Real-world data from the Japanese National Health Insurance System enable fine phenotyping in a 14K-scale population-based study

Yayoi Otsuka-Yamasaki,[1,2,*] Naoyuki Nishiya,[1,3,*] Yoichi Sutoh,[1,2,*] Reiko Aoyama,[1] Kozo Tanno,[1,4] Motoki Nakao,[1,5,6] Shohei Komaki,[1,2] Shiori Minabe,[1,2] Hideki Ohmomo,[1,2] Koichi Asahi,[1,7] Yasushi Ishigaki,[1,8] Makoto Sasaki[1,9] and Atsushi Shimizu[1,2,†]

[1]Iwate Tohoku Medical Megabank Organization, Disaster Reconstruction Center, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[2]Division of Biomedical Information Analysis, Institute for Biomedical Sciences, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[3]Division of Integrated Information for Pharmaceutical Sciences, Department of Clinical Pharmacy, School of Pharmacy, Iwate Medical University 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[4]Department of Hygiene and Preventive Medicine, School of Medicine, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[5]Department of Pharmacology, School of Medicine, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[6]Department of Cardiovascular Medicine, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Kita 15, Nishi 7, Kita-ku, Sapporo 060-8638, Hokkaido,

Japan

[7]Division of Nephrology and Hypertension, Department of Internal Medicine, School of Medicine, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[8]Division of Diabetes, Metabolism and Endocrinology, Department of Internal Medicine, School of Medicine, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan

[9]Division of Ultrahigh field MRI, Institute for Biomedical Sciences, Iwate Medical University, 1-1-1, Idaidori, Yahaba 028-3694, Iwate, Japan.

* These authors should be considered joint first authors.

†Correspondence: Atsushi Shimizu (ashimizu@iwate-med.ac.jp)

Data Availability Statement

The GWAS summaries have been published in jMorp (https://jmorp.megabank.tohoku.ac.jp/gwas-studies/TGA000016). The datasets of individuals in the TMM CommCohort study analyzed here are not publicly available for ethical reasons, such as protecting personal information and preventing unintended personal identification. However, they are available from the corresponding author upon request, subject to approval from the Ethics Committee of Iwate Medical University and the Materials and Information Distribution Review Committee of the TMM Project.

Funding

This study was supported by the Tohoku Medical Megabank (TMM) Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and by the Japan Agency for Medical Research and Development (AMED) under grant number JP23tm0124006, as well as by research funding from IQVIA. YO-Y and AS were supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under grant number 23K09696.

Conflict of interest disclosure

The authors declare no competing interests.

Ethical approval

This study was approved by the Institutional Review Board of Iwate Medical University (Approval number HG2021-009).

Patient consent statement

All participants provided written informed consent and the study protocol was approved by the Institutional Review Board of Iwate Medical University (Approval number HG2021-

009)

Permission to reproduce material from other sources

No materials requiring permission from other sources have been used in this manuscript.

Clinical trial registration

This study does not involve interventional components and therefore was not registered as a clinical trial.

Abstract

The use of medical databases, known as real-world data (RWD), enables accurate and efficient estimation of disease prevalence in population-based studies, making it a potential game-changer in epidemiology. However, the lack of standardized data formats across hospitals complicates integration across institutions. In Japan, health insurance claims data are standardized under a nationally unified format, providing a reliable source of structured RWD. We evaluated the utility of insurance claims data in epidemiological research. Incorporating both diagnosis and prescription information into case definitions resulted in four- and six-fold increases in the estimated prevalence of Alzheimer's disease (AD) and Parkinson's disease (PD), respectively, compared with conventional self-reported definitions. Subsequent genome-wide association studies (GWAS) for AD showed increased model log-likelihood and identified a characteristic *APOE* signal, findings observed only with extended case definitions. The *APOE* effect size was consistent with large case–control studies, while standard errors remained comparable to smaller studies. These results indicate that claims-based phenotyping improves case identification without loss of accuracy and supports scalable approaches for genomic epidemiology and public health surveillance.

Keywords: real-world data, health insurance claims, genome-wide association study, Alzheimer's disease, phenotyping

Introduction

The accurate estimation of disease prevalence within a community is crucial for public health. However, challenges remain. Recent advances in medical databases—such as electronic health records and health insurance claims data, collectively referred to as real-world data (RWD)—provide opportunities to enhance the accuracy, efficiency, and frequency of prevalence estimation in population-based studies. These developments also enable the scale-up of surveys and inform public health decision-making (Liu & Panagiotakos, 2022; Swift et al., 2018). Furthermore, the integration of RWD with large-scale genome cohorts increasingly contributes to the generation of epidemiological evidence that informs genomic research (Briggs et al., 2022). Nonetheless, technical challenges remain. Although diagnoses recorded in electronic health records are generally reliable, the lack of standardized data formats across hospitals in many regions limit the integration of data across institutions (Hiramatsu et al., 2021).

Insurance claims data offer a standardized format that facilitates integration across institutions, particularly in countries with universal health coverage, such as Japan. In Japan, national health insurance claims are standardized across diverse healthcare providers and administered by local governments (Shida et al., 2023). The Japanese health insurance system consists of two major schemes: the National Health Insurance (NHI), which covers individuals aged ≤74 years who are not enrolled in employee insurance, and the Late-stage Medical Care System for the Elderly (LsMCSE), which covers individuals aged ≥75 years (Ministry of Health, Labour and Welfare., 2025). Individuals are typically insured through their local municipal governments. Both systems provide structured diagnostic and prescription claims data, consequently enabling the retrieval and integration of health data from multiple regions under appropriate legal and ethical frameworks (Fujinaga & Fukuoka, 2022).

The Tohoku Medical Megabank Community Cohort (TMM CommCohort) is a large-scale population-based cohort study established to monitor public health in areas affected

125 by the 2011 Great East Japan Earthquake (Kuriyama et al., 2016). This cohort obtained
126 health insurance claims data through agreements with participating local governments. In
127 this context, diagnosis and prescription data from claims can supplement medical histories
128 based on self-administered questionnaires (SAQs) and serve as proxies for clinical
129 diagnoses.

130 In this study, we evaluated the incorporation of diagnosis and prescription data
131 from insurance claims into conventional case–control diagnostic definitions. The prevalence
132 of hypertension, dyslipidemia, diabetes, Parkinson's disease (PD), and Alzheimer's disease
133 (AD) is expected to increase in aging societies. Therefore, we aimed to determine whether
134 incorporating diagnosis and prescription data from insurance claims could improve case–
135 control definitions for these common diseases. We further sought to evaluate the accuracy
136 of phenotyping using an extended case definition incorporating claims-based diagnosis and
137 prescription records in the context of genome-wide association studies (GWAS). These
138 findings may facilitate the development of more scalable strategies in genomic epidemiology
139 and public health research. In addition, they highlight the potential value of a feedback
140 cycle in which claims data inform both public health action and evidence-based policy
141 development for insured individuals.
142

Materials and Methods

Population

The design of the TMM CommCohort study was described previously (Kuriyama et al., 2016). In the present study, we included participants from the TMM CommCohort study who held insurance from the NHI or LsMCSE and resided in Iwate Prefecture. Individuals with missing genotypic, phenotypic, and claims data, or other data necessary for analysis, were excluded from the study.

Participants were recruited during specific health checkups or through visits to cohort assessment centers between May 2013 and March 2016. Medical history (e.g., hypertension, dyslipidemia, diabetes, AD, and PD diagnoses) and treatment status for selected conditions (hypertension, dyslipidemia, and diabetes) were collected using SAQs. Systolic and diastolic blood pressure were measured twice at recruitment, and the average was used for analysis. Hemoglobin A1c (HbA1c), high- (HDL) and low-density lipoprotein (LDL) cholesterol, and triglyceride levels were measured in blood samples collected at recruitment.

National health insurance claims data, including medical, pharmacy, and Diagnosis Procedure Combination receipts, were collected annually by electronically collating with participant information from the TMM CommCohort study in Iwate. This process enabled the construction of a claims-based dataset for TMM participants in Iwate. Personally identifiable information and non-research related elements were excluded to ensure strict data minimization and privacy protection. The claims data are structured in a standard format used across Japan's public health insurance systems. Clinical endpoints were defined by a combination of disease and medication.

This study adhered to the principles of the Declaration of Helsinki, and written informed consent was obtained from all participants. All individual data were anonymized before analysis. The study was approved by the Institutional Review Board of Iwate Medical University (Approval number HG H25-2). Analyses involving personal genotypic or

170 phenotypic information were performed on a standalone computing system—the Tohoku
171 University Tohoku Medical Megabank Organization supercomputer.
172

173 ## Genotyping and GWAS

174 Genotyping and genotype imputation were performed as previously described
175 (Minegishi et al., 2019; Sutoh et al., 2025; Yasuda et al., 2019). In the hard-call genotype
176 data, variants that met any of the following criteria were excluded: low call rate (<0.99),
177 low heterozygous allele count (<5), deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-}$
178 ^{6}), low minor allele frequency (<0.01), or location within regions of strong linkage
179 disequilibrium (Price et al., 2008). Variant filtering was conducted using PLINK v1.9
180 (version b5.1) (Chang et al., 2015). Linkage disequilibrium pruning was performed with
181 PLINK v2.0 (version a3LM) using the option "--indep-pairwise 1500 150 0.03".
182 Principal component analysis was conducted in PLINK v2.0 using the "--pca
183 approx allele-wts" option. One individual from each kinship pair (kinship > 0.0884) was
184 excluded based on the KING method to account for relatedness (Manichaikul et al., 2010).
185 The resulting principal components were then projected onto all individuals, including
186 those excluded based on the kinship threshold.
187 Case–control GWAS were conducted using SAIGE v0.44.6.5 (Zhou et al., 2018).
188 Individuals with a low call rate (<0.99) or extreme genetic ancestry, defined as values of
189 PC1 or PC2 deviating by ≥4 standard deviations from the mean, were excluded. In addition,
190 imputed variants with low imputation quality ($R^2 < 0.3$) or allele frequency (<0.01) were
191 excluded from the analysis.
192

193 ## Statistical analysis

194 Statistical analyses were performed using R versions 4.1.0 and 4.3.2. Meta-
195 analysis was performed with the R package "meta" (version 8.0-2) (Schwarzer et al., 2015).
196

Use of artificial intelligence tools

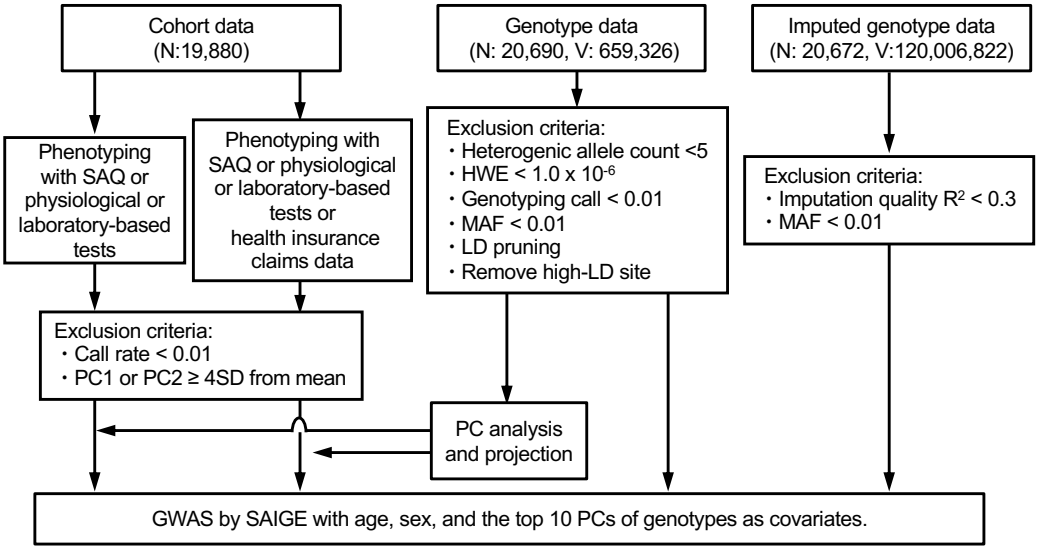
During the preparation of this work, the authors used ChatGPT 4o (OpenAI, version 4o, including Deep Research mode) to assist in reviewing and summarizing previous studies and improve the computer scripts for data analysis. The authors subsequently verified and edited all content generated by this tool and take full responsibility for the integrity and accuracy of the final manuscript.

204 # Results

205 ## Retrieval of health insurance claims data and phenotyping

206 A workflow of the study is shown in Figure 1.

207



208

209 **Figure 1. The workflow in this study.**

210 The study participants had hard-called and imputed genotype data, as well as
211 cohort information including health insurance claims, demographic characteristics (age
212 and sex), laboratory test results, and responses to pre-specified items in a self-
213 administered questionnaire (SAQ). The real-world data used in this study included
214 disease diagnoses and prescription histories derived from medical, pharmacy, and
215 Diagnosis Procedure Combination claims data. Phenotyping was performed using either
216 conventional definitions based on SAQs and laboratory tests or extended definitions
217 incorporating health insurance claims data. HWE, Hardy-Weinberg equilibrium; LD,
218 linkage disequilibrium; MAF, minor allele frequency; PC, principal component; SD,
219 standard deviation; V, variants.

220

221 Health insurance claims data on diagnosis and drug prescriptions were obtained for
222 19,880 individuals (7,638 men and 12,242 women) (Figure 1). Of these, 14,241 participants
223 met the cohort information and genotype data requirements. Claims data for hypertension,
224 diabetes, and dyslipidemia were retrieved from the NHI system and covered the period
225 between June 2013 and December 2019. The claims data for AD and PD were collected
226 between June 2013 and December 2020 and included data from both the NHI and LsMCSE.
227 Participants were classified into case and control groups using two criteria: (1) the
228 conventional case definition based on SAQs and laboratory tests used in the cohort study,
229 and (2) an extended case definition incorporating both disease names recorded in claims
230 and drug prescription data, in addition to (1) (Table 1).
231

232 **Table 1. Case definition and prescription drugs for target outcomes.**

Phenotype	Conventional case definition	Prescription drugs
Hypertension	SAQ (medical history or treatment status[a]) or SBP ≥ 140 mmHg or DBP ≥ 90 mmHg (JSH2019)[b]	Calcium antagonists, angiotensin II receptor blockers, diuretics, alpha blockers / beta blockers, angiotensin-converting-enzyme inhibitors, circulatory system hormones
Diabetes	SAQ (medical history or treatment status[a]) or HbA1c ≥ 6.5% (NGSP)[c]	Dipeptidyl peptidase-4 inhibitors, sulfonylureas, biguanide α-glucosidase inhibitors, insulin preparations, thiazolidine (TZD) derivative (pioglitazone hydrochloride), glinides, sodium glucose transporter 2

		inhibitors, glucagon-like peptide-1 receptor agonists
Dyslipidemia	SAQ (medical history or treatment status[a]) or HDL < 40 mg/dL or triglyceride ≥ 150 mg/dL or LDL ≥ 140 mg/dL[d]	3-hydroxy-3-methylglutaryl-CoA reductase inhibitors, fibrates Ezetimibe, polyunsaturated fatty acids, nicotinates, anion exchange resins, probucol
Alzheimer's disease, (Dementia)	SAQ (medical history)	Donepezil hydrochloride Galantamine hydrobromide Rivastigmine Memantine hydrochloride
Parkinson's disease,	SAQ (medical history)	Anticholinergics Dopamine receptor agonists / dopamine release enhancer (amantadine hydrochloride) Levodopa-containing agents / levodopa activator (zonisamide) Monoamine oxidase inhibitors

	Catechol-O-methyltransferase inhibitor (entacapone)
	Noradrenaline precursor (doroxidopa)
	Adenosine A2A receptor antagonist (istradefylline)

a) The response to the question regarding whether the participant was under treatment was used only for the conventional case definition, which did not incorporate prescription data.

b) Based on the Guidelines for the Management of Hypertension 2019 (JSH2019).

c) Defined according to the standards of the National Glycohemoglobin Standardization Program (NGSP).

d) Diagnostic criteria for arteriosclerosis, as defined by the Japan Atherosclerosis Society (2022).

Abbreviations: DBP, diastolic blood pressure; Hb, hemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SAQ, self-administered questionnaire; SBP, systolic blood pressure.

The incorporation of claims data into phenotyping increased the prevalence of all target outcomes by 1.0–6.3% (Table 2). Common lifestyle-related diseases such as hypertension, diabetes, and dyslipidemia showed modest increases of 4.0%, 1.0%, and 6.3%, respectively. In contrast, the prevalence of AD and PD increased by more than six- and four-fold, respectively.

Table 2. Characteristics and prevalence by phenotype.

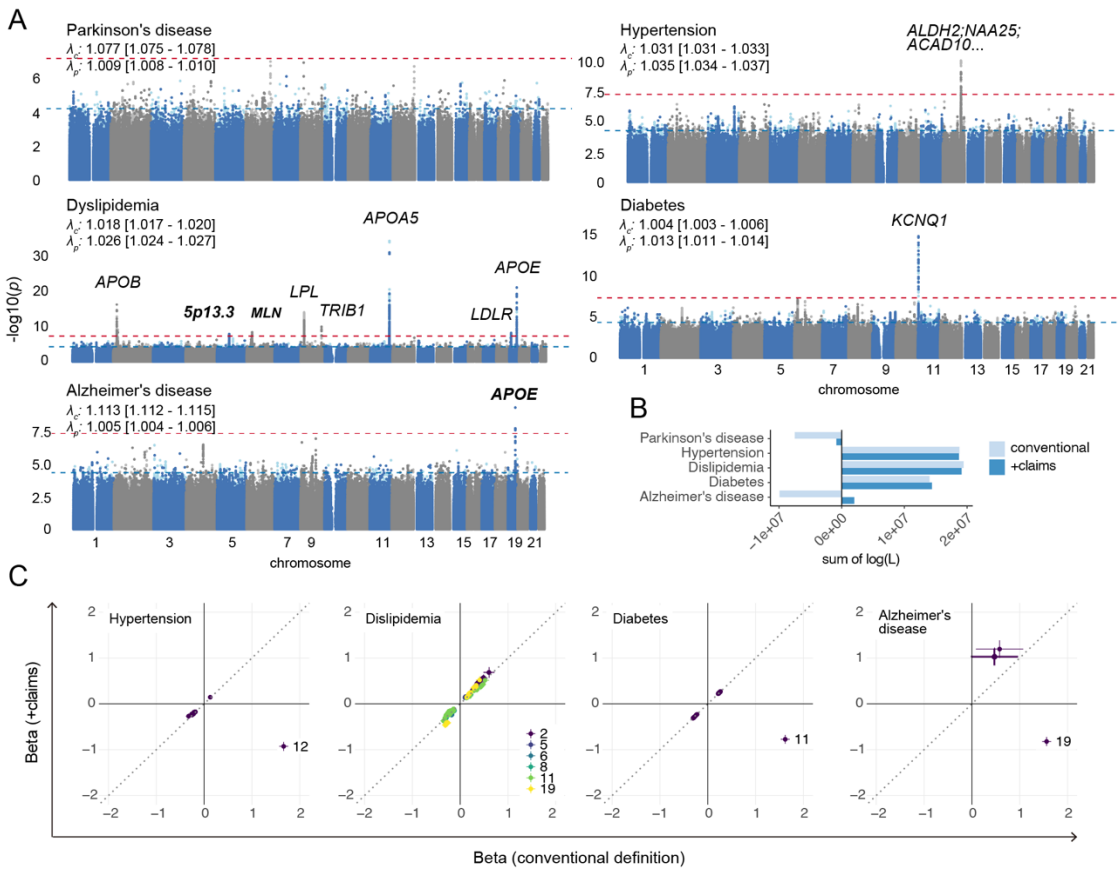
N	14,241[a]
Female (%)	61.5

Age	61.9 +/- 10.0	
	Conventional case definition	+claims data
Hypertension (% (n))	46.7 (6,644)	50.6 (7,211)
Diabetes (% (n))	11.2 (1,599)	12.2 (1,740)
Dyslipidemia (% (n))	53.6 (7,362)	59.9 (8,531)
Alzheimer's disease (% (n))	0.2 (22)	1.3 (183)
Parkinson's disease (% (n))	0.2 (34)	0.7 (106)

a) One participant was missing from the analysis of Alzheimer's disease, Parkinson's disease, and dyslipidemia.

GWAS

The effects of different phenotyping methods on GWAS results were confirmed (Figure 2A).



260

261 **Figure 2. Genome-wide association studies (GWAS) using two case definitions.**

262 GWAS were performed using two different case definitions: (1) the conventional
263 method based on SAQs and laboratory tests (light-colored Manhattan plot; most of the
264 plot is hidden), and (2) the extended method incorporating health insurance claims data
265 (dark-colored Manhattan plot) (A). Gene names near genome-wide significant variants
266 are shown in italics; those identified only in the GWAS using definition (2) are shown in
267 bold italics. The red and blue dashed horizontal lines indicate the thresholds for genome-
268 wide significance ($P = 5 \times 10^{-8}$) and suggestive significance ($P = 5 \times 10^{-5}$), respectively.
269 $\lambda c/\lambda p$ denote genomic inflation factors for the GWAS based on case definitions (1) and (2),
270 respectively; 95% confidence intervals are shown in square brackets. Comparison of the
271 sum of log-likelihoods (log L) from GWAS summary statistics between the two case
272 definitions (B). The number of variants and sample sizes were identical between the two
273 analyses. Effect sizes and directions of significant variants identified in GWAS using the
274 two different case definitions were compared (C). Variants are color-coded by chromosome.
275 Bars and points represent the standard errors and mean of the beta values, respectively.
276 The dotted line represents the diagonal.

277

278 Several genome-wide significant signals ($P < 5 \times 10^{-8}$) were detected for hypertension,
279 dyslipidemia, diabetes, and AD (Figure 2A and Supplementary Tables 1–7). *ALDH2* and
280 *KCNQ1* were identified in the GWAS for hypertension and diabetes. Both genes have been
281 reported for blood pressure and HbA1c, respectively, in previous GWAS (Sakaue et al., 2021).
282 Known lipid-associated loci such as *APOB, LPL, TRIB1, APOA5, LDLR*, and *APOE* were
283 replicated in the GWAS for dyslipidemia, consistent with prior findings on LDL HDL, total
284 cholesterol, and triglyceride levels. In contrast, association signals at 5p13.3 and the motilin
285 gene locus were detected only in the GWAS based on the extended case definition (Figure

286 2A). These two genes have been detected in the LDL cholesterol or triglyceride GWAS of the
287 TMM CommCohort study, which included 67K participants (Study ID: TGA000007,
288 https://jmorp.megabank.tohoku.ac.jp/gwas-studies/) and Asian populations (Chen et al.,
289 2023; Hsu et al., 2025; Ismail Umlai et al., 2025). The lead variant rs4704208 in the 5p13.3
290 region identified in this study is an expression quantitative trait locus for the 3-hydroxy-3-
291 methylglutaryl-CoA reductase gene (Figure 2A), as determined according to the GTEx
292 database (https://www.gtexportal.org/home/). This enzyme is the pharmacological target of
293 statins, which are widely prescribed for the treatment of dyslipidemia.

294 The GWAS for AD using the extended case definitions revealed significant signals
295 in *APOE*, whereas these signals were not detected in the GWAS based on conventional
296 phenotyping (Figure 2A and Supplementary Table 7). The log-likelihood values, which
297 indicate model fit, increased with the use of the extended definition for AD (Figure 2B). The
298 effect sizes and directions of significant variants were consistent across the different
299 phenotyping definitions (Figure 2C). For AD, using the extended case definition yielded
300 smaller standard errors of effect size estimates than the conventional definition, indicating
301 improved estimation precision.

302

303 **Comparison with previous studies**

304 The rs429358 C allele, a missense variant in *APOE* that results in a cysteine-to-
305 arginine substitution, is classified as a risk and/or pathogenic allele for AD (ClinVar
306 accession: VCV000017864.36). This variant was replicated in the present GWAS for AD
307 (Figure 2A and Supplementary Table 7). In total, 976 studies reported rs429358-associated
308 hits in the GWAS catalog (Cerezo et al., 2025). Of these, seven that reported associations
309 with AD (Moreno-Grau et al., 2019; Wang et al., 2021; Nazarian et al., 2019; Hong et al.,
310 2020; Akgun et al., 2024; Harper et al., 2022; Li et al., 2021) (Supplementary Table 8),
311 mostly based on clinically diagnosed cases using established guidelines, were comparable
312 to the present study (Figure 3).

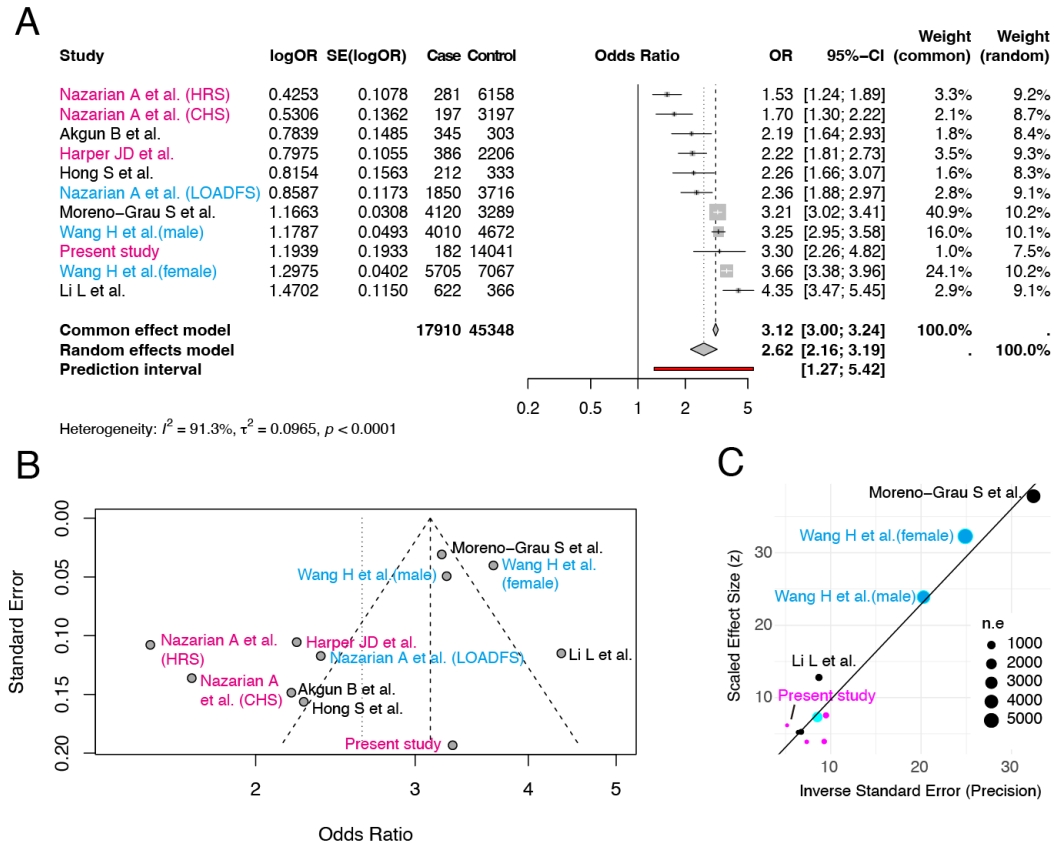


Figure 3. Comparison with previous results.

Effect sizes of rs429358 were compared with those reported in previous studies registered in the GWAS Catalog (Cerezo et al., 2025) (A). Studies associated with the trait "Alzheimer's disease" were selected. The effect size for each cohort (or subgroup) was used when available for analysis. The study design is indicated by color: magenta for population-based studies, black for case–control studies, and cyan for other designs (including meta-analyses and family-based studies). A funnel plot (B) and Egger's test (C) were used to assess potential publication bias. The linear regression test for funnel plot asymmetry yielded an intercept of -3.49 ± 1.52 (P= 0.0484). HRS, University of Michigan Health and Retirement Study; CHS, Cardiovascular Health Study; LOADFS, Late-Onset Alzheimer's Disease Family Study.

The odds ratio estimated in our study was 3.30 (95% CI: 2.26–4.82). This is consistent with estimates from larger-scale studies (number of cases > 4000; 3.21 (3.02–3.41) reported by Moreno-Grau et al (Moreno-Grau et al., 2019) and Wang et al (Wang et al., 2021) (3.25 (2.94–3.58) and 3.66 (3.37–3.97)) (Figure 3A, B). Moreover, the precision in our study was similar to that in studies with smaller cases (number of cases < 500), such as those by Nazarian et al (Nazarian et al., 2019) and Hong et al (Hong et al., 2020) (Figure 3C). Egger's test revealed a significant asymmetry (-3.49 ± 1.52; $P = 0.0484$) in the effect size estimates of these GWAS.

Discussion

According to the NIA-AA research framework (Jack et al., 2018), a biological definition of AD requires costly and time-consuming assessments, such as biomarker testing using positron emission tomography or cerebrospinal fluid analysis. However, increasing case numbers and frequency of follow-up limit the practical application of these approaches. The more than six-fold increase in the estimated prevalence of AD demonstrates the effectiveness of RWD-based phenotyping in community-based cohort studies.

In the present study, the estimated prevalence of AD in the cohort was 1.3%. By 2025, the prevalence of dementia is estimated to be 1.10% (95% CI: 1.07–1.17) in men and 3.47% (95% CI: 3.41–3.52) among women aged 60–74 years in Japan (Kasajima et al., 2022). As women comprised 61.5% of our dataset, the expected overall prevalence of dementia was approximately 2.56%, calculated as $0.615 \times 0.0347 + 0.386 \times 0.011$. Assuming that AD accounts for approximately two-thirds of all dementia cases in Japan (Ikeda et al., 2021; Montgomery et al., 2018), its estimated prevalence would be approximately 1.7%, resulting in a smaller gap with the estimate derived using RWD. However, our dataset included claims data from the LsMCSE, thereby covering individuals aged 75 years and older. Because the prevalence of AD increases substantially with age (Kasajima et al., 2022)—especially among individuals aged ≥75—the inclusion of older adults may increase the overall prevalence. Nevertheless, given that the participants were followed up for 7 years, and most of the cohort remained under 75 years old, the observed prevalence of 1.3% may still represent a reasonable estimate.

Case–control classification based on RWD also affected the GWAS results for AD. The increased log-likelihood values suggest that this updated phenotyping approach may reduce the effect of unmeasured confounders that are not included in the regression models. The replicated detection of the *APOE* variant rs429358, which has been identified as an AD risk allele in several studies, further supports the validity of this approach. The odds ratio

364 estimated in the present study was similar to those reported in large-scale clinical or meta-
365 analytic GWAS with more than 4,000 cases. However, the standard error remained
366 comparable to those from studies with similar case numbers. These findings suggest that
367 the revised case definition improved case identification while maintaining accuracy.

368 The prevalence of PD also increased when claims data were included. However, no
369 significant improvement was observed in the GWAS results. Although a few studies have
370 reported the prevalence of PD in Japan, regional estimates for East Asia suggested an all-
371 age prevalence of 0.357% (95% CI: 0.287–0.428) in 2021, which is estimated to increase to
372 0.821% (95% CI: 0.674–0.995) by 2050 (Su et al., 2025). The estimated prevalence of 0.7%
373 in the present cohort, which is predominantly older, is consistent with these projections.
374 A limitation of this study is the potential selection bias due to the characteristics of the
375 insured population. This study included claims data from the NHI system, which covers
376 approximately 30% of the Japanese population, and from the LsMCSE system for those
377 aged 75 years and older. Therefore, it does not represent the entire population (Shigeoka,
378 2014). Additionally, we could not directly compare the phenotypes estimated from
379 prescription data with clinical diagnoses made in hospitals. This was due to ethical
380 constraints and the substantial human effort required to collect diagnosis data from the
381 hospitals where cohort participants were treated. However, we plan to address this
382 limitation in future research.

383 In conclusion, this study demonstrates that incorporating RWD into case
384 definitions can improve the efficiency of epidemiological research in population-based
385 studies while maintaining accuracy. These approaches may improve public health
386 surveillance and contribute to the advancement of precision medicine.

387

388

389

References

Akgun, B., Feliciano-Astacio, B. E., Hamilton-Nelson, K. L., Scott, K., Rivero, J., Adams, L. D., et al. (2024). Genome-wide association analysis and admixture mapping in a Puerto Rican cohort supports an Alzheimer disease risk locus on chromosome 12. Frontiers in Aging Neuroscience, 16, 1459796. *doi*: 10.3389/fnagi.2024.1459796

Briggs, S. E. W., Law, P., East, J. E., Wordsworth, S., Dunlop, M., Houlston, R., et al. (2022). Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. BMJ, 379, e071707. *doi*: 10.1136/bmj-2022-071707

Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K. O., et al. (2025). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Research, 53, D998–D1005. *doi*: 10.1093/nar/gkae1070

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4, 7. *doi*: 10.1186/s13742-015-0047-8

Chen, C. Y., Chen, T. T., Feng, Y. A., Yu, M., Lin, S. C., Longchamps, R. J., et al. (2023). Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. Cell Genomics, 3, 100436. *doi*: 10.1016/j.xgen.2023.100436

Fujinaga, J., & Fukuoka, T. (2022). A review of research studies using data from the Administrative Claims Databases in Japan. Drugs - Real World Outcomes, 9, 543–550. *doi*: 10.1007/s40801-022-00331-5

Harper, J. D., Fan, K.-H., Aslam, M. M., Snitz, B. E., DeKosky, S. T., Lopez, O. L., et al. (2022). Genome-wide association study of incident dementia in a community-based sample of older subjects. Journal of Alzheimer's Disease, 88, 787–798. *doi*: 10.3233/JAD-220293

Hiramatsu, K., Barrett, A., Miyata, Y., & PhRMA Japan Medical Affairs Committee Working Group 1. (2021). Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs - Real World Outcomes, 8, 459–480. *doi*: 10.1007/s40801-021-00266-3

Hong, S., Prokopenko, D., Dobricic, V., Kilpert, F., Bos, I., Vos, S. J. B., et al. (2020). Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset. Translational Psychiatry, 10, 403. *doi*: 10.1038/s41398-020-01074-z

Hsu, L. A., Teng, M. S., Duan, D. M., Yeh, K. H., Wu, S., & Ko, Y. L. (2025). 3-Hydroxy-3-methylglutaryl-CoA reductase variants strongly associated with low-density lipoprotein cholesterol levels and diabetes mellitus risk in a Taiwanese population: a Mendelian randomization study. Tzu Chi Medical Journal, 37, 181–188. *doi*: 10.4103/tcmj.tcmj_247_24

Ikeda, S., Mimura, M., Ikeda, M., Wada-Isoe, K., Azuma, M., Inoue, S., et al. (2021). Economic burden of Alzheimer's disease dementia in Japan. Journal of Alzheimer's Disease, 81, 309–319. *doi*: 10.3233/JAD-210075

Ismail Umlai, U. K., Toor, S. M., Al-Sarraj, Y. A., Mohammed, S., Al Hail, M. S. H., Ullah, E., et al. (2025). A multi-ancestry genome-wide association study and evaluation of polygenic scores of LDL-C levels. Journal of Lipid Research, 66, 100752. *doi*: 10.1016/j.jlr.2025.100752

Jack, C. R., Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14, 535–562. *doi*: 10.1016/j.jalz.2018.02.018

Kasajima, M., Eggleston, K., Kusaka, S., Matsui, H., Tanaka, T., Son, B. K., et al. (2022). Projecting prevalence of frailty and dementia and the economic cost of care in Japan from 2016 to 2043: a microsimulation modelling study. The Lancet Public Health, 7, e458–e468. *doi*: 10.1016/S2468-2667(22)00044-5

Kuriyama, S., Yaegashi, N., Nagami, F., Arai, T., Kawaguchi, Y., Osumi, N., et al. (2016). The Tohoku Medical Megabank Project: design and mission. Journal of Epidemiology, 26, 493–511. *doi*: 10.2188/jea.JE20150268

Li, L., Yang, Y., Zhang, Q., Wang, J., Jiang, J., & Alzheimer's Disease Neuroimaging Initiative. (2021). Use of deep-learning genomics to discriminate healthy individuals from those with Alzheimer's disease or mild cognitive impairment. Behavioural Neurology, 14, 3359103. *doi*: 10.1155/2021/3359103

Liu, F., & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology, 22, 287. *doi*: 10.1186/s12874-022-01768-6

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26, 2867–2873. *doi*: 10.1093/bioinformatics/btq559

Minegishi, N., Nishijima, I., Nobukuni, T., Kudo, H., Ishida, N., Terakawa, T., et al. (2019).

Biobank establishment and sample management in the Tohoku Medical Megabank Project. The Tohoku Journal of Experimental Medicine, 248, 45–55. *doi*: 10.1620/tjem.248.45

Ministry of Health, Labour and Welfare. (2025). Overview of medical service regime in Japan. Retrieved from https://www.mhlw.go.jp/bunya/iryouhoken/iryouhoken01/dl/01_eng.pdf

Montgomery, W., Ueda, K., Jorgensen, M., Stathis, S., Cheng, Y., & Nakamura, T. (2018). Epidemiology, associated burden, and current clinical practice for the diagnosis and management of Alzheimer's disease in Japan. ClinicoEconomics and Outcomes Research, 10, 13–28. *doi*: 10.2147/CEOR.S146788

Moreno-Grau, S., de Rojas, I., Hernandez, I., Quintela, I., Montrreal, L., Alegret, M., et al. (2019). Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project. Alzheimer's & Dementia, 15, 1333–1347. *doi*: 10.1016/j.jalz.2019.06.4950

Nazarian, A., Yashin, A. I., & Kulminski, A. M. (2019). Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex disparities. Alzheimer's Research & Therapy, 11, 5. *doi*: 10.1186/s13195-018-0458-8

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., et al. (2008). Long-range LD can confound genome scans in admixed populations. American Journal of Human Genetics, 83, 132–135. *doi*: 10.1016/j.ajhg.2008.06.005

Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nature Genetics, 53, 1415–1424. *doi*: 10.1038/s41588-021-00931-x

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Meta-analysis with R. Springer International Publishing. *doi*: 10.1007/978-3-319-21416-0

Shida, H., Kajiyama, K., Sawada, S., Ishiguro, C., Kubo, M., Kimura, R., et al. (2023). Use of National Database of Health Insurance Claims and Specific Health Checkups for examining practical utilization and safety signal of a drug to support regulatory assessment on postmarketing drug safety in Japan. Frontiers in Medicine, 10, 1096992. *doi*: 10.3389/fmed.2023.1096992

Shigeoka, H. (2014). The effect of patient cost sharing on utilization, health, and risk protection. American Economic Review, 104, 2152–2184. Retrieved from https://www.jstor.org/stable/42920882

Su, D., Cui, Y., He, C., Yin, P., Bai, R., Zhu, J., et al. (2025). Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. BMJ, 388, e080952. *doi*: 10.1136/bmj-2024-080952

Sutoh, Y., Hachiya, T., Otsuka-Yamasaki, Y., Komaki, S., Minabe, S., Ohmomo, H., et al. (2025). Healthy lifestyle practice correlates with decreased obesity prevalence in individuals with high polygenic risk: TMM CommCohort study. Journal of Human Genetics, 70, 9–15. *doi*: 10.1038/s10038-024-01280-3

Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., et al. (2018). Innovation at the intersection of clinical trials and real-world data science to advance patient care. Clinical and Translational Science, 11, 450–460. *doi*:10.1111/cts.12559.

Wang, H., Lo, M. T., Rosenthal, S. B., Makowski, C., Andreassen, O. A., Salem, R. M., et al. (2021). Similar genetic architecture of Alzheimer's disease and differential *APOE* effect between sexes. Frontiers in Aging Neuroscience, 13, 674318. *doi*: 10.3389/fnagi.2021.674318

Yasuda, J., Kinoshita, K., Katsuoka, F., Danjoh, I., Sakurai-Yageta, M., Motoike, I. N., et al. (2019). Genome analyses for the Tohoku Medical Megabank Project towards establishment of personalized healthcare. Journal of Biochemistry, 165, 139–158. *doi*: 10.1093/jb/mvy096

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics, 50, 1335–1341. *doi*: 10.1038/s41588-018-0184-y

Supplementary information

Supplemental_Tables.xlsx. The Excel file includes Supplementary Tables 1–8.

Acknowledgements

This research utilized the supercomputer system and dbTMM provided by the Tohoku Medical Megabank Project, supported by AMED under grant numbers JP21tm0424601 and JP21tm0124005.

Author contributions

Conceptualization: YO-Y, YS, AS; Methodology: YO-Y, YS, AS; Investigation: NN, RA, KT; Formal analysis: YO-Y, NN, YS; Data curation: YO-Y, NN, YS, KT, AS; Writing—original

554 draft: YO-Y, NN, YS; Writing—review & editing: YO-Y, NN, YS, KT, MN, SK, SM, HO, KA,
555 YI, MS, AS.
556

557 Figure Legends

558 **Figure 1. The workflow in this study.**

559 The study participants had hard-called and imputed genotype data, as well as
560 cohort information including health insurance claims, demographic characteristics (age
561 and sex), laboratory test results, and responses to pre-specified items in a self-
562 administered questionnaire (SAQ). The real-world data used in this study included
563 disease diagnoses and prescription histories derived from medical, pharmacy, and
564 Diagnosis Procedure Combination claims data. Phenotyping was performed using either
565 conventional definitions based on SAQs and laboratory tests or extended definitions
566 incorporating health insurance claims data. HWE, Hardy-Weinberg equilibrium; LD,
567 linkage disequilibrium; MAF, minor allele frequency; PC, principal component; SD,
568 standard deviation; V, variants.

569

570 **Figure 2. Genome-wide association studies (GWAS) using two case definitions.**

571 GWAS were performed using two different case definitions: (1) the conventional
572 method based on SAQs and laboratory tests (light-colored Manhattan plot; most of the
573 plot is hidden), and (2) the extended method incorporating health insurance claims data
574 (dark-colored Manhattan plot) (A). Gene names near genome-wide significant variants
575 are shown in italics; those identified only in the GWAS using definition (2) are shown in
576 bold italics. The red and blue dashed horizontal lines indicate the thresholds for genome-
577 wide significance ($P = 5 \times 10^{-8}$) and suggestive significance ($P = 5 \times 10^{-5}$), respectively.

578 *λc/λp* denote genomic inflation factors for the GWAS based on case definitions (1) and (2),
579 respectively; 95% confidence intervals are shown in square brackets. Comparison of the
580 sum of log-likelihoods (log L) from GWAS summary statistics between the two case
581 definitions (B). The number of variants and sample sizes were identical between the two
582 analyses. Effect sizes and directions of significant variants identified in GWAS using the
583 two different case definitions were compared (C). Variants are color-coded by chromosome.
584 Bars and points represent the standard errors and mean of the beta values, respectively.
585 The dotted line represents the diagonal.

586

587 **Figure 3. Comparison with previous results.**

588 Effect sizes of rs429358 were compared with those reported in previous studies registered
589 in the GWAS Catalog (Cerezo et al., 2025) (A). Studies associated with the trait
590 "Alzheimer's disease" were selected. The effect size for each cohort (or subgroup) was used
591 when available for analysis. The study design is indicated by color: magenta for
592 population-based studies, black for case–control studies, and cyan for other designs
593 (including meta-analyses and family-based studies). A funnel plot (B) and Egger's test (C)
594 were used to assess potential publication bias. The linear regression test for funnel plot
595 asymmetry yielded an intercept of -3.49 ± 1.52 (P = 0.0484). HRS, University of Michigan
596 Health and Retirement Study; CHS, Cardiovascular Health Study; LOADFS, Late-Onset
597 Alzheimer's Disease Family Study.

598