frontiers in Neurology

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fneur.2021.787107

Physician-Confirmed and Administrative Definitions of Stroke in UK Biobank Reflect the Same Underlying Genetic Trait

Kristiina Rannikmäe[1†], Konrad Rawlik[2†], Amy C. Ferguson[1], Nikos Avramidis[3], Muchen Jiang[4], Nicola Pirastu[5], Xia Shen[5,6,7], Emma Davidson[8], Rebecca Woodfield[9], Rainer Malik[10], Martin Dichgans[10,11,12], Albert Tenesa[1,2,13‡] and Cathie Sudlow[1,14‡]*

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom, [2] Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom, [3] School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, [4] Medical School, University of Edinburgh, Edinburgh, United Kingdom, [5] Usher Institute, University of Edinburgh, Edinburgh, United Kingdom, [6] Biostatistics Group, Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China, [7] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [8] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom, [9] Department of Medicine for the Elderly, Western General Hospital, Edinburgh, United Kingdom, [10] Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany, [11] Munich Cluster for Systems Neurology, Munich, Germany, [12] German Center for Neurodegenerative Diseases (DZNE), Munich, Germany, [13] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom, [14] BHF Data Science Centre, Health Data Research UK, London, United Kingdom

OPEN ACCESS

Edited by:
Cheng-Yang Hsieh, Sin-Lau Christian Hospital, Taiwan

Reviewed by:
Edward Lai, National Cheng Kung University, Taiwan
Baptiste Couvy-Duchesne, INSERM U1127 Institut du Cerveau et de la Moelle épinière (ICM), France

***Correspondence:**
Kristiina Rannikmäe
kristiina.rannikmae@ed.ac.uk

†These authors have contributed equally to this work and share first authorship
‡These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Stroke, a section of the journal Frontiers in Neurology

Received: *01 October 2021*
Accepted: *13 December 2021*
Published: *02 February 2022*

Citation:
Rannikmäe K, Rawlik K, Ferguson AC, Avramidis N, Jiang M, Pirastu N, Shen X, Davidson E, Woodfield R, Malik R, Dichgans M, Tenesa A and Sudlow C (2022) Physician-Confirmed and Administrative Definitions of Stroke in UK Biobank Reflect the Same Underlying Genetic Trait. Front. Neurol. 12:787107. doi: 10.3389/fneur.2021.787107

Background: Stroke in UK Biobank (UKB) is ascertained *via* linkages to coded administrative datasets and self-report. We studied the accuracy of these codes using genetic validation.

Methods: We compiled stroke-specific and broad cerebrovascular disease (CVD) code lists (Read V2/V3, ICD-9/-10) for medical settings (hospital, death record, primary care) and self-report. Among 408,210 UKB participants, we identified all with a relevant code, creating 12 stroke definitions based on the code type and source. We performed genome-wide association studies (GWASs) for each definition, comparing summary results against the largest published stroke GWAS (MEGASTROKE), assessing genetic correlations, and replicating 32 stroke-associated loci.

Results: The stroke case numbers identified varied widely from 3,976 (primary care stroke-specific codes) to 19,449 (all codes, all sources). All 12 UKB stroke definitions were significantly correlated with the MEGASTROKE summary GWAS results (rg.81-1) and each other (rg.4-1). However, Bonferroni-corrected confidence intervals were wide, suggesting limited precision of some results. Six previously reported stroke-associated loci were replicated using ≥1 UKB stroke definition.

Conclusions: Stroke case numbers in UKB depend on the code source and type used, with a 5-fold difference in the maximum case-sample size. All stroke definitions are significantly genetically correlated with the largest stroke GWAS to date.

Keywords: stroke, genetic correlation, routinely collected health data, validation, accuracy

INTRODUCTION

UK Biobank (UKB) is a prospective population-based cohort study with extensive phenotype and genotype information on >500,000 participants from England, Scotland, and Wales (www.ukbiobank.ac.uk). It is an open-access resource, established to facilitate research into the determinants of a wide range of health outcomes, particularly those relevant in middle and older age (1). An example of such a disease is stroke, the second most common cause of death worldwide and a major global cause of disability (2).

Disease outcomes in UKB are ascertained chiefly *via* linkages to routinely collected, coded, national administrative health datasets. In addition, data on self-reported medical conditions were collected at recruitment. However, to use these data appropriately, researchers need to select which particular disease codes to use for their study and have an understanding of their accuracy. For example, to identify stroke cases, existing codes can be divided into those that are stroke-specific and those that fall under the broad cerebrovascular disease (CVD) category. Stroke-specific codes are used to code acute stroke events where the clinician is confident about the diagnosis and can usually assign a subtype. In contrast, broad CVD codes also capture cases with: (i) phenotypes that pose a high risk for a subsequent stroke (e.g., a code for a transient ischemic attack, an unruptured aneurysm, or carotid artery stenosis); (ii) a past history of stroke with residual symptoms (e.g., a code for sequelae of cerebral infarction); (iii) events where there may be some diagnostic uncertainty (e.g., a code for unspecified cerebrovascular disease); and (iv) intracranial hemorrhages other than intracerebral or subarachnoid hemorrhage (e.g., extradural or subdural hemorrhages, which most clinicians consider different from a stroke). Including codes from the broad CVD category will therefore significantly increase the overall number of cases identified, but while this is likely to include at least some misclassified true acute stroke cases, non-stroke cases will also be included.

In a systematic review of studies validating stroke code accuracy from case-note review, the overall positive predictive value (proportion of true-positive cases among all identified cases) for identifying acute stroke cases was consistently >70% for stroke-specific codes, dropping to <50% in many studies when broad CVD codes were included (3, 19). For self-reported stroke events, the positive predictive value ranged from 22 to 87% across different studies, making it hard to draw firm conclusions (19). While the case-note review for code validation is often considered a gold standard, this method also has its limitations. It is time-consuming and labor-intensive, so can only be achieved in relatively small numbers of cases, with limited precision of the results. In addition, the results rely on: (i) accessing the complete relevant medical record; (ii) the detail and quality of the medical record; (iii) the qualification of the person reviewing the notes; (iv) the inter-adjudicator agreement, which we know is not perfect even between highly specialized clinicians; and (v) the consistency of results across different healthcare settings/providers (4).

We set out to supplement current knowledge about the accuracy of stroke codes with a method making use of large-scale genetic data, which we refer to as 'genetic validation'. The fundamental idea is to use existing knowledge of genetic associations with a disease (in this case acute stroke), to assess how well various potential code lists capture people who truly have this disease, which in turn could be used to harmonize disease definitions across cohorts and health systems (5). If the code list captures true-positive cases, we would expect the genetic associations that result from stroke cases identified through coded data to closely mirror the genetic association results from previous studies of stroke.

METHODS

Study Setting

We included all 408,210 UKB white British ancestry participants in this study. We restricted our analyses to this ancestry subgroup because it covers 94% of the UKB participants and allowed us to achieve a good balance between attaining sufficient case numbers while reducing population stratification and analytic complexity. As part of the UK Biobank recruitment process, informed consent was obtained from all individual participants included in the study. At the time of the study, UKB had linked hospital admissions and death registry administrative coded data available for all participants, and primary care administrative coded data for 47% of the cohort (191,146), covering the time period up to March and September 2019, respectively (**Supplementary Table S1**). In addition, all participants self-reported pre-existing health conditions during an interview at recruitment. The subset of the cohort with primary care data available was similar to the whole cohort with respect to age at recruitment, sex, and Townsend deprivation index (**Supplementary Table S2**).

Identifying Stroke Cases in UKB

We compiled stroke-specific and broad CVD code lists for each medical setting (hospital admission, death record, primary care) and self-report. This process was informed by previously published codes where available (3, 19), supplemented by the selection of additional codes by expert clinicians (authors KR, CLMS, ED, RW) on discussion and mutual agreement (further detail is provided in Supplementary Methods). This resulted in a total of eight code lists, covering the ICD-9/ICD-10, Read Version 2, Clinical Terms Version 3 (Read Version 3), and UKB self-report illness coding systems (**Supplementary Table S3**).

Next, we identified all participants with a relevant code from any of the code lists and created 12 different ways of defining stroke cases in UKB based on the code type (stroke-specific, broad CVD) and source (hospital admission, death record, primary care, self-report). This resulted in 12 partially overlapping case-control groups, where cases were all the individuals with a stroke code for the particular stroke definition, and all the remaining participants acted as controls. A specific UKB participant could therefore be a stroke case for one definition and control for another definition.

Genome-Wide Association Studies

We performed 12 genome-wide association studies (GWASs), one for each case-control set (i.e., for each definition of stroke cases and their controls). We applied a linear mixed model method using the BoltLMM software package (v2.3.4) software (6). We included the following as covariates: genotyping array, UKB assessment center, sex, age at recruitment, and principal components one to ten. We filtered the results for single nucleotide polymorphisms (SNPs) with an imputation quality INFO score $\geq$0.9 and minor allele frequency $\geq$1%. After filtering the results for SNP imputation quality and minor allele frequency, we included 9,524,428 SNPs. For further analyses, we converted the linear mixed model effects to logistic regression-comparable odds ratios, betas, and standard errors using the R code provided in https://shiny.cnsgenomics.com/LMOR/ (7). We also estimated stroke heritability in each of the 12 GWAS, converting heritability on the observed scale to heritability on the liability scale based on the prevalence in the study sample (8). We then assessed stroke heritability in the MEGASTROKE European sample, using the BLD LDAK model (9) for comparison.

Analyses of GWAS Results

We compared summary results from our 12 GWASs against the largest published stroke GWAS meta-analysis project-the MEGASTROKE study. The MEGASTROKE study is a meta-analysis of 29 stroke GWASs (17 including individuals of European ancestry) and does not include UKB data. Almost all studies included in MEGASTROKE (covering >95% included cases) required the stroke diagnosis to be confirmed by a medical professional or required evidence of stroke from >1 source, even if the initial case ascertainment included using administrative codes (10). All analyses were done using R software version 3.6.2.

Genetic Correlation With the MEGASTROKE Study Results

We applied a high-definition likelihood method using the HDL software (11) to assess the genetic correlation between our GWAS results using the 12 stroke definitions, and the MEGASTROKE study GWAS summary results for any stroke subtype in European samples. Genetic correlation (rg) is the proportion of variance that two stroke definitions share due to genetic causes. A genetic correlation of 0 implies that the genetic effects on one definition are independent of the other, while a correlation of one implies that all of the genetic influences on the two definitions are identical. We assessed if the correlation was significantly different from 0 and 1, setting the *p*-value significance threshold to 0.0042 after a Bonferroni correction for the 12 tests. We used the LD matrix calculated from the UKB for the reference panel provided as part of the HDL package, therefore restricting analyses to 1,029,876 QCed imputed HapMap3 SNPs. We displayed the results (correlation measured as rg) on a heatmap. We also display Bonferroni corrected confidence intervals to aid interpretation.

Genetic Correlation Within Our Study Definitions

We then used the HDL software to assess genetic correlations within our study across the 12 definitions. We set the significance threshold to 0.0024 after a Bonferroni correction for seven independent non-overlapping case-control definitions (definitions not in bold in **Table 1**), resulting in 21 correlation tests. We used the LD matrix calculated from the UKB for the reference panel provided as part of the HDL package, therefore restricting analyses to 1,029,876 QCed imputed HapMap3 SNPs. We also display Bonferroni corrected confidence intervals to aid interpretation.

Replicating the MEGASTROKE Study Stroke-Significant Loci

The MEGASTROKE study identified 32 genetic loci significantly associated with stroke. We identified these loci (the lead SNP for each locus) in our GWAS summary results and considered a locus to be replicated (i.e., also significantly associated with the respective stroke definition in our data) if the *p*-value of association in our GWAS was <0.00156 (Bonferroni corrected for 32 loci). We compared the number of replicated loci across our summary definitions. We compared the effect sizes of the associations between MEGASTROKE trans-ethnic and European ancestry GWASs and our GWAS summary results. Where the lead SNP was not available in our data, we identified SNPs in moderate LD (r2 > 0.7 in the 1,000 Genomes GBR population using the Ensembl LD calculator https://www.ensembl.org/Homo_sapiens/Tools/LD) with the lead SNP, and if any SNPs in LD available in our data were identified, we examined their associations instead. We displayed results for five of our summary definitions of stroke cases and their controls: stroke-specific code from any medical setting; broad CVD code from any medical setting; stroke-specific or broad CVD code from any medical setting; specific or non-specific self-reported stroke event; any code or self-reported event. We highlighted significantly associated (i.e., replicated) loci.

We also calculated our expected power to replicate the 32 loci using the Genetic Association Study (GAS) Power Calculator (http://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html), assuming a stroke prevalence of 2.26% and inputting the disease allele frequency and genotype relative risk estimates from the MEGASTROKE publication **Table 1** (10).

RESULTS

Stroke Cases in UKB

The number of relevant codes for identifying stroke cases varied widely depending on the coding system (ICD vs. Read vs. self-report) and code type (stroke-specific vs. broad CVD code)–from less than five codes for a specific self-reported stroke event, to >500 codes when including all possible codes across all coding systems. The stroke-specific and broad cerebrovascular disease (CVD) code lists for each medical setting and self-report are shown in the **Supplementary Table S3**.

The number of stroke cases identified among the 408,210 participants also varied widely depending on the code type and source used–from 3,976 cases in primary care when using stroke-specific codes, to 19,449 cases when including all possible code combinations (stroke-specific and broad CVD) across all sources

TABLE 1 | Number and demographic characteristics of stroke cases identified in UK Biobank (UKB).

Stroke definition	Number of cases	Number of controls	Mean (median) age at recruitment (years)	Mean (median) age; age range at stroke (years)	Sex (% female)
Stroke-specific code from hospital/death records	6,887	401,323	61 (63)	63 (64); 31 to 79	40
Stroke-specific code from primary care	3,976	404,234	61 (63)	59 (61); 1 to 79	41
Stroke-specific code from any medical setting	**8,665**	**399,545**	61 (63)	61 (63); 1 to 79	41
Broad CVD code from hospital/death records	5,725	402,485	62 (63)	64 (65); 31 to 79	43
Broad CVD code from primary care	4,003	404,207	62 (63)	62 (63); 1 to 79	41
Broad CVD code from any medical setting	**8,085**	**400,125**	62 (63)	63 (64); 1 to 79	44
Stroke-specific or broad CVD code from hospital/death records	12,612	395,598	61 (63)	63 (64); 31 to 79	41
Stroke-specific or broad CVD code from primary care	7,979	400,231	62 (63)	60 (62); 1 to 79	41
Stroke-specific or broad CVD code from any medical setting	**16,750**	**391,460**	62 (63)	62 (63); 1 to 79	42
Specific self-reported stroke event	5,915	402,295	61 (62)	53 (55); 0 to 70	41
Specific or non-specific self-reported stroke event	**7,536**	**400,674**	61 (63)	53 (55); 0 to 70	42
Any code or self-reported event	**19,449**	**388,761**	61 (63)	60 (61); 0 to 79	43
Across all UKB participants	408,210		57 (58)	Not applicable	54

We considered code and self-reported event age missing if it predated the date of birth, contained a negative value, or was recorded by UKB as unreliable or missing–this affected 0.3% to 2.1% cases depending on the stroke definition. CVD, cerebrovascular disease. Bold represents combined groups, e.g. broad CVD code from hospital/death + broad CVD from primary care = broad CVD code from any medical setting.

(hospital admission, death record, primary care, self-report) (**Table 1**).

The code source for cases with a stroke-specific code was: self-report only in 27%, primary care only for 9%, hospital/death record code only for 29%, and >1 source for 35% (**Supplementary Figure S1**). The code source for cases with either a stroke-specific or a broad CVD code was: self-report only in 14%, primary care only for 15%, hospital/death record code only for 34%, and >1 source for 37% (**Supplementary Figure S1**). These proportions are calculated based on the primary care data being currently available only for ∼50% of the participants, and so will change when primary care data for the whole cohort become available.

The overall proportion of prevalent codes (i.e., first code predates participant's recruitment to UKB) vs. incident codes (i.e., first code date occurs after participant's recruitment to UKB) was the same for stroke-specific and broad CVD categories: 38% prevalent vs. 62% incident codes. These proportions are dependent on the updates to different linked health datasets and the proportion of incident codes will continue to increase with increasing duration of follow-up (**Supplementary Table S1**).

Mean and median age at recruitment was higher among stroke cases (for all stroke definitions) than for the whole cohort of UKB participants (mean age 61 to 62 vs. 57 years; median age 62 to 63 years vs. 58 years). Mean and median age at the time of stroke (in case of multiple events, age at the earliest event was taken) was higher for coded diagnoses from the medical setting compared to self-reported events (mean age 62 vs. 53 years, median age 63 vs. 55 years, respectively). This is to be expected, considering that all self-reported events were recorded at the time of recruitment, whereas medical codes also capture diagnoses after recruitment during follow-up. The proportion of women was lower among stroke cases than across all UKB participants (43% for those with any medical setting or self-reported code vs. 54% for all UKB). This is to be expected as age-specific incidence rates are substantially lower in women than men in younger and middle-age groups, but these differences narrow down so that in the oldest age groups, incidence rates in women are approximately equal to or even higher than in men (12) (**Table 1**).

Analyses of GWAS Results

Manhattan plots, QQ plots, and genomic inflation factors (λ, lambda) are displayed in **Supplementary Figure S2**. Lambda remained 1.002 for all analyses, suggesting no significant inflation. Heritability measures across different stroke definitions ranged from 1.41% (for broad CVD code from primary care) to 5.69% (for stroke-specific code from primary care) (**Supplementary Table S4**). Heritability in the MEGASTROKE study was similar at 2.92%.

Genetic Correlation With the MEGASTROKE Study Results

All 12 UKB stroke definitions were significantly correlated with the MEGASTROKE summary GWAS results, with genetic correlations (rg) ranging from 0.81 to 1, and confidence intervals overlapping. The *p*-values for the difference from one were

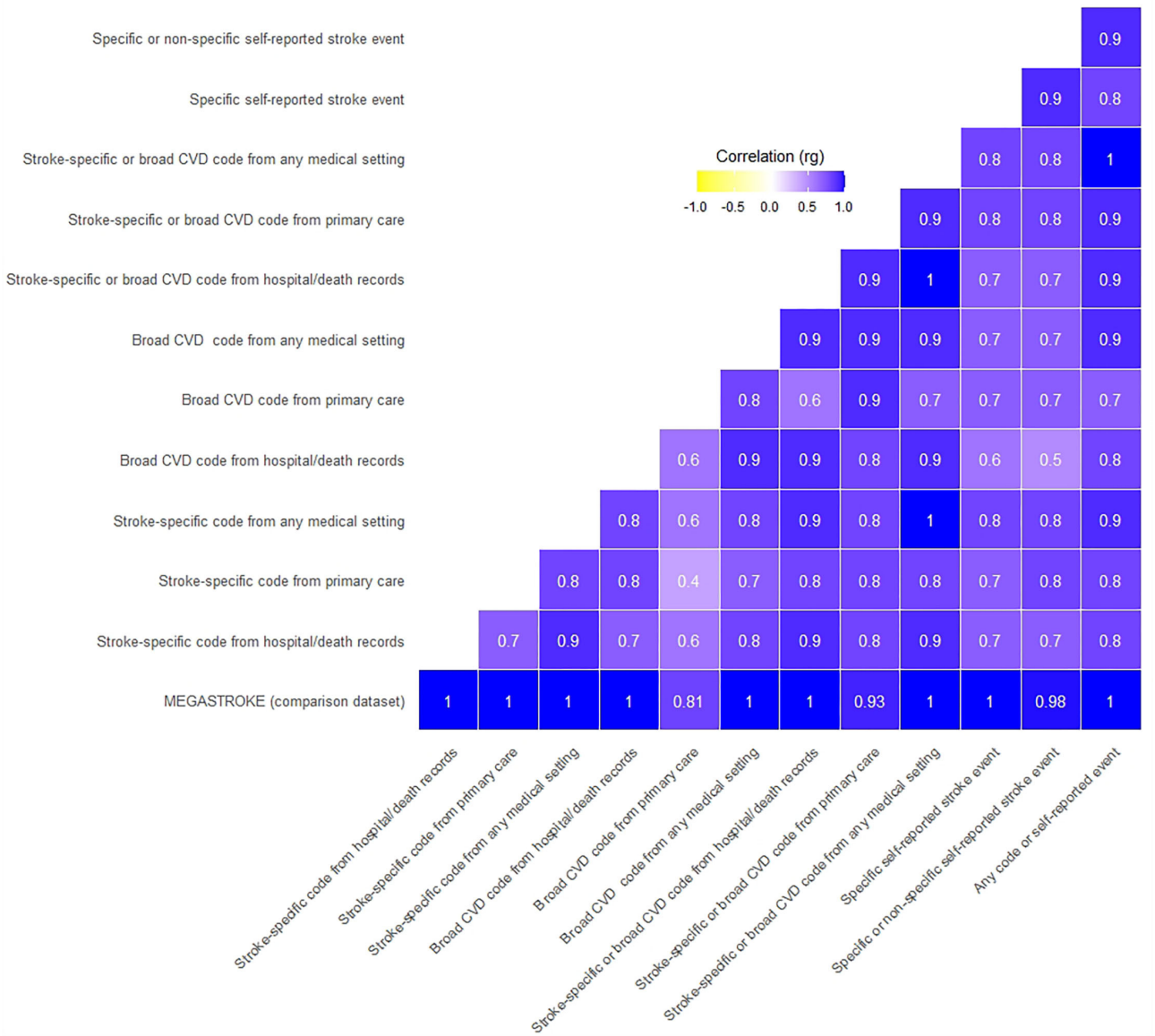

FIGURE 1 | Genetic correlation of UKB stroke definitions with MEGASTROKE and each other. Where the rg was >1, we rounded it to 1.

not significant, compatible with perfect correlation. However, the Bonferroni corrected CIs were wide, especially for five of the 12 tests, where the lower confidence limit was <0.7, limiting the precision of some of these results (**Figures 1**, **2**, **Supplementary Table S5**).

Genetic Correlation Within Our Study Definitions

The UKB summary definitions in our study were all significantly correlated with each other (all *p*-values significantly different from 0), with rg ranging from 0.4 to 1. Again, the Bonferroni corrected CIs were wide, with the lower confidence limit even suggesting the possibility of a negative correlation for two comparisons (**Figures 1**, **2**, **Supplementary Table S6**).

Considering the wide confidence intervals from the above genetic correlation analyses, we further explored this by calculating the effective sample size (Neff) (13). This ranged from 7,875 to 37,045 (**Supplementary Table S4**), which is significantly lower than the sample size used in the calculations by Ning et al. (11) when first describing the HDL method, and hence we would expect to see wider confidence intervals in our study.

Replicating the MEGASTROKE Study Stroke-Significant Loci

Within our GWASs, six of the 32 previously reported stroke-associated loci were replicated by one or more definitions. Analyses using stroke-specific codes and analyses using any code or self-reported event both replicated the

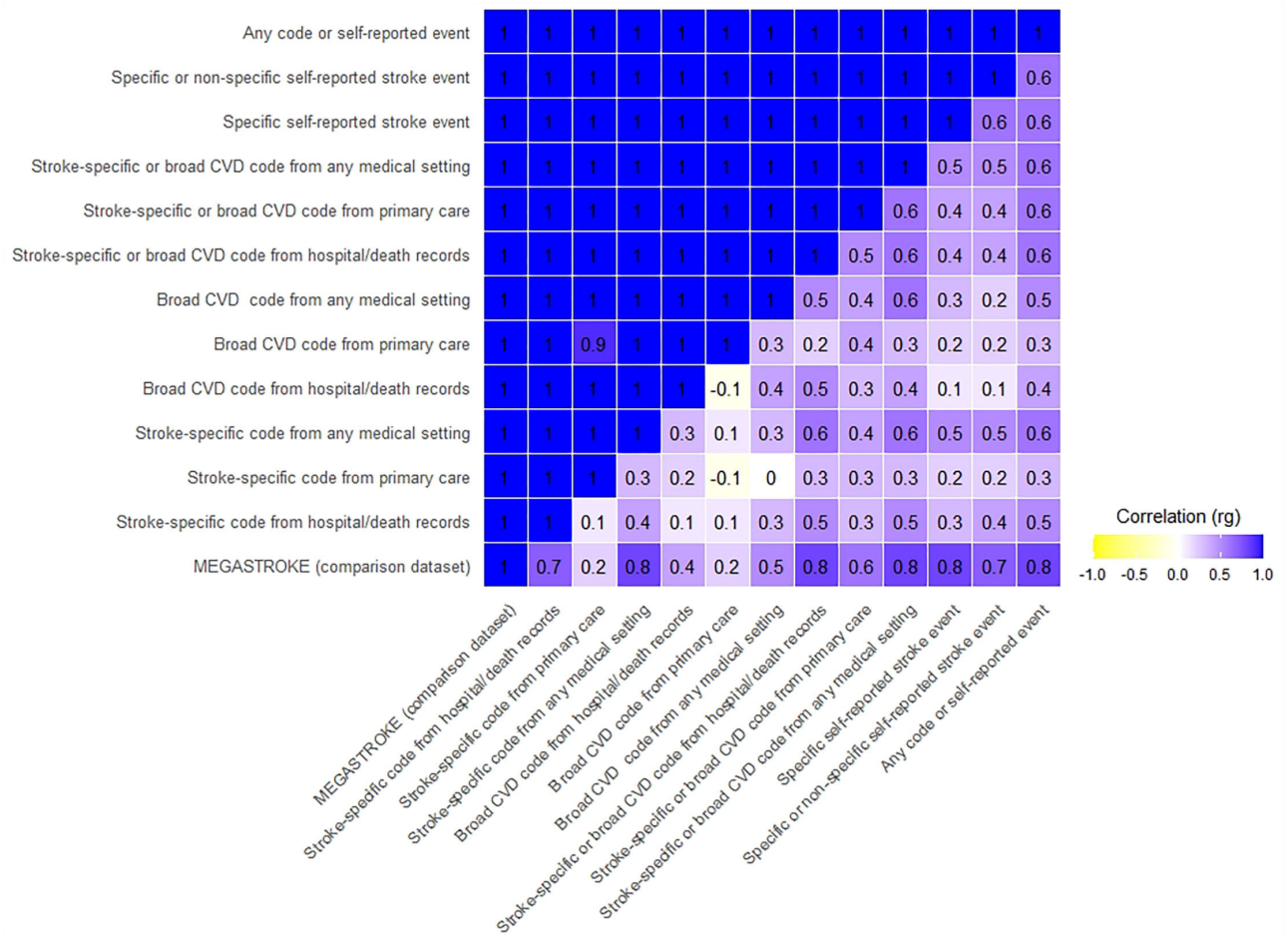

FIGURE 2 | Confidence intervals of genetic correlation of UKB stroke definitions with MEGASTROKE and each other. The upper triangle displays Bonferroni corrected upper confidence intervals, and the lower triangle displays Bonferroni corrected lower confidence intervals. Where the rg was >1, we rounded it to 1.

biggest number of known stroke loci (five of 32). The power from additional cases for the latter category did not result in replicating more loci than stroke-specific codes alone. However, for three of the five replicated loci, the *p*-values were smaller in the larger dataset (analyses using any code or self-reported event) suggesting a more robust replication when using the broadest definition of stroke in UKB. Within our data, effect sizes (expressed as odds ratios) were similar across the stroke definitions, with overlapping Bonferroni corrected confidence intervals (**Supplementary Figure S3**).

For two of the six replicated loci (PITX2 and HDAC9–TWIST1), the effect size of the association (odds ratio) was bigger in the MEGASTROKE dataset than in our data (across all five summary stroke definitions). These two loci are known to be associated with particular stroke subtypes–PITX2 with cardioembolic and HDAC9–TWIST1 with large artery stroke [10] (**Figure 3**, **Table 2**, **Supplementary Table S7**).

Power calculations suggested we had ≥80% power to replicate all 32 loci for any code or self-reported event definition while having ≥80% power for only 11/32 loci for the definition including a stroke-specific code from any medical setting (**Supplementary Table S8**).

DISCUSSION

Our analyses show, that depending on the code source and type used for identifying stroke cases in the UKB, the currently achieved maximum case-sample size can range from ∼4,000 to ∼20,000–a remarkable 5-fold difference. We go on to demonstrate, that regardless of the code source and type used, the resulting GWAS summary results are significantly genetically correlated with the largest stroke GWAS to date, with similar (albeit low) heritability estimates. Finally, when we try to replicate known stroke-significant loci in our data, both stroke-specific codes from any medical setting as well as a broad definition including any code or self-reported event, replicate five of the 32 loci. Replication generated broadly similar effect sizes for all but two stroke subtype-specific loci, which is likely explained by our dataset including a mix of stroke subtypes. Another possible explanation is the "winner's curse" phenomenon (i.e., the

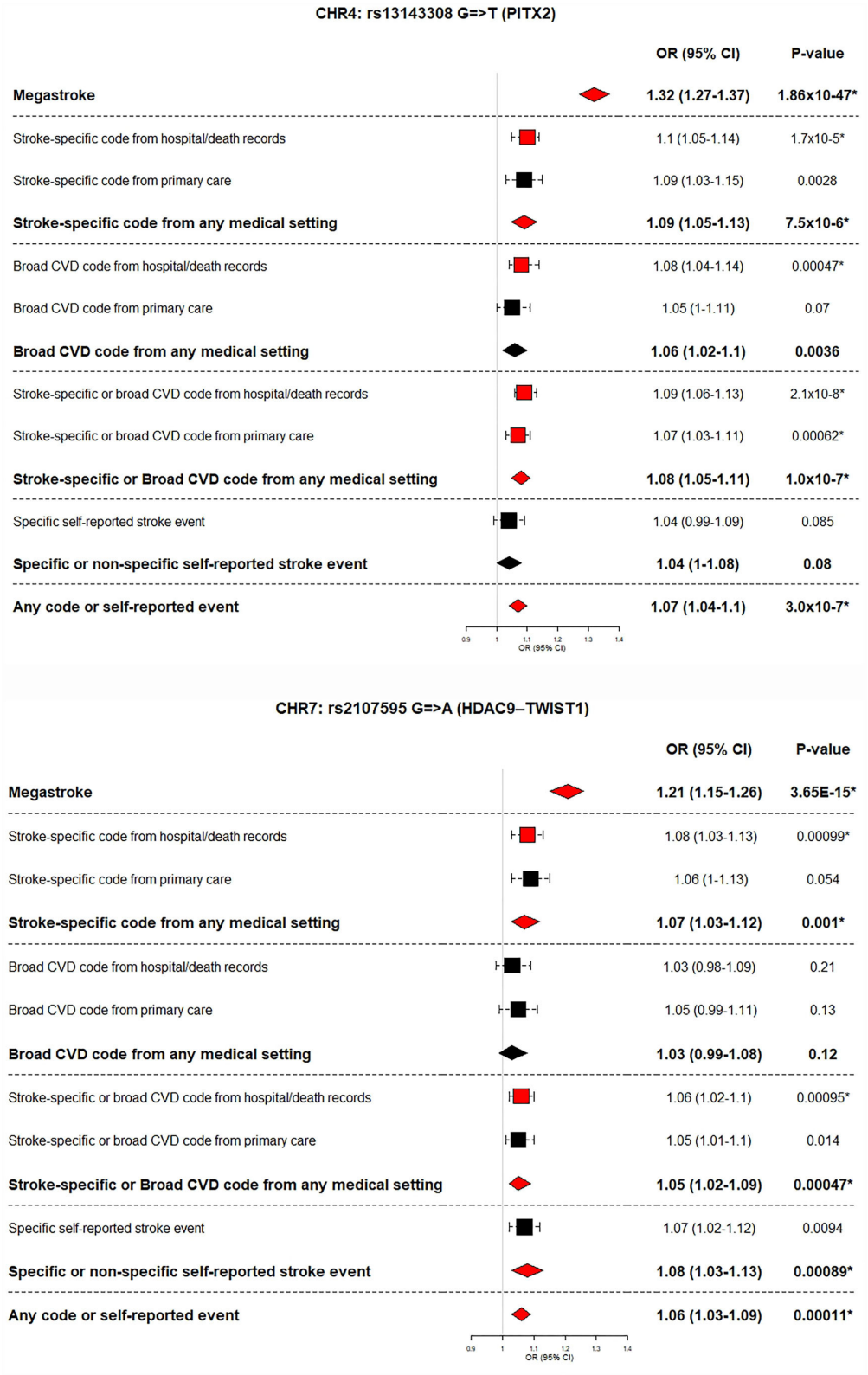

FIGURE 3 | MEGASTROKE stroke subtype-significant loci replicated using UK Biobank stroke definitions. MEGASTROKE odds ratio and *p*-value is shown for the analyses (European or trans-ethnic) showing the lowest *p*-value.

TABLE 2 | MEGASTROKE stroke-significant loci replicated using UKB stroke definitions.

Known stroke genetic locus	Significantly associated stroke type in MEGASTROKE	MEGASTROKE European meta-analyses *N* = 40,585*	MEGASTROKE transethnic meta-analyses *N* = 67,162†	Stroke-specific code from any medical setting *N* = 8,665	Broad CVD code from any medical setting *N* = 8,085	Stroke-specific or Broad CVD code from any medical setting *N* = 16,750	Specific or non-specific self-reported stroke event *N* = 7,536	Any code or self-reported event *N* = 19,449
CHR4: rs13143308 (PITX2)	Cardio-embolic stroke	**1.34 (1.28–1.40) 5.2×10^{-41}**	**1.32 (1.27–1.37) 1.9×10^{-47}**	**1.09 (1.05–1.13) 7.5×10^{-6}**	1.06 (1.02–1.1) 0.0036	**1.08 (1.05–1.11) 1×10^{-7}**	1.04 (0.995–1.08) 0.08	**1.07 (1.04–1.1) 3×10^{-7}**
CHR7: rs2107595 (HDAC9–TWIST1)	Large-vessel stroke	**1.27 (1.19–1.35) 1.4×10^{-13}**	**1.21 (1.15–1.26) 3.7×10^{-15}**	**1.07 (1.03–1.12) 1×10^{-3}**	1.03 (0.99–1.08) 0.12	**1.05 (1.02–1.09) 4.7×10^{-4}**	**1.08 (1.03–1.13) 8.9×10^{-4}**	**1.06 (1.03–1.09) 1.1×10^{-4}**
CHR10: rs2295786 (SH3PXD2A)	All stroke	1.05 (1.03–1.07) 1.4×10^{-7}	**1.05 (1.04–1.07) 1.8×10^{-10}**	**1.07 (1.04–1.11) 1.5×10^{-5}**	1.01 (0.98–1.04) 0.59	**1.04 (1.02–1.07) 4×10^{-4}**	1.04 (1.01–1.08) 0.021	**1.05 (1.03–1.07) 5.1×10^{-6}**
CHR12: rs3184504 (SH2B3)	All ischaemic stroke	**1.08 (1.06–1.10) 1.2×10^{-14}**	**1.08 (1.06–1.10) 2.2×10^{-14}**	1.04 (1.01–1.07) 0.0096	1 (0.97–1.03) 0.83	1.02 (1–1.05) 0.04	**1.06 (1.03–1.1) 1.9×10^{-4}**	1.03 (1.01–1.05) 0.0068
CHR9: rs635634 (ABO)	All ischaemic stroke	**1.08 (1.05–1.11) 9.2×10^{-9}**	1.07 (1.04–1.10) 6.9×10^{-3}	**1.08 (1.03–1.12) 1.8×10^{-4}**	1 (0.96–1.04) 0.86	1.04 (1.01–1.07) 0.0043	**1.1 (1.05–1.14) 5.8×10^{-6}**	**1.05 (1.02–1.07) 4.3×10^{-4}**
CHR9: rs7859727 (Chr9p21)	All stroke	1.05 (1.03–1.07) 7.2×10^{-8}	**1.05 (1.03–1.07) 4.2×10^{-10}**	**1.06 (1.03–1.09) 8.1×10^{-5}**	1.04 (1.01–1.07) 0.018	**1.05 (1.03–1.07) 5.3×10^{-6}**	1.02 (0.98–1.05) 0.32	**1.04 (1.02–1.06) 6.3×10^{-5}**
Summary: number replicated loci				**5/32**	**0/32**	**4/32**	**3/32**	**5/32**

Significant associations are in bold in shaded boxes. CVD, cerebrovascular disease; CHR, chromosome; N, number.

† MEGASTROKE transethnic meta-analyses included 60,341 ischaemic stroke cases; 9,006 cardio-embolic stroke cases; 6,688 large-vessel stroke cases. MEGASTROKE European meta-analyses included 34,217 ischaemic stroke cases; 7,193 cardio-embolic stroke cases; 4,373 large-vessel stroke cases.

estimated effect of a marker allele from the initial study reporting the marker-allele association is often exaggerated relative to the estimated effect in follow-up studies).

The correlation of all definitions with the MEGASTROKE study results suggests one or more of the following: (i) all definitions retrieve true-positive acute stroke cases, meaning that broad CVD codes include additional true-positive cases not identified by stroke-specific codes; (ii) cases coded with a broad CVD code have not necessarily suffered an acute stroke, but represent a range of phenotypes with a similar genetic architecture to acute stroke [e.g., previous research has shown at least one overlapping locus for carotid artery disease and acute stroke (10)]; (iii) the MEGASTROKE study includes some misclassified broad CVD cases as false-positive acute stroke cases. It is most likely that a combination of these factors is contributing to our findings, but we are unable to dissect their separate contributions in the current study.

Previous case-note validation studies suggested that broad CVD codes are better at identifying the broad conditions they signify as opposed to ascertaining acute stroke cases (14), supporting a role for option two above. An example of this would be a case-note review of patients with a code for an unruptured intracranial aneurysm or carotid artery stenosis confirming that the diagnosis was also most likely an unruptured intracranial aneurysm or carotid artery stenosis, rather than the reviewing clinician deciding it was an acute stroke that had been miscoded as an unruptured intracranial aneurysm or carotid artery stenosis.

Despite the definition using any code or self-reported event increasing the sample size by more than 2-fold compared to the definition using only stroke-specific codes from a medical setting, it did not replicate a higher number of known stroke-associated loci. This could suggest that there is still insufficient power to replicate additional loci using any of our definitions despite power calculations suggesting ≥80% power for all loci. Also, associations for nine of the 32 loci in MEGASTROKE were only found for specific stroke subtypes and 11 of the 32 loci were significant in analyses including only ischemic stroke cases, the proportions of which are unlikely to be identical between the two datasets. For example, the MEGASTROKE study sample included 90% confirmed ischemic stroke cases. The stroke subtype breakdown among the UKB participants is available only for stroke-specific codes from the hospital, death record, and self-reported data (UK Biobank data fields "42009," "42011," and "42013") and shows a proportion of confirmed ischemic stroke cases of 47%, with 10% cases being intracerebral hemorrhage and 11% subarachnoid hemorrhage and the remainder of unspecified stroke subtype. Furthermore, case-note validation suggests that while ischemic stroke cases can be identified with good accuracy using stroke-specific codes, further work is needed to understand the accuracy of hemorrhagic stroke codes (3). Alternatively, it could also suggest that the additional cases identified by using any code or self-report are not true-positive stroke cases or that some of these known stroke-associated loci are false-positive findings. Finally, the absence of primary care data for half of the cohort will have reduced the negative predictive value of the Read Version 2 and 3 code lists in this study, and hence to an extent reduced our power to replicate known stroke-associated loci.

Self-reported cases (both stroke-specific and broad CVD) also showed a close genetic correlation with the MEGASTROKE study, supporting the use of self-report as a means of identifying additional stroke cases in the UKB. This was so despite the highly variable results from the previous case-note-based validation studies of self-report for ascertaining stroke cases. Studying this by case-note validation in UKB itself would be challenging, given the difficulties accessing NHS records which predate recruitment by many years and the fact that participants may have moved between UK regions during their life-course. Other studies have

also reported a close genetic correlation between a wide range of self-reported diseases and medical setting diagnoses. Examples include both acute and chronic conditions (e.g., depression, myocardial infarction, rheumatoid arthritis) (15–17).

We used the HDL method for assessing genetic correlations to fully account for LD across the genome and improve precision in genetic correlation estimation. Compared to the LD Score Regression method, HDL reduces the variance of genetic correlation estimates by about 60%, equivalent to a 2.5-fold increase in sample size (11). For some of our definitions, the rg was >1. The estimated rg is a combination of the true rg and variation. When the true rg is close to the boundary (-1 or 1) and/or variation is large, the estimated rg can go beyond the boundary (11). In rg estimation, some common reasons for generating large variation are: (i) at least one of the h2 estimates is very low; (ii) small sample size; (iii) many SNPs in the reference panel are absent in one of the two GWASs; (iv) there is a severe mismatch between the GWAS population and the population for computing reference panel. We can exclude the last two options, and the small sample size is, therefore, the likely explanation.

In our analyses, the case-control groups were partially overlapping and a specific UKB participant could therefore be a stroke case for one definition and control for another definition. We used this study design to mimic the "real world" situation, creating binary case-control definitions based on each code list. In theory, this could reduce the power of some of the analyses, since it means controls can end up including some true-positive stroke cases. However, in reality, it is unlikely to have a significant effect given the overall large number of controls. For example, for analyses using stroke-specific codes from any medical setting, just over 2% of controls have a broad CVD code and/or have self-reported a stroke event.

We used BoltLMM for running the GWAS (6). Our case fraction ranged from 1 to 5% depending on the case definition and we limited our analyses to SNPs with a minor allele frequency of at least 1%. Based on simulations done using BoltLMM, the authors of the software suggest that with this case-fraction and minor allele frequency parameters, they did not find statistically significant inflation of the type I error rates [**Supplementary Table 8** in (5)].

The strengths of our study are: (i) we included-and have made available to re-use-a clinically informed, comprehensive set of codes across all relevant coding systems; (ii) we compared our results against the largest stroke GWAS to date; (iii) we used multiple methods for comparison accounting for both GWAS significant loci but also SNPs across the whole genome–i.e., correlation and replication; (iv) we have added novel data to what is already known from case-note validation.

Our study also has some limitations: (i) some of our definitions included relatively small case numbers compared to the MEGASTROKE study, reducing our power to replicate known loci; (ii) uneven numbers across definitions not allowing direct comparisons, but rather reflecting the real-world situation; (iii) our definitions included the subarachnoid hemorrhage stroke subtype codes, whereas the MEGASTROKE study did not, resulting in a slightly different mix of stroke cases; (iv) the UKB participants' demographic characteristics differ from those of the UK general population with evidence of a healthy-volunteer selection bias, which needs to be considered when extrapolating these results to other settings (18); and, (v) some controls are likely to experience a stroke during follow up in the future, which may have reduced study power.

We have shown that the selection of codes and code sources used to ascertain stroke cases has a major impact on the overall stroke case numbers in the UKB. Given the close genetic correlation between stroke cases identified using broad CVD codes, self-report, and physician-confirmed stroke cases, we suggest that for studies accepting more crude stroke and cerebrovascular disease outcomes, researchers may wish to include all codes and self-reported events for increased power. Alternatively, this information is also helpful in informing the selection of controls for various studies. Including a large number of broad CVD coded cases among controls might weaken any association seen for certain study designs. However, since we cannot exclude the effects of shared genetic control of broad CVD phenotypes and acute stroke, this evidence is not sufficient to support using broad CVD codes in studies that need to define acute stroke outcomes very accurately (e.g., clinical trials).

Further research is needed: to better understand the underlying reasons for the close genetic correlation between stroke-specific and broad CVD codes; to dissect the underlying explanation for our results with targeted case-note validation; to replicate our results in other datasets. In addition, more data is needed on the accuracy of different coding systems for identifying specific pathological stroke subtypes (ischemic stroke vs. intracerebral hemorrhage vs. subarachnoid hemorrhage) and etiological stroke subtypes (e.g., small vessel disease vs. large artery disease vs. cardioembolic stroke vs. other/unknown cause).

AUTHOR'S NOTE

This work has been published as a preprint on MEDRXIV (MEDRXIV/2021/264348-Version 1).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UK Biobank project 2532 approval. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, drafting the work or revising it critically

for important intellectual content, final approval of the version to be published, and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

KR was funded by Health Data Research UK Rutherford fellowship MR/S004130/1. AF was funded by BHF award RE/18/5/34216 and MR/S004130/1. This work was supported by the Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund. AT was funded by HDR-UK awards HDR-9004 and HDR-9003. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. The MEGASTROKE project received funding from sources specified at https://www.megastroke.org/acknowledgements.html and the author list for the MEGASTROKE consortium is added in **Appendix 1**.

ACKNOWLEDGMENTS

This work was undertaken under a UKB project 2532 UK Biobank Stroke Study (UKBiSS): Developing an in-depth understanding of the determinants of stroke and its subtypes. The authors acknowledge Dr. Spiros Denaxas, UCL, for his valuable comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2021.787107/full#supplementary-material

REFERENCES

1. Sudlow CLM, Gallacher J, Allen N, Beral V, Burton B, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. (2015) 31:12(3):e1001779. doi: 10.1371/journal.pmed.1001779
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *Lancet*. (2010) 380:2095–128. doi: 10.1016/S0140-6736(12)61728-0
3. Rannikmäe K, Ngoh K, Bush K, Al-Shahi Salman R, Doubal F, Flaig R, et al. Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology*. (2020) 95:e697–707. doi: 10.1212/WNL.0000000000009924
4. Liberman AL, Rostanski SK, Ruff IM, Meyer AND, Maas MB, Prabhakaran S. Inter-rater agreement for the diagnosis of stroke versus stroke mimic. *Neurologist*. (2018) 23:118–21. doi: 10.1097/NRL.0000000000000187
5. Manolio TA, Goodhand P, Ginsburg G. The international hundred thousand plus cohort consortium: integrating large-scale cohorts to address global scientific challenges. *Lancet Digital Health*. (2020) 2:e567–8. doi: 10.1016/S2589-7500(20)30242-9
6. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. (2015) 47:284–90. doi: 10.1038/ng.3190
7. Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio. *Genetics*. (2018) 208:1397–408. doi: 10.1534/genetics.117.300360
8. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*. (2011) 88:294–305. doi: 10.1016/j.ajhg.2011.02.002
9. Speed D, Holmes J, and Balding DJ. Evaluating and improving heritability models using summary. *Statistics*. (2020) 52:458–462. doi: 10.1038/s41588-020-0600-y
10. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet*. (2018) 50:524–37. doi: 10.1038/s41588-018-0058-3
11. Ning Z, Pawitan Y, Shen X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat Genet*. (2020) 52:859–64. doi: 10.1038/s41588-020-0653-y
12. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics - 2020 update: a report from the american heart association. *Circulation*. (2020) 141:e139. doi: 10.1161/CIR.0000000000000746
13. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*. (2014) 9:1192–212. doi: 10.1038/nprot.2014.071
14. McCormick N, Bhole V, Lacaille D, Avina-Zubieta JA. Validity of diagnostic codes for acute stroke in administrative databases: a systematic review. *PLoS ONE*. (2015) 10:e0135834. doi: 10.1371/journal.pone.0135834
15. Wray NR, Ripke S, Mattheien M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*. (2018) 50:668–81.
16. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*. (2019) 22:343–52.
17. DeBoever C, Tanigawa Y, Aguirre M, McInnes G, Lavertu A, Rivas MA. Assessing digital phenotyping to enhance genetic studies of human diseases. *Am J Hum Genet*. (2020) 106:611–22. doi: 10.1016/j.ajhg.2020.03.007
18. Fry A, Littlejohns TJ, Sudlow CLM, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *Am J Epidemiol*. (2017) 186:1026–34. doi: 10.1093/aje/kwx246
19. Woodfield R, Grant, I. UK Biobank Stroke Outcomes group, UK Biobank Follow-Up and Outcomes Working Group, Sudlow CLM. Accuracy of Electronic Health Record Data for Identifying Stroke Cases in Large-Scale Epidemiological Studies: A Systematic Review from the UK Biobank Stroke Outcomes Group. *PLoS One*. (2015) 10:e0140533. doi: 10.1371/journal.pone.0140533

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.