Identification of the putative causal risk factors and biomarkers of stroke using large-scale genome-wide studies

Tania Islam[1], Luis M García-Marín[2,3], Miguel E. Rentería[2,3], Gabriel Cuellar-Partida[4], Asaduzzaman Khan[1], Mohammad Ali Moni[1,*]

[1] School of Health and Rehabilitation Sciences, Faculty of Health, The University of Queensland, Brisbane, QLD 4072, Australia

[2]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

[3]School of Biomedical Sciences, Faculty of Medicine, The University of Queensland, Brisbane, QLD, Australia

[4]Diamantina Institute, The University of Queensland & Translational Research Institute, Woolloongabba QLD, Australia

[*]Corresponding author:

Dr. Mohammad Ali Moni
School of Health and Rehabilitation Sciences,
Faculty of Health, The University of Queensland,
Brisbane, QLD 4072, Australia
E-mail: m.moni@uq.edu.au

Abstract

Stroke is a complex neurological disorder, and the risk factors and genetic biomarkers associated with stroke development are not completely understood. This study aims to identify putative causal traits and their biomarkers that influence the risk of stroke. Here the latent causal variable (LCV) method has been used to investigate the potential causal genetic relationships between large-scale genome-wide association studies (GWAS) data of 1504 complex traits from UK Biobank and stroke. Generalised Mendelian randomisation (GSMR) method has also been further used to examine causal inference. These analyses suggest 14 causal traits associated with stroke risk (|GCP|> 0.60; FDR < 0.05), including atrial fibrillation, deep venous thrombosis, gamma-glutamyl transferase, and platelet crit. Gene-based analysis has revealed shared genes, providing novel insights into the genetic biomarkers of the causal traits on stroke risk. Functional enrichment analyses of the shared genes have provided biological pathways underlying biological mechanisms to stroke risk, including “oxidative damage”, “platelet activation”, “cell aging”, and others. This study provides causal evidence of cardiovascular, metabolic, and blood clot-related traits increasing stroke risk. The identified shared gene biomarkers provide valuable insights into the shared genetic biomarkers and underlying mechanisms linking causal traits to stroke risk.

Keywords: Stroke, latent causal variable method, summary statistic for complex traits, and genome-wide association studies (GWAS).

Introduction

Stroke is the second largest cause of mortality and disability worldwide [1]. Stroke affects around 13.7 million individuals, and about 5.5 million die yearly [2]. Stroke is a neurological condition characterised by blockage of a blood vessel. For instance, blood clots can form in the brain, obstructing blood flow and causing blood vessels to rupture, resulting in haemorrhage in the brain. When the brain arteries rupture during a stroke event, the lack of oxygen causes the brain cells to die[2].

Clinical and epidemiological studies have repeatedly discovered a hereditary component to stroke vulnerability along with more typical risk factors, including smoking, hypertension, and diabetes mellitus [3]. Twins' studies and family history also provide evidence that genetic susceptibility is significant in stroke aetiology [4]. The estimated heritability was 37.9% for all ischemic strokes [5]. In genome-wide association studies (GWAS), several significant risk loci have been identified as associated with stroke [6]. For instance, cardioembolic stroke was linked to two atrial fibrillation-related loci (*PITX2* and *ZFHX3*), whereas large-vessel stroke was linked to a locus on chromosome 9p21 that was previously associated with coronary artery disease [7]. In addition, a multi-ancestry GWAS analysis revealed 35 risk loci associated with the aetiology of stroke and its subtypes [8,9]. Recently, cross-ancestry meta-analyses of GWAS summary data on stroke identified a total of 89 independent loci significantly associated with stroke and stroke subtypes [10]. Mendelian randomisation (MR) was used genetic data to understand the causal role of traits on stroke risk [11]. For example, blood pressure, atrial fibrillation, venous thrombosis, circulating lipoproteins, LDL (low-density lipoprotein) cholesterol, obesity, type 2 diabetes, insulin resistance, hyperglycaemia, education, BMI, physical activity, alcohol consumption and smoking have been associated with increased risk of stroke using MR [11,12]. Mendelian randomisation (MR) and latent causal variable (LCV) approaches examine the influence of genetic liability for a specific trait on the outcome [13,14].

Interestingly, the LCV method represents some advantages over MR methods. Firstly, in contrast to conventional MR methods, LCV can accurately differentiate between genetic correlation and full or partial genetic causality. In addition, positive LCV outcomes are more likely to represent actual causal effects. Secondly, the genetic causality proportion (*gcp-value*) can be easily estimated by quantifying the degree of causation, while MR methods, especially MR Egger can be easily confounded by genetic correlations [15]. In addition, the instrumental variables defined by the conventional MR method may exhibit putative pleiotropic effect (SNP instruments have an impact on both risk factors and diseases), which may violate the MR method assumption. The recently developed generalised summary data-based Mendelian randomisation (GSMR) approach uses more SNPs as instrumental variables and has a strong statistical power than the conventional MR approach. Furthermore, the GSMR approach used the HEIDI outlier method, which eliminates pleiotropic outliers, therefore GSMR results were not influenced by the pleiotropic effect of risk factors and disease and offered unbiased estimates of the causal influence.

A genetic correlation might be explained by pleiotropic effects, which can be vertical or horizontal [16]. When genetic variants directly influence trait A and trait B, this type of effect is called horizontal pleiotropy. On the other hand, vertical pleiotropy may be seen as a causal cascade where the influence of a genetic variant on a single trait is explained by its impact on another trait [13]. Horizontal pleiotropy could lead to high-false positive results in genetic epidemiology to assess the causality [15,17].

Due to the massive public health concerns with severe socio-economic consequences, there is a growing interest in exploring the potential causal genetic risk factors of stroke. Therefore, in the present study, we perform a hypothesis-free exploratory large-scale genetic screening for risk factors of stroke using the LCV method, which is less susceptible to confounding by horizontal pleiotropic effects [15]. Our findings support some of the causal associations

hypothesised by observational studies and shed light on the relationship between stroke, lifestyle factors, diseases, and health conditions. Finally, we performed gene-based and pathway-based analyses to discover shared gene biomarkers and underlying pathways for the causal traits and stroke.

Materials and Methods

The overall analytical approach that we employed in this study is shown in Figure 1.

Stroke data

We obtained GWAS summary statistics for stroke from a meta-analysis of GWAS data from MEGASTROKE consortium (http://megastroke.org/). This data is generated based on the fixed-effects meta-analysis of GWAS in 40,585 stroke cases and 406,111 controls of European ancestry. The details of the study have been described in [18].

UK biobank cohort datasets

We used 1504 phenotypes available at the Complex Traits Genetics Virtual Lab (CTG-VL) web-portal (https://genoma.io) [19], which is generated based on UK Biobank datasets (http://www.nealelab.is/uk-biobank). As a result, the majority of these GWAS summary results represent people of European ancestry, which eliminates potential biases due to genetic ancestry variations. This data was corrected for age, age-squared, inferred sex and genetic ancestry [19].

Latent causal variable

The latent causal variable (LCV) is a model used to assess the causal association of two genetically correlated traits. LCV considers a latent variable (L) that has a causal effect on each trait [15]. The GCP value indicates if the genetic correlation is due mainly to horizontal or

vertical pleiotropy and is calculated by determining the correlations between latent variable L with trait A and trait B respectively [15]. A GCP value of 0 indicates that the genetic correlation is mediated by horizontal pleiotropic effects, suggesting the absence of causal genetic effects. On the other hand, |GCP| = 1 denotes the presence of vertical pleiotropic effects as well as complete genetic causality. A |GCP| > 0.6 is regarded as a robust indicator of potential vertical pleiotropic effects [15]. To evaluate LCV outcomes, three things must be considered: the amount of genetic correlation, the magnitude of the GCP estimation, and the directionality of the GCP estimate. The GCP value does not represent the magnitude of prospective causal effects; rather, it represents the proportion of a genetic correlation that could be described by potential causal effects. For example, associations with a low genetic correlation (|rG| < 0.30) and a high |GCP| value near 1 indicate that all the genetic elements overlap between two traits, despite its small genetic correlation, which is likely due to vertical pleiotropic effects. Furthermore, a negative GCP value between stroke and trait A indicates that trait A may have causal genetic effects on stroke, whereas a positive GCP value between stroke and trait A indicates that stroke may have causal genetic effects on trait A.

The phenome-wide analytic framework was built by using an R script for the LCV approach that the original authors of the method made accessible on GitHub [15]. We uploaded GWAS summary statistics for stroke to perform the analysis on the CTG-VL web platform. The phenome-wide analytic pipeline was then utilised to perform LD score regression[20] and LCV analyses[15] to assess genetic correlations and potential putative causal links, respectively. Full instructions for utilising and understanding the publicly accessible phenome-wide analysis methodology can be found elsewhere [21]. The LCV was performed on all traits that showed a genetic correlation with stroke (false discovery rate, FDR<0.05). Similarly, we considered all GCP estimate values that were significant after multiple testing corrections (FDR<0.05).

Mendelian Randomization Analysis

To validate the findings of LCV analyses, we performed generalised Mendelian randomisation (GSMR) analyses. Mendelian randomisation method uses genetic variants (SNPs) as instrumental variables to identify causal relationships between exposure and outcome [22]. GSMR method uses all the top associated genome-wide significant level SNPs as an instrumental variable for the exposures to examine the causality [23]. GSMR analysis results are not influenced by the risk factors and disease. In short, GSMR used the MR framework to perform summary-based Mendelian randomisation (SMR) analysis on each individual SNP instrument. This analysis considered the variability in sampling for each SNP and the linkage disequilibrium (LD) between SNPs. It then combined the causal estimates from all SNP instruments using generalised least squares method. GSMR estimates the causal relationship between causal traits (exposure) and stroke (outcome). Additionally, GSMR method employed HEIDI outlier method (*p-value* < 0.01) to exclude the putative pleiotropic outlier SNPs and unbiasedly, estimates the causal effect of exposures on outcomes [23].

Gene-based association analysis

We used fastBAT – a fast set-based association analysis method to identify potential causal genes. The fastBAT method uses GWAS summary statistics and linkage-disequilibrium reference to summarise genetic associations with a trait of interest at the gene level [24]. We performed fastBAT analysis for stroke, atrial fibrillation, self-reported pulmonary embolism, blood clot in the lung, cardiac arrhytmias, blood clot in the leg diagnosed, medication: Gliclazide, insulin and piolitazone, deep venous thrombosis, atrial fibrillation and flutter (ICD10), deep venous thrombosis of lower extremities and pulmonary embolism, platelet crit and gamma-glutamyl transferase to identify potential causal disease associated genes. The fastBAT method found the association between 24273 genes. We then adjusted the *P-value*

using Benjamini-Hochberg method (Adjusted *p-value* > 0.05*)* for all traits to identify shared genes between causal traits and stroke.

Functional enrichment analysis of shared genes between complex traits and stroke

We performed functional enrichment analysis of shared genes of the causal traits (i.e, cardiovascular traits, blood clot traits (blood in lung, blood clot in leg, DVT, pulmonary embolism), metabolic traits (gamma glutamyl-transferase), blood trait (platelet crit)) and stroke using ConsensusPathDB bioinformatics tool [25], accessed on 20 July 2023, for identification of significant molecular pathways, gene ontology such as biological processes and molecular functions. ConcensusPathDB is the largest repository for functional enrichment analysis of genes. We performed pathway enrichment analyses of the shared genes using database for pathways such as KEGG, BioCarta, Reactome and, Wikipathways. We also performed gene ontology (GO) enrichment analysis. In the enrichment analysis, the p-value generated from the hypergeometric test was adjusted for multiple testing corrections. We considered the significant pathways and GO terms which passed the adjusted *p-value* < 0.05.

Results

Identification of causal traits associated with the risk of stroke using latent causal variable method

In order to identify potential causal traits of stroke, we have performed phenome-wide screening using latent causal variable (LCV) analysis by leveraging 1,504 complex traits GWAS datasets from UK biobank and stroke GWAS from MEGASTROKE consortium. Of the 1,504 complex traits, we identified 262 traits genetically correlated with stroke risk (FDR<0.05; Supplementary Table (S1). Among those genetically correlated traits, 14 could be

explained via robust vertical pleiotropic effects with stroke (|GCP|> 0.60; FDR <0.05), while seven showed limited partial genetic causality (|GCP|< 0.60; FDR < 5%; (Supplementary Table S2 & S3 respectively).

We observed the putative causal risk factors of stroke including, cardiovascular traits (atrial fibrillation, cardiac arrythmias), blood cell trait (platelet crit), liver enzymes (gamma-glutamyl transferase), physical health conditions such as pulmonary embolism, blood clot in the lung, blood clot in the leg, deep venous thrombosis of lower extremities. Our findings suggest that these traits may be causally associated with the risk of stroke (Table 1).

Exploring the relationships between causal traits and stroke using generalised summary-based Mendelian randomisation

To further validate the causal relationships identified by LCV method, we performed generalised summary-based Mendelian randomisation (GSMR) analysis of the causal traits. To assess causal roles, the causal traits were considered as exposure and stroke was considered as outcome in GSMR analysis. GSMR analysis showed that putative risk factors suggested by the LCV method were also consistently causal for stroke. The GSMR suggested cardiovascular traits (atrial fibrillation, cardiac arrythmias), physical health conditions (pulmonary embolism, blood clot in the lung, blood clot in the leg, deep venous thrombosis of lower extremities), blood cell traits (platelet crit) and liver enzyme (gamma-glutamyl transferase) were found causally associated with the increase the risk of stroke (Table 2). Due to small sample size, GSMR could not be performed on the traits, weight change during the worst episode of depression, and diabetes proxy trait (use of pioglitazone medication).

Gene-based association analysis of causal traits and stroke

To identify genes associated with causal traits and stroke, we performed fastBAT analysis and identified 97 genes for atrial fibrillation, 202 genes for atrial fibrillation and flutter, 85 genes

for blood clot in leg, 58 genes for blood clot in lung, 180 genes for cardiac arrhymias, 118 genes for DVT of lower extremities and pulmonary embolism, 52 genes for pulmonary embolism, 10515 gene for platelet crit, 12734 genes for gamma-glutamyl transferase, 85 genes for stroke (*adjusted p-value <0.05)* (Supplementary Table S4 to S12). Next, Venn analysis demonstrated several overlapping genes between causal traits and stroke Figure 3 and (Supplementary Table S13). We found 4 genes (*ZFHX3, HCCAT5, OBFC1, SH3PXD2A)* were shared among atrial fibrillation, cardiac arrhythmias and stroke. *FGA, FGG, ILF3, FGB, ILF3-AS1, SH2B3, PLRG1, ARHGAP1* were overlapped in blood clot in leg, deep venous thrombosis in lower extremities & pulmonary embolism and stroke. *ARMS2, ACAD10, CDKN2B-AS1, FOXF2, ZFHX3,* among other genes, were shared between DVT and stroke Figure 3 and (Supplementary Table S13).

Functional enrichment analysis

To gain a better understanding of the potential biological functions and molecular pathways that underlie shared genetic overlap between causal traits and stroke, we performed functional enrichment analysis of those shared candidate genes. Our functional enrichment analysis revealed some important pathways and gene ontology terms shared between causal traits and stroke. We found some significant pathways, such as “intrinsic and extrinsic prothrombin activation pathway”, “fibrinolysis pathway” and “invadopodia formation” pathways were shared between atrial fibrillation, cardiac arrhythmias, and stroke (Figure 4(A)). In addition, we also found “blood clotting cascade”, “platelet aggregation”, “fibrin clot formation”, “regulation of TLR by endogenous ligand” pathways were shared between traits i.e., blood clot in lung, blood clot in leg, deep venous thrombosis, pulmonary embolism, and stroke (Figure 4(B)). Furthermore, we also found “STAT5 activation downstream of FLT3ITD mutants”, “FLT3 signalling in disease”, “oxidative damage” and among other pathways were shared between platelet crit and stroke (Figure 4(C)), and “chromic myeloid leukemia”, “glioma” and

"nervous system development pathways were shared by gamma glutamyl-transferase and stroke (Figure 4(D)) (Supplementary Table S14 to S22).

The shared genes were also enriched in some gene ontology pathways, particularly involved in biological process and molecular functions. We identified some significant GO terms such as "blood coagulation, fibrin clot formation", "apoptotic signalling pathway", "defence response" were shared by pulmonary embolism and stroke. Furthermore, we also found "cell aging", "regulation of cell adhesion" and "regulation of haemostasis" were shared between deep venous thrombosis and stroke. The enriched gene ontology terms are presented in Figure 5(A-F).

Discussion

This study advances our understanding of stroke aetiology by shedding light on the causal architecture of stroke. We leveraged GWAS summary data to examine the potential causal association between 1,504 complex traits and stroke. Based on genetic evidence, we discovered 14 traits potentially causally associated with stroke. Consistent with the LCV results, GSMR analyses also provided significant evidence for the causal genetic influence of those causal traits in increasing the risk of stroke. Our findings suggest that certain cardiovascular traits, pathophysiological conditions, vascular disorders, liver enzymes (gamma glutamyl-transferase), and blood cell traits (platelet crit) may increase the risk of stroke. Furthermore, gene-based and pathway-based analyses revealed their shared genes and underlying biological pathways between causal traits and stroke, which may suggest novel molecular insights into stroke aetiology.

Our study revealed that the causal genetic effects of certain traits may influence the risk of stroke, such as atrial fibrillation (AF). Previous studies reported that AF contributes to the pathogenesis of stroke by causing stasis in the left atrium and subsequent brain embolism [26]. A study showed that about 15–38% of people who suffered an ischemic stroke had AF [27], and it also enhanced stroke risk five-fold compared with healthy people [28]. However, apart from triggering stroke, AF may be associated with other risk factors that may influence stroke. For instance, age, sex, coronary artery disease, high blood pressure, diabetes, inflammatory disorders, and cardiac embolism are all potential risk factors for both stroke and AF [29]. Our study highlighted that the genetic predisposition of AF contributes to stroke risk. Venous thromboembolism (VET), also known as deep vein thrombosis (DVT) and pulmonary embolism (PE), may have a role in circulatory collapse and short- and long-term serious effects on quality of life [30]. VTE is a complex disorder, with increasing age and obesity recognised as prevalent atherosclerotic risk factors for ischemic stroke [31]. The incidence of DVT is higher in aged people with age ≥ 65 years [32]. DVT happens when a thrombus (blood clot) forms in the deep veins, particularly in the lower extremities (legs), which leads to pain, cramping, and swelling in the lower extremities, and DVT is common among patients with stroke [32]. Our findings demonstrated that genetic variants of VTE and PE make a strong contribution to the risk of stroke. Moreover, weight change (weight gain) after depression is also a serious public health concern, and depressed individuals have a 58% chance of developing obesity [33]. Research conducted in 2021 revealed that patients who had a history of adolescent overweight and obesity as measured by BMI had a higher risk of developing an early-onset ischemic stroke [34]. Our study found the genetic variation of weight gain after depression may increase the risk of stroke. Our study highlights the genetic variants of the liver enzyme “gamma-glutamyl transferase” that were casually associated with stroke risk. Gamma-glutamyl transferase is a hepatic biomarker that is linked to liver dysfunction, physiology, and

metabolism [35]. Genetic variants associated with liver enzymes have previously been found to be genetically correlated with increased risk of cardiovascular diseases, and lipid and glucose metabolism [36]. The gamma-glutamyl transferase is secreted by the liver in response to various liver diseases[36], which could lead to increased metabolic disruption or lipid metabolism alterations ultimately contributing to stroke risk. Our study provides novel insights into the causal role of gamma-glutamyl transferase in the increased risk of stroke. Our study demonstrated that the blood cell trait i.e., platelet crit, was causally associated with stroke risk. Platelets are an essential component of blood and have an immense role in physiology. The total volume of platelets in the blood is regulated under normal physiological conditions. The total mass of platelets in the blood is called platelet crit. Platelet activation may narrow the blood vessels, obstructing the normal blood flow, which is called atherosclerosis [37]. Atherosclerosis may increase the risk of stroke by impairing the circulation of blood through the blood vessels. Consistently, our ontology and pathway analysis also showed platelet activation and platelet coagulations were enriched in stroke and platelet crit traits.

Previous studies reported that lifestyle and vascular factors may increase the risk of ischemic and intracerebral haemorrhagic stroke by 80–90% [38]. MR studies revealed that genetically predicted type 2 diabetes was associated with ischemic heart disease, stroke, and cardiovascular diseases [39]. We found the use of medication GWAS (pioglitazone, insulin, gliclazide) for type 2 diabetes were causally associated with stroke risk. In these cases, type 2 diabetes would be the risk factor for stroke. Use of medication GWAS contribute to understand the aetiology and biological mechanisms of the lifestyle related risk factors [40]. It could be one of the explanations that power of these medication use GWAS were more robust than the disease associated GWAS (e.g., type 2 diabetes).

Further, gene-based association analyses revealed shared genetic architecture between stroke and causal traits. Gene based association analysis is an alternative and complementary to single

marker (SNPs), providing an opportunity to discover gene level associations for respective traits [41]. Therefore, in this study, we performed the gene-based associations analysis for stroke and causal traits. Our findings suggest many expected overlapped genes between stroke and causal traits. For instance, *ILF3* was found overlapping among stroke and deep venous thrombosis, blood clot in lung and leg, platelet crit, and pulmonary embolisms. *ILF3* is a protein coding gene, and its over expression was reported to be associated with the calcification of human aortic vascular smooth muscle cells and induced the formation of atherosclerotic plaques [42]. Likewise, we observed *ARMS2* was overlapped between stroke and deep venous thrombosis, platelet crit. *ARMS2* encodes the mitochondrial protein age-related maculopathy susceptibility protein 2 which was reported to be associated with oxidative damage, apoptosis, and aging process [43]. Furthermore, *OBFC1*, *SH3PXD2A*, *HCCAT5* were overlapped between stroke and atrial fibrillation, cardiac arrhythmias. Our previous study demonstrated *SH3PXD2A*, *OBFC1*, among others were causal risk biomarkers of stroke [44].

Our study demonstrated some important biological insights and molecular pathways enriched by the shared genes, providing novel insights into aetiology and underlying biological mechanisms implicated in the causal traits and stroke, including “oxidative damage”, “mitochondrial fatty acid beta-oxidation”, “platelet activation and aggregation”, “regulation of hemostasis”, “cell aging”, “fibrinolysis pathway”. The interruption of blood circulation in the brain is the hallmark of stroke, and have negatively impacted cell homeostasis, resulting in inflammation, oxidative damage, ionic imbalance, excitotoxicity, and ultimately induced neuronal cell death [45]. The production of high volume of intracellular molecules particularly, Ca2+, Na+, and adenosine diphosphate (ADP) causes oxidative stress that has a negative impact on mitochondrial function, as mitochondria are the fundamental energy generation stations of cells, ischemic stroke has a significant impact on them [46]. When necessary resources, particularly glucose, are inaccessible to brain cells, the mitochondria then produce

energy by consuming 15% more oxygen, produces more superoxides and less ATP via fatty acid oxidation (β-oxidation) process [47]. Our pathway-based enrichment analysis suggested the causal influence of mitochondrial fatty acid beta oxidation pathway in the progression of stroke. By inhibiting the mitochondrial fatty acid beta oxidation pathway could be a potential target for the treatment of stroke. Prior study has reported that oxidative stress is associated with aging process that influence the risk of stroke [48]. Another pathway, we found in our study is platelet activation and aggregation, platelet is the primary components of thrombi and have potential role in the formation of thrombus in both healthy and disease blood vessels. In hemostasis, during vascular injury, circulating platelet become struck in the injured site and stimulate the activation and aggression of hemostatic thrombi, which stop further bleeding. However, in thrombosis, abnormal platelet activation results in uncontrollable thrombus formation, which in turn may embolise or block the blood vessel, potentially resulting downstream ischemic events [49]. We also found fibrinolysis pathway which is involved in blood coagulation process and contributes to the progression of arterial and venous thrombosis [50] and ultimately induced the risk of stroke.

We acknowledged several limitations in this present study. Firstly, the GWAS summary statistics used in this study were obtained from the MEGASTROKE and UK Biobank cohort, where the participants are all from European ancestry. Thus, our findings should not be consistent with other ancestries unless they are verified using data from diverse populations. Secondly, our studies included 1504 traits, but other causal relationships may still exist with other traits. Thirdly, by definition, the LCV method evaluates the causal relationship between two traits and assumes no bidirectional causality [15]. Consequently, it is still debatable whether the LCV method would be able to determine the bidirectional causality between the traits. This assumption must be systematically evaluated in the future. Null findings in this study might be explained in a way that there is no causal association between the phenotypes

or there may exist bidirectional causality. In the current study, caution must be taken when interpreting null findings.

In summary, our study provides evidence for potential causal genetic relationships between stroke and other complex traits. Our findings suggest that the role of several cardiovascular traits (such as atrial fibrillation, cardiac arrhythmias), blood cell traits (such as platelet crits), and blood clot-related traits, and liver enzyme traits increase the risk of stroke. Our study also revealed the shared genetic architecture, genes, pathways implicated in stroke and causal traits. Overall, our findings support the evidence of previous epidemiological studies and provide novel insights into the causal genetic architecture, shared genes and pathways of complex traits and stroke, which in turn could be used as testable hypotheses to improve the development of future studies, treatments, and preventative strategies.

Data Availability Statement

All data here analysed is publicly available on MEGASTROKE consortium and the UK biobank summary data is available on CTG-VL.

References

[1] Organization, W.H. (2010). *World health statistics 2010*, World Health Organization.

[2] Kuriakose, D. and Xiao, Z. (2020). Pathophysiology and treatment of stroke: present status and future perspectives. *International journal of molecular sciences*.

[3] Boehme, A.K. et al. (2017). Stroke risk factors, genetics, and prevention. *Circulation research*.

[4] Dichgans, M. (2007). Genetics of ischaemic stroke. *The Lancet Neurology*.

[5] Bevan, S. et al. (2012). Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke*.

[6] Gretarsdottir, S. et al. (2008). Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Annals of neurology*.

[7] Traylor, M. et al. (2012). Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *The Lancet Neurology*.

[8] Malik, R. et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics*.

[9] Malik, R. et al. (2018). Genome-wide meta-analysis identifies 3 novel loci associated with stroke. *Annals of neurology*.

[10] Mishra, A. et al. (2022). Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature*.

[11] Georgakis, M.K. and Gill, D. (2021). Mendelian randomization studies in stroke: exploration of risk factors and drug targets with human genetic data. *Stroke*.

[12] Liu, J. et al. (2018). Causal impact of type 2 diabetes mellitus on cerebral small vessel disease: a Mendelian randomization analysis. *Stroke*.

[13] García-Marín, L.M. et al. (2021). Phenome-wide analysis highlights putative causal relationships between self-reported migraine and other complex traits. *The Journal of Headache and Pain*.

[14] Siewert, K.M. et al. (2020). Cross-trait analyses with migraine reveal widespread pleiotropy and suggest a vascular component to migraine headache. *International journal of epidemiology*.

[15] O'Connor, L.J. and Price, A.L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature genetics*.

[16] Van Rheenen, W. et al. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*.

[17] Koellinger, P.D. and De Vlaming, R. (2019). Mendelian randomization: the challenge of unobserved environmental confounds. *International journal of epidemiology*.

[18] Malik, R. et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics*.

[19] Cuellar-Partida, G. et al. (2019). Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses. *BioRxiv*.

[20] Bulik-Sullivan, B.K. et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*.

[21] Haworth, S. et al. (2021). Assessment and visualization of phenome-wide causal relationships using genetic data: an application to dental caries and periodontitis. *European Journal of Human Genetics*.

[22] Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*.

[23] Zhu, Z. et al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications*.

[24] Bakshi, A. et al. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific reports*.

[25] Kamburov, A. et al. (2013). The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research*.

[26] Kamel, H. et al. (2016). Atrial fibrillation and mechanisms of stroke: time for a new model. *Stroke*.

[27] Santos, J.V. et al. (2017). Atrial fibrillation as an ischemic stroke clinical and economic burden modifier: a 15-year nationwide study. *Value in Health*.

[28] Son, M.K. et al. (2017). Risk of ischemic stroke after atrial fibrillation diagnosis: A national sample cohort. *PloS one*.

[29] Kamel, H. et al. (2016). Atrial fibrillation and mechanisms of stroke: time for a new model. *Stroke*.

[30] Heit, J.A. (2005). Venous thromboembolism: disease burden, outcomes and risk factors. *Journal of Thrombosis and Haemostasis.*

[31] Glynn, R.J. and Rosner, B. (2005). Comparison of risk factors for the competing risks of coronary heart disease, stroke, and venous thromboembolism. *American journal of epidemiology*.

[32] Liu, Z. et al. (2021). Incidence and Risk Factors of Lower-Extremity Deep Vein Thrombosis After Thrombolysis Among Patients with Acute Ischemic Stroke. *Pharmacogenomics and Personalized Medicine*.

[33] Blasco, B.V. et al. (2020). Obesity and depression: Its prevalence and influence as a prognostic factor: A systematic review. *Psychiatry investigation*.

[34] Bardugo, A. et al. (2021). Body mass index in 1.9 million adolescents and stroke in young adulthood. *Stroke*.

[35] Pazoki, R. et al. (2021). Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nature communications*.

[36] Pazoki, R. et al. (2021). Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nature communications*.

[37] Mohamed, A.-A.B. et al. (2019). The mean platelet volume and plateletcrit as predictors of short-term outcome of acute ischemic stroke. *The Egyptian journal of neurology, psychiatry and neurosurgery*.

[38] Hauer, A.J. et al. (2017). Age-Specific vascular risk factor profiles according to stroke subtype. *Journal of the American Heart Association*.

[39] Huang, M. et al. (2022). Causal Association of Type 2 diabetes mellitus and glycemic traits with cardiovascular diseases and lipid traits: a Mendelian randomization study. *Frontiers in Endocrinology*.

[40] Wu, Y. et al. (2019). Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nature communications*.

[41] Bakshi, A. et al. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific reports*.

[42] Xie, F. et al. (2021). ILF3 is responsible for hyperlipidemia-induced arteriosclerotic calcification by mediating BMP2 and STAT1 transcription. *Journal of Molecular and Cellular Cardiology*.

[43] Zhang, J. et al. (2021). Meta-analysis of the pharmacogenetics of ARMS2 A69S polymorphism and the response to advanced age-related macular degeneration. *Ophthalmic research*.

[44] Islam, T. et al. (2023). Integration of Mendelian randomisation and systems biology models to identify novel blood-based biomarkers for stroke. *Journal of Biomedical Informatics*.

[45] Doyle, K.P. et al. (2008). Mechanisms of ischemic brain damage. *Neuropharmacology*.

[46] Mavroudakis, L. and Lanekoff, I. (2023). Ischemic Stroke Causes Disruptions in the Carnitine Shuttle System. *Metabolites*.

[47] Schönfeld, P. and Reiser, G. (2013). Why does brain metabolism not favor burning of fatty acids to provide energy?-Reflections on disadvantages of the use of free fatty acids as fuel for brain. *Journal of Cerebral Blood Flow & Metabolism*.

[48] Islam, T. et al. (2023). Integration of Mendelian randomisation and systems biology models to identify novel blood-based biomarkers for stroke. *Journal of Biomedical Informatics*.

[49] Rana, A. et al. (2019). Shear-dependent platelet aggregation: mechanisms and therapeutic opportunities. *Frontiers in cardiovascular medicine*.

[50] Grover, S.P. and Mackman, N. (2019). Intrinsic pathway of coagulation and thrombosis: Insights from animal models. *Arteriosclerosis, thrombosis, and vascular biology*.

Acknowledgments

We thank the MEGASTROKE project for making stroke GWAS summary statistics data publicly available. The author list of MEGASTROKE is available in Supplementary Table 23. T.I. is supported by a Research Training scholarship from The University of Queensland, Australia.

Author contributions: All authors contributed to the design and conception of the study. T.I collected datasets and performed the analyses supervised by M.A.M. L.M.G provided important insights into the interpretation of the results. TI wrote the manuscript and all authors commented on the manuscript. All authors read and approved the final manuscript. T.I is a recipient of an RTP scholarship from the University of Queensland, Australia.

Competing Interests: "Authors declare no competing of interest".

Figure legends:

Figure 1. The schematic diagram illustrates the genetic relationships between 1504 complex traits and stroke. The genetic correlation was consistent across the phenotypes, demonstrating that genetic variants specific to each of the 1504 complex traits had consistent positive or negative correlations with the genetic variants associated with stroke. Gene-based and pathway-based analyses revealed shared genetic architecture and molecular mechanisms shared between causal traits and stroke. This figure was generated from Biorender.com.

Figure 2. Share genes between causal traits and stroke. Right side represents causal traits and stroke, and left side represents shared genes between causal traits and stroke.

Figure 3. Common pathways shared between complex traits and stroke. (A) pathways shared between atrial fibrillation, atrial fibrillation and flutter, and cardiac arrhythmias and stroke; (B) pathways shared between blood clot in lung, blood clot in leg, deep venous thrombosis, pulmonary embolisms, and stroke; (C) pathways shared between platelet crit and stroke; (D) pathways shared between gamma glutamyl-transferase and stroke.

Figure 4. Common gene ontology (GO) terms shared between complex traits and stroke. (A) shared GO term between pulmonary embolisms and stroke; (B) common GO term between deep venous thrombosis and stroke; (C) shared GO between blood clot in lung and stroke; (D) shared GO term between blood clot in leg and stroke; (E) shared GO between platelet crit and stroke; (F) shared GO term between gamma-glutamyl transferase and stroke.

Table legends:

Table 1. Traits with potential causal association with stroke. This table represents the phenotypes with FDR<.05 and robust genetic causal proportions (|GCP| > 0.60) with stroke.

Table 2. Traits with potential causal association with stroke using generalised Mendelian randomisation analysis.

Table 1. Traits with potential causal association with stroke. This table represents the phenotypes with FDR<.05 and robust genetic causal proportions (|GCP| > 0.60) with stroke.

Trait	GCP	GCPse	GCPpval	rG	rGse	rGpval
Self-reported atrial fibrillation	-0.97	0.03	8.97×10^{-202}	0.58	0.18	9.51×10^{-4}
Self-reported pulmonary embolism	-0.79	0.15	1.37×10^{-7}	0.45	0.13	3.20×10^{-4}
Blood clot in the lung diagnosed by a doctor	-0.75	0.17	2.46×10^{-5}	0.44	0.13	3.41×10^{-4}
Cardiac arrhytmias	-0.72	0.19	2.99×10^{-4}	0.44	0.08	1.76×10^{-7}
Blood clot in the leg diagnosed by a doctor	-0.74	0.18	5.25×10^{-5}	0.43	0.11	1.96×10^{-4}
Medication: Gliclazide	-0.66	0.23	3.75×10^{-3}	0.42	0.09	2.33×10^{-5}
Self-reported deep venous thrombosis	-0.74	0.18	4.23×10^{-5}	0.40	0.11	2.60×10^{-4}
Atrial fibrillation and flutter (ICD10)	-0.80	0.14	2.10×10^{-8}	0.40	0.08	1.91×10^{-6}
Medication: Insulin	-0.84	0.12	1.43×10^{-11}	0.38	0.14	6.43×10^{-3}
Medication: Pioglitazone	-0.74	0.17	1.72×10^{-5}	0.33	0.12	8.45×10^{-3}
Weight change during worst episode of depression: Gained weight	-0.64	0.23	4.03×10^{-3}	0.32	0.12	1.01×10^{-2}
Deep venous thrombosis of lower extremities and pulmonary embolism	-0.78	0.14	1.28×10^{-8}	0.24	0.08	6.00×10^{-3}
Gamma glutamyltransferase (female)	-0.69	0.15	4.13×10^{-6}	0.16	0.04	3.23×10^{-4}
Platelet crit	-0.67	0.20	7.61×10^{-4}	0.14	0.04	1.32×10^{-3}
Gamma glutamyltransferase (male)	-0.64	0.19	7.61×10^{-4}	0.14	0.05	6.60×10^{-3}

This table represents the traits with significant false discovery rate (FDR < .05); GCP: genetic causal proportion; GCP se: genetic causal proportion standard deviation; GCP pval: genetic causal proportion unadjusted p-value; rG: genetic correlation; rG se: genetic correlation standard deviation; rG pval: genetic correlation unadjusted p-value.

Table 2. Traits with potential causal association with stroke using Generalised Mendelian Randomisation analysis.

Exposure	**Outcome**	**Effect**	**SE**	***P***	**HEIDI test *P***	**SNPs #**
Self-reported atrial fibrillation	Stroke	13.80	1.37	1.2×10^{-23}	0.28	13
Self-reported pulmonary embolism		20.06	2.75	3.2×10^{-13}	0.31	4
Blood clot in the lung diagnosed by a doctor		17.20	2.69	1.6×10^{-10}	0.12	3
Cardiac arrhytmias		7.00	0.65	1.4×10^{-26}	0.06	24
Blood clot in the leg diagnosed by a doctor		4.72	1.13	2.9×10^{-5}	0.1	9
Medication: Gliclazide		5.23	2.22	0.01	0.08	7
Self-reported deep venous thrombosis		4.75	1.14	3.2×10^{-5}	0.09	9
Atrial fibrillation and flutter (ICD10)		7.40	0.66	6.1×10^{-29}	0.12	29
Medication: Insulin		1.89	0.89	0.02	0.2	10
Deep venous thrombosis of lower extremities and pulmonary embolism		15.21	2.18	3.49×10^{-12}	0.26	5
Gamma glutamyltransferase		0.07	0.01	8.23×10^{-7}	0.05	654
Platelet crit		1.72	0.37	4.71×10^{-6}	0.06	616

This table represents the traits with significant false discovery rate (FDR < .05); SE: standard error; *P*: *p*-value; HEIDI test *P*: heterogeneity in dependent instruments test. SNPs#: number of single nucleotide polymorphism

Figure 1.

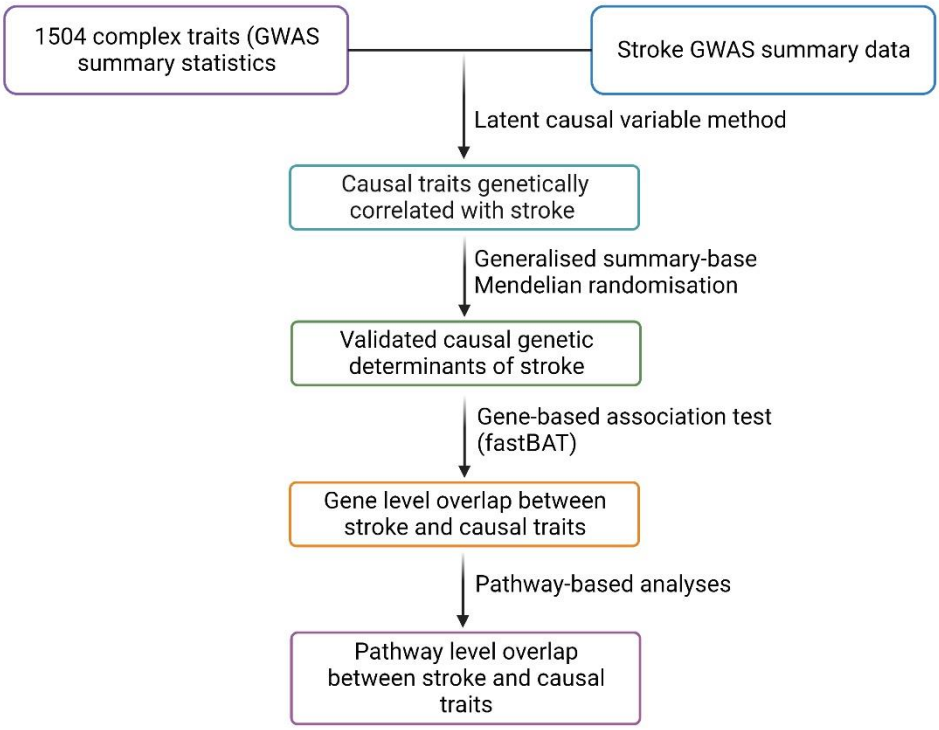

Figure 2.

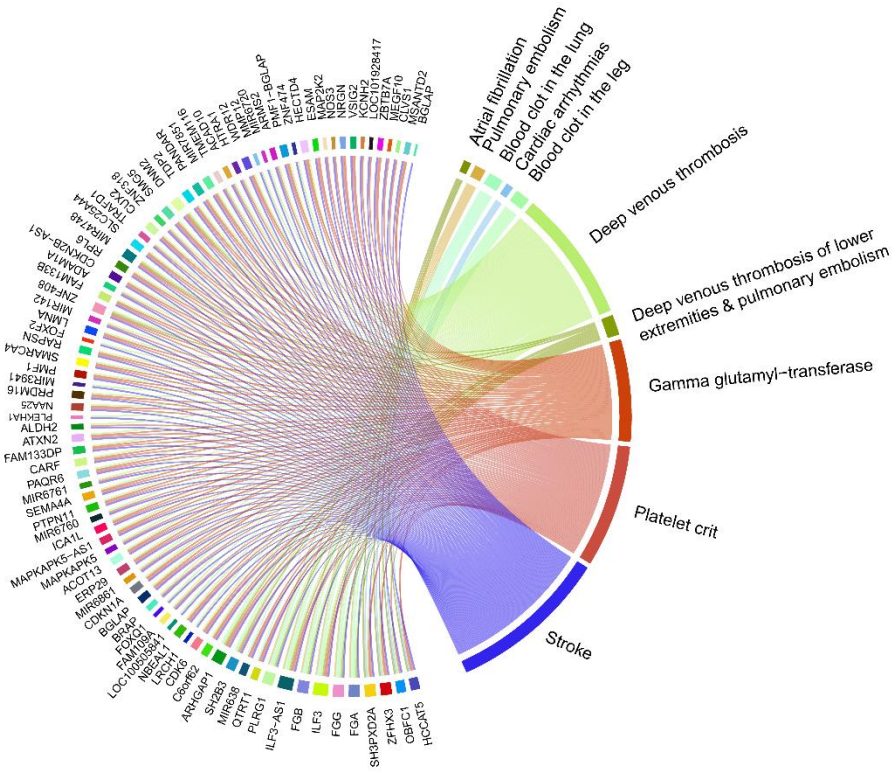

Figure 3.

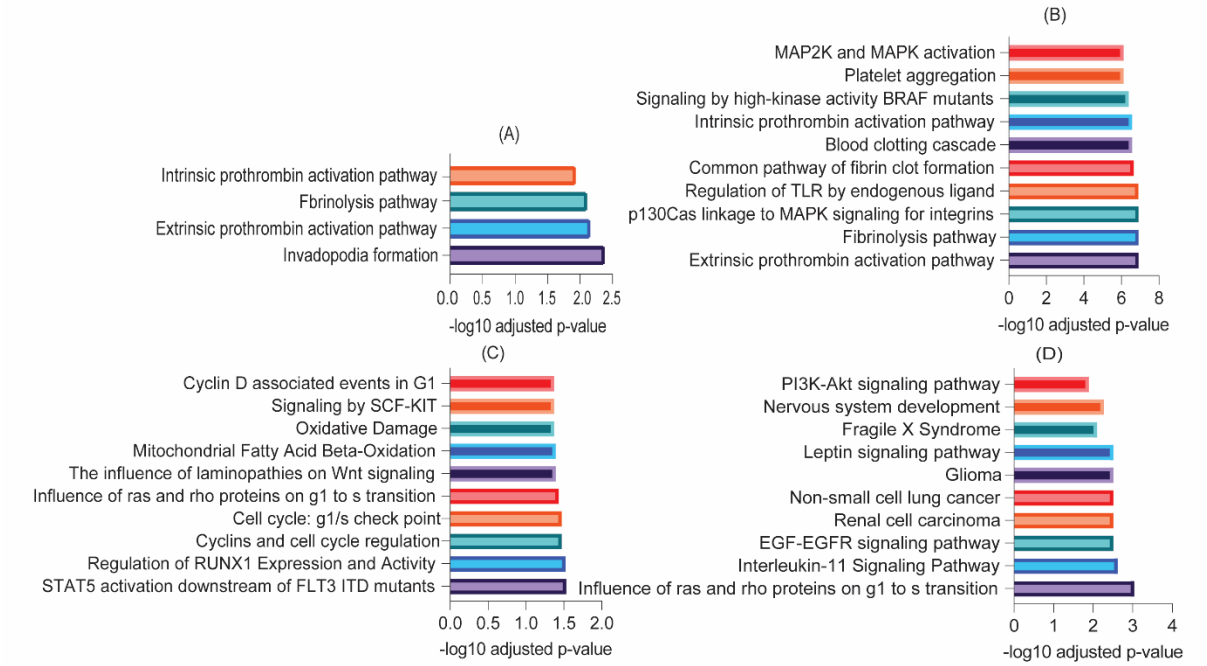

Figure 4.