Real-world data from the Japanese National Health Insurance System enables 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 contributed equally to this work

[†]Correspondence: ashimizu@iwate-med.ac.jp

Abstract

The use of medical databases, known as real-world data (RWD), facilitates accurate, efficient, and frequent estimation of disease prevalence in population-based studies. Thus, it is expected to be a game-changer in epidemiology and public health research. However, the lack of standardized data formats across hospitals makes it challenging to integrate data across institutions. In contrast, health insurance claims data in Japan are standardized under a nationally unified format, making them a reliable source of structured RWD.

We aimed to evaluate the utility of insurance claims data in epidemiological research. Incorporating both diagnosis and prescription information from claims into case definitions resulted in four- and six-fold increases in the estimated prevalence of Alzheimer's disease (AD)

and Parkinson's disease, respectively, compared to conventional definitions based on self-administered questionnaires. Subsequent genome-wide association studies (GWAS) for AD demonstrated an increased log-likelihood of the model and identified a characteristic signal at the *APOE* locus. These changes were exclusively observed when using claims-enriched case definitions. The effect size of the *APOE* variant was further compared with previous GWAS findings. Our results were consistent with case–control studies involving over 4,000 cases, while the standard error remained comparable to that of smaller studies. These findings suggest that incorporating claims-based clinical data—diagnoses and prescriptions—into phenotype definitions improves case identification while maintaining accuracy. This approach may support scalable strategies in genomic epidemiology and public health surveillance.

Keywords: Real-world data, Health insurance claims data, Genome-wide association studies, Alzheimer's disease, Phenotyping

Introduction

The accurate estimation of disease prevalence within a community is crucial in public health. However, challenges remain. Recent advances in medical databases—such as electronic health records (EHRs) 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.[1,2] Furthermore, the integration of RWD with large-scale genome cohorts increasingly contributes to the generation of epidemiological evidence that informs genomic research.[3] Nonetheless, technical challenges remain. Although diagnoses recorded in EHRs are generally reliable, the lack of standardized data formats across hospitals in many regions limit the integration of data across institutions.[4]

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.[5] 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.[6] 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.[7]

The Tohoku Medical Megabank Community Cohort (TMM CommCohort) is a large-scale population-based cohort study established to monitor public health in areas affected by the 2011 Great East Japan Earthquake.[8] This cohort obtained health insurance claims data through agreements with participating local governments. In this context, diagnosis and

89 prescription data from claims can supplement medical histories based on self-administered
90 questionnaires (SAQs) and serve as proxies for clinical diagnoses.

91 In this study, we evaluated the incorporation of prescription and diagnosis data from
92 insurance claims into conventional case/control diagnostic definitions. The prevalence of
93 hypertension, dyslipidemia, diabetes, Parkinson's disease (PD), and Alzheimer's disease (AD)
94 is expected to increase in aging societies. Therefore, we aimed to evaluate whether
95 incorporating diagnosis and prescription data from insurance claims could improve
96 case/control definitions for these common diseases. We further sought to evaluate the accuracy
97 of phenotyping using claims-enriched case definitions in the context of genome-wide
98 association studies (GWAS). These findings may contribute to more scalable strategies in
99 genomic epidemiology and public health research. In addition, they highlight the potential
100 value of a feedback cycle in which claims data inform both public health action and evidence-
101 based policy development for insured individuals.

Materials and Methods

Population

The design of the TMM CommCohort study was described previously.[8] 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 SAQ. 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 (DPC) 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. Personal 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 defined by the Japanese National Health Insurance System. 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 phenotypic information were performed on a standalone computing system—the Tohoku University Tohoku Medical Megabank Organization supercomputer.

Genotyping and GWAS

Genotyping and genotype imputation were performed as previously described.[9-11] In the hard-call genotype data, variants that met any of the following criteria were excluded: low call rate (<0.99), low heterozygous allele count (<5), deviation from Hardy–Weinberg equilibrium ($p < 1 \times 10^{-6}$), low minor allele frequency (<0.01), or location within regions of strong linkage disequilibrium (LD).[12] Variant filtering was conducted using PLINK v1.9 (version b5.1).[13] LD pruning was performed with PLINK v2.0 (version a3LM) using the option "--indep-pairwise 1500 150 0.03".

Principal component analysis (PCA) was conducted in PLINK v2.0 using the "--pca approx allele-wts" option. One individual from each kinship pair (kinship > 0.0884) was excluded based on the KING method to account for relatedness.[14] The resulting principal components were then projected onto all individuals, including those excluded based on the kinship threshold.

Case-control GWAS were conducted using SAIGE v0.44.6.5.[15] Individuals with a low call rate (<0.99) or extreme genetic ancestry, defined as values of PC1 or PC2 deviating by ≥4 standard deviations from the mean, were excluded. In addition, imputed variants with low imputation quality ($R^2 < 0.3$) or allele frequency (<0.01) were excluded from the analysis.

Statistical analysis

Statistical analyses were performed using R versions 4.1.0 and 4.3.2. Meta-analysis was performed with the R package "meta" (version 8.0-2).[16]

Results

Retrieval of health insurance claims data and phenotyping

A workflow of the study is shown in Figure 1.

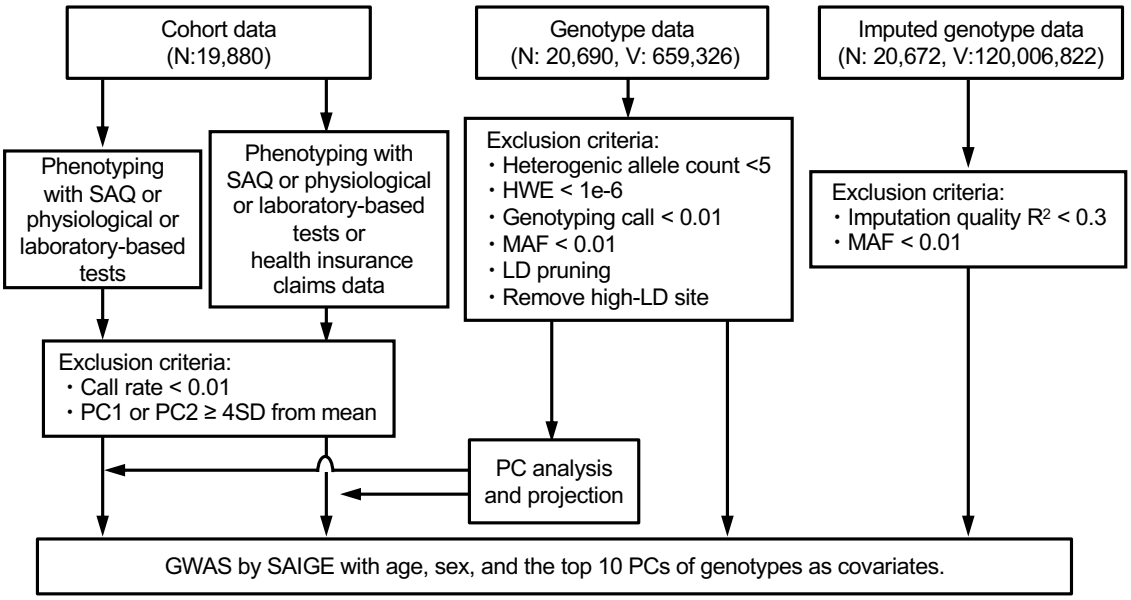

Figure 1. The workflow in this study.

The study participants had hard-called and imputed genotype data, as well as cohort information including health insurance claims, demographic characteristics (age and sex), laboratory test results, and responses to pre-specified items in a self-administered questionnaire (SAQ). The RWD used in this study included disease diagnoses and prescription histories derived from medical, pharmacy, and DPC claims data. Phenotyping was performed using either conventional definitions based on SAQs and laboratory tests or

164 expanded definitions incorporating insurance claims data. HWE, Hardy-Weinberg
165 equilibrium; LD, linkage disequilibrium; MAF, minor allele frequency; PC, principal
166 component; V, variants.

167

168 Health insurance claims data on diagnosis and drug prescriptions were obtained for
169 19,880 individuals (7,638 men and 12,242 women) (Figure 1). Of these, 14,241 participants
170 met the cohort information and genotype data requirements. Claims data for hypertension,
171 diabetes, and dyslipidemia were retrieved from the NHI system and covered the period
172 between June 2013 and December 2019. The claims data for AD and PD were collected
173 between June 2013 and December 2020 and included data from both the NHI and LsMCSE.
174 Participants were classified into case and control groups using two criteria: (1) the
175 conventional case definition based on SAQs and laboratory tests used in the cohort study, and
176 (2) an expanded definition incorporating both disease names recorded in claims and drug
177 prescription data, in addition to (1) (Table 1).

178

179 **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]	HMG-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).

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

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 (GWAS)	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 in 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).

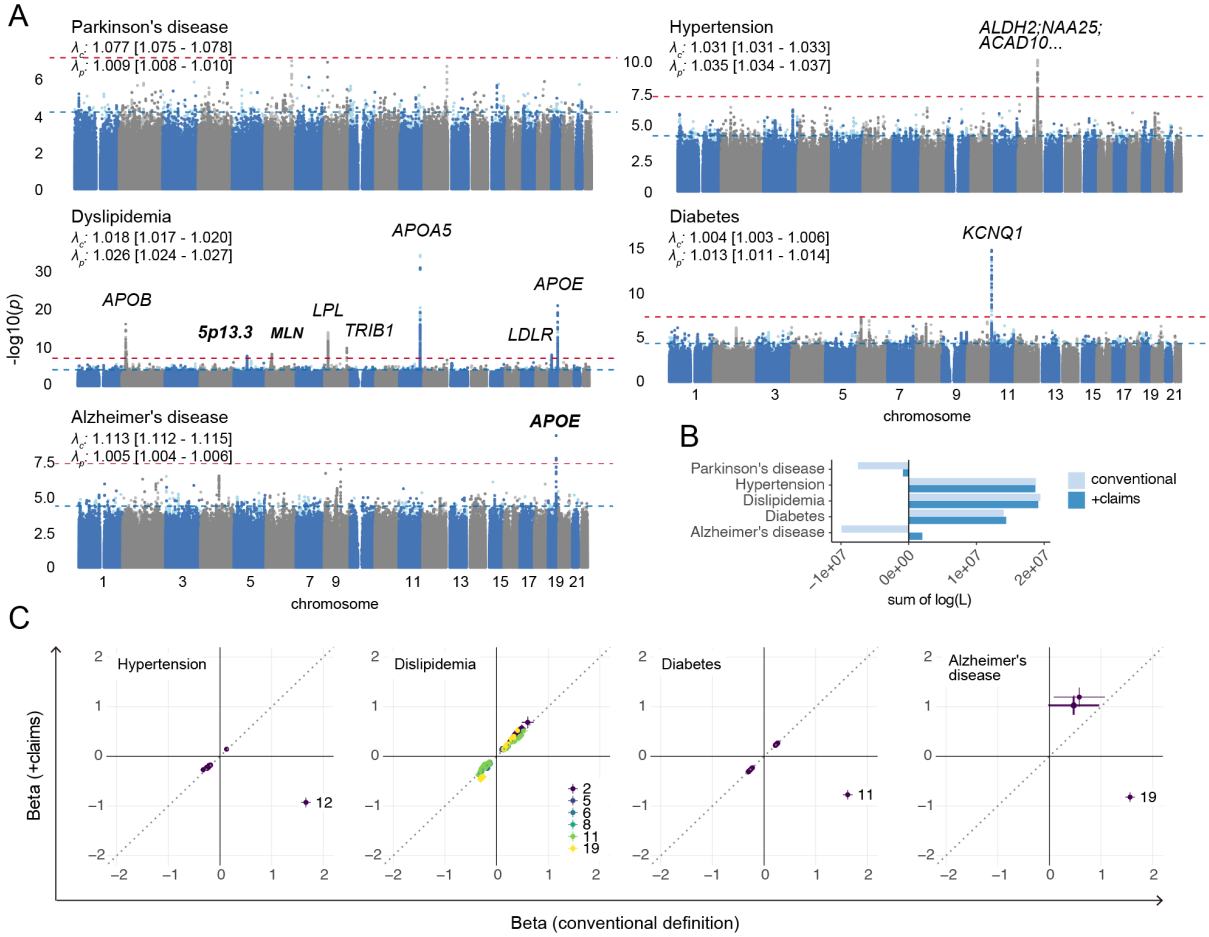

Figure 2. Genome-wide association studies (GWAS) using two case definitions.

GWAS were performed using two different case definitions: (1) the conventional method based on self-administered questionnaires (SAQ) (light-colored Manhattan plot; most of the plot is hidden), and (2) the conventional method combined with claims data (dark-colored Manhattan plot) (A). Gene names near genome-wide significant variants are shown in italics; those identified only in the GWAS using definition (2) are shown in bold italics. The red and blue dashed horizontal lines indicate the thresholds for genome-wide significance ($p = 5 \times 10^{-8}$) and suggestive significance ($p = 5 \times 10^{-5}$), respectively. $\lambda c/\lambda p$ denote genomic inflation factors for the GWAS based on case definitions (1) and (2), respectively; 95% confidence intervals are shown in square brackets. Comparison of the sum of log-likelihoods

216 (log L) from GWAS summary statistics between the two case definitions (B). The number of
217 variants and sample sizes were identical between the two analyses. Effect sizes and
218 directions of significant variants identified in GWAS using the two different case definitions
219 were compared (C). Variants are color-coded by chromosome. Bars and points represent the
220 standard errors and mean of the beta values, respectively. The dotted line represents the
221 diagonal.

222 Several genome-wide significant signals ($p < 5 \times 10^{-8}$) were detected for hypertension,
223 dyslipidemia, diabetes, and AD (Supplementary Table 1). *ALDH2* and *KCNQ1* were identified
224 in the GWAS for hypertension and diabetes. Both genes have been reported for blood pressure
225 and HbA1c, respectively, in previous GWAS (Figure 2A).[17] Known lipid-associated loci such as
226 *APOB, LPL, TRIB1, APOA5, LDLR*, and *APOE* were replicated in the GWAS for dyslipidemia,
227 consistent with prior findings on LDL and HDL, total cholesterol, and triglyceride levels.[17] In
228 contrast, association signals at 5p13.3 and the motilin gene locus were detected only in the
229 GWAS using claims-enriched case definitions. These two genes have been detected in the LDL
230 cholesterol or triglyceride GWAS of the TMM CommCohort study, which included 67K
231 participants (Study ID: TGA000007, https://jmorp.megabank.tohoku.ac.jp/gwas-studies/) and
232 Asian populations.[18,19,20] The lead variant rs4704208 in the 5p13.3 region identified in this
233 study is an expression quantitative trait locus for the HMGCR gene (Supplementary Table 1),
234 as determined according to the GTEx database (https://www.gtexportal.org/home/). This
235 enzyme is the pharmacological target of statins, which are widely prescribed for the treatment
236 of dyslipidemia.

237 The GWAS for AD using claims-enriched case definitions revealed significant signals
238 in *APOE*, whereas these signals were not detected in the GWAS based on conventional
239 phenotyping. The log-likelihood values, which indicate model fit, increased with the use of
240 claims-enriched phenotyping for AD (Figure 2B). The effect sizes and directions of significant
241 variants were consistent across the different phenotyping definitions (Figure 2C). The

242 standard errors of effect size estimates were smaller for AD when using the claims-enriched
243 case definition, indicating improved estimation precision.
244

245 Comparison with previous studies

246 The rs429358 C allele, a missense variant in *APOE* that results in a cysteine-to-
247 arginine substitution, is classified as a risk and/or pathogenic allele for AD (ClinVar accession:
248 VCV000017864.36). This variant was replicated in the present GWAS for AD (Figure 2). In
249 total, 976 studies reported rs429358-associated hits in the GWAS catalog.[21] Of these, seven
250 that reported associations with AD (Supplementary Table 1), mostly based on clinically
251 diagnosed cases using established guidelines, were comparable to the present study (Figure
252 3).

253 The odds ratio estimated in our study was 3.30 (95% CI: 2.26–4.82). This is consistent
254 with estimates from larger-scale studies (number of cases > 4000; 3.21 (3.02–3.41) reported
255 by Moreno-Grau et al.[22] and Wang et al.[23] (3.25 (2.94–3.58) and 3.66 (3.37–3.97)) (Figure 3A,
256 B). Moreover, the precision in our study was similar to that in studies with smaller cases
257 (number of cases < 500), such as those by Nazarian et al.[24] and Hong et al.[25] (Figure 3C).
258 Egger's test revealed a significant asymmetry (-3.49 ± 1.52; $p = 0.0484$) in the effect size
259 estimates of these GWAS.

260
261

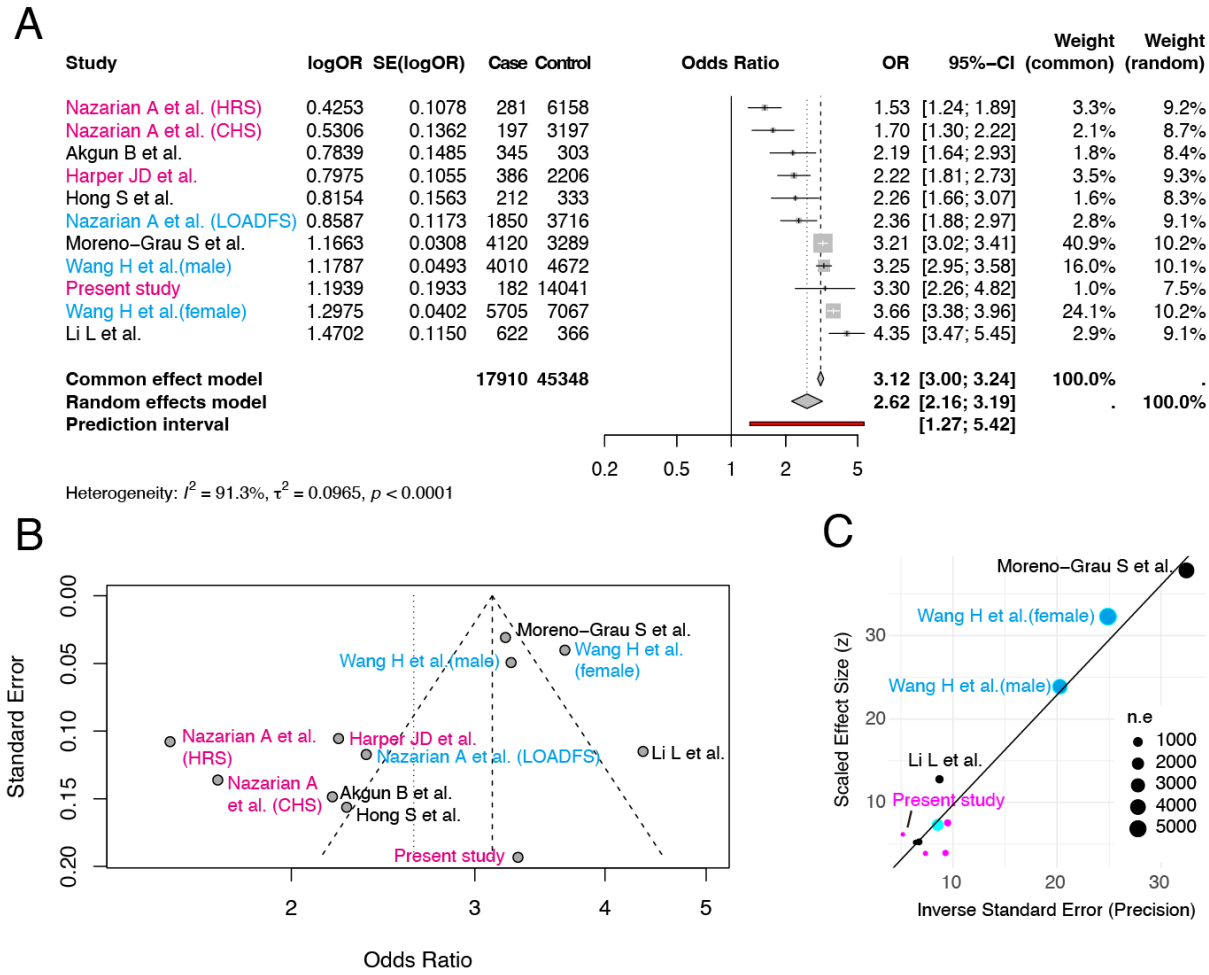

Figure 3. Comparison with previous results.

Effect sizes of rs429358 were compared with those reported in previous studies registered in the GWAS Catalog (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.

Discussion

According to the NIA-AA research framework,[26] 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.[27] 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,[28,29] 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[27]—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 estimated in the present study was similar to those reported in large-scale clinical or meta-

analytic GWAS with more than 4,000 cases. However, the standard error remained comparable to those from studies with similar case numbers. These findings suggest that the revised case definition improved case identification while maintaining accuracy.

The prevalence of PD also increased when claims data were included. However, no significant improvement was observed in the GWAS results. Although a few studies have reported the prevalence of PD in Japan, regional estimates for East Asia suggested an all-age prevalence of 0.357% (95% CI: 0.287–0.428) in 2021, which is estimated to increase to 0.821% (95% CI: 0.674–0.995) by 2050.[30] The estimated prevalence of 0.7% in the present cohort, which is predominantly older, is consistent with these projections.

A limitation of this study is the potential selection bias due to the characteristics of the insured population. The claims data used in this study included data from the NHI system, which covers approximately 30% of the Japanese population, and from the LsMCSE system for those aged 75 years and older; therefore, it does not represent the entire population.[31] Additionally, we could not directly compare the phenotypes estimated from prescription data with clinical diagnoses made in hospitals. This was due to ethical constraints and the substantial human effort required to collect diagnosis data from the hospitals where cohort participants were treated. However, we plan to address this limitation in future research.

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

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 owing to ethical

reasons, such as the protection of personal information and prevention of unintended personal identification. However, they are accessible 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. Those interested in obtaining the datasets for individuals may contact the corresponding author.

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22:287. https://doi.org/10.1186/s12874-022-01768-6.
2. Swift B, Jain L, White C, Chandrasekaran V, Bhandari A, Hughes DA, *et al.* Innovation at the intersection of clinical trials and real-world data science to advance patient care. Clin Transl Sci. 2018;11:450–460. https://doi.org/10.1111/cts.12559.
3. Briggs SEW, Law P, East JE, Wordsworth S, Dunlop M, Houlston R, *et al.* Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. BMJ 2022;379:e071707. 10.1136/bmj-2022-071707.
4. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group 1. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes 2021;8:459–480. https://doi.org/10.1007/s40801-021-00266-3.
5. Shida H, Kajiyama K, Sawada S, Ishiguro C, Kubo M, Kimura R, *et al.* Use of National Database of Health Insurance Claims and Specific Health Checkups for

348 examining practical utilization and safety signal of a drug to support regulatory
349 assessment on postmarketing drug safety in Japan. Front Med (Lausanne)
350 2023;10:1096992. https://doi.org/10.3389/fmed.2023.1096992.
351 6. Ministry of Health, L.a.W. Overview of medical service regime in Japan.
352 https://www.mhlw.go.jp/bunya/iryouhoken/iryouhoken01/dl/01_eng.pdf. (Accessed
353 25 Sept 2025).
354 7. Fujinaga J, Fukuoka T. A review of research studies using data from the
355 Administrative Claims Databases in Japan. Drugs Real World Outcomes
356 2022;9:543–550. https://doi.org/10.1007/s40801-022-00331-5.
357 8. Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, Osumi N, *et al.* The
358 Tohoku Medical Megabank Project: design and mission. J Epidemiol. 2016;26:493–
359 511. https://doi.org/10.2188/jea.JE20150268.
360 9. Minegishi N, Nishijima I, Nobukuni T, Kudo H, Ishida N, Terakawa T, *et al.*
361 Biobank establishment and sample management in the Tohoku Medical Megabank
362 Project. Tohoku J Exp Med. 2019;248:45–55. https://doi.org/10.1620/tjem.248.45.
363 10. Sutoh Y, Hachiya T, Otsuka-Yamasaki Y, Komaki S, Minabe S, Ohmomo H, *et al.*
364 Healthy lifestyle practice correlates with decreased obesity prevalence in
365 individuals with high polygenic risk: TMM CommCohort study. J Hum Genet.
366 2025;70:9–15. https://doi.org/10.1038/s10038-024-01280-3.
367 11. Yasuda J, Kinoshita K, Katsuoka F, Danjoh I, Sakurai-Yageta M, Motoike IN, *et*
368 *al.* Genome analyses for the Tohoku Medical Megabank Project towards
369 establishment of personalized healthcare. J Biochem. 2019;165:139–158.
370 https://doi.org/10.1093/jb/mvy096.
371 12. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, *et al.* Long-
372 range LD can confound genome scans in admixed populations. Am J Hum Genet.

373 2008;83:132–135; author reply 135–139. https://doi.org/10.1016/j.ajhg.2008.06.005.
374 13. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-
375 generation PLINK: rising to the challenge of larger and richer datasets.
376 Gigascience 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8.
377 14. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust
378 relationship inference in genome-wide association studies. Bioinformatics
379 2010;26:2867–2873. https://doi.org/10.1093/bioinformatics/btq559.
380 15. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, *et al.*
381 Efficiently controlling for case-control imbalance and sample relatedness in large-
382 scale genetic association studies. Nat Genet. 2018;50:1335–1341.
383 https://doi.org/10.1038/s41588-018-0184-y.
384 16. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. Springer
385 International Publishing, Cham, 2015. https://doi.org/10.1007/978-3-319-21416-0.
386 17. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, *et al.* A cross-
387 population atlas of genetic associations for 220 human phenotypes. Nat Genet.
388 2021;53:1415–1424. https://doi.org/10.1038/s41588-021-00931-x.
389 18. Chen CY, Chen TT, Feng YA, Yu M, Lin SC, Longchamps RJ, *et al.* Analysis across
390 Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci
391 for 36 quantitative traits. Cell Genom. 2023;3:100436.
392 https://doi.org/10.1016/j.xgen.2023.100436.
393 19. Hsu LA, Teng MS, Duan DM, Yeh KH, Wu S, Ko YL. 3-Hydroxy-3-methylglutaryl-
394 CoA reductase variants strongly associated with low-density lipoprotein
395 cholesterol levels and diabetes mellitus risk in a Taiwanese population: A
396 Mendelian randomization study. Tzu Chi Med J. 2025;37:181–188.
397 https://doi.org/10.4103/tcmj.tcmj_247_24.

398 20. Ismail Umlai UK, Toor SM, Al-Sarraj YA, Mohammed S, Al Hail MSH, Ullah E, *et*
399 *al.* A multi-ancestry genome-wide association study and evaluation of polygenic
400 scores of LDL-C levels. J Lipid Res. 2025;66:100752.
401 https://doi.org/10.1016/j.jlr.2025.100752.
402 21. Cerezo M, Sollis E, Ji Y, Lewis E, Abid A, Bircan KO, *et al.* The NHGRI-EBI GWAS
403 Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Res.
404 2025;53:D998–D1005. https://doi.org/10.1093/nar/gkae1070.
405 22. Moreno-Grau S, de Rojas I, Hernandez I, Quintela I, Montrreal L, Alegret M, *et al.*
406 Genome-wide association analysis of dementia and its clinical endophenotypes
407 reveal novel loci associated with Alzheimer's disease and three causality networks:
408 The GR@ACE project. Alzheimers Dement. 2019;15:1333–1347.
409 https://doi.org/10.1016/j.jalz.2019.06.4950.
410 23. Wang H, Lo MT, Rosenthal SB, Makowski C, Andreassen OA, Salem RM, *et al.*
411 Similar genetic architecture of Alzheimer's disease and differential APOE effect
412 between sexes. Front Aging Neurosci. 2021;13:674318.
413 https://doi.org/10.3389/fnagi.2021.674318.
414 24. Nazarian A, Yashin AI, Kulminski AM. Genome-wide analysis of genetic
415 predisposition to Alzheimer's disease and related sex disparities. Alzheimers Res
416 Ther. 2019;11:5. https://doi.org/10.1186/s13195-018-0458-8.
417 25. Hong S, Prokopenko D, Dobricic V, Kilpert F, Bos I, Vos SJB, *et al.* Genome-wide
418 association study of Alzheimer's disease CSF biomarkers in the EMIF-AD
419 Multimodal Biomarker Discovery dataset. Transl Psychiatry 2020;10:403.
420 https://doi.org/10.1038/s41398-020-01074-z.
421 26. Jack CR Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, *et al.*
422 NIA-AA Research Framework: Toward a biological definition of Alzheimer's

423 disease. Alzheimers Dement. 2018;14:535–562.
424 https://doi.org/10.1016/j.jalz.2018.02.018.
425 27. Kasajima M, Eggleston K, Kusaka S, Matsui H, Tanaka T, Son BK, *et al.* Projecting
426 prevalence of frailty and dementia and the economic cost of care in Japan from
427 2016 to 2043: a microsimulation modelling study. Lancet Public Health
428 2022;7:e458–e468. https://doi.org/10.1016/S2468-2667(22)00044-5.
429 28. Ikeda S, Mimura M, Ikeda M, Wada-Isoe K, Azuma M, Inoue S, *et al.* Economic
430 burden of Alzheimer's disease dementia in Japan. J Alzheimers Dis. 2021;81:309–
431 319. https://doi.org/10.3233/JAD-210075.
432 29. Montgomery W, Ueda K, Jorgensen M, Stathis S, Cheng Y, Nakamura T.
433 Epidemiology, associated burden, and current clinical practice for the diagnosis
434 and management of Alzheimer's disease in Japan. Clinicoecon Outcomes Res.
435 2018;10:13–28. https://doi.org/10.2147/CEOR.S146788.
436 30. Su D, Cui Y, He C, Yin P, Bai R, Zhu J, *et al.* Projections for prevalence of
437 Parkinson's disease and its driving factors in 195 countries and territories to 2050:
438 modelling study of Global Burden of Disease Study 2021. BMJ 2025;388:e080952.
439 https://doi.org/10.1136/bmj-2024-080952.
440 31. Shigeoka H. The effect of patient cost sharing on utilization, health, and risk
441 protection. Am Econ Rev. 2014;104:2152–2184.
442

443 ## Acknowledgements

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

Author contributions

Conceptualization: Y.O-Y., Y.S., A.S.; Methodology: Y.O-Y., Y.S., A.S.; Investigation: N.N., R.A., K.T.; Formal analysis: Y.O-Y., N.N., Y.S.; Data curation: Y.O-Y., N.N., Y.S., K.T., A.S.; Writing—original draft: Y.O-Y., N.N., Y.S.; Writing—review & editing: Y.O-Y., N.N., Y.S., K.T., M.N., S.K., S.M., H.O., K.A., Y.I., M.S., A.S.

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. Y.O-Y. and A.S. were supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under grant number 23K09696.

Ethical approval

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

Declaration of interests

The authors declare no competing interests.

Declaration of generation AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4o, including their Deep Research mode, in order to review and summarize previous papers and improve the computer script for analysis and the literature. After using this tool/service, the authors reviewed and

edited the content as needed and take full responsibility for the content of the publication.

Supplementary information

Supplemental_Table_1.xlsx. The Excel file includes Supplementary Table S1.