OPEN

Genetic Predisposition to Ischemic Stroke
A Polygenic Risk Score

Tsuyoshi Hachiya, PhD; Yoichiro Kamatani, MD; Atsushi Takahashi, PhD; Jun Hata, MD; Ryohei Furukawa, PhD; Yuh Shiwa, PhD; Taiki Yamaji, MD; Megumi Hara, MD; Kozo Tanno, MD; Hideki Ohmomo, PhD; Kanako Ono, BSc; Naoyuki Takashima, MD; Koichi Matsuda, MD; Kenji Wakai, MD; Norie Sawada, MD; Motoki Iwasaki, MD; Kazumasa Yamagishi, MD; Tetsuro Ago, MD; Toshiharu Ninomiya, MD; Akimune Fukushima, MD; Atsushi Hozawa, MD; Naoko Minegishi, MD; Mamoru Satoh, MD; Ryujin Endo, MD; Makoto Sasaki, MD; Kiyomi Sakata, MD; Seiichiro Kobayashi, MD; Kuniaki Ogasawara, MD; Motoyuki Nakamura, MD; Jiro Hitomi, MD; Yoshikuni Kita, PhD; Keitaro Tanaka, MD; Hiroyasu Iso, MD; Takanari Kitazono, MD; Michiaki Kubo, MD; Hideo Tanaka, MD; Shoichiro Tsugane, MD; Yutaka Kiyohara, MD; Masayuki Yamamoto, MD; Kenji Sobue, MD; Atsushi Shimizu, PhD

Background and Purpose—The prediction of genetic predispositions to ischemic stroke (IS) may allow the identification of individuals at elevated risk and thereby prevent IS in clinical practice. Previously developed weighted multilocus genetic risk scores showed limited predictive ability for IS. Here, we investigated the predictive ability of a newer method, polygenic risk score (polyGRS), based on the idea that a few strong signals, as well as several weaker signals, can be collectively informative to determine IS risk.

Methods—We genotyped 13 214 Japanese individuals with IS and 26 470 controls (derivation samples) and generated both multilocus genetic risk scores and polyGRS, using the same derivation data set. The predictive abilities of each scoring system were then assessed using 2 independent sets of Japanese samples (KyushuU and JPJM data sets).

Results—In both validation data sets, polyGRS was shown to be significantly associated with IS, but weighted multilocus genetic risk scores was not. Comparing the highest with the lowest polyGRS quintile, the odds ratios for IS were 1.75 (95% confidence interval, 1.33–2.31) and 1.99 (95% confidence interval, 1.19–3.33) in the KyushuU and JPJM samples, respectively. Using the KyushuU samples, the addition of polyGRS to a nongenetic risk model resulted in a significant improvement of the predictive ability (net reclassification improvement=0.151; $P<0.001$).

Conclusions—The polyGRS was shown to be superior to weighted multilocus genetic risk scores as an IS prediction model. Thus, together with the nongenetic risk factors, polyGRS will provide valuable information for individual risk assessment and management of modifiable risk factors. **(*Stroke*. 2017;48:253-258. DOI: 10.1161/STROKEAHA.116.014506.)**

Key Words: genome-wide association study ■ genotype ■ risk assessment ■ stroke

Received June 23, 2016; final revision received November 21, 2016; accepted November 28, 2016.

From the Division of Biomedical Information Analysis (T.H., R.F., Y.S., H.O., K. Ono, M. Satoh, A.S.), Division of Biobank and Data Management (T.H., Y.S., M. Satoh), Division of Clinical Research and Epidemiology (K. Tanno, K. Sakata), Division of Innovation and Education (A.F.), Division of Community Medical Supports and Health Record Informatics (M. Satoh), and Division of Public Relations and Planning (R.E.), Iwate Tohoku Medical Megabank Organization (M. Sasaki, S.K., K. Ogasawara, M.N., J. Hitomi, K. Sobue), Iwate Medical University, Japan; Laboratory for Statistical Analysis (Y. Kamatani, A.T.), RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan (M.K.); Laboratory for Omics Informatics, Omics Research Center, National Cerebral and Cardiovascular Center, Osaka, Japan (A.T.); Department of Environmental Medicine (J. Hata), Department of Medicine and Clinical Science (J. Hata, T.A., T.K.), and Center for Cohort Studies (J. Hata, T.N., T.K.), Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; Epidemiology and Prevention Group, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan (T.Y., N.S., M.I., S.T.); Department of Preventive Medicine, Faculty of Medicine, Saga University, Japan (M.H., K. Tanaka); Department of Public Health, Shiga University of Medical Science, Japan (N.T., Y. Kita); Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, Japan (K.M.); Department of Preventive Medicine (K.W.) and Department of Epidemiology (H.T.), Nagoya University Graduate School of Medicine, Japan; Department of Public Health Medicine, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan (K.Y.); Department of Preventive Medicine and Epidemiology (A.H.), Department of Biobank (N.M.), and Department of Integrative Genomics (M.Y.), Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan; Faculty of Nursing Science, Tsuruga Nursing University, Fukui, Japan (Y. Kita); Public Health, Department of Social Medicine, Osaka University Graduate School of Medicine, Japan (H.I.); Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan (H.T.); and Hisayama Research Institute for Lifestyle Diseases, Fukuoka, Japan (Y. Kiyohara).

The online-only Data Supplement is available with this article at http://stroke.ahajournals.org/lookup/suppl/doi:10.1161/STROKEAHA.116.014506/-/DC1.

Correspondence to Tsuyoshi Hachiya, PhD, Iwate Tohoku Medical Megabank Organization, Iwate Medical University, 2-1-1 Nishitokuta, Yahaba, Shiwa, Iwate 028-3694, Japan, E-mail thachiya@iwate-med.ac.jp or Atsushi Shimizu, PhD, Iwate Tohoku Medical Megabank Organization, Iwate Medical University, 2-1-1 Nishitokuta, Yahaba, Shiwa, Iwate 028-3694, Japan, E-mail ashimizu@iwate-med.ac.jp

Ischemic stroke (IS) is a leading cause of death and long-term disability in the world.[1] Although a large proportion of IS events could be prevented by appropriate management of modifiable risk factors, such as high blood pressure and tobacco use,[2] the burden attributable to these modifiable risk factors remains problematic.[3] To lower this burden, it is important to apply both population and high-risk approaches.[4] Genetic information can be a useful tool for the identification of high-risk individuals.

The effects of individual genetic markers are relatively small for common polygenic disorders, and therefore, a well-studied approach, weighted multilocus genetic risk score (wGRS), typically integrates tens of weak genetic markers into a single risk score, based on summary statistics from genome-wide association (GWA) studies.[5] For hypertension and coronary artery disease, wGRSs have been derived from GWA data of the target traits.[5] However, previous IS GWA studies have identified only a few replicable susceptibility loci,[6–9] possibly because of the etiologic heterogeneity of IS. Therefore, wGRSs for IS in previous studies have been derived from GWA data on hypertension,[10] atrial fibrillation,[11] and coronary artery disease,[12–14] and their predictive abilities have been limited.[10–14]

We hypothesized that the lower predictive abilities of wGRSs are related to the polygenic nature of IS. A previous analysis on international IS GWA data inferred that genetic predispositions to IS are shared among the different subtypes,[15] which may be related to susceptibilities to arteriosclerosis, hypertension, hyperlipidemia, and their combinations. Thus, we sought to develop a statistical model to predict the genetic predispositions shared among IS subtypes, rather than identify genetic markers specific to IS subtypes. To accomplish this, we created a polygenic risk score (polyGRS) based on the assumption that, in addition to a few genome-wide signals obtained from IS GWA data, numerous weaker signals can be collectively informative for predicting IS incidence. In previous studies, polyGRSs showed remarkable predictive abilities for schizophrenia, bipolar disorder, hypertension, and coronary artery disease.[16,17] However, the predictive ability of polyGRS for IS remains to be determined. The complexity of IS subtypes requires the validation of the genetic model; therefore, we derived both a wGRS and polyGRS from the same large-scale derivation samples and compared their predictive abilities in 2 sets of independent samples.

Methods

Cohorts and Case Definition

In this study, we used 3 sets of Japanese samples. The first set (derivation data set) was used to derive a wGRS and polyGRS. The second set (KyushuU data set) was used to assess the predictive abilities of the 2 GRSs with detailed clinical information. The third set (JPJM data set) was used as an additional data set for the validation of the predictive abilities. All 3 sets of samples were independent from each other.

For the derivation data set, patients with IS were recruited by the BioBank Japan Project from 2003 to 2008.[18] All participants provided written informed consent, as approved by the ethical committees of the BioBank Japan Project and the University of Tokyo. Clinical information on the subjects was collected from medical charts, neuroimaging results (including computed tomography and magnetic resonance imaging), and self-reported questionnaires. Controls in the derivation data set were enrolled from participants in Japanese prospective cohort studies, including the Tohoku Medical Megabank Project,[19] the Japan Public Health Center–based prospective study,[20] and the Japan Multi-Institutional Collaborative Cohort Study.[21] Details of the study design and recruitment methods of the 3 cohort studies were described previously.[19–21]

For the KyushuU data set, details of the recruitment methods and diagnostic criteria were described previously.[6,22] Briefly, affected individuals with IS were recruited from 7 hospitals affiliated with Kyushu University in 2004. For all cases, diagnoses of IS and its subtypes were made by stroke neurologists from the affiliated hospitals by referencing clinical presentation and ancillary laboratory examinations—namely, cerebral angiography, brain imaging, echocardiography, and carotid duplex imaging. Participants in the Hisayama study were enrolled as control subjects. The Hisayama study is a population-based cohort study established in 1961.[23,24] Of 3328 Hisayama residents aged ≥40 years who consented to participate in the Hisayama study between 2002 and 2003, we selected age-matched (within 5 years) and sex-matched control subjects by 1:1 matching using random numbers, after excluding subjects with a history of stroke or coronary heart disease. For the subjects in the KyushuU data set, hypertension was defined as systolic blood pressure ≥140 mm Hg and diastolic blood pressure ≥90 mm Hg on at least 3 different occasions or as current treatment with antihypertensive drugs.[25] Diabetes mellitus was determined by a 75-g oral glucose tolerance test, casual blood glucose levels (>11.1 mmol/L), or a medical history of diabetes mellitus. Hyperlipidemia was defined as a total cholesterol level ≥5.69 mmol/L or current treatment with a cholesterol-lowering drug. Atrial fibrillation was diagnosed based on electrocardiographic findings.

For the JPJM data set, nested case–control subjects were enrolled from the participants of the Japan Public Health Center and Japan Multi-Institutional Collaborative Cohort studies in the Saga and Takashima regions.[20,21,26] IS cases were confirmed by imaging studies, and age- and sex-matched controls were extracted from a pool of individuals with no history of stroke.[27]

Genotyping

All subjects from the 3 data sets were genotyped using a HumanOmniExpressExome BeadChip array (Illumina, Inc, San Diego, CA).

Quality Control Filters for Derivation Samples

Samples with low call rate (<0.98), single-nucleotide polymorphisms (SNPs) with low call rate (<0.99), and close relationships characterized by the identity-by-state method were excluded, as well as subjects whose estimated ancestries outside of the Hondo cluster of the Japanese population[28] by PCA.[29,30] Variants with a Hardy–Weinberg equilibrium exact test P value of $<1\times10^{-6}$ and a minor allele frequency <0.01 were also excluded. Ultimately, 39 684 individuals (Table 1) with 537 999 autosomal SNPs were included in our analyses.

Genetic Risk Score Derivation

Our statistical analysis workflow is shown in Figure 1. The wGRS included 5 SNPs selected at the end of the replication and exploratory analyses from the derivation samples.

Additional quality control filters were applied to the derivation data set to generate the polyGRS: (1) samples with a call rate of <0.99, (2) SNPs with a Hardy–Weinberg equilibrium exact test P value of <0.05, (3) SNPs with a P value in the test for nonignorable difference between cases and controls of <0.05 were excluded, according to a previous study.[31] Ultimately, no individuals were excluded by these filters, and 357 367 autosomal SNPs were retained.

According to a previous study,[31] polyGRS by genotyped data shows higher predictive ability than the model using all imputed data. Therefore, we also used the genotyped variants (357 367 variants) to generate the polyGRS. As such, we used a dual-formula technique to minimize overfitting to the derivation data set. The polyGRS was generated via 2 models: (1) restricted maximum likelihood and (2) best

Table 1. Age and Sex Distributions of the Derivation, KyushuU, and JPJM Samples

	Derivation		KyushuU		JPJM	
	Cases	Controls	Cases	Controls	Cases	Controls
Study design	Case–Population		Case–Control		Nested Case–Control	
Subjects	13214	26470	1097	1097	336	336
Age, y*, mean±SD	69±10	56±10	70±10	70±10	59±7	59±7
Women, %	35.7	60.6	39.1	39.1	37.5	37.5

*Age at enrollment.

linear unbiased prediction. A mixed linear model was assumed in both cases,[31] which accounts for genotype data via the genetic relationship matrix, leaving only the variances of genetic and nongenetic effects as free parameters. More specifically, the restricted maximum likelihood model estimates the 2 free parameters based on derivation samples (N=39684), whereas the best linear unbiased prediction model converts the variance parameter estimates into the weight parameters for the 357367 SNPs in conjunction with the genotype data of derivation samples. In total, only 2 parameters were estimated through the restricted maximum likelihood and best linear unbiased prediction steps. Further details are provided in the online-only Data Supplement.

Predictive Ability Assessment

Because previous studies on wGRS typically evaluated the odds ratio (OR) of the highest score quintile versus the lowest score quintile,[11,12] we estimated the OR for each score quintile using Fisher exact test. In addition, the OR per 1 SD, the corresponding 95% confidence interval (CI), and the overall *P* value were estimated by the conditional logistic regression analysis. To compare the predictive abilities of the 2 GRSs, a continuous version of net reclassification improvement (NRI), an integrated discrimination improvement, and the C-index were calculated. The NRI, integrated discrimination improvement, and C-index were also calculated to assess the improvements in predictive ability obtained by adding the 2 GRSs to a nongenetic risk model.

Results

Derivation of Genetic Risk Scores

Of the 6 well-studied variants, the associations with IS were nominally significant regarding 4 variants (rs6843082 [*PITX2*]; rs2383207 [*CDKN2B-CDKN2A*]; rs2107595 [*HDAC9*]; and rs879324 [*ZFHX3*]) in our derivation samples (Tables I and II in the online-only Data Supplement).

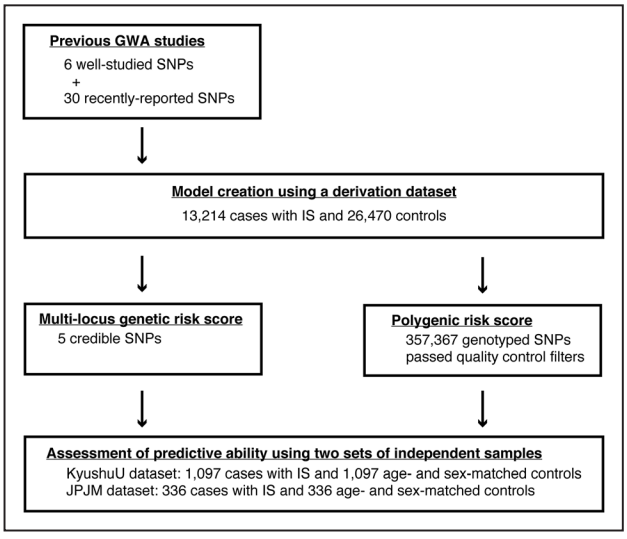


Figure 1. Statistical analysis flow diagram. GWA indicates genome-wide association; IS, ischemic stroke; and SNP, single-nucleotide polymorphism.

For exploratory identification of IS-susceptibility loci, based on our derivation samples, the associations between 6204347 imputed genetic variants and IS were subjected to a genome-wide analysis (Figure I and Table III in the online-only Data Supplement). The results showed genome-wide significance ($P<5\times10^{-8}$) for one variant, rs1275923, which is located in an intron of *KCNK3* (Figure II in the online-only Data Supplement) and is reportedly associated with blood pressure.[32]

Based on these results, we chose 5 variants as model variables for our wGRS (Table IV in the online-only Data Supplement). Additionally, from the same derivation data set, we created a polyGRS, based on the assumption that, in addition to several genome-wide signals, numerous weaker signals can be collectively informative for the prediction of IS incidence. All genotyped variants that passed our quality control filters (357367 variants) were used to derive the polyGRS.

Predictive Ability of Genetic Risk Scores

The predictive abilities of the wGRS and polyGRS were assessed using the KyushuU and JPJM samples. The power to detect significance with an OR of 1.2 per 1 SD was 99% and 66% for the KyushuU and JPJM data sets, respectively (Table 2). This was comparable to the OR of 1.17 per SD at P_T=1 by the score-profiling model (Table V in the online-only Data Supplement), indicating that all genotyped variants in the polyGRS contributed useful signal to the score.

Although the polyGRS significantly associated with IS in both validation data sets, this was not observed with wGRS (Table 2). Moreover, the OR for the highest polyGRS quintile compared with the lowest polyGRS quintile was 1.75 (95% CI, 1.33–2.31) and 1.99 (95% CI, 1.19–3.33) in the KyushuU and JPJM samples, respectively (Figure 2), with a significant improvement observed in the KyushuU samples (NRI=0.179; Figure 3; Table VI in the online-only Data Supplement).

Predictive Ability of Genetic Risk Scores for Each Etiologic Subtype

In the KyushuU samples, the predictive abilities of the 2 GRSs were investigated for each etiologic subtype: large-vessel disease, small-vessel disease, and cardioembolic stroke (Figure 3). The wGRS failed to associate with any subtype, whereas the polyGRS was significantly associated with all 3. Furthermore, the predictive ability of the subtype-mixture

Table 2. Predictive Ability of the Multilocus and the Polygenic Risk Scores in the KyushuU and JPJM Samples

Validation Samples	Model	Q1 OR (95% CI)	Q2 OR (95% CI)	Q3 OR (95% CI)	Q4 OR (95% CI)	Q5 OR (95% CI)	OR per SD* (95% CI)	Overall* *P* Value	C-index (95% CI)
KyushuU (N=2194)	wGRS	Reference	1.09 (0.83–1.43)	1.08 (0.82–1.42)	1.03 (0.78–1.35)	1.17 (0.89–1.54)	1.04 (0.96–1.14)	0.313	0.510 (0.486–0.534)
	polyGRS	Reference	1.08 (0.82–1.42)	1.10 (0.84–1.45)	1.41 (1.08–1.86)†‡	1.75 (1.33–2.31)†‡	1.20 (1.10–1.31)†‡	<0.001‡	0.555 (0.531–0.579)†‡
JPJM (N=672)	wGRS	Reference	1.37 (0.82–2.28)	2.02 (1.21–3.39)†‡	1.39 (0.83–2.32)	1.45 (0.87–2.42)	1.11 (0.96–1.29)	0.172	0.530 (0.487–0.574)
	polyGRS	Reference	1.96 (1.18–3.29)†‡	1.69 (1.01–2.83)‡	1.33 (0.80–2.23)	1.99 (1.19–3.33)†‡	1.20 (1.01–1.41)‡	0.033‡	0.536 (0.492–0.580)

CI indicates confidence interval; OR, odds ratio; polyGRS, polygenic risk score; Q1–Q5, quantiles 1–5; and wGRS, weighted multilocus genetic risk score.
*Considering the genetic risk scores as continuous variables.
†Significant after multiple corrections.
‡Results are nominally significant ($P<0.05$).

polyGRS was higher than that of the subtype-specific model (Table VII in the online-only Data Supplement).

Integration of Genetic Risk Scores Into a Nongenetic Risk Model

Based on the KyushuU samples, the wGRS or polyGRS was added to a nongenetic risk model that included hypertension, diabetes mellitus, hyperlipidemia, and atrial fibrillation as model variables. Notably, the polyGRS showed an improved predictive ability, whereas the wGRS did not (Figure 4; Table VIII in the online-only Data Supplement). For all IS cases, the polyGRS NRI was estimated to be 0.151 (95% CI, 0.068–0.235), the integrated discrimination improvement was 0.004 (95% CI, 0.001–0.006), and the ΔC-index was 0.700 (95% CI, 0.679–0.722). When stratified by subtype, the NRI was significant for large-vessel disease and small-vessel disease but not cardioembolic stroke.

Discussion

Based on large-scale genotyping of Japanese samples (N>40 000 in total), we showed that the polyGRS has a superior ability to predict IS compared with the wGRS. Given that the estimates of ORs of the highest score quintile compared with the lowest score quintile were at most ≈1.3 in previous studies,[8–12] the higher OR for polyGRS (>1.75) indicated that the polygenic approach was well suited to derive genetic risk scores for IS. The difference in incident risk between the highest and lowest score quintiles (>75% difference)—which was consistently estimated from both retrospective (case–control) and prospective (nested case–control) designs—supports the utility of polyGRS methodology for healthcare and preventive purposes. Moreover, the polyGRS significantly associated with all etiologic subtypes, suggesting that it effectively predicts the genetic predispositions shared among IS subtypes. Furthermore, the significant NRI estimates for all IS cases indicated that the integration polyGRS into nongenetic risk models would be valuable, whereas the nonsignificant NRI for cardioembolic stroke suggested that the adjustment by atrial fibrillation and other nongenetic factors attenuated the predictive ability of the polyGRS.

One aspect of the clinical utility of the polyGRS is the individualization of clinical criteria. Japanese clinical guidelines for the management of hypertension published in 2014 recommended the grouping of patients with hypertension into risk strata based on blood pressure levels and other cardiovascular risk factors, including age, smoking, dyslipidemia, obesity, diabetes mellitus, and family history of young-onset cardiovascular disease.[33] For each risk stratum, a distinct therapeutic strategy was recommended. Here, we showed that polyGRS may be a valuable predictor of IS. Accordingly, hypertensive patients whose risks have been underestimated without the additional genetic information would be reclassified into higher risk strata after including polyGRS as a part of the clinical criteria.

To advance further, prospective cohort studies on the predictive ability of the polyGRS would be essential. It would also be interesting to investigate whether the polyGRS derived from Japanese samples can predict IS in other East Asians and other ethnicities. Additionally, ethical, legal, social, and policy issues, including the responsibility for the management of the genetic information, should be discussed in future studies.

The difference between the predictive abilities of polyGRS and wGRS elucidates the polygenic nature of IS. An important methodological difference between wGRS and polyGRS

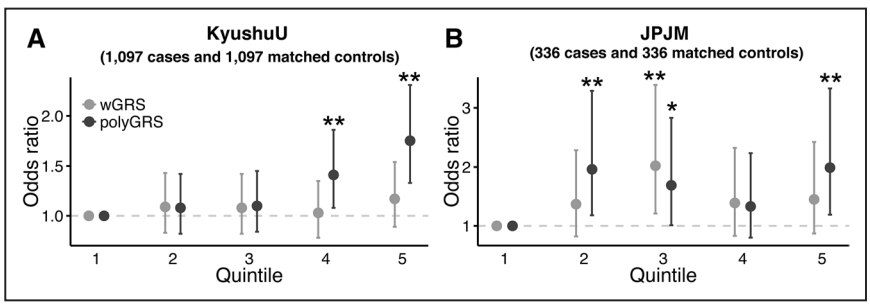


Figure 2. Predictive ability of the weighted multilocus genetic risk scores (wGRS) and polygenic risk scores (polyGRS) in the (**A**) KyushuU and (**B**) JPJM samples. Odds ratio for each score quintile compared with the lowest score quintile. *Nominal significant odds ratios ($P<0.05$). **Significant odds ratios after multiple corrections.

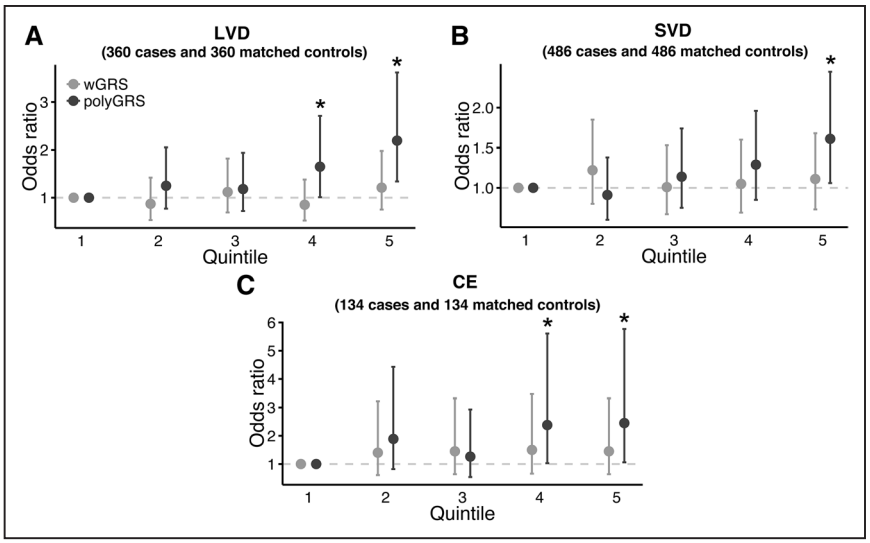


Figure 3. Predictive ability of the weighted multilocus genetic risk scores and polygenic risk scores for each etiologic subtype in the KyushuU samples. Odds ratio for each score quintile compared with the lowest score quintile. **A**, Large-vessel disease (LVD). **B**, Small-vessel disease (SVD). **C**, Cardioembolic stroke (CE). *Significant odds ratios.

is that the wGRS only included 5 credible SNPs as model variables, whereas polyGRS used all genotyped SNPs. The superior predictive ability of polyGRS demonstrates the validity of our assumption that, in addition to a few genome-wide signals, numerous weaker signals are collectively informative for predicting IS. Furthermore, in the score-profiling model, the predictive ability improved as P_T threshold increased. This result implies that usage of more SNPs is essential for improvement of the predictive ability. Our results suggest that a large number of IS-susceptibility loci with small effect size have yet to be discovered and that larger derivation and replication of GWA data sets would be advantageous for discovering novel susceptibility loci in future studies. Furthermore, given that we are currently unable to identify all IS-susceptibility variants with small effect size, polyGRS approach may represent a fascinating method to use the valuable information contained in weak GWA signals to predict IS.

Conclusions

We demonstrated that the polyGRS approach is superior to that of wGRS as a method of choice for the assessment of IS genetic risks. This is clinically important because the polyGRS approach is promising for the individualization of clinical criteria in the era of precision medicine.

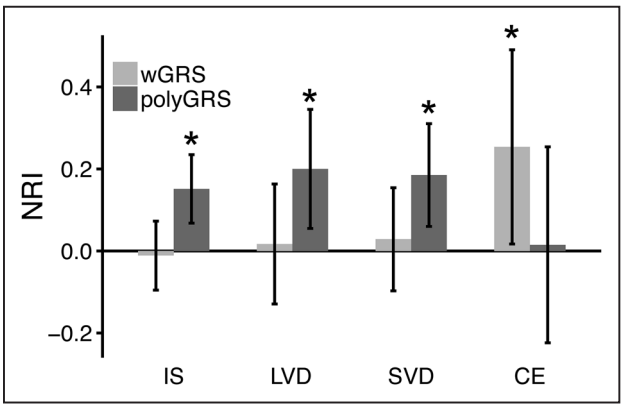


Figure 4. Predictive ability improvement offered by the addition of weighted multilocus genetic scores (wGRS) or polygenic risk score (polyGRS) to a nongenetic risk model in the KyushuU samples. CE indicates cardioembolic stroke; IS, ischemic stroke; LVD, large-vessel disease; NRI, net reclassification improvement; and SVD, small-vessel disease. *Significant differences.

Acknowledgments

We are grateful to the patients, participants, research co-ordinators, scientists, and physicians participating in the Tohoku Medical Megabank Project, the Japan Public Health Center (JPHC)–based study, the Japan Multi-Institutional Collaborative Cohort (J-MICC) study, the Hisayama study, and the BioBank Japan project and providing the phenotype data and DNA samples used in this study. We thank Dr Yukihide Momozawa, Kyota Ashikawa, and the staff of the Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, for genotyping samples from Biobank Japan and the Hisayama Study. We thank Dr Hideki Sugihara, Dr Masaharu Ichikawa, Dr Yutaka Morita, Dr Tanvir Chowdhury, Dr Aya Kadota, Dr Yuichiro Nishida, and Dr Chisato Shimanoe for their contributions to recruitment, enrollment, and data collection for J-MICC samples. We thank Dr Yoshihiro Kokubo (National Cerebral and Cardiovascular Center), Prof Dr Hiroshi Yatsuya (Fujita Health University) for contributing to data collection for JPHC samples.

Sources of Funding

The BioBank Japan Project was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government. The Japan Multi-Institutional Collaborative Cohort (J-MICC) study was supported by the Grants-in-Aid for Scientific Research (B), grant numbers 17390186, 20390184, 24390165, priority area grant number 17015018, and innovative area grant number 221S0001 of the MEXT and the Japan Society for the Promotion of Science. The Japan Public Health Center (JPHC)–based Prospective study was supported by National Cancer Center Research and Development Fund (23-A-31[toku] and 26-A-2; since 2011) and a Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare of Japan (from 1989 to 2010). This work was supported by a grant of the Tohoku Medical Megabank Project from the MEXT of Japan; Ministry of Health, Labour and Welfare of Japan; Japan Agency for Medical Research and Development.

Disclosures

None.

References

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries

in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380:2197–2223. doi: 10.1016/S0140-6736(12)61689-4.
2. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al; INTERSTROKE Investigators. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet*. 2010;376:112–123. doi: 10.1016/S0140-6736(10)60834-3.
3. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380:2224–2260. doi: 10.1016/S0140-6736(12)61766-8.
4. Bousser MG. Stroke prevention: an update. *Front Med*. 2012;6:22–34. doi: 10.1007/s11684-012-0178-6.
5. Smith JA, Ware EB, Middha P, Beacher L, Kardia SL. Current applications of genetic risk scores to cardiovascular outcomes and subclinical phenotypes. *Curr Epidemiol Rep*. 2015;2:180–190. doi: 10.1007/s40471-015-0046-4.
6. Kubo M, Hata J, Ninomiya T, Matsuda K, Yonemoto K, Nakano T, et al. A nonsynonymous SNP in PRKCH (protein kinase C eta) increases the risk of cerebral infarction. *Nat Genet*. 2007;39:212–217. doi: 10.1038/ng1945.
7. Bellenguez C, Bevan S, Gschwendtner A, Spencer CC, Burgess AI, Pirinen M, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet*. 2012;44:328–333.
8. Holliday EG, Maguire JM, Evans TJ, Koblar SA, Jannes J, Sturm JW, et al; Australian Stroke Genetics Collaborative; International Stroke Genetics Consortium; Wellcome Trust Case Control Consortium 2. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet*. 2012;44:1147–1151. doi: 10.1038/ng.2397.
9. Traylor M, Farrall M, Holliday EG, Sudlow C, Hopewell JC, Cheng YC, et al; Australian Stroke Genetics Collaborative, Wellcome Trust Case Control Consortium 2 (WTCCC2); International Stroke Genetics Consortium. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2012;11:951–962. doi: 10.1016/S1474-4422(12)70234-X.
10. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. 2011;478:103–109.
11. Tada H, Shiffman D, Smith JG, Sjögren M, Lubitz SA, Ellinor PT, et al. Twelve-single nucleotide polymorphism genetic risk score identifies individuals at increased risk for future atrial fibrillation and stroke. *Stroke*. 2014;45:2856–2862. doi: 10.1161/STROKEAHA.114.006072.
12. Yiannakouris N, Katsoulis M, Dilis V, Parnell LD, Trichopoulos D, Ordovas JM, et al. Genetic predisposition to coronary heart disease and stroke using an additive genetic risk score: a population-based study in Greece. *Atherosclerosis*. 2012;222:175–179. doi: 10.1016/j.atherosclerosis.2012.02.033.
13. Lövkvist H, Sjögren M, Höglund P, Engström G, Jern C, Olsson S, et al. Are 25 SNPs from the CARDIoGRAM study associated with ischaemic stroke? *Eur J Neurol*. 2013;20:1284–1291. doi: 10.1111/ene.12183.
14. Tragante V, Doevendans PA, Nathoe HM, van der Graaf Y, Spiering W, Algra A, et al; SMART Study Group. The impact of susceptibility loci for coronary artery disease on other vascular domains and recurrence risk. *Eur Heart J*. 2013;34:2896–2904. doi: 10.1093/eurheartj/eht222.
15. Holliday EG, Traylor M, Malik R, Bevan S, Falcone G, Hopewell JC, et al; Australian Stroke Genetics Collaborative; Wellcome Trust Case Control Consortium 2; International Stroke Genetics Consortium. Genetic overlap between diagnostic subtypes of ischemic stroke. *Stroke*. 2015;46:615–619. doi: 10.1161/STROKEAHA.114.007930.
16 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511:421–427.
17. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*. 2013;9:e1003264. doi: 10.1371/journal.pgen.1003264.
18. Nakamura Y. The BioBank Japan Project. *Clin Adv Hematol Oncol*. 2007;5:696–697.
19. Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, Osumi N, et al. The Tohoku Medical Megabank project: design and mission. *J Epidemiol*. 2016;26:493–511. doi: 10.2188/jea.JE20150268.
20. Tsugane S, Sawada N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn J Clin Oncol*. 2014;44:777–782. doi: 10.1093/jjco/hyu096.
21. Hamajima N; J-MICC Study Group. The Japan Multi-Institutional Collaborative Cohort study (J-MICC study) to detect gene-environment interactions for cancer. *Asian Pac J Cancer Prev*. 2007;8:317–323.
22. Kubo M, Kiyohara Y, Ninomiya T, Tanizaki Y, Yonemoto K, Doi Y, et al. Decreasing incidence of lacunar vs other types of cerebral infarction in a Japanese population. *Neurology*. 2006;66:1539–1544. doi: 10.1212/01.wnl.0000216132.95207.b4.
23. Kiyohara Y, Kubo M, Kato I, Tanizaki Y, Tanaka K, Okubo K, et al. Ten-year prognosis of stroke and risk factors for death in a Japanese community: the Hisayama study. *Stroke*. 2003;34:2343–2347. doi: 10.1161/01.STR.0000091845.14833.43.
24. Kubo M, Kiyohara Y, Kato I, Tanizaki Y, Arima H, Tanaka K, et al. Trends in the incidence, mortality, and survival rate of cardiovascular disease in a Japanese community: the Hisayama study. *Stroke*. 2003;34:2349–2354. doi: 10.1161/01.STR.0000090348.52943.A2.
25. Hagiwara N, Kitazono T, Kamouchi M, Kuroda J, Ago T, Hata J, et al. Polymorphisms in the lymphotoxin alpha gene and the risk of ischemic stroke in the Japanese population. The Fukuoka Stroke Registry and the Hisayama Study. *Cerebrovasc Dis*. 2008;25:417–422. doi: 10.1159/000121342.
26. Hara M, Higaki Y, Imaizumi T, Taguchi N, Nakamura K, Nanri H, et al. Factors influencing participation rate in a baseline survey of a genetic cohort in Japan. *J Epidemiol*. 2010;20:40–45.
27. Iso H, Noda H, Ikeda A, Yamagishi K, Inoue M, Iwasaki M, et al. The impact of C-reactive protein on risk of stroke, stroke subtypes, and ischemic heart disease in middle-aged Japanese: the Japan public health center-based study. *J Atheroscler Thromb*. 2012;19:756–766.
28. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, et al. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet*. 2008;83:445–456. doi: 10.1016/j.ajhg.2008.08.019.
29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–909. doi: 10.1038/ng1847.
30. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2:e190. doi: 10.1371/journal.pgen.0020190.
31. 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.
32. Kato N, Loh M, Takeuchi F, Verweij N, Wang X, Zhang W, et al; BIOS-consortium; CARDIo GRAMplusCD; LifeLines Cohort Study; InterAct Consortium. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet*. 2015;47:1282–1293. doi: 10.1038/ng.3405.
33. Shimamoto K, Ando K, Fujita T, Hasebe N, Higaki J, Horiuchi M, et al; Japanese Society of Hypertension Committee for Guidelines for the Management of Hypertension. The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2014). *Hypertens Res*. 2014;37:253–390. doi: 10.1038/hr.2014.20.

Genetic Predisposition to Ischemic Stroke: A Polygenic Risk Score

Tsuyoshi Hachiya, Yoichiro Kamatani, Atsushi Takahashi, Jun Hata, Ryohei Furukawa, Yuh Shiwa, Taiki Yamaji, Megumi Hara, Kozo Tanno, Hideki Ohmomo, Kanako Ono, Naoyuki Takashima, Koichi Matsuda, Kenji Wakai, Norie Sawada, Motoki Iwasaki, Kazumasa Yamagishi, Tetsuro Ago, Toshiharu Ninomiya, Akimune Fukushima, Atsushi Hozawa, Naoko Minegishi, Mamoru Satoh, Ryujin Endo, Makoto Sasaki, Kiyomi Sakata, Seiichiro Kobayashi, Kuniaki Ogasawara, Motoyuki Nakamura, Jiro Hitomi, Yoshikuni Kita, Keitaro Tanaka, Hiroyasu Iso, Takanari Kitazono, Michiaki Kubo, Hideo Tanaka, Shoichiro Tsugane, Yutaka Kiyohara, Masayuki Yamamoto, Kenji Sobue and Atsushi Shimizu

Stroke. 2017;48:253-258; originally published online December 29, 2016;
doi: 10.1161/STROKEAHA.116.014506
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

SUPPLEMENTAL MATERIAL

Genetic predisposition to ischemic stroke: a polygenic risk score

Supplementary Methods

Genotype imputation for the derivation samples

Genotype imputation was performed using MaCH[1] and MiniMac2[2] software based on the 1000 Genomes reference panel.[3] After imputation, variants with an $R^2 < 0.5$ and a MAF < 0.01 were excluded, resulting in 6,204,347 autosomal variants.

Genome-wide association tests using derivation samples

To investigate the associations of the 6,204,347 imputed variants with IS using derivation samples, *P*-values were calculated by a mixed linear model association (MLMA) method[4] with age- and sex-adjustment. In comparison with principal component-based adjustment methods, the MLMA method has advantages for the adjustment of population structure in derivation samples.[4] As expected, a quantile-quantile (QQ) plot showed that the inflation of the association *P*-values was well controlled in our genome-wide association tests (Supplementary Figure IV).

Coefficients estimated from the MLMA method cannot be converted to odds ratios (ORs) because the interpretation of the coefficients of mixed linear models is similar to that of linear regression models, and is different from that of logistic regression models. Accordingly, we estimated ORs for the 6,204,347 variants by an additive effect model and a logistic regression analysis with age- and sex-adjustment. We did not adjust using principal components for the calculation of ORs.

Imputation of genotypes for the validation samples

The KyushuU and JPJM samples were genotyped at the Center for Integrative Medical Science (RIKEN) and at the Iwate Medical Megabank Organization (Iwate Medical University), respectively. Both samples were genotyped using a HumanOmniExpressExome BeadChip array.

Samples with a call rate < 0.95 and SNPs with a call rate < 0.99 were excluded. Variants with a Hardy-Weinberg equilibrium exact test *P*-value < 1×10^{-6} and a minor allele frequency < 0.01 were also excluded.

Subsequently, we imputed genotypes of six SNPs included in our multi-locus genetic risk score. For each target SNP, genotyped SNPs within 500,000 base-pairs distance from the target SNP were extracted, and the target SNP was imputed by SHAPIT[5] and Minimac3[2] software based on the 1000 Genomes reference panel.[3]

Derivation of a multi-locus risk score

A weighted multi-locus genetic risk score (wGRS) was calculated by multiplying the number of risk alleles for each SNP by the weight for that SNP, and then taking the sum across the five SNPs, according to the following formula:

$$wGRS_i = \sum_{j=1}^{5} w_j X_{ij}$$

where i is an individual, $wGRS_i$ is the wGRS for the individual i, j is a SNP, w_j is the weight for the SNP j, and X_{ij} is the number of risk alleles for the individual i and for the SNP j. Note that X_{ij} takes the value 0, 1, or 2 for directly genotyped SNPs, and takes a continuous value ranging from 0 to 2 for imputed SNPs. The weights for the five variants were estimated from the derivation samples (Supplementary Table IV).

Derivation of a polygenic risk score

Based on the derivation datasets that had been subjected to the quality control filters, the polyGRS was derived via two steps: (i) restricted maximum likelihood (REML) and (ii) best linear unbiased prediction (BLUP). In both steps, we assumed the following mixed linear model (MLM)[4,6]:

$$y_i = \mu + g_i + e_i,$$

where y_i represents the dichotomous disease status ($y_i = 0$: no disease; $y_i = 1$: diseased status) of the i-th individual, μ is the proportion of the cases in derivation samples, and g_i and e_i are normally distributed random variables representing genetic and nongenetic effects, respectively, i.e., $g_i \sim N(0,\sigma_g^2\mathbf{A})$ and $e_i \sim N(0,\sigma_e^2\mathbf{I})$, where $\mathbf{A}$ is a genetic relationship matrix (GRM), $\mathbf{I}$ is the identity matrix, and σ_g^2 and σ_e^2 are the variances of genetic and nongenetic effects, respectively. The GRM was calculated from SNP data as:

$$A(i,j) = \frac{1}{M}\sum_k \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1-p_k)},$$

where M represents the number of SNPs after quality control filterings, x_{ik} is the genotype (0, 1, or 2) of the i-th individual and k-th SNP, and p_k is the allele frequency of the k-th SNP. Estimates of σ_g^2 and σ_e^2 (denoted as $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$, respectively) were determined by solving the REML equation with the average information (AI-REML) algorithm.[7,8] To avoid inflation of the estimate of σ_g^2, we included the top 20 principal components (PCs) as covariates of the above MLM.

Given the REML estimates of $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$, the BLUP method calculates the genetic effect $\hat{\mathbf{g}}$,[9] where $\hat{\mathbf{g}}$ represents the N-dimensional vector of the estimates of genetic effects, and N is the number of individuals in the derivation dataset. The weight parameters for the 357,367 genotyped variants $\hat{\mathbf{u}}$ (the M-dimensional vector) were calculated from estimates of genetic effects $\hat{\mathbf{g}}$ and genotype data of the derivation dataset by the equation: $\hat{\mathbf{u}} = \frac{1}{N}\mathbf{W}_t'\mathbf{A}^{-1}\hat{\mathbf{g}}$.

Here, $\mathbf{W}_t$ is the genotype data of the derivation dataset ($N \times M$-dimensional matrix), which was normalized so that the average and standard deviation of the column vectors became 0 and 1, respectively. $\mathbf{W}_t'$ represents the transposition of the matrix $\mathbf{W}_t$.

Note that the MLM takes into account genotype data via the GRM, and only the variance parameters, σ_g^2 and σ_e^2, were free parameters in the MLM. The REML step estimates the two free parameters based on derivation samples ($N = 39{,}684$). In the BLUP step, no parameters were estimated. The BLUP step converts the estimates of the variance parameters ($\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$) into the weight parameters for the 357,367 SNPs in conjunction with the information of genotype data of derivation samples ($\mathbf{A}$, and $\mathbf{W}_t$). In total, only two parameters were estimated through the REML and BLUP steps. This mathematical technique is called a 'dual formula', or 'kernel trick'.[10]

Based on the weight parameters $\hat{\mathbf{u}}$, the polyGRS was calculated by multiplying the number of risk alleles for each SNP by the weight for that SNP, and then taking the sum across the 357,367 SNPs, according to the following equation:

$$polyGRS_i = \sum_{j=1}^{357,367} w_j X_{ij}$$

Subtype-specific model

The polyGRS was used in the subtype-specific models with 639 cases and 26,470 controls in the LVD-specific model, 2,214 cases and 26,470 controls in the SVD-specific model, and 310 cases and 26,470 controls in the CE-specific model as training dataset for each IS subtype.

Score-profiling model

Of the 537,999 genotyped SNPs, variants with an associated *P* value lower than a given P_T threshold were included in the model. Coefficients (β) for the variants estimated from the MLM association method were used as model parameters. Let x_{ik} be the genotype (0, 1, or 2) of the *i*-th individual and *k*-th SNP, P_k the *P* value for the *k*-th SNP and β_k the coefficients for the *k*-th SNP. Then, the score-profiling-based genetic risk score of the *i*-th validation subjects could be calculated as follows:

$$\mathrm{GRS}_i^{score-profiling}(P_T) = \sum_{k|P_k<P_T} \beta_k x_k \ .$$

Supplementary Table I. Association signals for six well-studied SNPs in derivation samples

SNP	Chr	Candidate gene(s)	RA[a]	RAF	R_{sq}	*P*[e]	OR[f] (95% CI)	PMID
rs2107595	7	*HDAC9*	A[b]	0.35	0.98	**0.02**	**1.03 (0.99–1.07)**	26089329, 25031287, 23041239
rs6843082	4	*PITX2*	G[b]	0.68	0.99	**3.8×10^{-4}**	**1.09 (1.04–1.13)**	26089329, 25031287, 23041239
rs879324	16	*ZFHX3*	A[b]	0.40	1.00[d]	**0.04**	**1.06 (1.02–1.10)**	26089329, 25031287, 23041239
rs2383207	9	*CDKN2B*, *CDKN2A*	G[b]	0.65	1.00[d]	**6.2×10^{-4}**	**1.08 (1.04–1.12)**	22306652, 26089329, 23041239
rs11833579	12	*NINJ2*	G[c]	0.57	0.96	0.12	1.04 (1.00–1.08)	19369658, 22306652, 26089329, 23041239
rs2230500	14	*PRKCH*	G[c]	0.78	1.00	0.60	1.01 (0.97–1.06)	17206144, 23041239

SNP indicates single nucleotide polymorphism; Chr, chromosome; RA, risk allele; RAF, risk allele frequency; R_{sq}, R-squared value of the locus imputation quality; OR, odds ratio; and CI, confidence interval; PMID, PubMed ID.

[a]The risk allele was determined based on our derivation samples. [b]The direction of the effects was consistent with previous studies. [c]The direction of the effects was inconsistent with previous studies. [d]These variants were directory genotyped using the SNP array. [e]*P*-values were calculated by a mixed linear model association method. [f]ORs were calculated from a logistic regression analysis.

Results listed in bold are nominally significant ($P < 0.05$).

Supplementary Table II. Association signals for 30 recently reported SNPs in derivation samples

SNP	Chr	Candidate gene	RA[a]	RAF	R_{sq}	*P*[c]	OR[d] (95% CI)	PMID
rs225132	1	*ERRFI1*	T	0.44	1.00[b]	0.53	1.00 (0.96–1.03)	23041239
rs1937787	1	*ELTD1*	C	0.08	1.00[b]	0.11	1.02 (0.96–1.09)	26089329
rs11681884	2	*IL1RN, IL1F10, IL36RN*	C	0.78	1.00	0.88	0.97 (0.93–1.02)	26089329
rs16851055	3	*SPSB4*	G	0.81	0.99[b]	0.42	1.03 (0.99–1.08)	23041239
rs6763538	3	*OXNAD1*	C	0.96	1.00	0.12	**1.12 (1.03–1.23)**	23041239
rs7432308	3	*SATB1, KCNH8*	T	0.25	0.99	0.64	**1.05 (1.01–1.10)**	23041239
rs2930144	3	*SLC6A11*	C	0.14	0.99	0.47	**1.06 (1.00–1.11)**	25031287
rs704341	3	*PTPRG*	G	0.98	0.73	0.07	1.06 (0.91–1.23)	26089329
rs17007400	4	*ILI5*	G	0.01	0.94	0.61	1.10 (0.93–1.29)	25031287
rs12646447	4	*PITX2*	C	0.45	0.98	0.06	**1.05 (1.01–1.09)**	26089329, 25031287
rs7705819	5	*MSX2, NKX2-5*	T	0.44	0.65	0.38	0.99 (0.95–1.04)	26089329
rs4867766	5	*SUMO2P6, GAPDHP71*	A	0.12	0.99	0.21	1.00 (0.95–1.06)	26089329
rs556621	6	*CDC5L, SUPT3H*	T	0.59	1.00[b]	0.55	1.04 (1.00–1.08)	26089329, 22941190, 23041239
rs9345396	6	*TSG1*	T	0.22	0.99	0.86	1.01 (0.96–1.06)	26089329
rs17347800	7	*HDAC9*	G	0.87	1.00[b]	0.30	1.01 (0.95–1.06)	26089329
rs4875812	8	*ARHGEF10*	G	0.64	0.89	0.75	1.02 (0.98–1.07)	23041239
rs2281673	10	*PLEKHA1*	T	0.83	0.98	0.75	1.01 (0.96–1.06)	25031287
rs7937106	11	*SPSB4*	C	0.08	1.00[b]	0.12	1.06 (0.99–1.13)	23041239
rs2084637	11	*UBASH3B*	T	0.67	0.92	0.22	**1.05 (1.01–1.09)**	26089329
rs2238151	12	*ALDH2*	T	0.07	1.00[b]	0.46	0.96 (0.90–1.04)	23041239
rs10744777	12	*ALDH2*	T	0.07	1.00[b]	0.46	0.96 (0.90–1.04)	25031287
rs4597201	13	*NDF1P2*	C	0.24	1.00	0.49	1.02 (0.98–1.07)	25031287
rs10400694	14	–	C	0.59	1.00	0.99	1.01 (0.97–1.05)	26089329
rs7156510	14	*FLRT2*	T	0.58	0.98	0.83	1.01 (0.98–1.05)	26089329
rs1564060	14	–	G	0.64	0.98	0.30	1.02	26089329

							(0.98–1.06)	
rs4471613	15	*ALDH1A2*	A	0.04	0.97	0.62	**1.11 (1.01–1.22)**	26089329
rs7407640	18	*AFG3L2*	A	0.41	1.00	0.88	1.00 (0.96–1.03)	23041239
rs8113518	19	*KRTDAP*	T	0.73	1.00	0.59	1.02 (0.98–1.06)	25031287
rs5752326	22	*HPS4*	T	0.85	1.00[b]	0.49	1.03 (0.98–1.09)	26089329

SNP indicates single nucleotide polymorphism; Chr, chromosome; RA, risk allele; RAF, risk allele frequency; R_{sq}, R-squared value of the imputation quality on the locus; OR, odds ratio; and CI, confidence interval; PMID, PubMed ID.

[a]The risk allele was determined based on our derivation samples. [b]These variants were directory genotyped using the SNP array. [c]*P*-values were calculated by a mixed linear model association method. [d]ORs were calculated from a logistic regression analysis.

Results listed in bold are nominally significant ($P < 0.05$).

Supplementary Table III. Novel association signals discovered from derivation samples with a suggested significance level ($P < 1 \times 10^{-5}$)

SNP	Chr	Candidate gene(s)	RA	RAF	R_{sq}	P^a	OR[b] (95% CI)
rs1275923	**2**	***KCNK3***	**C**	**0.75**	**0.86**	**4.8×10^{-8}**	**1.15 (1.10–1.21)**
rs149493746	1	*SFPQ, ZMYM4*	G	0.07	0.81	1.0×10^{-6}	1.24 (1.15–1.34)
rs142873372	1	*ERI3, RNF220*	C	0.01	0.51	2.4×10^{-7}	1.72 (1.40–2.12)
rs203801	1	*ADCY10*	A	0.16	0.96	4.0×10^{-6}	1.18 (1.12–1.24)
rs13014165	2	*INSIG2, EN1*	T	0.15	0.97	1.4×10^{-6}	1.14 (1.08–1.20)
rs788159	2	*METAP1D, DLX1*	G	0.70	1.00[c]	1.3×10^{-6}	1.12 (1.08–1.17)
rs3774430	3	*CACNA1D*	G	0.09	0.99	3.0×10^{-7}	1.24 (1.16–1.32)
rs814778	5	*ADAMTS16*	G	0.43	0.93	7.4×10^{-6}	1.11 (1.06–1.15)
rs150112040	5	*PRLR*	C	0.82	0.88	7.8×10^{-6}	1.15 (1.09–1.21)
rs3783923	14	*ITPK1*	C	0.11	0.98	3.4×10^{-6}	1.20 (1.13–1.27)
rs140161480	15	*NEO1, HCN4*	A	0.95	0.96	8.4×10^{-6}	1.14 (1.05–1.25)
rs2047221	15	*PCSK6*	A	0.40	0.96	2.5×10^{-6}	1.11 (1.07–1.15)
rs144695373	17	*KCNJ2, CASC17*	A	0.02	0.54	5.7×10^{-6}	1.47 (1.23–1.76)
rs7222752	17	*NTPX1, RPTOR*	G	0.03	0.54	3.2×10^{-6}	1.49 (1.30–1.71)
rs4371240	18	*WDR7, LINC-ROR*	C	0.89	0.98	4.9×10^{-6}	1.18 (1.11–1.25)

SNP indicates single nucleotide polymorphism; Chr, chromosome; RA, risk allele; RAF, risk allele frequency; R_{sq}, R-squared value of the imputation quality; OR, odds ratio; and CI, confidence interval.

[a]*P*-values were calculated by a mixed linear model association method. [b]ORs were calculated from a logistic regression analysis.

Results listed in bold are genome-wide significance ($P < 5 \times 10^{-8}$).

Supplementary Table IV. Weight parameters for our multi-locus genetic risk score

SNP	Chr	Candidate gene(s)	RA	Weight per allele[a]	Derivation (*N* = 39,684)		KyushuU (*N* = 2,194)		JPJM (*N* = 672)	
					OR (95% CI)[b]	*P*[c]	OR (95% CI)[b]	*P*[d]	OR (95% CI)[b]	*P*[d]
rs2107595	7	*HDAC9*	A	0.00662955	**1.03 (0.99–1.07)**	**0.02**	**1.13 (1.00–1.28)**	**0.05**	0.82 (0.66–1.01)	0.06
rs6843082	4	*PITX2*	G	0.01113360	**1.09 (1.04–1.13)**	**3.8×10^{-4}**	1.04 (0.92–1.17)	0.57	1.05 (0.85–1.30)	0.63
rs879324	16	*ZFHX3*	A	0.00601980	**1.06 (1.02–1.10)**	**0.04**	1.03 (0.92–1.16)	0.61	1.04 (0.84–1.28)	0.71
rs2383207	9	*CDKN2B*, *CDKN2A*	G	0.01047050	**1.08 (1.04–1.12)**	**6.2×10^{-4}**	1.03 (0.91–1.17)	0.61	1.03 (0.83–1.29)	0.78
rs1275923	2	*KCNK3*	C	0.01882280	**1.15 (1.10–1.21)**	**4.8×10^{-8}**	0.98 (0.85–1.14)	0.79	**1.35 (1.03–1.76)**	**0.03**

SNP indicates single nucleotide polymorphism; Chr, chromosome; RA, risk allele; OR, odds ratio; and CI, confidence interval.
[a]Weight parameters were estimated from the derivation samples. [b]ORs were calculated by a linear regression analysis in the derivation samples. [c]*P*-values were calculated by a mixed linear model association analysis in the derivation samples. [d]*P*-values and ORs were calculated by a conditional logistic regression analysis in the KyushuU and JPJM samples.
Results listed in bold are nominally significant ($P < 0.05$).

Supplementary Table V. Predictive ability of the polygenic risk scores with varying P_T threshold in the KyushuU samples (N = 2,194)

P_T	#. SNPs	Q1 OR (95% CI)	Q2 OR (95% CI)	Q3 OR (95% CI)	Q4 OR (95% CI)	Q5 OR (95% CI)	OR per 1 SD[a] (95% CI)	Overall[a] *P*-value	C-index (95% CI)
1.0×10^{-5}	12	Reference	0.83 (0.63–1.09)	1.03 (0.79–1.34)	1.10 (0.84–1.45)	0.82 (0.63–1.08)	0.95 (0.87–1.03)	0.221	0.506 (0.482–0.530)
1.0×10^{-4}	49	Reference	0.77 (0.59–1.02)	0.98 (0.74–1.29)	0.95 (0.72–1.24)	0.86 (0.66–1.14)	0.98 (0.90–1.07)	0.655	0.494 (0.470 –0.518)
0.001	565	Reference	0.89 (0.68–1.17)	1.05 (0.80–1.38)	1.27 (0.96–1.67)	1.09 (0.83–1.43)	1.07 (0.98–1.16)	0.129	0.522 (0.498–0.546)
0.01	5,304	Reference	**1.57 (1.19–2.07)**	**1.44 (1.09–1.89)**	**1.60 (1.21–2.10)**	**1.55 (1.18–2.05)**	**1.15 (1.06–1.25)**	**0.001**	**0.538 (0.514–0.562)**
0.05	26,327	Reference	**1.43 (1.09–1.88)**	1.24 (0.94–1.63)	1.20 (0.91–1.58)	**1.73 (1.32–2.28)**	**1.15 (1.06–1.25)**	**0.001**	**0.538 (0.514–0.562)**
0.1	52,910	Reference	1.07 (0.81–1.40)	1.23 (0.93–1.62)	1.21 (0.92–1.59)	**1.54 (1.17–2.02)**	**1.14 (1.05–1.24)**	**0.003**	**0.539 (0.515–0.563)**
0.2	106,326	Reference	1.12 (0.85–1.47)	1.13 (0.86–1.49)	1.22 (0.93–1.61)	**1.58 (1.20–2.08)**	**1.15 (1.06–1.25)**	**0.001**	**0.543 (0.519–0.567)**
0.5	268,120	Reference	1.21 (0.92–1.59)	1.22 (0.93–1.60)	1.14 (0.86–1.50)	**1.75 (1.33–2.31)**	**1.17 (1.08–1.27)**	**<0.001**	**0.545 (0.521–0.569)**
1.0	537,999	Reference	1.23 (0.94–1.62)	**1.32 (1.00–1.74)**	1.09 (0.83–1.43)	**1.82 (1.38–2.40)**	**1.17 (1.08–1.27)**	**<0.001**	**0.546 (0.522–0.570)**

SNP indicates single nucleotide polymorphism; OR indicates odds ratio; CI, confidence interval; and Q1–Q5, quantiles 1–5.
[a]Considering the genetic risk scores as continuous variables.
Results listed in bold are nominally significant ($P < 0.05$).

Supplementary Table VI. Improvement in predictive ability by substituting the multi-locus genetic risk score with the polygenic risk score in the KyushuU and JPJM samples

Validation samples	Subtype	NRI (95% CI)	IDI (95% CI)	ΔC-index
KyushuU	all IS *N* = 2,194	**0.179 (0.095–0.262)** ***P* < 0.001***	**0.008 (0.004–0.012)** ***P* < 0.001***	**0.045** ***P* = 0.007***
	LVD *N* = 720	**0.206 (0.060–0.351)** ***P* = 0.006**	**0.013 (0.004–0.021)** ***P* = 0.004**	0.054 *P* = 0.066
	SVD *N* = 972	**0.185 (0.060–0.310)** ***P* = 0.003**	**0.007 (0.002–0.012)** ***P* = 0.011**	**0.049** ***P* = 0.044**
	CE *N* = 268	0.194 (-0.044–0.432) *P* = 0.110	0.015 (0.000–0.030) *P* = 0.056	0.053 *P* = 0.250
JPJM	all IS *N* = 672	0.071 (-0.080–0.223) *P* = 0.354	0.003 (-0.004–0.010) *P* = 0.384	0.006 *P* = 0.854

NRI indicates net reclassification improvement; IDI, integrated discrimination improvement; CI, confidence interval; IS, ischemic stroke; LVD, large-vessel disease; SVD, small-vessel disease; and CE, cardioembolic stroke.
Results listed in bold are nominally significant ($P < 0.05$).
*Significant after multiple corrections.

Supplementary Table VII. Predictive ability of the multi-locus and the polygenic risk scores for each etiological subtype in the KyushuU samples

Subtype	Model	Q1 OR (95% CI)	Q2 OR (95% CI)	Q3 OR (95% CI)	Q4 OR (95% CI)	Q5 OR (95% CI)	OR per 1 SD[a] (95% CI)	Overall[a] *P*-value	C-index (95% CI)
LVD *N* = 720	wGRS	Reference	0.87 (0.53–1.42)	1.12 (0.69–1.82)	0.85 (0.52–1.38)	1.21 (0.75–1.98)	1.06 (0.92–1.23)	0.434	0.514 (0.472–0.556)
	polyGRS (subtype-mixture)	Reference	1.25 (0.77–2.05)	1.18 (0.72–1.94)	**1.65 (1.01–2.71)**	**2.19 (1.34–3.62)**	**1.26 (1.09–1.47)**	**0.002**	**0.568 (0.527–0.610)**
	polyGRS (subtype-specific)	Reference	1.18 (0.72–1.93)	0.97 (0.60–1.59)	1.21 (0.75–1.98)	1.25 (0.77–2.04)	1.08 (0.93–1.26)	0.302	0.522 (0.479–0.564)
SVD *N* = 972	wGRS	Reference	1.22 (0.80–1.85)	1.01 (0.67–1.53)	1.05 (0.69–1.60)	1.11 (0.73–1.68)	1.00 (0.89–1.14)	0.944	0.501 (0.465–0.537)
	polyGRS (subtype-mixture)	Reference	0.91 (0.60–1.38)	1.14 (0.75–1.74)	1.29 (0.85–1.96)	**1.61 (1.06–2.45)**	**1.18 (1.04–1.34)**	**0.012**	**0.550 (0.514–0.586)**
	polyGRS (subtype-specific)	Reference	0.89 (0.59–1.36)	0.91 (0.60–1.38)	1.14 (0.75–1.74)	1.25 (0.83–1.90)	1.09 (0.96–1.23)	0.197	0.526 (0.490–0.562)
CE *N* = 268	wGRS	Reference	1.40 (0.61–3.22)	1.45 (0.64–3.32)	1.50 (0.66–3.47)	1.45 (0.64–3.32)	1.08 (0.85–1.37)	0.542	0.527 (0.458–0.597)
	polyGRS (subtype-mixture)	Reference	1.89 (0.82–4.43)	1.26 (0.54–2.93)	**2.38 (1.03–5.61)**	**2.45 (1.06–5.77)**	**1.28 (1.00–1.62)**	**0.047**	**0.580 (0.512–0.649)**
	polyGRS (subtype-specific)	Reference	0.89 (0.39–2.04)	1.44 (0.64–3.31)	2.06 (0.89–4.85)	0.80 (0.35–1.83)	0.94 (0.75–1.18)	0.601	0.492 (0.422–0.562)

OR indicates odds ratio; CI, confidence interval; SD, standard deviation; wGRS, weighted multi-locus genetic risk score; polyGRS, polygenic risk score; IS, ischemic stroke; LVD, large-vessel disease; SVD, small-vessel disease; CE, cardioembolic stroke; and Q1–Q5, quantiles 1–5.

[a]Considering the genetic risk scores as continuous variables.

Results listed in bold are nominally significant ($P < 0.05$).

Supplementary Table VIII. Improvement in predictive ability gained by adding the multi-locus or polygenic risk scores to a non-genetic risk model in the KyushuU samples (*N* = 2,194)

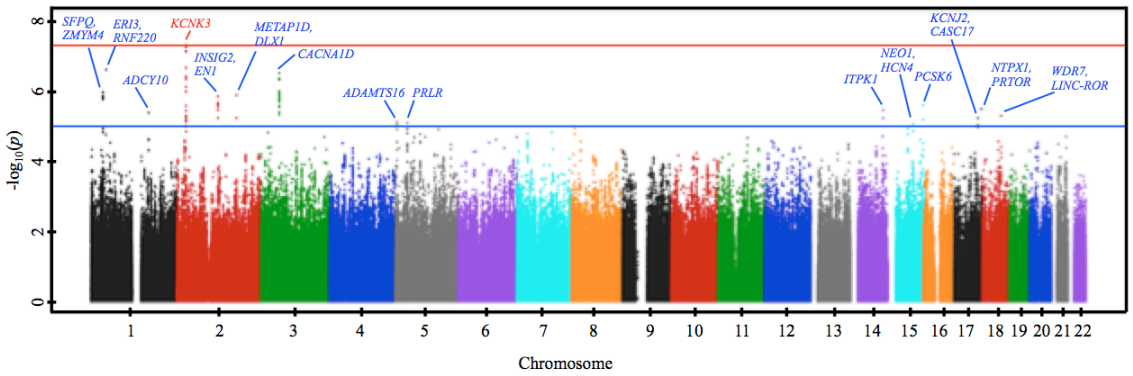
Subtype	C-index without GRS[a] (95% CI)	Model	NRI (95% CI)	IDI (95% CI)	C-index with GRS[b] (95% CI)	ΔC-index
all IS *N* = 2,194	**0.694** **(0.672–0.715)**	wGRS	-0.011 (-0.095–0.073) *P* = 0.797	0.000 (0.000–0.001) *P* = 0.479	**0.695** **(0.673–0.717)**	0.001 *P* = 0.487
		polyGRS (P_T = 1)	0.075 (-0.009–0.158) *P* = 0.080	**0.004 (0.001–0.006)** ***P* = 0.004**	**0.700** **(0.679–0.722)**	0.007 *P* = 0.067
		polyGRS G-BLUP	**0.151 (0.068–0.235)** ***P* < 0.001**	**0.004 (0.001–0.006)** ***P* = 0.003**	**0.700** **(0.679–0.722)**	0.007 *P* = 0.080
LVD *N* = 720	**0.696** **(0.659–0.734)**	wGRS	0.017 (-0.129–0.163) *P* = 0.823	0.001 (-0.001–0.002) *P* = 0.629	**0.701** **(0.663–0.739)**	0.005 *P* = 0.156
		polyGRS (P_T = 1)	-0.022 (-0.168–0.124) *P* = 0.766	0.002 (-0.001–0.006) *P* = 0.178	**0.702** **(0.664–0.740)**	0.006 *P* = 0.131
		polyGRS G-BLUP	**0.200 (0.055–0.345)** ***P* = 0.007**	**0.007 (0.001–0.012)** ***P* = 0.026**	**0.707** **(0.669–0.745)**	**0.011** ***P* = 0.045**
SVD *N* = 972	**0.675** **(0.642–0.709)**	wGRS	0.029 (-0.097–0.154) *P* = 0.653	0.000 (0.000–0.000) *P* = 0.980	**0.676** **(0.642–0.710)**	0.001 *P* = 0.846
		polyGRS (P_T = 1)	**0.128 (0.002–0.253)** ***P* = 0.046**	**0.007 (0.001–0.012)** ***P* = 0.014**	**0.687** **(0.654–0.721)**	0.012 *P* = 0.262
		polyGRS G-BLUP	**0.185 (0.060–0.310)** ***P* = 0.004**	**0.004 (0.000–0.009)** ***P* = 0.047**	**0.688** **(0.654–0.721)**	0.012 *P* = 0.214
CE *N* = 268	**0.917** **(0.881–0.953)**	wGRS	**0.254 (0.017–0.490)** ***P* = 0.035**	0.002 (-0.003–0.008) *P* = 0.421	**0.922** **(0.886–0.957)**	0.004 *P* = 0.539
		polyGRS (P_T = 1)	-0.015 (-0.254–0.224) *P* = 0.903	0.000 (-0.002–0.002) *P* = 0.651	**0.920** **(0.885–0.956)**	0.003 *P* = 0.527
		polyGRS G-BLUP	0.015 (-0.224–0.254) *P* = 0.903	0.000 (0.000–0.000) *P* = 0.943	**0.918** **(0.882–0.954)**	0.000 *P* = 0.883

GRS indicates genetic risk score; IS, ischemic stroke; LVD, large-vessel disease; SVD, small-vessel disease; CE, cardioembolic stroke; NRI, net reclassification improvement; IDI, integrated discrimination improvement; and CI, confidence interval.

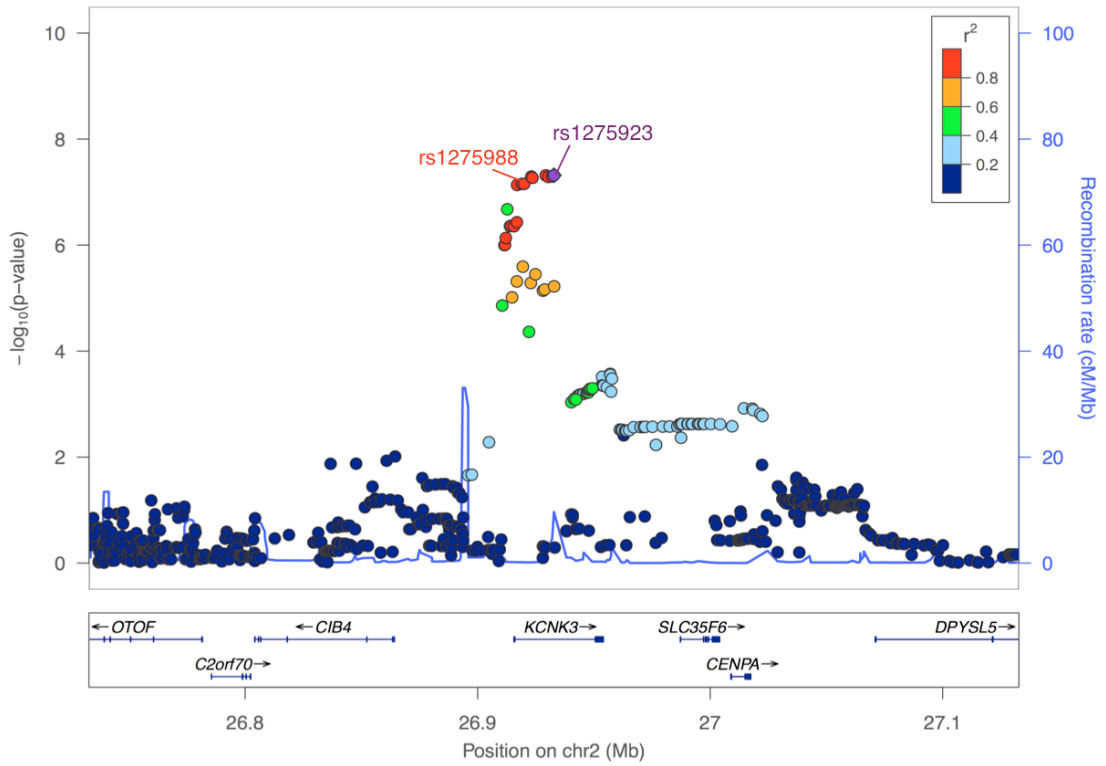
[a]A non-genetic risk model, which includes hypertension, diabetes mellitus, hyperlipidemia, and atrial fibrillation as model variables.

[b]An integrated risk model, which includes hypertension, diabetes mellitus, hyperlipidemia, atrial fibrillation, and a genetic risk score as model variables.

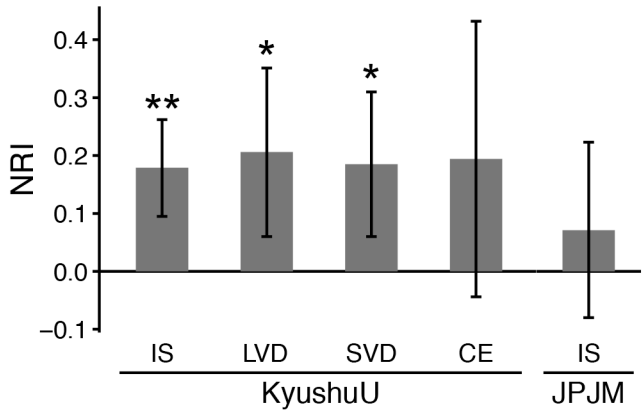
Results listed in bold are nominally significant ($P < 0.05$).



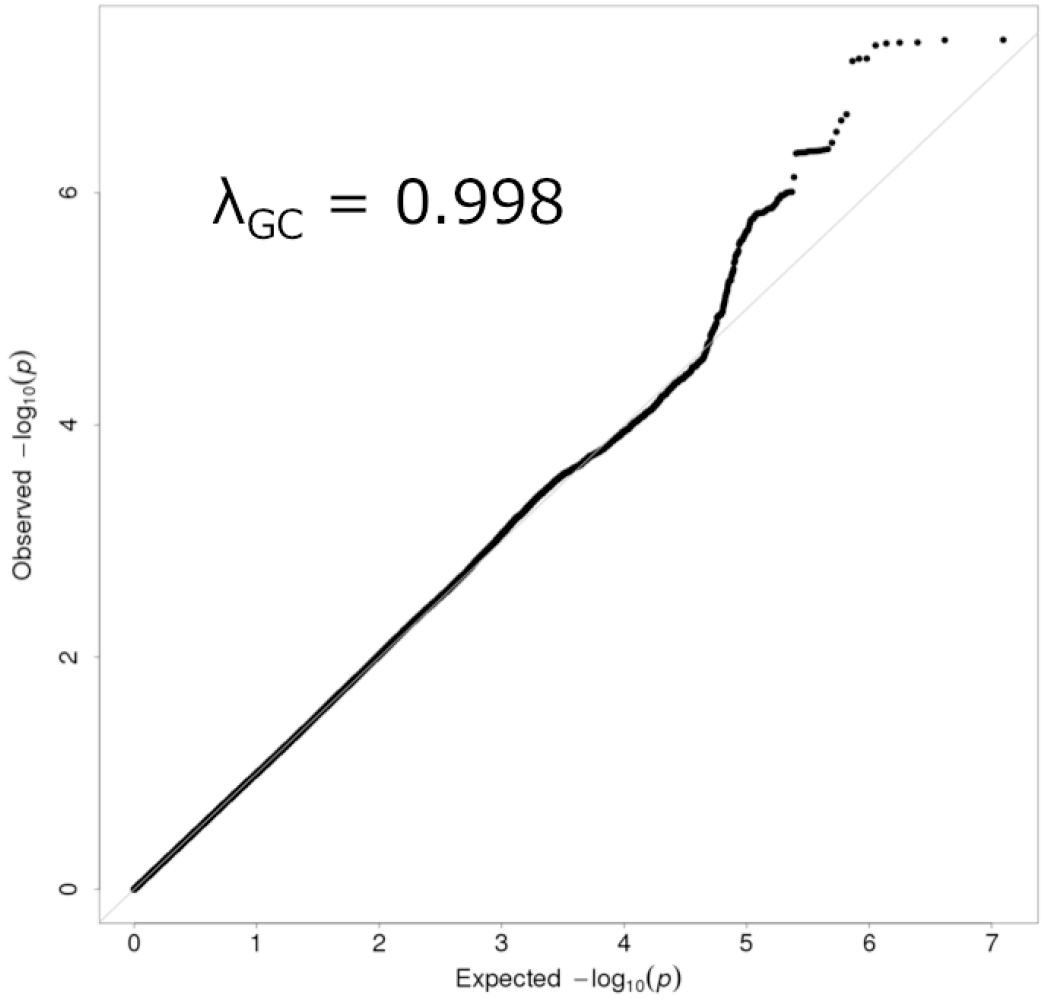
Supplementary Figure I. Genome-wide association signals from the Japanese derivation samples. The *x*-axis represents chromosomal positions and the *y*-axis represents $-\log_{10} P$-values calculated by a mixed linear model association analysis. The red and blue horizontal lines indicate the genome-wide significance level ($P = 5 \times 10^{-8}$) and the suggestive significance level ($P = 1 \times 10^{-5}$), respectively. Candidate genes near genome-wide and suggestive signals are shown in red and blue, respectively.



Supplementary Figure II. Association signals around the *KCNK3* gene. The *x*-axis represents chromosomal positions near the *KCNK3* gene, and the *y*-axis represents $-\log_{10}$ *P*-values. The top signal in this locus (rs1275923) is shown in purple. Dot color for a variant represents the degree of linkage disequilibrium (R^2) estimates between the variant and rs1275923. In this region, a variant (rs1275988) related to blood pressure was harbored. The rs1275988 variant, shown in red, is apart from ~18,000 base-pairs from the rs1275923 variant. The rs1275923 and rs1275988 variants were located within the same linkage disequilibrium block.



Supplementary Figure III. Predictive ability improvement by substituting the multi-locus genetic risk score with the polygenic risk score in the KyushuU and JPJM samples. NRI indicates net reclassification improvement; IS, ischemic stroke; LVD, large-vessel disease; SVD, small-vessel disease; and CE, cardioembolic stroke. Nominal significant differences ($P < 0.05$) are indicated with asterisks. Significant difference after multiple corrections is indicated with double-asterisk.



Supplementary Figure IV. A quantile-quantile plot of the *P*-values from the genome-wide association tests. The *x*-axis indicates the expected -$\log_{10}$ *P*-values under the null hypothesis. The *y*-axis shows the observed -$\log_{10}$ *P*-values calculated by a mixed linear model association method.[7] The grey line represents $y = x$, which corresponds to the null hypothesis. λ_{GC} (the inflation factor of the genomic control method) is the median of the observed test statistics divided by the median of the expected test statistics.

References

1 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010;34:816–834.

2 Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* 2015;31:782–784.

3 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65.

4 Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet.* 2014;46:100–106.

5 Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011;9:179–181.

6 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.

7 Corbeil R, Searle S. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics.* 1976;18:31–38.

8 Gilmour A, Thompson R, Cullis B. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics.* 1995;51:1440–1450.

9 Robinson GK. That BLUP is a good thing: The estimation of random effects. *Statistical Science.* 1991;6:15–32.

10 Bishop CM. Pattern recognition and machine learning. New York: *Springer-Verlag*, 2006.