations with the risk factor are derived in

205 one cohort (e.g. pQTLs from INTERVAL) and the association of these genetic variants with the

206 outcome is tested in a second cohort (*e.g.* stroke GWAS from MEGASTROKE). Two-sample MR

207 allows evaluation of causal effects using summary genetic association data, negating the need for

208 individual participant data. The MR approach was based on the following assumptions: (i) the genetic

209 variants used as instrumental variable (IV) are associated with target exposure, *i.e.*, protein levels; (ii)

210 there are no unmeasured confounders of the associations between genetic variants and outcome; (iii)

211 the genetic variants are associated with the outcome only through changes in the exposure, *i.e.*, no

212 pleiotropy.

213 After extracting the association estimates between the variants and the exposures or the outcomes, we 214 harmonized the direction of estimates by effect alleles, and applied the Wald's ratio method to 215 estimate the causal effects when there was only one IV available for target exposure. If more than one 216 IV was available, we applied the inverse-variance weighting (IVW) method, either in a fixed-effect 217 framework (IVs \leq 3) or in a multiplicative random-effect meta-analysis framework (IVs > 3) ³⁴. We 218 chose 3 as a cut-off for the random effects model, as with >3 variants, there is potential for some 219 heterogeneity within instrumental variables. (The multiplicative random-effects model allows for 220 heterogeneity between causal estimates targeted by the genetic variants by allowing over-dispersion

221 the regression model.) We also performed several sensitivity analyses to assess the robustness of our 222 results to potential violations of the MR assumptions given these analyses have different assumptions 223 for validity: (i) heterogeneity was estimated by Cochran Q test ³⁴; (ii) horizontal pleiotropy was 224 estimated using MR-Egger's intercept³⁸; (iii) influential outlier IVs due to pleiotropy were identified 225 using MR Pleiotropy Residual Sum and Outlier (MR-PRESSO)³⁹; (iv) reverse MR was used to 226 eliminate spurious results due to reverse causation. Additionally, the contamination mixture method 227 40 , which can explicitly model multiple potential causal estimates and therefore infer multiple causal 228 mechanisms associated with the same risk factor that affect the outcome to different degrees, was also 229 used to calculate the MR estimates. Although these methods may have differing assumptions and 230 statistical power, the rationale for using them is that if they give a similar conclusion, this provides 231 greater certainty in inferring that any positive results are unlikely to be driven by violation of the MR 232 assumptions. 233 Effects on binary outcomes (*i.e.*, stroke, AF, T2D, smoking initiation/cessation) are reported as odds 234 ratios (ORs) with their 95% confidence intervals (CIs) scaled to a one standard deviation (SD) higher 235 plasma protein level. Effects on quantitative outcomes (i.e., BP, WMH, BMI) are reported as the 236 effect size (95% CI) scaled to a 1-SD higher plasma protein levels. All statistical tests were two-sided 237 and considered statistically significant at $P_{CausalEstimate} \leq 1.62 \times 10^{-4}$ (Bonferroni-adjusted for 308) proteins: $0.05/308 = 1.62 \times 10^{-4}$), $P_{O-stat} \ge 0.05$, $P_{Egeer-Intercept} \ge 0.05$ and $P_{GlobalTest} \ge 0.05$. The MR 238 analyses were conducted using *MendelianRandomization* (Version: 0.4.2)³⁵, *TwoSampleMR* 239 (Version: 0.4.22)³⁶ and *MR-PRESSO* (Version: 1.0)³⁹ packages in R 3.5.1 (R Foundation, <u>www.R-</u> 240 241 project.org). Plots were generated using various R packages including ggplot2 (Version: 3.2.0), 242 forestplot (Version: 1.9), and PheWAS (Version: 0.99.5-4). We employed the same statistical analysis 243 framework, incorporating the sensitivity analyses for all MR analyses. 244 245 **Multi-trait colocalization analyses**

- 246 As the instruments used in the current setting were identified based on their statistical associations
- 247 with the protein level, we conducted another sensitivity analysis colocalization, to investigate

248 whether the genetic associations with both protein and phenotypes shared the same causal variants. 249 We conducted colocalization analysis for each potential causal protein across multiple traits, including 250 protein level and five stroke outcomes, to estimate the posterior probability (PP) of multiple traits 251 sharing the same causal SNP simultaneously using a multi-trait colocalization (HyPrColoc) method ⁴¹. HyPrColoc extends the established coloc methodology ⁴² by approximating the true posterior 252 253 probability of colocalization with the posterior probability of colocalization at a single causal variant 254 and a small number of related hypotheses. If all traits do not share a causal variant, HyPrColoc 255 employs a novel branch-and-bound selection algorithm to identify subsets of traits that colocalize at 256 distinct causal variants at the locus. We used uniform priors for the primary analysis. We also 257 performed sensitivity analysis with non-uniform priors to access the choice of priors, which used a 258 conservative trait-level prior structure with $P=1\times10^{-4}$ (prior probability of a SNP being associated 259 with one trait) and $\gamma=0.98$ (1-prior probability of a SNP being associated with an additional trait given 260 that the SNP is associated with at least one other trait), *i.e.*, 1 in 500,000 variants are expected to be 261 causal for two traits. 262 Variants within a ± 1 Mb window around the pQTL with the smallest *P*-value, with imputation 263 (INFO)-score ≥ 0.8 and minor allele frequency (MAF) ≥ 0.01 were included. All variants across each

of the datasets were harmonized to the same effect alleles prior to colocalization analyses. We

265 conducted the colocalization analysis using the 'HyPrColoc' R package ⁴¹.

266 Results

267 Genetically determined plasma protein levels and risk of stroke

- 268 Three hundred and eight plasma proteins were tested for causal associations with stroke outcomes
- 269 (Figure 1). As cis-pQTLs were considered to have a more direct and specific biological effect upon
- the protein (compared to trans-pQTLs)⁴³, we first performed MR analyses using only cis-pQTLs as
- 271 instrumental variables and identified six putatively causal proteins with at least one stroke outcome (P
- 272 $\leq 1.62 \times 10^{-4} = 0.05/308$ proteins; Table 2, Figure 2 & Figure 3, Supplementary Figure 1): TFPI
- 273 (Tissue Factor Pathway Inhibitor), TMPRSS5 (Transmembrane Serine Protease 5), CD40 (B Cell
- 274 Surface Antigen CD40), MMP12 (Matrix Metallopeptidase 12), IL6RA (Interleukin 6 Receptor), and
- 275 CD6 (T-Cell Differentiation Antigen CD6). TFPI, CD40, IL6RA, and MMP12 were significantly
- associated with lower risk of any stroke and any ischemic stroke, while TMPRSS5 and CD6 was
- 277 significantly associated with higher risk of any stroke. Among the ischemic stroke subtypes, genetic
- 278 predisposition to upregulated TMPRSS5 was associated with higher risk of any ischemic stroke (OR
- per-1-SD higher plasma protein level $[95\%CI]=1.059[1.038, 1.08]; P=1.36\times10^{-8}$ and Cardioembolic
- stroke (OR[95%CI]=1.089[1.045, 1.134]; $P=5.33\times10^{-5}$). Higher genetically predicted levels of both
- 281 MMP12 (OR[95%CI]=0.793[0.73, 0.861]; *P*=3.53×10⁻⁸) and CD40 (OR[95% CI]= 0.795[0.723,
- $282 \quad 0.874$]; *P*= 2.09×10^{-6}) were associated with lower risk of Large-artery stroke. Higher genetically
- 283 predicted soluble IL6RA (and lower IL6R signalling ⁴⁴) was associated with lower risk of Small-
- 284 vessel stroke (OR[95% CI]= 0.939[0.909, 0.970]; $P=1.60\times10^{-4}$).
- 285 We extended the MR analyses to include trans-pQTLs as instrumental variables and identified nine
- additional proteins significantly associated with at least one stroke outcome ($P \le 1.62 \times 10^{-4}$;
- 287 Supplementary Table 3). However, seven proteins (VSIG2, EPHB4, Gal4, ICAM2, LIFR, SELE, and
- 288 vWF), included instrumental variables from the ABO locus, which is well known to have pleiotropic
- 289 effects. We note that the ABO protein has previously been identified as a genetic risk factor for
- stroke⁴⁵. Interestingly, both Bone Morphogenetic Protein 6 (BMP6) and Growth Differentiation
- Factor 2 (GDF2, also known as BMP9) were instrumented by *trans*-pQTLs located in the genetic
- regions of KNG1 (Kininogen 1) and F11 (Coagulation Factor XI). Both genes are essential for blood

coagulation and the latter has previously been reported to be a causal risk factor for stroke ⁴⁶. GDF2 293 has also been found to have a causal role in pulmonary artery hypertension (PAH)⁴⁷. We therefore 294 295 focused further analyses on the proteins with *cis* pQTL only (*i.e.*, TFPI, TMPRSS5, CD40, MMP12, 296 IL6RA, CD6), as these associations with stroke are unlikely to be due to pleiotropy. 297 Results of sensitivity analyses confirmed the robustness of the primary MR analyses. There was no 298 evidence for heterogeneity in the association of any of the six proteins in **Supplementary Table 3** as 299 measured by Cochran Q statistics ($P_{Q-stat} > 0.05$), and no indication that the instrumental variables had 300 horizontal pleiotropy as assessed by MR-Egger intercept ($P_{Egger-Intercept} > 0.05$) or MR-PRESSO global 301 pleiotropy test ($P_{GlobalTest} > 0.05$). All MR causal effect estimates adjusting for correlation of IVs' 302 were consistent with the primary analyses (Supplementary Table 10). Moreover, MR causal 303 estimates using IVs derived from conditionally independent variants and credible sets of variants from

304 fine-mapping showed consistent results (Supplementary Table 11 & 12). There was no evidence of

305 reverse causations (Supplementary Table 13).

306

307 Co-localization

308 We formally tested whether the associations of the variant with the protein levels used as IVs and the 309 stroke outcome are shared for the six proteins using statistical colocalization analysis. We applied a 310 Bayesian algorithm, Hypothesis Prioritization in multi-trait Colocalization (HyPrColoc), which allows 311 for the assessment of colocalization across multiple complex traits simultaneously (Methods), to test 312 whether the protein associations and stroke associations are shared. The association of the genetic 313 variants selected as instrumental variables for four proteins (TFPI, TMPRSS5, CD40, and CD6) 314 colocalized with the stroke associations (posterior probability (PP) ≥ 0.7) (Supplementary Table 4, 315 Supplementary Figure 2) *i.e.*, the associations in these regions were likely due to the same 316 underlying causal variants. The colocalization suggested the genetic variants associated with TFPI 317 (pQTLs) were due to the same genetic variants underlying the association with all-stroke. Similarly, 318 CD6 pQTLs colocalized with all-stroke genetic associations; CD40 pQTLs colocalized with the 319 genetic associations for all-stroke, ischemic-stroke and large-artery-stroke; TMPRSS5 pQTLs

- 13 -

320	colocalized with all-stroke, ischemic-stroke and cardioembolic-stroke genetic associations. Notably,
321	we found for TFPI, CD40, and CD6 that >80% of the posterior probability of colocalization of the
322	primary genetic association with stroke and the respective protein levels were explained by a single
323	variant (rs67492154, rs4810485, and rs2074227 for TFPI, CD40, and CD6, respectively). The
324	colocalization evidence at MMP12, was less strong than with the other proteins, with colocalization
325	PP>0.6 and there was no colocalization evidence for IL6RA with stroke, which could be due to
326	violation of the single causal variant assumption of the HyprColoc method.
327	
328	Characterizing the potential causal effects of stroke risk factors on stroke
329	To understand potential causal mechanisms between plasma proteins and stroke, we conducted
330	mediation MR analyses for conventional stroke risk factors. First, we performed two-sample MR
331	analyses to characterize the causal relationship of the stroke risk factors with all stroke outcomes.
332	Second, we assessed the causal effects of the proteins on the highlighted risk factors.
333	For each of the six stroke risk factors we considered (<i>i.e.</i> , blood pressure (BP), atrial fibrillation (AF),
334	type 2 diabetes (T2D), white matter hyper-intensity (WMH), body mass index (BMI), smoking
335	behaviours and alcohol consumption), instrumental variables were derived from published GWAS
336	summary statistics restricted to European populations (Table 1 & Supplementary Table 5). AF,
337	T2D, smoking, increased systolic BP, diastolic BP, pulse pressure, WMH, and BMI significantly
338	increased the risk of any stroke ($P \le 0.05/7 = 0.007$, Bonferroni adjusted for seven risk factors; Figure
339	4, Supplementary Table 6 & Supplementary Figure 3). As expected, systolic BP exhibited the
340	strongest effect of all the risk factors on any ischemic stroke and LAS (OR per-1-SD [95%
341	CI]=1.68[1.57, 1.80] and 2.58[2.21, 3.01], respectively) and AF had a positive association with CES
342	(OR[95% CI]: 2.04[1.92, 2.16]; <i>P</i> =2.72×10 ⁻¹²⁵). WMH increased risk of any stroke and SVS (1-SD
343	increased in WMH was associated with 49% higher odds for SVS (OR[95% CI]=1.49[1.17, 1.9];
344	P=0.00147). Both genetically determined higher T2D risk and smoking initiation were associated
345	with increased risk of LAS and SVS; and genetically determined higher BMI was associated with

346 higher risk of LAS. No significant association was observed for alcohol consumption with any of the

347 stroke outcomes (P>0.05).

348

349 Associations of genetically determined plasma protein levels with stroke risk factors

350 We performed MR of all 308 plasma proteins with the highlighted stroke risk factors (excluding

alcohol consumption which was not associated with increased stroke risk in the above MR analyses).

352 After multiple testing correction, 39 proteins instrumented with cis-pQTLs were significantly

associated with at least one stroke risk factor ($P \le 1.62 \times 10^{-4}$): 5 with Systolic BP; 7 with Diastolic BP;

354 7 with Pulse Pressure; 6 with AF; 4 with T2D; 9 with BMI; 3 with WMH; and 8 with smoking. There

355 was no evidence of horizontal pleiotropy, and sensitivity analyses yielded consistent causal effect

356 estimates (Supplementary Table 14).

357 Among the six stroke-associated proteins, three proteins were found to be significantly associated

358 with one or more of the stroke risk factors (Figure 5; Table 3; Supplementary Figure 4). Of note,

359 we found genetically predicted higher TFPI level was associated with lower WMH and lower BMI (a

360 0.06 SD lower WMH β [95% CI]= -0.06[-0.08, -0.04]; *P*=7.15×10⁻¹⁰ and a 0.013 SD lower BMI

361 β[95% CI]= -0.013[-0.019, -0.007]; $P=3.56\times10^{-5}$ per-SD higher TFPI; **Supplementary Table 7**). We

thus inferred that the association between TFPI and stroke could be partially mediated by BMI and

363 WMH. Genetically determined higher TMPRSS5 levels were associated with higher risk of AF

364 (OR[95% CI]: 1.03[1.016, 1.045]; $P=2.15\times10^{-5}$). Together with the causal relationship of AF and

365 cardioembolic stroke, we can also infer that AF is a possible mediator on the effect of TMPRSS5 on

366 cardioembolic stroke. Genetically higher IL6RA levels were associated with a 4.1% lower risk of AF

367 (OR[95% CI]: 0.96[0.95, 0.97]; $P=2.55\times10^{-18}$). All the effect directions of these associations of

368 proteins with risk factors were consistent with those of the proteins with stroke, indicating that these

369 risk factors may be potential mediators of the protein-stroke associations.

Among the 39 proteins that were associated with at least one stroke risk factor, 36 were found to be

associated with the risk factors but not stroke outcome (Supplementary Table 14). For example,

372	genetically determined Fibroblast Growth Factor 5 (FGF5) level was associated with higher risk of
373	AF (OR=1.056 per SD higher FGF5); each SD higher genetically determined Glypican 5 (GPC5) was
374	associated with higher risk of T2D (OR=1.02); each SD higher in genetically determined Scavenger
375	Receptor Class F Member 2 (SCARF2) was associated with a 0.062-SD higher WMH. We found that
376	higher genetically determined Alpha-L-Iduronidase (IDUA) and Sialic Acid-Binding Ig-Like Lectin 9
377	(SIGLEC9) were both associated with lower BMI. Higher genetically determined Serine Protease 27
378	(PRSS27) was associated with higher SBP, higher DBP and higher PP, while higher genetically
379	determined levels of Neurocan (NCAN) were associated with lower risk of T2D (OR=0.76) and 0.07-
380	SD lower SBP.

381

382 Phenome-wide MR (Phe-MR) analysis of stroke-associated proteins in UK Biobank

383 To assess whether the six stroke-associated proteins have either beneficial or deleterious effects for

384 other indications, we performed a broader MR screen of 784 diseases and traits in UK Biobank

385 (Supplementary Table 8). Our Phe-MR results can be interpreted as a per-SD increase in genetically

386 determined plasma protein level that leads to an higher or lower odds of a given disease or trait. If the

387 effect direction of the protein on the disease or trait is the same as on stroke, the effect is considered

388 "beneficial" and "deleterious" otherwise. Overall, 34 significant associations were identified ($P \le$

 $0.05/6/784 = 1.06 \times 10^{-5}$), of which 21 (61.7%) were in the same direction as the stroke association

390 (Supplementary Table 9).

391 Notably, genetically higher levels of plasma TFPI were not only associated with lower risk of stroke,

392 but also lower risk of other diseases involving the circulatory system (cerebrovascular disease, other

disorders of arteries), metabolic traits (hyperlipidemia and hypercholesterolemia, disorders of lipid

394 metabolism) and digestive system disorders (acute gastritis); however, they were also associated with

395 higher risk of excessive or frequent menstruation (Figure 6 & Supplementary Figure 5).

396 Genetically higher levels of plasma TMPRSS5 were associated with higher risk of cardioembolic

- 397 stroke, as well as protein–calorie malnutrition (metabolic trait) (Figure 6 & Supplementary Figure
- 398 5). All the significant associations for CD40, including haemoptysis and abnormal sputum

- 16 -

399	(respiratory system) were consistent with the effect direction of that with stroke. Effects of IL6RA on
400	risk of diseases on circulatory system disorders (ischemic heart disease, cardiac dysrhythmias, atrial
401	fibrillation and flutter, coronary atherosclerosis, angina pectoris, abdominal aortic aneurysm) and
402	musculoskeletal disease (other inflammatory spondylopathies) were consistent with that on risk of
403	stroke; but had inverse effects on dermatologic symptoms (e.g. cellulitis and abscess of arm/foot),
404	digestive system (e.g. cholelithiasis) and chronic renal failure [CKD] (Supplementary Figure 6 &
405	Supplementary Table 9). Genetically predicted CD6 was associated with alcoholic liver damage
406	and degeneration of intervertebral disc (musculoskeletal system) but in the inverse direction to stroke.
407	Summary results of the primary and sensitivity MR analyses for all the 784 phenotypes are provided
408	in Supplementary Table 9.

409

410

411 Discussion

412	Based on genetic data for 308 proteins involved in cardiovascular disease, inflammation and
413	neurological processes from ~5000 individuals ¹⁶ , our study provides robust evidence that six proteins
414	(TFPI, TMPRSS5, CD40, MMP12, IL6RA, and CD6) are causally associated with stroke. We
415	showed that AF, systolic and diastolic BP, BMI, T2D, WMH and smoking were causally associated
416	with risk of any stroke (and some ischemic stroke subtypes), demonstrating a key role of the risk
417	factors in the pathogenesis of stroke consistent with classical epidemiological data ⁴⁸⁻⁵⁶ . We found the
418	associations of TFPI, IL6RA, and TMPRSS5 with stroke were likely to be mediated by one or more
419	of these risk factors. In addition, we showed that 36 additional proteins were causal for these risk
420	factors. Finally, the Phe-MR highlighted additional beneficial indications of therapeutically targeting
421	the six stroke-associated proteins and importantly, indicated few potential safety concerns. Although,
422	as many of the phenotypes tested are not independent, the definition of significance used here might
423	be too conservative (Bonferroni-corrected P-value adjusted for the number of proteins tested (six) and
424	the total number of phenotypes (784) ($P=0.05/6/784=1.06\times10^{-5}$).
425	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an
425 426	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of
425 426 427	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent
425 426 427 428	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding
425 426 427 428 429	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally
425 426 427 428 429 430	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart
425 426 427 428 429 430 431	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, $P=5.23 \times 10^{-5}$;
425 426 427 428 429 430 431 432	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, $P=5.23 \times 10^{-5}$; excessive or frequent menstruation in females, $P=2.70 \times 10^{-10}$). We also showed that higher levels of
425 426 427 428 429 430 431 432 433	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, $P=5.23 \times 10^{-5}$; excessive or frequent menstruation in females, $P=2.70 \times 10^{-10}$). We also showed that higher levels of TFPI were associated with lower BMI and WMH (Figure 5), and lower risk of hyperlipidemia,
425 426 427 428 429 430 431 432 433 434	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, $P=5.23 \times 10^{-5}$; excessive or frequent menstruation in females, $P=2.70 \times 10^{-10}$). We also showed that higher levels of TFPI were associated with lower BMI and WMH (Figure 5), and lower risk of hyperlipidemia, specifically hypercholesterolemia (Figure 6), suggesting that the pathways through which TFPI
425 426 427 428 429 430 431 432 433 434 435	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, $P=5.23 \times 10^{-5}$; excessive or frequent menstruation in females, $P=2.70 \times 10^{-10}$). We also showed that higher levels of TFPI were associated with lower BMI and WMH (Figure 5), and lower risk of hyperlipidemia, specifically hypercholesterolemia (Figure 6), suggesting that the pathways through which TFPI influences stroke risk might go beyond anticoagulation, <i>e.g.</i> , inflammation or atherosclerotic changes.
425 426 427 428 429 430 431 432 433 434 435 436	Tissue factor pathway inhibitor (TFPI) is primarily secreted by endothelial cells and is an anticoagulant that acts on the clotting cascade ⁵⁷ . Observational studies showed that lower levels of free TFPI were associated with higher risk of ischemic stroke ⁵⁸ and higher risk of first and recurrent venous thrombosis ⁵⁹ , while inhibition of TFPI showed to be an effective treatment of bleeding associated with hemophilia ⁶⁰ . Consistent with this, we provided genetic evidence for directionally consistent effects of TFPI on multiple ischemic traits, such as ischemic stroke and ischemic heart disease, and opposite effects on haemorrhagic traits (<i>e.g.</i> , gastrointestinal haemorrhage, <i>P</i> =5.23×10 ⁻⁵ ; excessive or frequent menstruation in females, <i>P</i> =2.70×10 ⁻¹⁰). We also showed that higher levels of TFPI were associated with lower BMI and WMH (Figure 5), and lower risk of hyperlipidemia, specifically hypercholesterolemia (Figure 6), suggesting that the pathways through which TFPI influences stroke risk might go beyond anticoagulation, <i>e.g.</i> , inflammation or atherosclerotic changes. Animal studies ^{60,61} provide supporting evidence that TFPI has a role in attenuating arterial thrombus

focusing on the role of TFPI activity and different TFPI isoforms in the development of atherogenesiscould provide further insights.

440	TMPRSS5 (Transmembrane Protease Serine 5, also known as Spinesin) is a member of the Type II
441	Transmembrane Serine Protease Family (TTSPs) ⁶² . For example, TMPRSS10 (Corin), one member
442	of the TTSPs, has been reported to be involved in cardiac conduction and myometrial relaxation and
443	contraction pathways in regulating blood pressure and promoting natriuresis, diuresis and vasodilation
444	⁶³ . Unlike Corin, the function of TMPRSS5 on cardiovascular systems is poorly understood. Human
445	TMPRSS5 mRNA has been shown to be expressed in the brain and the protein is predominantly
446	expressed in neurons, in their axons in the spinal cord ⁶⁴ . A mouse model with mutant TMPRSS5 had
447	reduced proteolytic activity and suggested a role in hearing loss ⁶⁵ . We were unable to find other
448	studies that implicate TMPRSS5 in cardiovascular disease, both for any ischemic stroke and
449	cardioembolic stroke, an effect that might be mediated by risk of atrial fibrillation (Figure 4).
450	Furthermore, Phe-MR analysis revealed suggestive additional beneficial effects when targeted at
451	TMPRSS5, <i>e.g.</i> , reduced risk of Parkinson's disease ($P=2.15\times10^{-5}$) and left bundle branch block
452	$(P=1.43\times10^{-5})$. Taken together, TMPRSS5 represents a potentially promising therapeutic target for
453	atrial fibrillation and cardioembolic stroke, and further research is warranted to decipher the
454	mechanism through which it protects against cardiovascular and neurological diseases.
455	In addition, we have identified CD6, a lymphocyte surface receptor, associated with increased risk of
456	any stroke. CD6 is a pan T cell marker ^{66,67} , and involved in T cell proliferation and activation
457	through its interaction with ALCAM (activated leukocyte cell adhesion molecule) ⁶⁸ . The interaction
458	of CD6 and ALCAM is required to promote an inflammatory T cell response ⁶⁹ . Interestingly,
459	Smedbakken L et al ⁷⁰ found that acute ischemic stroke patients with upregulated ALCAM at
460	admission had a significantly poorer survival rate ($P < 0.001$). Given this interaction and that the
461	recruitment of leukocytes and platelets is widely regarded as a pivotal step in the inflammatory
462	response associated with cerebral ischemia ^{71,72} , together with our finding that CD6 is associated
463	with stroke, further investigation of CD6 in the context of stroke is justified.

464 Our study not only identified potentially novel targets (*i.e.* TFPI, TMPRSS5 and CD6) for stroke, but 465 also validated proteins that had been identified as causally associated with cardiovascular diseases in 466 previous proteome MR studies ^{11,46,73}, e.g. CD40, MMP12 and IL6RA. Genetic variants in the *IL6R* region are associated with risk of inflammatory related diseases ⁴⁴, including coronary heart disease ⁷⁴, 467 stroke ⁷³, atrial fibrillation ⁷⁵ and rheumatoid arthritis ⁷⁶. Moreover, IL6R is the target of an FDA-468 469 approved therapy (Tocilizumab) for the treatment of several diseases, e.g., rheumatoid arthritis and 470 systemic juvenile idiopathic arthritis. Phase II clinical trials testing tocilizumab for the therapy of Non-ST Elevation Myocardial Infarction have reported promising results ⁷⁷ and a phase III clinical 471 472 trial testing Ziltivekimab in cardiovascular disease and chronic kidney disease has recently started 473 (NCT05021835). 474 To avoid violating the MR assumptions, we performed various sensitivity analyses. We used LD 475 clumping at R^2 0.1 for pQTLs with P5.0×10⁻⁸ to choose instruments for each of plasma protein level. 476 However, concern ⁷⁸ has been raised about the independency of the variants used as instrumental 477 variables leading to violation of the InSIDE (instrument strength independent of direct effect) 478 assumption of the MR-Egger method used. Therefore, we performed several sensitivity analyses to 479 validate the robustness of the instrumental variables used in the MR analysis. Firstly, we performed 480 MR analyses adjusting for the correlation of the variants used and obtained consistent and similar 481 causal effect estimates to those obtained without adjusting for the correlation (Supplementary Table 482 **10**). Secondly, we performed conditional analysis and fine-mapping analysis to obtain instrumental 483 variables for the six potential causal proteins and we obtained consistent MR results (Supplementary 484
 Table 12 & Supplementary Figure 7).
 Finally, colocalization analyses across the genetic
 485 associations with protein levels and stroke outcomes showed they were likely to have shared causal 486 variants across these traits, supporting the validity of instrumental variables and the causal protein 487 associations (Supplementary Table 4). 488 The Olink assay¹⁹ used in our study measures the bulk concentration of protein in plasma. However,

489 because this assay cannot distinguish free from bound protein or active from inactive, only limited

490 mechanistic insights can be made. Due to the limited capture of human proteome (1.5% of all known

491	proteins), we could not evaluate the effects of all proteins within the same family or all proteins
492	encoded within the same genomic region. For instance, we found that TMPRSS5 was a potential
493	novel drug target for cardioembolic stroke, while other proteins in the Type II Transmembrane Serine
494	Protease Family (TTSPs) that play crucial roles in cardiac functions ^{62,79} could not be evaluated.
495	Thus, a targeted study of the TTSP family is warranted to comprehensively evaluate their effects in
496	cardiovascular and neurological traits.
497	Our results highlight potential targets of future therapies for stroke outcomes and illustrates the
498	relevance of proteomics in identifying drug targets. Further research is necessary to assess the
499	viability of the six identified proteins as drug targets for stroke treatment. Additional drug targets
500	may be uncovered as more comprehensive proteomics platforms become available and more diverse
501	non-European ancestry populations are increasingly studied. Finally, there is an increasing need for
502	similarly comprehensive proteomics across different tissues and organs to evaluate tissue- or organ-
503	specific protein effects.

504

505 Funding

- 506 This work and LC were funded by a program grant from the British Heart Foundation
- 507 (RG/16/4/32218). PS is supported by a Rutherford Fund Fellowship from the Medical Research
- 508 Council (MR/S003746/1). BP and SK are funded by a British Heart Foundation Programme grant
- 509 (RG/18/13/33946). EY-D was funded by the Isaac Newton Trust / Wellcome Trust ISSF / University
- 510 of Cambridge Joint Research Grants Scheme. SB is supported by a Sir Henry Dale Fellowship jointly
- 511 funded by the Wellcome Trust and the Royal Society (204623/Z/16/Z). JMMH was funded by the
- 512 NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. CML is funded by the NIHR
- 513 Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's
- 514 College London. HSM is supported by a NIHR Senior Investigator award, and his work is supported
- 515 by the Cambridge Universities NIHR Comprehensive Biomedical Research Centre. JD holds a
- 516 British Heart Foundation Professorship and a NIHR Senior Investigator Award [*].
- *The views expressed are those of the author(s) and not necessarily those of the NIHR or the
- 518 Department of Health and Social Care.
- 519
- 520

521 Acknowledgements

- 522 Participants in the INTERVAL randomised controlled trial were recruited with the active
- 523 collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field
- 524 work and other elements of the trial. DNA extraction and genotyping were co-funded by the National
- 525 Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the
- 526 NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. The Olink® Proteomics
- 527 assays were funded by Biogen, Inc. (Cambridge, MA, US). The academic coordinating centre for
- 528 INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in
- 529 Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council
- 530 (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and
- 531 NIHR Cambridge BRC (BRC-1215-20014) [*] and funding from the EC-Innovative Medicines

- 532 Initiative (BigData@Heart). A complete list of the investigators and contributors to the INTERVAL
- 533 trial is provided in reference [**]. The academic coordinating centre would like to thank blood donor
- centre staff and blood donors for participating in the INTERVAL trial.
- 535 This work was supported by Health Data Research UK, which is funded by the UK Medical Research
- 536 Council, Engineering and Physical Sciences Research Council, Economic and Social Research
- 537 Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish
- 538 Government Health and Social Care Directorates, Health and Social Care Research and Development
- 539 Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation
- 540 and Wellcome.
- *The views expressed are those of the author(s) and not necessarily those of the NIHR or the
- 542 Department of Health and Social Care.
- 543 **Di Angelantonio E, Thompson SG, Kaptoge SK, Moore C, Walker M, Armitage J, Ouwehand WH,
- 544 Roberts DJ, Danesh J, INTERVAL Trial Group. Efficiency and safety of varying the frequency of
- 545 whole blood donation (INTERVAL): a randomised trial of 45 000 donors. Lancet. 2017 Nov
- 546 25;390(10110):2360-2371.
- 547

548 Author contributions

- 549 L.C. performed the analyses and wrote the initial draft of the manuscript. J.M.M.H. designed and
- supervised the project. J.E.P, B.P, E.P provided data and analytical support. All authors contributed
- to the data preparation and critically reviewed the manuscript.
- 552
- 553

554 **Competing interests**

- 555 JMMH, LC, MT and EY-D became full time employees of Novo Nordisk Ltd during the drafting of
- this manuscript. JD reports grants, personal fees and non-financial support from Merck Sharp &
- 557 Dohme (MSD), grants, personal fees and non-financial support from Novartis, grants from Pfizer and
- 558 grants from AstraZeneca outside the submitted work. JD sits on the International Cardiovascular and

- 559 Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since
- 560 2011); the MRC International Advisory Group (ING) member, London (since 2013); the MRC High
- 561 Throughput Science 'Omics Panel Member, London (since 2013); the Scientific Advisory Committee
- 562 for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development
- 563 Portfolio Committee for Novartis; and the Astra Zeneca Genomics Advisory Board (2018). ASB
- 564 reports institutional grants from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Merck, Novartis
- and Sanofi and personal fees from Novartis. PB is a full-time employee of Biogen Inc. CML sits on
- 566 the Scientific Advisory Board for Myriad Neuroscience.
- 567

568 **References**

569 570	1.	Feigin, V.L. <i>et al.</i> Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. <i>The</i>
571		Lancet Neurology 18 , 459-480 (2019).
572	2.	Ettehad, D. et al. Blood pressure lowering for prevention of cardiovascular disease
573		and death: a systematic review and meta-analysis. The Lancet 387 , 957-967 (2016).
574	3.	Hankey, G.J. Stroke. The Lancet 389 , 641-654 (2017).
575	4.	Santos, R. et al. A comprehensive map of molecular drug targets. Nat Rev Drug
576		Discov 16, 19-34 (2017).
577	5.	Olszewski, A.J. & Szostak, W.B. Homocysteine content of plasma proteins in ischemic
578		heart disease. Atherosclerosis 69, 109-13 (1988).
579	6.	Robins, S.J., Lyass, A., Brocia, R.W., Massaro, J.M. & Vasan, R.S. Plasma lipid transfer
580		proteins and cardiovascular disease. The Framingham Heart Study. Atherosclerosis
581		228 , 230-6 (2013).
582	7.	Goetzl, E.J. et al. Altered lysosomal proteins in neural-derived plasma exosomes in
583		preclinical Alzheimer disease. <i>Neurology</i> 85 , 40-7 (2015).
584	8.	Feldreich, T. et al. Circulating proteins as predictors of cardiovascular mortality in
585		end-stage renal disease. Journal of nephrology 32 , 111-119 (2019).
586	9.	Hauser, A.S. et al. Pharmacogenomics of GPCR Drug Targets. Cell 172, 41-54 e19
587		(2018).
588	10.	Ursu, O., Glick, M. & Oprea, T. Novel drug targets in 2018. Nat Rev Drug Discov
589		(2019).
590	11.	Sun, B.B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73-79
591		(2018).
592	12.	Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively
593		causal genes and pathways for cardiovascular disease. Nat Commun 9, 3268 (2018).
594	13.	Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to
595		disease. Science 361 , 769-773 (2018).
596	14.	Smith, G.D. Mendelian randomization for strengthening causal inference in
597		observational studies: application to gene× environment interactions. Perspectives
598		on Psychological Science 5 , 527-545 (2010).
599	15.	Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal
600		inference in epidemiological studies. <i>Human molecular genetics</i> 23 , R89-R98 (2014).
601	16.	Di Angelantonio, E. et al. Efficiency and safety of varying the frequency of whole
602		blood donation (INTERVAL): a randomised trial of 45 000 donors. The Lancet 390 ,
603		2360-2371 (2017).
604	17.	Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects
605		identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 50, 524-537
606		(2018).
607	18.	Denny, J.C. et al. Systematic comparison of phenome-wide association study of
608		electronic medical record data and genome-wide association study data. Nat
609		Biotechnol 31 , 1102-10 (2013).
610	19.	Assarsson, E. et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity,
611		specificity, and excellent scalability. PloS one 9, e95192 (2014).
612	20.	Durbin, R. Efficient haplotype matching and storage using the positional Burrows-
613		Wheeler transform (PBWT). Bioinformatics 30, 1266-1272 (2014).

614	21.	Astle, W.J. et al. The allelic landscape of human blood cell trait variation and links to
615		common complex disease. <i>Cell</i> 167 , 1415-1429. e19 (2016).
616	22.	Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint
617		method for genome-wide association studies by imputation of genotypes. Nature
618		Genetics 39 , 906-913 (2007).
619	23.	Chang, C.C. et al. Second-generation PLINK: rising to the challenge of larger and
620		richer datasets. <i>Gigascience</i> 4 , 7 (2015).
621	24.	Pierce, B.L., Ahsan, H. & Vanderweele, T.J. Power and instrument strength
622		requirements for Mendelian randomization studies using multiple genetic variants.
623		Int J Epidemiol 40 , 740-52 (2011).
624	25.	Palmer, T.M. et al. Using multiple genetic variants as instrumental variables for
625		modifiable risk factors. Statistical Methods in Medical Research 21, 223-242 (2012).
626	26.	Benner, C. et al. FINEMAP: efficient variable selection using summary data from
627		genome-wide association studies. <i>Bioinformatics</i> 32 , 1493-1501 (2016).
628	27.	Surendran, P. et al. Discovery of rare variants associated with blood pressure
629		regulation through meta-analysis of 1.3 million individuals. Nature Genetics (2020).
630	28.	Nielsen, J.B. et al. Biobank-driven genomic discovery yields new insight into atrial
631		fibrillation biology. <i>Nat Genet</i> 50 , 1234-1239 (2018).
632	29.	Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using
633		high-density imputation and islet-specific epigenome maps. Nat Genet 50, 1505-
634		1513 (2018).
635	30.	Persyn, E. et al. Genome-wide association study of MRI markers of cerebral small
636		vessel disease in 42,310 participants. Nature Communications 11, 2175 (2020).
637	31.	Pulit, S.L. et al. Meta-analysis of genome-wide association studies for body fat
638		distribution in 694 649 individuals of European ancestry. Hum Mol Genet 28, 166-174
639		(2019).
640	32.	Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights
641		into the genetic etiology of tobacco and alcohol use. Nat Genet 51, 237-244 (2019).
642	33.	Zhou, W. et al. Efficiently controlling for case-control imbalance and sample
643		relatedness in large-scale genetic association studies. Nat Genet 50, 1335-1341
644		(2018).
645	34.	Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian Randomization Analysis
646		With Multiple Genetic Variants Using Summarized Data. Genetic Epidemiology 37,
647		658-665 (2013).
648	35.	Yavorska, O.O. & Burgess, S. MendelianRandomization: an R package for performing
649		Mendelian randomization analyses using summarized data. International Journal of
650		Epidemiology 46 , 1734-1739 (2017).
651	36.	Hemani, G. et al. The MR-Base platform supports systematic causal inference across
652		the human phenome. <i>Elife</i> 7 (2018).
653	37.	Lawlor, D.A. Commentary: Two-sample Mendelian randomization: opportunities and
654		challenges. International journal of epidemiology 45 , 908 (2016).
655	38.	Bowden, J., Smith, G.D. & Burgess, S. Mendelian randomization with invalid
656		instruments: effect estimation and bias detection through Egger regression.
657		International Journal of Epidemiology 44 , 512-525 (2015).
658	39.	Verbanck, M., Chen, Cy., Neale, B. & Do, R. Detection of widespread horizontal
659		pleiotropy in causal relationships inferred from Mendelian randomization between
660		complex traits and diseases. Nature genetics 50, 693-698 (2018).

661	40.	Burgess, S., Foley, C.N., Allara, E., Staley, J.R. & Howson, J.M.M. A robust and
662		efficient method for Mendelian randomization with hundreds of genetic variants.
663		Nature Communications 11 , 376 (2020).
664	41.	Foley, C.N. et al. A fast and efficient colocalization algorithm for identifying shared
665		genetic risk factors across multiple traits. Nat Commun 12, 764 (2021).
666	42.	Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic
667		association studies using summary statistics. <i>PLoS Genet</i> 10 , e1004383 (2014).
668	43.	Zheng, J. <i>et al.</i> Phenome-wide Mendelian randomization mapping the influence of
669		the plasma proteome on complex diseases. <i>Nat Genet</i> 52 , 1122-1131 (2020).
670	44.	Ferreira, R.C. <i>et al.</i> Functional IL6R 358Ala Allele Impairs Classical IL-6 Receptor
671		Signaling and Influences Risk of Diverse Inflammatory Diseases. PLOS Genetics 9.
672		e1003444 (2013).
673	45.	Williams, F.M. <i>et al.</i> Ischemic stroke is associated with the ABO locus: the EuroCLOT
674		study. Ann Neurol 73 , 16-31 (2013).
675	46	Chong, M. <i>et al.</i> Novel Drug Targets for Ischemic Stroke Identified Through
676	10.	Mendelian Randomization Analysis of the Blood Proteome <i>Circulation</i> (2019)
677	47	Hodgson L et al. Characterization of GDE2 Mutations and Levels of BMP9 and
678	47.	BMP10 in Pulmonary Arterial Hypertension Am Respir Crit Care Med 201 575-585
679		
680	18	Wolf P.A. Abbott R.D. & Kannel W.B. Atrial fibrillation as an independent risk
681	40.	factor for stroke: the Framingham Study. Stroke 22 , 983-988 (1991)
682	10	Vang $X = M$ at al. Atrial fibrillation known before or detected after stroke share
683	49.	similar rick of ischamic stroke recurrence and death. Stroke 50 , 1124, 1120 (2010)
684	50	Similar fisk of ischemic stroke recurrence and death. Stroke 50 , 1124-1129 (2015).
685	50.	overview of published reviews Stroke 35 , 776-785 (2004)
686	51	Kannel W.B. Wolf P.A. Verter I.& McNamara P.M. Enidemiologic assessment of
687	51.	the role of blood pressure in stroke: the Framingham study. <i>Jama</i> 276 , 1260-1278
688		(1006)
680	50	(1990). Mäntulä P. at al. Magnotic reconance imaging white matter hyperintensities and
600	52.	machanism of icchamic stroke. Stroke 20 , 2052, 2059 (1000)
601	ED	Mitchell A. P. et al. Obesity Increases Bick of Ischemic Stroke in Young Adults. Stroke
602	55.	AC 1600 1602 (2015)
602	E /	40, 1050-1052 (2015). Kivimäki M. et al. Overweight, obesity, and rick of cardiometabolic multimerhidity:
604	54.	NVIIIIdki, W. <i>et al.</i> Over weight, obesity, and fisk of caldiometabolic multimorbidity.
694		from the USA and Europe. The Langest Dublic Logith 2 , a227, a285 (2017)
093		from the USA and Europe. The Lancet Public Health 2, e277-e285 (2017).
696	55.	Jangnorbani, IVI. <i>et al.</i> Prospective Study of Type 1 and Type 2 Diabetes and Risk of
697	50	Stroke Subtypes. The Nurses' Health Study 30 , 1730-1735 (2007).
698	56.	Rost, N.S. <i>et al.</i> white matter hyperintensity volume is increased in small vessel
699 700		stroke subtypes. Neurology 75 , 16/0-16/7 (2010).
/00	57.	Broze Jr, G.J. Tissue factor pathway inhibitor. <i>Thrombosis and haemostasis</i> 73 , 090-
701		093 (1995).
702	58.	He, M. et al. Observation on tissue factor pathway and some other coagulation
703		parameters during the onset of acute cerebrocardiac thrombotic diseases.
704		Thrombosis research 107 , 223-228 (2002).
705	59.	Hoke, M. et al. Tissue factor pathway inhibitor and the risk of recurrent venous
706		thromboembolism. <i>Thrombosis and haemostasis</i> 94 , 787-790 (2005).

707	60.	Waters, E.K. et al. Aptamer ARC19499 mediates a procoagulant hemostatic effect by
708		inhibiting tissue factor pathway inhibitor. <i>Blood</i> 117 , 5514-5522 (2011).
709	61.	Westrick, R.J. et al. Deficiency of tissue factor pathway inhibitor promotes
710		atherosclerosis and thrombosis in mice. Circulation 103, 3044-3046 (2001).
711	62.	Bugge, T.H., Antalis, T.M. & Wu, Q. Type II transmembrane serine proteases. J Biol
712		Chem 284 , 23177-81 (2009).
713	63.	Knappe, S., Wu, F., Masikat, M.R., Morser, J. & Wu, Q. Functional analysis of the
714		transmembrane domain and activation cleavage of human corin design and
715		characterization of a soluble corin. Journal of Biological Chemistry 278, 52363-52370
716		(2003).
717	64.	Yamaguchi, N., Okui, A., Yamada, T., Nakazato, H. & Mitsui, S. Spinesin/TMPRSS5, a
718		novel transmembrane serine protease, cloned from human spinal cord. J Biol Chem
719		277 , 6806-12 (2002).
720	65.	Guipponi, M. et al. An integrated genetic and functional analysis of the role of type II
721		transmembrane serine proteases (TMPRSSs) in hearing loss. Hum Mutat 29, 130-41
722		(2008).
723	66.	Carrasco, E. et al. Human CD6 Down-Modulation following T-Cell Activation
724		Compromises Lymphocyte Survival and Proliferative Responses. Front Immunol 8,
725		769 (2017).
726	67.	Hernández, P., Moreno, E., Aira, L.E. & Rodríguez, P.C. Therapeutic Targeting of CD6
727		in Autoimmune Diseases: A Review of Cuban Clinical Studies with the Antibodies
728		IOR-T1 and Itolizumab. <i>Curr Drug Targets</i> 17 , 666-77 (2016).
729	68.	Zimmerman, A.W. et al. Long-term engagement of CD6 and ALCAM is essential for T-
730		cell proliferation induced by dendritic cells. <i>Blood</i> 107 , 3212-3220 (2006).
731	69.	Gimferrer, I. et al. Relevance of CD6-Mediated Interactions in T Cell Activation and
732		Proliferation. The Journal of Immunology 173 , 2262-2270 (2004).
733	70.	Smedbakken, L. et al. Activated Leukocyte Cell Adhesion Molecule and Prognosis in
734		Acute Ischemic Stroke. <i>Stroke</i> 42 , 2453-2458 (2011).
735	71.	Jin, R., Yang, G. & Li, G. Inflammatory mechanisms in ischemic stroke: role of
736		inflammatory cells. <i>J Leukoc Biol</i> 87, 779-89 (2010).
737	72.	Elkind, M.S. Inflammatory mechanisms of stroke. <i>Stroke</i> 41 , S3-8 (2010).
738	73.	Georgakis, M.K. <i>et al.</i> Genetically determined levels of circulating cytokines and risk
739		of stroke: role of monocyte chemoattractant protein-1. <i>Circulation</i> 139 , 256-268
740		(2019).
741	74.	Swerdlow, D.I. <i>et al.</i> The interleukin-6 receptor as a target for prevention of coronary
742		heart disease: a mendelian randomisation analysis. <i>Lancet</i> 379 , 1214-24 (2012).
/43	75.	Schnabel, R.B. <i>et al.</i> Large-scale candidate gene analysis in whites and African
/44		Americans identifies IL6R polymorphism in relation to atrial fibrillation: the National
/45		Heart, Lung, and Blood Institute's Candidate Gene Association Resource (CARE)
/40	70	project. <i>Circulation: Caralovascular Genetics</i> 4 , 557-564 (2011).
/4/	76.	Eyre, S. <i>et al.</i> High-density genetic mapping identifies new susceptibility loci for
/48 740		meumatoio arthritis. <i>Noture genetics</i> 44 , 1336-1340 (2012).
/49 750	//.	Deland, I. <i>et al.</i> Serum PCSK9 is modified by interleukin-6 receptor antagonism in
/30 751		patients with hypercholesterolaemia following non-ST-elevation myocardial
131	70	Initial Culon. Open Ineurity, e000705 (2018).
132 752	/ð.	Plump, A. & Davey Smith, G. Identifying and Validating New Drug Targets for Stroke
155		and Beyond: Can Mendelian Randomization Help? (Am Heart Assoc, 2019).

754 79. Szabo, R. et al. Type II transmembrane serine proteases. Thromb Haemost 90, 185-

755 93 (2003).

756

757

758	Figures (in separate file)
759	Figure 1. Overview of MR analyses.
760	Figure 2. Venn diagram of identified potential causal proteins for stroke subtypes.
761	Figure 3. Effects of six potential causal proteins on stroke subtypes.
762	Figure 4. Causal effects of risk factors on stroke subtypes.
763	Figure 5. Effect sizes (Z-score) of six potential causal proteins on stroke subtypes and causal risk
764	factors for stroke.
765	Figure 6. Forest plots illustrating the potential on-target side-effects associated with causal proteins
766	revealed by Phe-MR analysis for TFPI (A) and TMPRSS5 (B).
767 768 769	Tables (in separate file)
770	Table 1. Data sources for the Mendelian Randomization analysis in current study.
771	Table 2. Summary of significant proteins (biomarkers) representing potential causal factors for stroke
772	and subtypes.
773	Table 3. Stroke risk factors are potential mediators of stroke associated proteins and stroke.
774 775	Supplementary Figures (in separate file)
776 777	Supplementary Tables (in separate file)

778

Figure 1. Overview of MR analyses.



Four O-link panels were used to measure plasma proteins in a subset of ~5000 samples from the INTERVAL study ¹⁶. Genetic variants associated with plasma protein levels were identified based on results from their corresponding GWAS. These genetic variants were then used as proxies for the protein level and tested their relationship with stroke was tested used data from the MEGASTROKE consortium ¹⁷ for stroke outcomes (Primary MR), with conventional stroke risk factors (Secondary MR), and with 784 phenotypes (Phe-MR) in UK Biobank to test a broad spectrum of potential effects of hypothetical therapeutic agents for stroke.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.22.21265375; this version posted October 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license. Figure 2. Venn diagram of identified potential causal proteins for stroke subtypes.





OR: Odds ratio; *I*²: heterogeneity. CD40: B Cell Surface Antigen CD40; TFPI: Tissue Factor Pathway Inhibitor; MMP12: Matrix Metallopeptidase 12; IL6RA: Interleukin 6 Receptor Subunit Alpha; TMPRSS5: Transmembrane Serine Protease 5; CD6: T-Cell Differentiation Antigen CD6.



*I*²: heterogeneity. SBP = Systolic Blood Pressure; AF = Atrial Fibrillation; WMH = White Matter Hyperintensity; T2D = Type 2 Diabetes; BMI = Body Mass Index; Smoking = Smoking Initiation.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.22.21265375; this version posted October 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND.4.0 International license. Figure 5. Effect sizes (Z-score) of six potential causal proteins on stroke subtypes and causal risk

factors for stroke.



Colours in each lattice of the heatmap represent the effect size (Z-score), with genetically predicted increased protein level associated with higher risk of outcomes coloured in brown and lower risk of outcomes coloured in blue. The darker the colour the larger the effect size.

* indicates that the causal association is significant, which passed Bonferroni correction of $P_{causalEstimate_IVW} \le 0.05/308 = 1.61 \times 0^{-4}$ and passed sensitivity tests with $P_{Qstat} \ge 0.05$ and $P_{EggerIntercept} \ge 0.05$.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.22.21265375; this version posted October 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.
Figure 6. Forest plots illustrating the potential on-target side-effects associated with causal proteins

revealed by Phe-MR analysis for TFPI (A) and TMPRSS5 (B).

A.



OR of diseases per-SD increase in TMPRSS5 level

In general, results can be perceived as the effects of per SD higher circulating protein level on each phenotype. If the effect direction of the target protein on the phenotype is consistent with that on stroke outcomes, it represents "beneficial" additional indications through intervention of circulating protein level. Conversely, opposing effect directions of the target protein on the phenotype and stroke represents "deleterious" side-effects. For example, a higher level of TFPI is associated with lower

medRxiv preprint doi: https://doi.org/10.1101/2021.10.22.21265375; this version posted October 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license . risk of ischemic stroke and so phenotypes with OR<1 represents "beneficial effects", OR>1 represents

"deleterious effects" when the hypothetical intervention increases TFPI levels. Only significant

associations that passed Bonferroni correction ($P \le 0.05/6/784 = 1.06 \times 10^{-5}$) were plotted. See

Supplementary Table 9 for more clinical information of the ICD code phenotypes.

Table 1. Data sources for the Mendelian Randomization analysis for current study.

Phenotype	Source	N (Total or Cases/Controls)	Imputation reference panel	Ancestry
Olink protein levels	INTERVAL study (unpublished data)	4,994	1000 Genomes Phase 3 +	European
Inflammation Panel (INF1)			UK10K	
Cardiovascular Panels				
(CVD2/CVD3)				
Neurology Panel (NEURO)				
Primary Outcomes				
Any stroke	17 studies (Malik et al) ¹⁷	40,585/406,111	1000 Genomes Phase 1	European
Ischemic stroke		34,217/406,111		
Large artery stroke		4,373/406,111		
Cardio-embolic stroke		7,193/406,111		
Small vessel stroke		5,386/406,111		
Secondary Outcomes				
Atrial fibrillation (AF)	6 Studies (Nielsen, et al) ²⁸	60,620/970,216	HRC [*]	European
Type 2 Diabetes (T2D)	32 Studies (Mahajan, et al) ²⁹	74,124/824,006	HRC	European
Body Mass Index (BMI)	GIANT + UK Biobank (Pulit, et al) 31	694,649	HRC	European
Tobacco and alcohol use	29 Studies (Liu, et al) 32		HRC	European
AgeSmk		341,427		
CigDay		337,334		
SmkCes		547,219		
SmkInit		1,232,091		
DrnkWk		941,280		
Blood pressure (BP)	UK Biobank (Surendran, et al.) ²⁷	445,415	HRC	European
Systolic BP				
Diastolic BP				
Pulse pressure (PP)				
White Matter Hyperintensity (WMH)	UK Biobank + CHARGE + study in stroke	42,310	HRC	Trans-ethnic, mainly
	patients (Persyn et al., 2020) ³⁰			European
On-target side effects evaluation	22			
784 Phenotypes	UK Biobank (Zhou, et al) 33	408,961	HRC	European

*HRC: The Haplotype Reference Consortium (HRC); AgeSmk: Age of Initiation of Regular Smoking; CigDay: Cigarettes per day; SmkCes: Smoking Cessation; SmkInit: Smoking Initiation; DrnkWk: Drinks per week.

Protein	N SNPs	Outcome	OR [95%CI] [#]	P value
TFPI	21	Stroke	0.949[0.926, 0.972]	2.01×10 ⁻⁵
	21	Ischemic-stroke	0.934[0.91, 0.959]	4.68×10 ⁻⁷
TMPRSS5	20	Stroke	1.058[1.039, 1.077]	1.17×10 ⁻⁹
	20	Ischemic-stroke	1.059[1.038, 1.08]	1.36×10 ⁻⁸
	20	Cardioembolic-stroke	1.089[1.045, 1.134]	5.33×10 ⁻⁵
CD40	10	Stroke	0.919[0.891, 0.949]	2.05×10 ⁻⁷
	10	Ischemic-stroke	0.913[0.882, 0.946]	3.61×10 ⁻⁷
	10	Large-artery-stroke	0.795[0.723, 0.874]	2.09×10 ⁻⁶
CD6	23	Stroke	1.039[1.021, 1.057]	1.26×10 ⁻⁵
IL6RA	39	Stroke	0.972[0.96, 0.984]	5.43×10 ⁻⁶
	39	Ischemic-stroke	0.969[0.955, 0.982]	7.99×10 ⁻⁶
	39	Small-vessel-stroke	0.939[0.909, 0.97]	1.60×10 ⁻⁴
MMP12	13	Stroke	0.922[0.901, 0.945]	3.42×10 ⁻¹¹
	13	Ischemic-stroke	0.918[0.895, 0.942]	9.07×10 ⁻¹¹
	12	Large-artery-stroke	0.793[0.73, 0.861]	3.53×10 ⁻⁸

Table 2. Pr	oteins represe	nting potential	causal factors	for stroke a	and subtypes.
-------------	----------------	-----------------	----------------	--------------	---------------

[#]OR [95%CI]=Odds ratio and its 95% confidence interval per 1-SD higher genetically-predicted plasma protein level

Table 3.

Risk Factors	Stroke outcome	N SNPs	OR [95%CI] *	P-value
AF	Stroke	130	1.21[1.18, 1.24]	2.23×10 ⁻³⁹
	Ischemic-stroke	130	1.22[1.19, 1.26]	1.11×10 ⁻³⁹
	Large-artery-stroke	131	1.04[0.97, 1.12]	0.315
	Cardioembolic-stroke	133	2.04[1.92, 2.16]	2.72×10 ⁻¹²⁵
	Small-vessel-stroke	131	0.99[0.92, 1.06]	0.714
BMI	Stroke	792	1.19[1.13, 1.25]	1.73×10 ⁻¹⁰
	Ischemic-stroke	791	1.17[1.11, 1.24]	3.25×10 ⁻⁰⁸
	Large-artery-stroke	794	1.29[1.12, 1.49]	3.24×10 ⁻⁰⁴
	Cardioembolic-stroke	792	1.15[1.04, 1.28]	7.49×10 ⁻⁰³
	Small-vessel-stroke	794	1.11[0.98, 1.25]	0.118
SBP	Stroke	701	1.67[1.57, 1.78]	1.37×10 ⁻⁵⁶
	Ischemic-stroke	704	1.68[1.57, 1.81]	3.49×10 ⁻⁴⁸
	Large-artery-stroke	710	2.58[2.21, 3.01]	3.69×10 ⁻³³
	Cardioembolic-stroke	705	1.36[1.2, 1.53]	8.31×10 ⁻⁰⁷
	Small-vessel-stroke	708	1.99[1.72, 2.29]	3.66×10 ⁻²¹
DBP	Stroke	661	1.5[1.4, 1.6]	1.04×10 ⁻³²
	Ischemic-stroke	667	1.5[1.4, 1.62]	1.67×10 ⁻²⁸
	Large-artery-stroke	674	1.72[1.46, 2.02]	5.01×10 ⁻¹¹
	Cardioembolic-stroke	673	1.26[1.12, 1.43]	2.11×10 ⁻⁰⁴
	Small-vessel-stroke	677	1.8[1.54, 2.1]	6.30×10 ⁻¹⁴
PP	Stroke	723	1.4[1.31, 1.49]	1.07×10 ⁻²⁵
	Ischemic-stroke	723	1.41[1.33, 1.51]	1.82×10 ⁻²⁵
	Large-artery-stroke	725	2.23[1.92, 2.59]	3.29×10 ⁻²⁶
	Cardioembolic-stroke	726	1.19[1.06, 1.33]	3.21×10 ⁻⁰³
	Small-vessel-stroke	725	1.53[1.34, 1.75]	4.17×10 ⁻¹⁰
T2D	Stroke	340	1.09[1.07, 1.12]	2.23×10 ⁻¹³
	Ischemic-stroke	341	1.1[1.08, 1.13]	3.96×10 ⁻¹⁵
	Large-artery-stroke	341	1.24[1.16, 1.32]	1.76×10 ⁻¹¹
	Cardioembolic-stroke	337	1.05[1, 1.1]	0.0352
	Small-vessel-stroke	342	1.17[1.1, 1.24]	1.05×10^{-07}
WMH	Stroke	16	1.21[1.07, 1.36]	2.21×10 ⁻⁰³
	Ischemic-stroke	17	1.15[1, 1.31]	0.046
	Large-artery-stroke	18	1.13[0.9, 1.43]	0.295
	Cardioembolic-stroke	18	1.14[0.94, 1.37]	0.183
	Small-vessel-stroke	15	1.49[1.17, 1.9]	1.47×10^{-03}
SmkInit	Stroke	365	1.24[1.16, 1.32]	1.80×10 ⁻¹¹
	Ischemic-stroke	364	1.22[1.15, 1.31]	6.34×10 ⁻¹⁰
	Large-artery-stroke	365	1.66[1.42, 1.93]	8.88×10 ⁻¹¹
	Cardioembolic-stroke	365	1.11[0.98, 1.25]	0.116
	Small-vessel-stroke	365	1.46[1.25, 1.71]	1.27×10^{-06}

a. MR results of risk factors and stroke outcomes.

Protein	N SNPs	Risk Factors	OR/β [95%CI] [#]	P value
TFPI	19	BMI	-0.013[-0.019, -0.007]	3.56×10 ⁻⁵
	19	WMH	-0.06[-0.08, -0.04]	7.15×10 ⁻¹⁰
TMPRSS5	21	AF	1.03[1.016, 1.045]	2.15×10 ⁻⁵
IL6RA	46	AF	0.959[0.95, 0.968]	2.55×10 ⁻¹⁸

b. MR results of stroke associated proteins and risk factors.

[#]OR/ β [95%CI]: Odds ratio/log(odds ratio) and its 95% confidence interval, indicates pe-1-SD increase in protein level and its effect on each outcome, if the outcome is a continuous trait, *e.g.* BMI, BP, we report the effect using β , otherwise, we use OR (for binary outcome). OR=Odds Ratio; CI=Confidence Interval; AF=Atrial Fibrillation; BMI=Body Mass Index; WMH=White Matter Hyperintensity; T2D=Type 2 Diabetes; SBP=Systolic Blood Pressure; DBP=Diastolic Blood Pressure; PP=Pulse Pressure; SmkInit=Smoking Initiation.