Shared Genetic Contributions to Atrial Fibrillation and Ischemic Stroke Risk

Sara L. Pulit[1,2,3], Lu-Chen Weng[3,4], Patrick F McArdle[5], Ludovic Trinquart[6,7], Seung Hoan Choi[7], Braxton D. Mitchell[5,8], Jonathan Rosand[3,9,10], Paul I W de Bakker[1,11], Emelia J Benjamin[6,12,13], Patrick T Ellinor[3,4,14], Steven J Kittner[15,16], Steven A Lubitz*[3,4,14], Christopher D Anderson*[3,9,10], on behalf of the Atrial Fibrillation Genetics Consortium and the International Stroke Genetics Consortium.

1. Department of Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
2. Li Ka Shing Centre for Health Information and Discovery, The Big Data Institute, University of Oxford, Oxford, UK
3. Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
4. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
5. Department of Medicine, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, Maryland, United States of America
6. National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA, USA
7. Department of Biostatistics, Boston University School of Public Health, Boston, MA
8. Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA
9. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
10. J.P. Kistler Stroke Research Center, Massachusetts General Hospital, Boston, MA, USA
11. Computational Genomics, Vertex Pharmaceuticals, Boston, MA, USA (current address)
12. Cardiology Preventive Medicine Sections, Evans Department of Medicine, Boston University School of Medicine, Boston, MA, USA
13. Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
14. Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA
15. Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA
16. Department of Neurology, Veterans Affairs Medical Center, Baltimore, MD, USA

* These authors contributed equally to supervision of this study

Corresponding authors

Dr. Steven A. Lubitz
Cardiovascular Research Center and Cardiac Arrhythmia Service
Massachusetts General Hospital
55 Fruit Street
GRB 109
Boston, MA, 02114, USA
slubitz@mgh.harvard.edu

Dr. Christopher D. Anderson
Center for Genomic Medicine
Massachusetts General Hospital
185 Cambridge Street, CPZN-6818
Boston, MA, 02114, USA
cdanderson@mgh.harvard.edu

Abstract

Atrial fibrillation is a prevalent arrhythmia associated with a five-fold increased risk of ischemic stroke, and specifically the cardioembolic stroke subtype. Genome-wide association studies of these traits have yielded overlapping risk loci, but genome-wide investigation of genetic susceptibility shared between stroke and atrial fibrillation is lacking. Comparing the genetic architectures of the two diseases could inform whether cardioembolic strokes are driven by inherited atrial fibrillation susceptibility, and may help elucidate ischemic stroke mechanisms. Here, we analyze genome-wide genotyping data and estimate SNP-based heritability in atrial fibrillation and cardioembolic stroke to be nearly identical (20.0% and 19.5%, respectively). Further, we find that the traits are genetically correlated (r=0.77 for SNPs with $p<4.4x10^{-4}$ in a previous atrial fibrillation meta-analysis). Clinical studies are warranted to assess whether genetic susceptibility to atrial fibrillation can be leveraged to improve the diagnosis and care of ischemic stroke patients.

Introduction

Atrial fibrillation affects nearly 34 million individuals worldwide[1] and is associated with a five-fold increased risk of ischemic stroke[2], a leading cause of death and long-term disability globally[3,4]. Strokes caused by atrial fibrillation are classified as cardioembolic in nature, as atrial fibrillation promotes the formation of blood clots in the heart which can embolize distally to the brain. Cardioembolic stroke is often, but not always, caused by atrial fibrillation. However, whether cardioembolic stroke results from genetic or non-genetic susceptibility to atrial fibrillation remains an open question.

Recent genome-wide association studies (GWAS) of atrial fibrillation[5] and ischemic stroke indicate that both traits are common complex disorders with polygenic architectures[6,7]. The top loci for cardioembolic stroke, on chromosome 4q25 upstream of *PITX2* and on 16q22 near *ZFHX3*, are both leading risk loci for atrial fibrillation[8–10]. Although the leading genetic loci for atrial fibrillation and cardioembolic stroke overlap, the genetic susceptibility to each is likely to involve the aggregate contributions of hundreds or thousands of genetic loci, consistent with other polygenic traits[11]. Understanding the degree to which the genetic architectures of these diseases overlap may elucidate the extent to which genetic risk factors for atrial fibrillation contribute to cardioembolic stroke risk.

We therefore leveraged genome-wide common variation to address two specific questions. First, we assessed the degree of overlap between the genetic architectures of atrial fibrillation in the general population and atrial fibrillation in patients with stroke. Second, we compared the genetic architectures of atrial fibrillation and cardioembolic stroke to ascertain the extent to which heritable risk of cardioembolic stroke is explained by genetic risk factors for atrial fibrillation. Using a sample comprised of 13,390 ischemic stroke cases and 28,026 referents with imputed genotyping data from the NINDS-Stroke Genetics Network (SiGN)[12], we have established that additive genetic variation accounts for ~20% of the variance in risk for both atrial fibrillation and cardioembolic stroke. Further, we find

that the genetic architectures of the two traits are highly correlated (r = 0.77 across ~1.9 million SNPs shown previously to associate to atrial fibrillation), and a genetic risk score constructed from SNPs predisposing to atrial fibrillation explained approximately one-fifth of the genetic variance in cardioembolic stroke risk.

Results

We began by testing our ability to rediscover known atrial fibrillation genetic associations in the SiGN dataset, assembled to study the genetics of ischemic stroke. Of the 13,390 European- and African-descent stroke cases in SiGN, we identified 3,190 cases with atrial fibrillation or paroxysmal atrial fibrillation, as well as other diagnoses suggestive of underlying atrial fibrillation, including atrial flutter[13], left atrial thrombus, and sick sinus syndrome[14] (**Table 1** and **Supplementary Table 1**). SiGN also includes 28,026 European- and African-descent controls, the majority of which are publicly-available referent sets with minimal medical history information.

We first performed a genome-wide association study (GWAS) of the 3,190 atrial fibrillation cases and 28,026 referents. The results were highly concordant with the latest GWAS in atrial fibrillation performed by the Atrial Fibrillation Genetics (AFGen) Consortium (**Supplementary Figure 1**). Specifically, the GWAS in SiGN identified SNPs at the *PITX2* and *ZFHX3* loci[8,10] (index SNPs rs6843082 and rs2106261, respectively) that were associated with atrial fibrillation beyond a genome-wide significance threshold ($p < 5 \times 10^{-8}$) and with similar effect sizes compared to those reported in the AFGen GWAS[5] (**Supplementary Table 2**). We additionally looked up 26 variants most recently reported by AFGen as being associated to atrial fibrillation[5]; of the 25 available in the SiGN data, 24 showed a consistent direction of effect (binomial $p = 7.7 \times 10^{-7}$), and 14 showed nominal association signal ($p < 0.05$; **Supplementary Table 2**). The only SNP without a consistent direction of effect in both analyses had originally been identified through an eQTL analysis[5].

Next, we evaluated heritability within cardioembolic stroke and atrial fibrillation, with the specific goal of evaluating the proportion of genetic variation in cardioembolic stroke that can be explained by genetic susceptibility to atrial fibrillation.

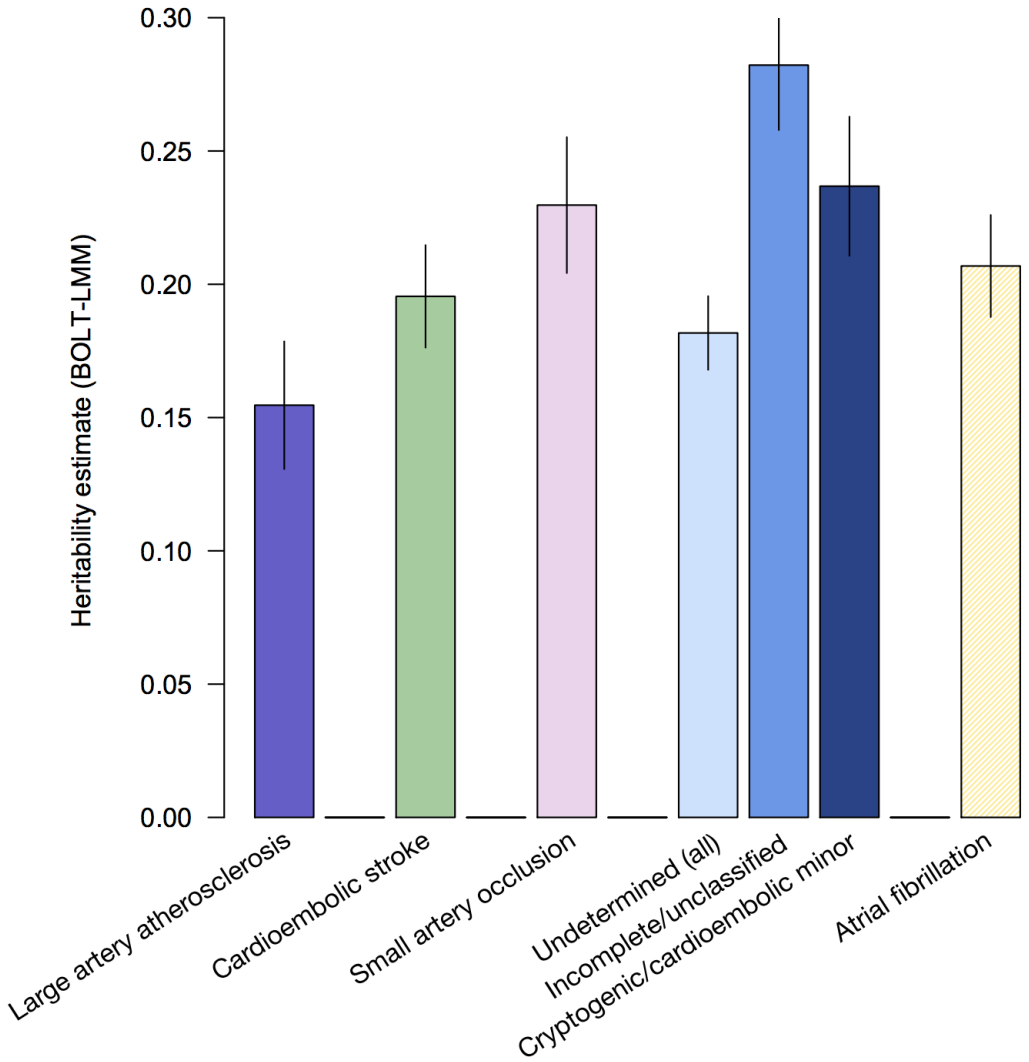
Table 1 | Atrial fibrillation and stroke cases in SiGN. Of the 13,390 stroke cases available in the SiGN dataset, a total of 3,190 cases had atrial fibrillation or other suggestive diagnoses. While the majority of these cases were subtyped as having a cardioembolic (CE) stroke, a fraction were distributed among the other stroke subtypes. Samples can appear more than once per row, but totals represent the number of unique atrial fibrillation samples in each stroke subtype. There are no subjects with atrial fibrillation or equivalent subtyped as "cryptogenic/CE minor" because such a diagnosis would remove them from this category.

	Phenotype	Total	*Ischemic stroke subtype*				
			Cardioembolic	Large artery atherosclerosis	Small artery occlusion	Undetermined: incomplete/ unclassified	Undetermined: cryptogenic/ CE minor
Atrial fibrillation diagnosis	Atrial fibrillation	1,751	1,495	63	32	151	0
	Paroxysmal atrial fibrillation	1,315	1,088	52	23	138	0
	Left atrial thrombus	48	37	3	3	4	0
	Sick sinus syndrome	79	65	5	3	4	0
	Atrial Flutter	106	90	4	2	10	0
	Total	3,190	2,684	123	61	298	0

As a basis for comparison, we also calculated heritability in the two additional primary stroke subtypes[15] (large artery atherosclerosis and small artery occlusion), as well as in 'undetermined' stroke, a highly heterogeneous subset of cases[16,17] that cannot be classified by standard subtyping systems[15,18]. This latter category could potentially contain misclassified cardioembolic strokes caused by occult atrial fibrillation[19]. We analyzed common variants (minor allele frequency > 1%) in SiGN to evaluate heritability of atrial fibrillation, followed by the major stroke subtypes (cardioembolic, large artery atherosclerosis, and small artery occlusion), and finally in strokes of undetermined cause (**Materials and Methods**). We performed restricted maximum likelihood analysis implemented in BOLT-LMM[20] (i.e., BOLT-REML) to calculate genetic heritability using a high-quality set of linkage disequilibrium (LD)-pruned imputed variants (imputation info score > 0.8, minor allele frequency >1%, LD r^2 threshold = 0.2) converted to best-guess genotypes (**Materials and Methods**). We additionally corrected the REML analyses for sex and the top ten principal components and converted the observed-scale heritability estimates produced by BOLT-REML to the liability scale (**Supplementary Information**). Using the 3,190 atrial fibrillation cases and 28,026 referents available in SiGN, we estimated heritability of atrial fibrillation at 20.0% (**Figure 1**), consistent with previous estimates in larger sample sizes[21,22].

We next evaluated SNP-based heritability in the stroke subtypes. We focus these and all subsequent analyses on phenotypes resulting from the CCS causative (CCSc) subtyping algorithm, one of the three stroke subtyping systems available in SiGN[23]; heritability results from all three subtyping systems are provided in **Supplementary Table 3**. Our analyses estimate cardioembolic stroke heritability at 19.5% (**Figure 1**), a heritability highly consistent with that of atrial fibrillation. We found large artery stroke and small artery occlusion to have similarly modest heritability (15.5% and 23.0%, respectively; **Figure 1** and **Supplementary Table 3**), consistent with the hypothesis that stroke risk is modulated partially by biological factors, but with a large contribution from non-genetic variables as well.

Figure 1 | Estimated heritability of ischemic stroke subtypes and atrial fibrillation. Using all available stroke cases in SiGN, we estimated SNP-based heritability of the ischemic stroke subtypes (as sub-typed by the CCS Causative subtyping system) and atrial fibrillation (using the subset of 3,190 cases with atrial fibrillation) using BOLT-LMM and a genetic relationship matrix of high-quality SNPs converted to best-guess genotypes (imputation quality > 0.8, minor allele frequency > 0.01, and pruned at a linkage disequilibrium threshold of 0.2). We assumed a trait prevalence of 1% for all phenotypes. We found heritability estimates in cardioembolic stroke (green) and atrial fibrillation (yellow) to be approximately similar.



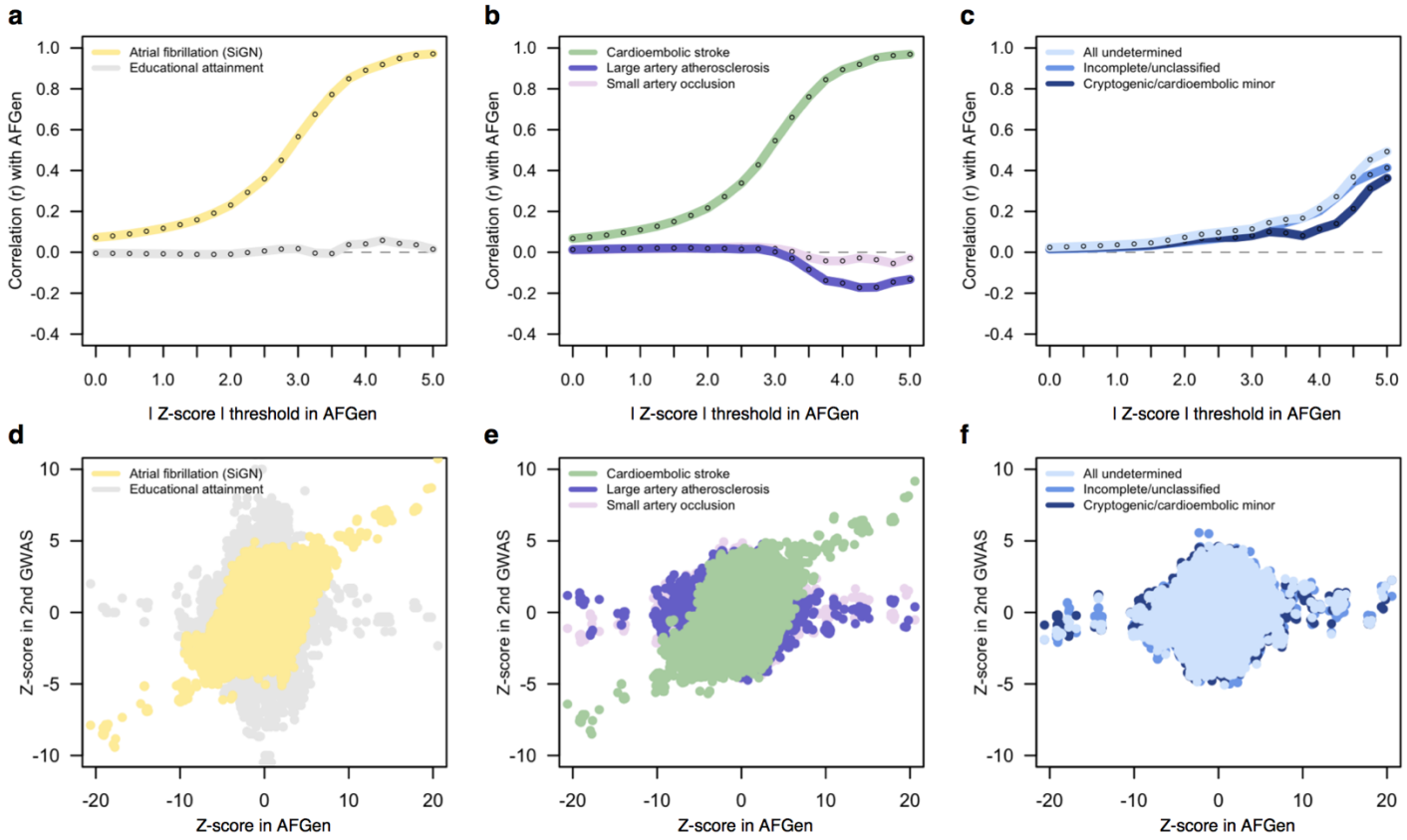
Despite their presumed etiologic and phenotypic heterogeneity, the undetermined categories of stroke showed heritability measures comparable to those of the major stroke subtypes (**Figure 1**, **Supplementary Figure 2** and **Supplementary Table 3**). Specifically, heritability in all undetermined strokes (N_{cases} = 4,574) was 18.2%. The mutually exclusive subsets of all undetermined stroke, categorized as incomplete/unclassified strokes (N_{cases} = 2,280) and cryptogenic/cardioembolic minor strokes (N_{cases} = 2,294), were also moderately heritable (28.2% and 23.7%, respectively; **Figure 1**). We repeated all of the heritability calculations in GCTA[24] to evaluate the stability of the estimates across methods (**Supplementary Information** and **Supplementary Table 3**); estimates from GCTA were primarily similar to those from BOLT (**Supplementary Figure 3**). These heritability estimates suggest that, despite the fact that recent GWAS of undetermined stroke have not revealed any robustly-associated loci[12,25], the phenotype(s) captured by this subtype are heritable, and likely represent either improperly-classified cases from the primary subtypes or novel stroke phenotypes awaiting further delineation.

Given the known phenotypic overlap of cardioembolic stroke and atrial fibrillation (89.5% of cardioembolic stroke cases in SiGN also have atrial fibrillation, **Supplementary Table 1**) as well as the similar heritable components of the two traits, we next aimed to evaluate how genetic susceptibility to atrial fibrillation might contribute to cardioembolic stroke. To measure correlation between the genetic architectures of atrial fibrillation and stroke, we first performed a series of GWAS in the SiGN data for each of the stroke subtypes as well as all stroke, using BOLT-LMM[20] (**Materials and Methods**). We then calculated the z-score (beta/standard error) of each SNP in each stroke subtype, all stroke, and from our analysis of atrial fibrillation described above. Next, we used summary-level results available from the most recent GWAS of atrial fibrillation[5] from AFGen and calculated the z-score for each SNP in that independent dataset. Measuring Pearson's correlation (r) between AFGen z-scores and z-scores from our own GWAS of atrial fibrillation, we found only a modest correlation between atrial fibrillation as measured in the two GWAS (r = 0.07 across ~7.8M SNPs, **Figure 2a**). However, when we iteratively subsetted the AFGen GWAS results by the (absolute values of)

z-scores of the SNPs, we found that correlation with our own atrial fibrillation GWAS in SiGN improved as the z-score threshold became more stringent. For example, for those ~4.5M SNPs with |z| > 1 in AFGen, correlation with the SiGN atrial fibrillation GWAS was 0.12; for those ~1.9M SNPs with |z| > 3.5 in AFGen, correlation with the SiGN atrial fibrillation GWAS rose to 0.77 (**Figure 2a** and **Supplementary Table 4**). Removing ±2Mb around the *PITX2* and *ZFHX3* loci only modestly impacted the correlation between AFGen and atrial fibrillation in SiGN (r = 0.63 for SNPs with |z| > 3.5; **Supplementary Figure 4** and **Supplementary Table 4**). As a null comparator, we performed correlations between the AFGen results with z-scores derived from the latest GWAS of educational attainment[26] and found that correlation remained at approximately zero regardless of the z-score threshold used (**Figure 2a** and **Supplementary Table 4**).

To investigate the overlap of atrial fibrillation with stroke and its subtypes, we repeated these same correlation calculations, comparing AFGen results to GWAS results from each of the primary stroke subtypes (cardioembolic stroke, large artery atherosclerosis, and small artery occlusion) and the undetermined subtypes (**Figure 2**). The correlation between cardioembolic stroke and AFGen was highly similar to that of the results with atrial fibrillation in SiGN (r = 0.76 for AFGen SNPs with |z| > 3.5, likely due to high concordance between the atrial fibrillation and cardioembolic stroke phenotypes in SiGN; **Figure 2** and **Supplementary Figure 5**) whereas correlation between AFGen and either large artery atherosclerosis or small artery occlusion was essentially zero (**Figure 2**). Notably, the correlation between atrial fibrillation and the undetermined stroke subtypes increased steadily as we partitioned the AFGen data by z-score (all undetermined vs. AFGen r = 0.04 for AFGen SNPs with |z| > 1 and r = 0.16 for AFGen SNPs with |z| > 3.5; **Figure 2** and **Supplementary Table 4**), indicating that genome-wide, there is residual genetic correlation between atrial fibrillation and the undetermined stroke categories, some of which could represent causal atrial fibrillation stroke mechanisms in that subgroup.

Figure 2 | Genetic correlation between atrial fibrillation in the Atrial Fibrillation Genetics (AFGen) Consortium meta-analysis and GWAS of atrial fibrillation and ischemic stroke subtypes analysed in SiGN. Pearson's r correlation between SNP z-scores in the AFGen GWAS of atrial fibrillation and in GWAS of selected traits performed in the SiGN data. (a) GWAS of atrial fibrillation in AFGen and in SiGN correlate with increasing strength as SNP z-scores in AFGen increase. Correlation with educational attainment (performed separately, shown here as a null comparator) remains approximately zero across all z-score thresholds. (b) SNP effects in AFGen also correlate strongly with cardioembolic stroke in SiGN, but not with the other primary stroke subtypes. (c) Undetermined subtypes of stroke also show modest correlation to the genetic architecture of atrial fibrillation in AFGen. Panels d-f show genome-wide z-score distributions underlying correlations.



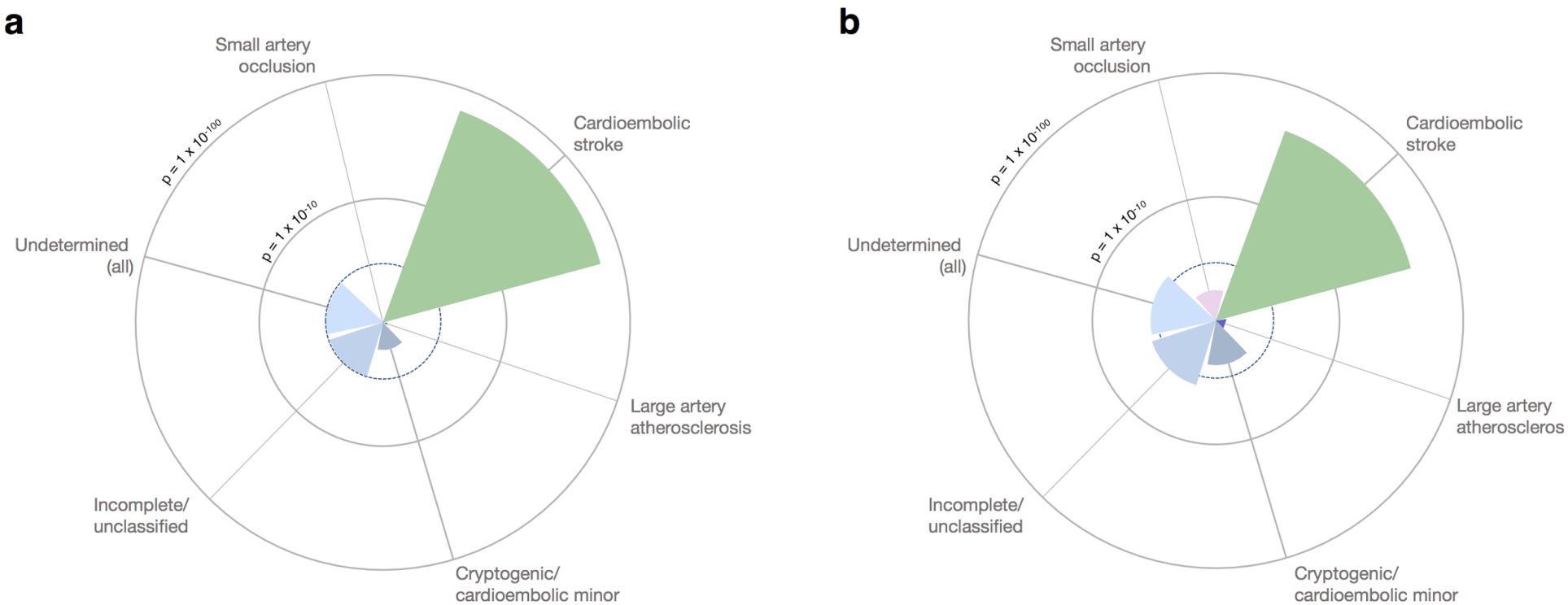
To explicitly test what proportion of cardioembolic stroke heritability could be explained by atrial fibrillation loci, we used SNPs from an atrial fibrillation genetic risk score (GRS) independently derived from the AFGen GWAS[5] (**Materials and Methods**). Of the 1,168 SNPs used to generate this pre-established GRS, we identified 934 in the SiGN dataset with imputation quality > 0.8 and minor allele frequency >1%. To compute the GRS per individual, we weighted the imputed dosage of the risk allele (a number between 0 and 2) by the effect of the SNP (i.e., the ß-coefficient) as reported in AFGen[5], repeated this weighting across all SNPs, and then summed these weighted effects.

We tested the association of the atrial fibrillation GRS with cardioembolic stroke, using a logistic regression and adjusting for the top ten principal components and sex (**Materials and Methods**). We found the GRS to be strongly associated with cardioembolic stroke (beta (s.e.) = 0.17 (0.01), $p = 1.01 \times 10^{-65}$; **Figure 3** and **Supplementary Table 5**), confirming the high genetic concordance of these phenotypes across SNPs which, individually, confer only a modest average association with atrial fibrillation. Expanding the set of SNPs used to construct the GRS to the original 934 SNPs ±25kb, ±50kb, and ±100kb (and subsequently linkage disequilibrium pruned at $r^2 = 0.2$ to ensure independence between markers) revealed a persistently strong though somewhat attenuated association between the GRS and cardioembolic stroke (GRS ± SNPs within 100Kb, $p = 4.47 \times 10^{-44}$, **Supplementary Table 6**). Additionally, adjusting the association with the original GRS for a set of clinical covariates (**Supplementary Table 7**) associated with atrial fibrillation including age, type 2 diabetes, cardiovascular disease, smoking, and presence of hypertension[27] only modestly reduced the persistent GRS signal in cardioembolic stroke (beta (s.e.) = 0.17 (0.01), $p = 1.45 \times 10^{-48}$; **Supplementary Table 6**).

None of the other subtypes we tested were significantly associated (Bonferroni-corrected $p = 5 \times 10^{-3}$, adjusting for five subtypes and the TOAST and CCS classification systems) to the GRS (all $p > 0.013$, **Figure 3** and **Supplementary Figure 6**).

As we had been leveraging referent samples for which atrial fibrillation status was mostly unknown, cryptic atrial fibrillation cases in the referent group were potentially negatively confounding associations between the AF GRS and stroke subtype. Of all available referents, 4,038 had information on atrial fibrillation diagnosis and of these samples, 3,861 are reported as not having atrial fibrillation. We therefore refined the set of referents by selecting only the 3,861 referents in whom atrial fibrillation status was non-missing and who did not have a reported atrial fibrillation diagnosis. Using this restricted set of samples, we observed a persistently strong association between the GRS and cardioembolic stroke (beta (s.e.) = 0.19 (0.01), $p = 1.66 \times 10^{-42}$; **Supplementary Table 5**). The lack of association between the GRS and large artery atherosclerotic stroke or small artery occlusion strokes also persisted when using these refined controls ($p = 0.60$ and $p = 0.17$, respectively). Interestingly, we found the atrial fibrillation risk score to be modestly associated ($p < 5 \times 10^{-3}$) with the overall undetermined subtype (beta (s.e.) = 0.036 (0.013), $p = 4.15 \times 10^{-3}$) (**Figure 3**). Further examination of the two mutually exclusive subgroups of the undetermined group revealed that the GRS associated significantly with the incomplete/unclassified categorization (beta (s.e.) = 0.046 (0.016), $p = 3.17 \times 10^{-3}$) (**Figure 3**) but not with cryptogenic/cardioembolic minor classification (beta (s.e.) = 0.030 (0.156), $p = 5.10 \times 10^{-2}$). Correcting for clinical covariates modestly changed the signal in the incomplete/unclassified phenotype ($p = 9.7 \times 10^{-3}$, **Figure 3**).

Figure 3 | Association of atrial fibrillation genetic risk score in ischemic stroke subtypes. We constructed an independent genetic risk score (GRS) from atrial fibrillation-associated SNPs identified in the AFGen GWAS, and tested associations between this GRS and ischemic stroke subtypes using (a) all available referents (N = 28,026) and (b) referents without atrial fibrillation (N = 3,861). The GRS strongly associated with cardioembolic stroke in both sets of samples. In the atrial fibrillation-free set of controls (panel b) we observed association of the GRS ($p < 5 \times 10^{-3}$, after adjusting for five subtypes and two sets of referents; indicated by the dashed dark blue line) with incomplete/unclassified stroke as well.



Lastly, we created a model in BOLT-LMM, fitting two genetic variance components: one component including SNPs for the genetic relationship matrix, and the second component including the original GRS SNPs from the atrial fibrillation GRS (including ±100kb around these SNPs, to include a sufficient number of markers to estimate variance explained). We found that the SNPs from the atrial fibrillation GRS explained 4.1% of the total 20.0% heritability in atrial fibrillation. In evaluating how much of cardioembolic stroke heritability that atrial fibrillation risk-increasing SNPs explain, we found a similar result: the component representing the atrial fibrillation risk score explained 4.5% (s.e. = 1.00%) of the total 19.5% genetic heritability in cardioembolic stroke. Thus, atrial fibrillation genetic risk accounts for 23.1%, or approximately one-fifth, of the total heritability of cardioembolic stroke.

Discussion

Although atrial fibrillation is an established causal risk factor for cardioembolic stroke, formal investigation of the overlapping biology of the two traits has been limited. Here, we leverage genome-wide common variation data from two large collaborative datasets to define locus-specific and genome-wide genetic signatures shared between the two phenotypes. Our analyses reveal that the genetic signature of atrial fibrillation in patients with stroke strongly parallels that of atrial fibrillation in the general population. Despite the fact that cardioembolic stroke often affects older adults with multiple clinical comorbidities[28] that could increase risk for atrial fibrillation due to non-genetic factors, we observed that nearly 20% of the variance in cardioembolic stroke risk can be explained by genetic risk factors for atrial fibrillation. Finally, our results from genome-wide correlation analysis and construction of a genetic risk score further indicate that cardioembolic stroke and atrial fibrillation share a highly similar genetic architecture, extending our understanding of the morbid consequences of heritable forms of the arrhythmia.

The implementation of genetic risk scores in complex traits has proved an efficient means of understanding the proportion of phenotypic variance explained by

common genetic markers, or how genetic predisposition to a disease can overlap with risk for other common disease(s). Given the onslaught of genotyping data made available in common disease over the last decade, and the resulting rapid increase in implicated common disease loci, genetic risk scores are beginning to be used as a means of stratifying patients by risk (e.g., in breast cancer[29,30]) or predicting outcome (e.g., in neuropsychiatric disease[29,31]). More recently, polygenic risk scores derived from millions of markers have been used to identify individuals in the general population with a four-fold risk for coronary disease[32] and to identify individuals for whom lifestyle changes or statin intervention would be particularly beneficial[33,34]. Future work in atrial fibrillation and cardioembolic stroke may reveal how genetic scores based on variants associated to atrial fibrillation could be used either to predict individuals at highest risk of cardioembolic stroke (and who therefore could be more closely monitored), or help distinguish between clinical causes of strokes.

Though our analysis was aimed at understanding the genetic overlap between cardioembolic stroke and atrial fibrillation, we additionally observed genome-wide correlation between atrial fibrillation and undetermined stroke and an association between the atrial fibrillation genetic risk score and stroke of unknown cause. Perhaps contrary to expectation, we specifically found the atrial fibrillation GRS to be more strongly associated with the subset of etiology-undetermined strokes with an incomplete clinical evaluation, as opposed to those with cryptogenic stroke of a presumed, but not demonstrated, embolic source. Our results therefore do not directly support the hypothesis that embolic strokes of undetermined cause[35] are largely composed of occult atrial fibrillation cases. These observations are consistent with the recent termination (due to futility) of a randomized trial of oral anticoagulation, a treatment typically reserved for patients with strokes caused by atrial fibrillation, in patients with embolic stroke of undetermined source.[36] Nevertheless, stroke diagnostic workups in our observational study rely in part on physician decisions, which could lead to biases in the level of evidence supporting stroke etiology in the undetermined classifications studied here. Our results could therefore represent an artifact of this decision-making process. In aggregate, we

submit that our findings indicate that atrial fibrillation genetic risk may augment clinical algorithms to determine stroke etiology.

While our work presented here has helped to define the overlapping common genetic risk factors for both atrial fibrillation and cardioembolic stroke, a number of limitations remain. First, stroke is a highly heterogeneous phenotype. Its onset is typically late, but can vary anywhere from neonates to the elderly. Therefore, genetic and environmental factors are not expected to carry a similar proportion of risk throughout an individual's lifespan. Furthermore, due to its late onset, stroke often co-occurs with a number of other common, complex traits[16] including hypertension, type 2 diabetes, and cardiovascular disease, the genetic contributions of which have not been explored herein. Further compounding the genetic study of ischemic stroke is the lifetime prevalence of the disease (>15%[37]), which increases the likelihood that referents studied here will have a stroke in the future; and sample size, which has lagged behind other GWAS efforts now interrogating tens or hundreds of thousands of samples[38–41]. The challenge of subtyping impacts sample size; nearly one-third of all cases fail to fall into one of the primary subtypes and are instead categorized as undetermined[23]. Further, while subtyping is key for genetic discovery, it also reduces sample sizes in stroke GWAS (as the full set of cases is often reduced substantially once subtyping is complete); reduced sample sizes impact power for discovery and make other analytic approaches — such as standard approaches for measuring trait correlation[20,24,42] — unfeasible. Finally, the most recent GWAS of atrial fibrillation revealed a role for rare variation, untested here due to limitations in reliably imputing rare variants in the SiGN samples[12]; similarly, we did not test other types of genetic variation, such as insertions and deletions or larger structural variants.

We have shown that the genetic liability to atrial fibrillation is the same in individuals presenting with stroke as in primary studies of incident and prevalent atrial fibrillation. Additionally, we have provided updated heritability estimates of ischemic stroke and its subtypes, and demonstrated that genetic loci that fall far below the genome-wide significance threshold in studies of atrial fibrillation carry a

substantial association with cardioembolic stroke risk. Although atrial fibrillation genetic risk did not associate with other primary stroke subtypes, we found an association with strokes of the 'undetermined' subtype, supporting the notion that undetected atrial fibrillation underlies a proportion of the stroke risk in these individuals. Given the increased prevalence of stroke in non-European populations, and especially in African Americans for whom risk of stroke is double that of individuals of European descent[43], future studies will be needed to extend our results into populations of non-European ancestry. Further work will also need to incorporate emerging discoveries of rare genetic variants in atrial fibrillation, and explore the potential for genotype-based tools to assist in the diagnostic workup of individuals with ischemic stroke.

Materials and Methods

Data access

For access to code, supporting data, and downloadable supplemental tables, please see this GitHub repository: https://github.com/UMCUGenetics/Afib-Stroke-Overlap. Please see the **Supplementary Information** for additional information regarding access to genotypes, dbGaP data identifiers, and DOIs where summary-level data from our genome-wide association studies of atrial fibrillation are available for download.

The Stroke Genetics Network (SiGN) and genome-wide association study of ischemic stroke subtypes

The Stroke Genetics Network (SiGN) was established with the aim of performing the largest genome-wide association study (GWAS) of ischemic stroke to date. The complete design of the study, including sample collection, phenotyping, and data processing, have been described previously[12]. Briefly, ischemic stroke cases were collected from cohorts located in the United States, Europe and Australia. Stroke cases included in SiGN were classified into stroke subtypes using the Causative Classification System (CCS), which subtypes cases through an automated, web-based system that accounts for clinical data, test results, and imaging information[18,44]. Within CCS, there are two sub-categories: CCS causative, which does not allow for competing subtypes in a single sample; and CCS phenotypic, which does.

SiGN is comprised of several case cohorts with pre-existing genotyping data. Newly-collected cases, as well as a small number of matched referents, were genotyped on the Illumina 5M array[12]. The majority of referents included were drawn from publicly-available genotyping data. Samples represent three continental populations (European-ancestry; African-ancestry; and non-European ancestry and non-African ancestry samples, primarily of admixed ancestry from Latin American

populations, labelled 'Hispanic'). For quality control (QC) and downstream association testing, cases and referents were matched by genotyping array and PCA-determined ancestry. In total, the study contains 13 case-referent analysis groups: 10 of European ancestry, two of African ancestry, and one Hispanic[12].

After QC, the SiGN dataset comprised 16,851 ischemic stroke cases and 32,473 stroke-free controls. European-ancestry samples were imputed with IMPUTE2[45] using a reference panel built from whole-genome sequence data collected by the 1000 Genomes Project (Phase 1)[46] and the Genome of the Netherlands[47] project; African-ancestry and Hispanic samples were imputed with the 1000 Genomes Project data only.[46] Due to data-sharing restrictions regarding the referents used for the Hispanic set of samples, only the European- and African-ancestry samples were analyzed here, totaling 13,390 cases and 28,026 referents distributed across 12 case-control analysis groups. A complete breakdown of case and referent information, including counts by stroke subtype, are included in **Supplementary Table 1**.

Phenotyping in SiGN

There are three primary subtype definitions of ischemic stroke: cardioembolic stroke, large artery atherosclerotic stroke, and small artery occlusion. The SiGN consortium used the CCS system to attempt to assign each case to one of these three categories. Additionally, ~74% of cases were also classified using the Trial of Org 10 172 in Acute Stroke Treatment (TOAST)[15,48] system, which classifies stroke cases based on clinical decision-making and clinically-ascertained information. The CCS and TOAST subtyping systems yield moderately-to-strongly correlated phenotyping results (**Supplementary Figure 5**)[49]. Use of these traits in a GWAS setting also yields concordant association results, as previously shown[12]. These subtypes are similarly defined in CCS and TOAST, though determined differently across the two subtyping systems.

In addition to the three primary subtypes, both the CCS and TOAST classification systems generate two additional subtypes: "undetermined" and "other." The "other" classification was small in sample size (N_{cases} = 595, 719 and 374 in CCS Causative, CCS Phenotypic and TOAST, respectively), and was therefore not included in the original SiGN GWAS and was not tested here[12]. The "undetermined" classification, though named the same in CCS and TOAST, is defined differently across the two subtyping systems[15,18]. In TOAST, patients with conflicting subtype classifications are placed in the undetermined category[12,15]. In contrast, the CCS undetermined classification includes patients with cryptogenic embolism, other cryptogenic cases, patients with an incomplete evaluation, or samples with competing subtypes[18].

Identifying atrial fibrillation cases and controls in SiGN

To perform association testing and downstream analyses, we sought to identify a set of atrial fibrillation cases and suitably-matched referents available in the SiGN dataset. We defined atrial fibrillation on the basis of five variables available in the CCS phenotyping system: (i) atrial fibrillation, (ii) paroxysmal atrial fibrillation, (iii) atrial flutter, (iv) sick sinus syndrome, and (v) atrial thrombus. This definition yielded 3,190 atrial fibrillation cases for analysis. For quality control purposes, we also ran an initial genome-wide association study in atrial fibrillation only (N = 1,751 cases) to ensure overlap with the broader atrial fibrillation definition.

Given that the vast majority of stroke-free referents in SiGN were collected from publicly-available sources, available atrial fibrillation information for these samples is extremely limited. From the 28,026 controls available for analysis, we established a set of 3,861 individuals for whom atrial fibrillation information was available and who did not have atrial fibrillation. For the remaining subjects, we assumed that individuals did not have atrial fibrillation.

Genome-wide association testing of ischemic stroke subtypes and atrial fibrillation in SiGN

To perform genome-wide association testing, we merged all genotype dosages available from the European-ancestry and African-ancestry analysis groups in SiGN. SNPs with imputation quality (info score) < 0.3 and minor allele frequency (MAF) < 1% were removed from the data. Additionally, for those SNPs that were genotyped in a subset of the SiGN study strata but imputed in others, we compared the frequency of the SNP across the various strata. We removed any SNP with a frequency difference > 15% within ancestral group or >50% across ancestral groups comparing imputed and genotyped data, likely induced by sequencing errors in the imputation reference panel(s).

We then implemented a linear mixed model to perform association testing using BOLT-LMM[20]. Linear mixed models can account for structure in the data, such as that due to (familial or cryptic) relatedness and population structure, while improving power for discovery[50–52]. Due to extensive structure in the SiGN data[12], induced by both study design and population ancestry, we adjusted the BOLT-LMM model for the top ten principal components (PCs) and sex, in addition to the genetic relationship matrix used as a random effect in the linear mixed model.[20] PCs were calculated in EIGENSTRAT[53] using the set of SNPs used in the genetic relationship (see next paragraph) but using a missingness threshold of 0.1%.

To construct the genetic relationship matrix (GRM) implemented in BOLT-LMM, we used SNPs that were (i) common (MAF > 5%), (ii) with missingness < 5%, (iii) linkage disequilibrium (LD) pruned at an r^2 threshold of 0.2, (iv) on the autosomal chromosomes only, (v) and not in stratified areas of the genome (i.e., not in the major histocompatibility complex (MHC), the inversions on chromosomes 8 and 17, or in the lactase (*LCT*) locus on chromosome 2; **Supplementary Information**). After association testing, we additionally removed SNPs with imputation quality (info score) < 0.8, due to excess inflation of the test statistic in those SNPs (**Supplementary Figure 1**).

Heritability calculations in ischemic stroke subtypes

To assess the aggregate genetic contributions of common variants to disease susceptibility, we calculated additive SNP-based heritability ($\square^2_{\square}$) estimates for ischemic stroke, within ischemic stroke subtypes, and for atrial fibrillation using restricted maximum likelihood as implemented in Genome-wide Complex Trait Analysis (GCTA)[24]. To construct the genetic relationship matrix (GRM) to use in heritability analyses, we first identified the set of SNPs with imputation quality > 0.8 and MAF > 1%. More than 5.5M SNPs passed these QC criteria, so we randomly selected 20% of the data (~1.1M SNPs) for computational efficiency in calculating the GRM. We then further identified SNPs outside the MHC and *LCT* regions, outside the inversions on chromosomes 8 and 17, and LD pruned (r^2 = 0.2). These filtering steps resulted in ~250,000 SNPs available to compute the GRM. We used Plink 1.9[54,55] to convert imputed dosages to best-guess genotypes and then compute the GRM (**Supplementary Information**).

We computed heritability estimates in BOLT-LMM[20] using BOLT-REML. We adjusted all heritability analyses for ten PCs and sex. To check the robustness of the heritability calculations to the SNPs included in the GRM, we calculated heritability using the GRM described above, as well as three additional GRMs: (i) using the ~1.1M SNPs with imputation quality > 0.8 and MAF > 1% (and without LD pruning); (ii) using the SNPs that were genotyped across all study strata (~155,000 SNPs); and (iii) the set of genotyped SNPs with the MHC, *LCT* locus, inversions on chromosomes 8 and 17 removed, and LD pruned at r^2 = 0.2. Additionally, we computed heritability in GCTA[24] using the same GRMs and assuming a trait prevalence of 1%. We compared the results to the BOLT-based $\square^2_{\square}$ estimates (**Supplementary Table 3** and **Supplementary Figures 2-3**). As genome-wide heritability estimates need a large number of SNPs to be accurate, we report in the paper all estimates using a GRM containing imputed, pruned SNPs. Estimates resulting from all GRMs are presented in the **Supplementary Information**.

Genetic correlation between atrial fibrillation and ischemic stroke subtypes

Due to small case sample sizes, we were unable to evaluate trait correlation through commonly-used methods. Instead, we used summary-level data from the latest Atrial Fibrillation Genetics (AFGen) Consortium meta-analysis of atrial fibrillation[5] to calculate a z-score for each SNP in that GWAS. Additionally, we calculated a z-score for each SNP in a GWAS of each stroke subtype in SiGN as well as in the GWAS of atrial fibrillation we performed in the SiGN data. Finally, as a null comparator, we downloaded SNP z-scores from a GWAS of educational attainment[26] available through LDHub (http://ldsc.broadinstitute.org/, accessed 11-1-2017). We aligned z-score signs based on the risk allele reported in each study. SNPs with an allele frequency difference >5% between AFGen and SiGN (all stroke analysis) were removed from the AFGen data (25,784 SNPs); similarly, SNPs with an allele frequency difference >5% between the educational attainment GWAS and SiGN (all stroke) were also removed (27,866 SNPs). We calculated Pearson's r between z-scores from two traits to evaluate correlation.

We also evaluated correlation between the phenotypic data and provide this in the **Supplementary Information**.

Constructing an atrial fibrillation genetic risk score

To construct an atrial fibrillation genetic risk score (GRS), we used SNPs from a previously-derived atrial fibrillation GRS[56]. Briefly, the GRS was derived using results from a recent GWAS of atrial fibrillation, comprised of 17,931 cases and 115,142 referents[5] and testing various sets of SNPs based on their p-value in the GWAS (varying from $p < 5 \times 10^{-8}$ to $p < 0.001$) and using varied linkage disequilibrium thresholds (0.1 - 0.9).[56] These sets of SNPs were used to generate various GRSs, which were then independently tested for association to atrial fibrillation in an independent sample from the UK Biobank; the best-performing GRS (defined as the GRS with the lowest Akaike's Information Criterion) comprised 1,168 SNPs with $p < 1 \times 10^{-4}$ in the atrial fibrillation GWAS and LD pruned at an r^2

threshold of 0.5.[56] Of these 1,168 SNPs, we identified 934 SNPs in the SiGN dataset with imputation info > 0.8 and MAF > 1%. We used these 934 SNPs to construct the atrial fibrillation GRS in the SiGN dataset by weighting the imputed number of risk-increasing alleles carried by an individual at a given SNP (0-2 risk-increasing alleles) and then weighting the dosage by the effect of the allele, as determined by the most recent GWAS[5]. We computed the final GRS for each individual by summing across all of the weighted genotypes.

Testing an atrial fibrillation genetic risk score in ischemic stroke subtypes

To test the association between the atrial fibrillation GRS and ischemic stroke subtypes, we performed a logistic regression implemented in R. Phenotypes were coded as 0 (unaffected) or 1 (affected), and we included sex and the top 10 PCs as additional covariates. We also optionally adjusted the association tests for the following clinical covariates: age, diabetes mellitus status, cardiovascular disease status, smoking status (current smoker, former smoker, or never smoked), and hypertension status. To ensure that our analyses of the GRS were robust, we additionally tested the GRS in the subset of European-ancestry samples only (**Supplementary Information**).

To determine the variance explained by the atrial fibrillation GRS in cardioembolic stroke, we constructed a model in BOLT-LMM that consisted of two variance components: (1) a variance component made up of SNPs for the genetic relationship matrix, and (2) a variance component made up of SNPs from the GRS. After computing the estimated variance explained for each component in BOLT-LMM, we converted the estimate to the liability score (**Supplementary Information**).

Acknowledgements

This work uses data from the National Institute of Neurological Disorders and Stroke - Stroke Genetics Network (NINDS-SiGN) Consortium and the Atrial Fibrillation Genetics (AFGen) Consortium. Members of NINDS-SiGN, the International Stroke Genetics Consortium, and AFGen are provided as an appendix to this paper.

Drs. Pulit, Mitchell, McArdle, and Kittner are supported by NIH grant R01NS100178. The NINDS-SiGN Consortium is supported by the NIH grants R01NS100178 and U01NS069208.

Dr. Anderson is supported in part by K23NS086873 and the Massachusetts General Hospital Center for Genomic Medicine Catalyst Award.

Dr. Lubitz is supported by NIH grants K23HL114724 and a Doris Duke Charitable Foundation Clinical Scientist Development Award 2014105.

Dr. Weng is supported by an American Heart Association Postdoctoral Fellowship Award 17POST33660226.

Drs. Ellinor and Benjamin are supported by the NIH grants R01HL092577 and R01HL128914. Dr. Benjamin is additionally supported by the NIH grants 1RC1HL101056 and 1R01HL102214. Dr. Ellinor is additionally supported by the NIH grants R01HL104156 and K24HL105780; the National Heart, Lung, and Blood Institute (NHLBI); American Heart Association Established Investigator Awards 13EIA14220013; and the Fondation Leducq 14CVD01.

Disclosures/Competing Financial Interests

Dr. Lubitz receives sponsored research support from Boehringer Ingelheim, Bristol Myers Squibb, Bayer HealthCare, Biotronik, and has consulted for St. Jude Medical / Abbott and Quest Diagnostics. Dr. Anderson has consulted for ApoPharma. Dr. Ellinor is a principal investigator on a Bayer HealthCare grant to the Broad Institute related to the genetics and development of new therapeutics for cardiovascular diseases.

Author Contributions

All named authors have contributed meaningfully to the present study. Specific contributions for each author are described below.

S.L. Pulit: Conception of study design, data analysis, drafting of manuscript, critical revision of manuscript
L.C. Weng: data analysis, critical revision of manuscript
P.F. McArdle: data acquisition, data analysis, critical revision of manuscript
L. Trinquart: data acquisition, critical revision of manuscript
S.H. Choi: data acquisition, critical revision of manuscript
B.D. Mitchell: data acquisition, study supervision, critical revision of manuscript
J. Rosand: data acquisition, study supervision, critical revision of manuscript
P.I.W. de Bakker: study supervision, critical revision of manuscript
E.J. Benjamin: data acquisition, study supervision, critical revision of manuscript
P.T. Ellinor: data acquisition, study supervision, critical revision of manuscript
S.J. Kittner: data acquisition, study supervision , critical revision of manuscript
S.A. Lubitz: Conception of research design, study supervision, drafting of manuscript, critical revision of manuscript
C.D. Anderson: Conception of research design, study supervision, drafting of manuscript, critical revision of manuscript

References

1. Chugh, S. S. *et al.* Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* **129,** 837–847 (2014).

2. Wolf, P. A., Abbott, R. D. & Kannel, W. B. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke* **22,** 983–988 (1991).

3. World Health Organization. WHO | The top 10 causes of death. (2014). Available at: http://www.who.int/mediacentre/factsheets/fs310/en/. (Accessed 1st September 2017).

4. Feigin, V. L., Lawes, C. M., Bennett, D. a., Barker-Collo, S. L. & Parag, V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurol.* **8,** 355–369 (2009).

5. Christophersen, I. E. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **57,** 289 (2017).

6. Bevan, S. *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genome-wide associations. *Stroke* **43,** 3161–3167 (2012).

7. Lubitz, S. A. *et al.* Genetics of atrial fibrillation: implications for future research directions and personalized medicine. *Circ. Arrhythm. Electrophysiol.* **3,** 291–299 (2010).

8. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448,** 353–357 (2007).

9. Gretarsdottir, S. *et al.* Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann. Neurol.* **64,** 402–409 (2008).

10. Gudbjartsson, D. F. *et al.* A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat. Genet.* **41,** 876–878 (2009).

11. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).

12. Pulit SL, *et al*. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *Lancet Neurology* **15,** 174–184 (2015).

13. Franz, M. R., Karasik, P. L., Li, C., Moubarak, J. & Chavez, M. Electrical remodeling of the human atrium: similar effects in patients with chronic atrial fibrillation and atrial flutter. *J. Am. Coll. Cardiol.* **30,** 1785–1792 (1997).

14. Lamas, G. A. *et al.* Ventricular pacing or dual-chamber pacing for sinus-node dysfunction. *N. Engl. J. Med.* **346,** 1854–1862 (2002).

15. Adams, H. P., Jr. *et al.* Classification of Subtype of Acute Ischemic Stroke: Definitions for Use in a Multicenter Clinical Trial. *Stroke* **24,** 35–41 (1993).

16. Petty, G. W. *et al.* Ischemic Stroke Subtypes. *Stroke* **30,** 2513–2516 (1999).

17. Kolominsky-Rabas, P. L., Weber, M., Gefeller, O., Neundoerfer, B. & Heuschmann, P. U. Epidemiology of Ischemic Stroke Subtypes According to TOAST Criteria. *Stroke* **32,** 2735–2740 (2001).

18. Ay, H. *et al.* A computerized algorithm for etiologic classification of ischemic stroke: The causative classification of stroke system. *Stroke* **38,** 2979–2984 (2007).

19. Sanna, T. *et al.* Cryptogenic stroke and underlying atrial fibrillation. *N. Engl. J. Med.* **370,** 2478–2486 (2014).

20. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).

21. Go, A. S. *et al.* Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA* **285,** 2370–2375 (2001).

22. Weng, L.-C. *et al.* Heritability of Atrial Fibrillation. *Circ. Cardiovasc. Genet.* **10,** e001838 (2017).

23. NINDS Stroke Genetics Network (SiGN), International Stroke Genetics Consortium. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *Lancet Neurol.* **15,** 4–7 (2015).

24. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

25. Malik, R. *et al.* Low-frequency and common genetic variation in ischemic stroke: The METASTROKE collaboration. *Neurology* **86,** 1217–1226 (2016).

26. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* (2016).

27. Alonso, A. *et al.* Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J. Am. Heart Assoc.* **2,** e000102 (2013).

28. Henninger, N., Goddeau, R. P., Jr, Karmarkar, A., Helenius, J. & McManus, D. D. Atrial Fibrillation Is Associated With a Worse 90-Day Outcome Than Other Cardioembolic Stroke Subtypes. *Stroke* **47,** 1486–1492 (2016).

29. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9,** 96 (2017).

30. Mavaddat, N. *et al.* Prediction of breast cancer risk based on profiling with

common genetic variants. *J. Natl. Cancer Inst.* **107,** (2015).

31. Vassos, E. *et al.* An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biol. Psychiatry* **81,** 470–477 (2017).

32. Khera, A. V. *et al.* Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. *bioRxiv* (2017).

33. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* **375,** 2349–2358 (2016).

34. Mega, J. L. *et al.* Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385,** 2264–2271 (2015).

35. Hart, R. G., Diener, H.-C. & Connolly, S. J. Embolic strokes of undetermined source: support for a new clinical construct–Authors' reply. *Lancet Neurol.* **13,** 967 (2014).

36. NAVIGATE ESUS Study Stopped Early Due to Comparable Efficacy in Treatment Arms | Johnson & Johnson. Available at: https://www.jnj.com/media-center/press-releases/navigate-esus-study-stopped-early-due-to-comparable-efficacy-in-treatment-arms. (Accessed: 18th December 2017)

37. Seshadri, S. *et al.* The Lifetime Risk of Stroke: Estimates From the Framingham Study. *Stroke* **37,** 345–350 (2006).

38. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).

39. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518,** 187–196 (2015).

40. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).

41. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

42. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **advance on,** 291–295 (2015).

43. Gorelick, P. B. Cerebrovascular Disease in African Americans. *Stroke* **29,** 2656–2664 (1998).

44. Arsava, E. M. *et al.* The causative classification of stroke system: An international reliability and optimization study. *Neurology* **75,** 1277–1284 (2010).

45. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast

and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).

46. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

47. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

48. Goldstein, L. B. *et al.* Improving the reliability of stroke subgroup classification using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria. *Stroke* **32,** 1091–1098 (2001).

49. McArdle, P. F. *et al.* Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. *Neurology* **83,** 1653–1660 (2014).

50. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–835 (2011).

51. Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.* **4,** 6874 (2014).

52. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46,** 100–106 (2014).

53. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

54. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 1–16 (2015).

55. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

56. Weng, L.-C. *et al.* Genetic Predisposition, Clinical Risk Factor Burden, and Lifetime Risk of Atrial Fibrillation. *Circulation* (2017).